Aehnlichkeitsmasse

4. Aehnlichkeitsmasse

Im vorigen Kapitel wurde gezeigt, dass die Wahl der Transformationsmethode einer Entscheidung bedarf, welche sowohl von der Datenstruktur als auch vom Untersuchungsziel abhängt. Analog verhält es sich mit der Anwendung der verschiedenen Aehnlichkeitsmasse, die zur Bestimmung des Zusammenhanges zweier Aufnahmen oder zweier Arten beizuziehen sind. Je nach Standpunkt des Vegetationskundlers ergeben sich unterschiedliche Betrachtungsweisen. Viele verschiedene mathematische Formulierungen können diesen Rechnung tragen. Im folgenden wird eine beschränkte Auswahl teils bekannter, häufig verwendeter, teils für Vegetations- und Standortsdaten bewährter Aehnlichkeitsmasse vorgestellt und diskutiert.

4.1 Die Euklidsche Distanz

Die Euklidsche Distanz als Mass für die Unähnlichkeit zweier Vegetationsaufnahmen oder zweier Arten ist geometrisch leicht interpretierbar, mithin anschaulich und in ihren Eigenschaften ohne weiteres überblickbar. Aus diesen Gründen ist sie trotz einiger noch zu erwähnender Nachteile schon sehr oft verwendet worden (vgl. z.B. KUHN 1983). Betrachten wir zunächst den Vergleich zweier Aufnahmen. Dazu werden die Arten als räumliche Achsen aufgefasst. Im einfachsten Falle besitzt jede Aufnahme zwei Arten:

	Aufnahmne A	Aufnahme B
Art 1	2	1
Art 2	2	3

Mit diesen Werten lassen sich die Aufnahmen als Punkte im zweidimensionalen Raum darstellen (Abb. 4.1). Die Distanz (Unähnlichkeit) ergibt sich nach dem pythagoräischen Lehrsatz als

$$d(A,B) = [\sum_{i}(x_{Ai} - x_{Bi})^2]^{1/2} .$$

Im vorliegenden Beispiel mit i = 2 Arten ergibt das:

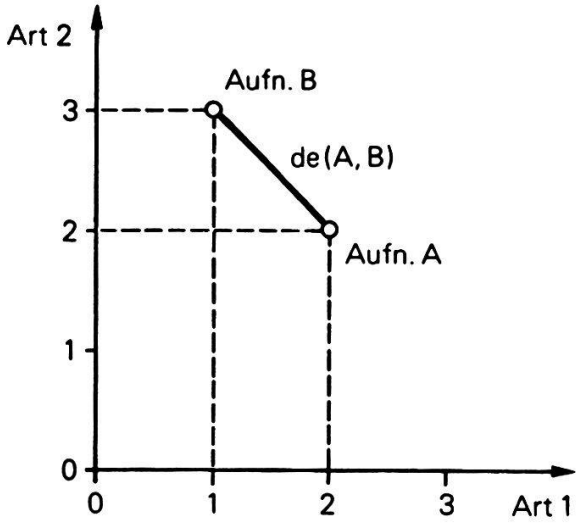

Abb. 4.1 Geometrische Bestimmung der Euklidschen Distanz zwischen zwei Aufnahmen A und B mit je 2 Arten.

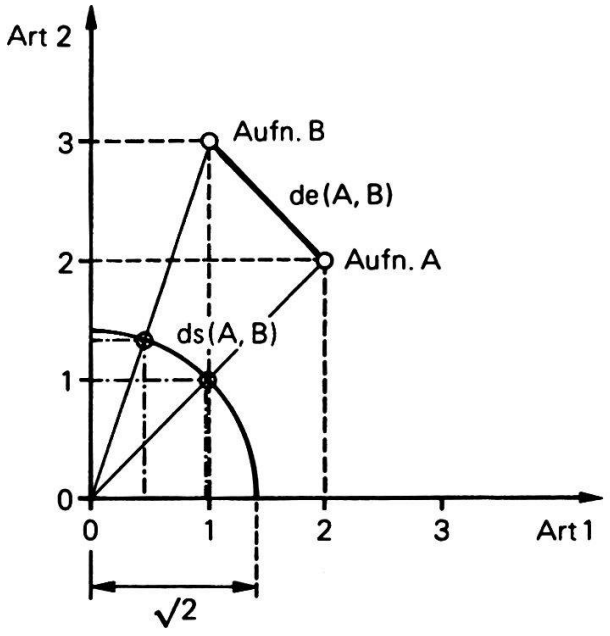

Abb. 4.2 Vergleich der Euklidschen Distanz de und der Sehnendistanz ds. Die ursprünglichen Abundanzen sind punktiert, die duch die Sehnendistanz implizit transformierten sind strichpunktiert.

$$d(A,B) = [(2 - 1)^2 + (2 - 3)^2]^{1/2} = 2^{1/2} = 1.41 .$$

Die Euklidsche Distanz beruht also grundsätzlich auf der Annahme, dass die Artmächtigkeiten als metrische Angaben aufgefasst werden können. Sie funktioniert aber ebenso für Präsenz-Absenz Daten. Für fehlende Arten wird dann mit Gewicht 0, für anwesende mit 1 gerechnet. Dabei tritt nun eine bedeutende Schwäche dieses sehr einfachen Masses besonders augenfällig in Erscheinung. In pflanzensoziologischen Tabellen kommt es sehr häufig vor, dass eine Art sowohl in einer Aufnahme A als auch in Aufnahme B fehlt. Diese Art trägt also nicht zur Unterschiedlichkeit von A und B bei. Artenarme Aufnahmen weisen deshalb meist eine kleinere Euklidsche Distanz auf als artenreiche und erscheinen daher oft ungerechtfertigterweise als relativ ähnlich. ORLOCI (1978), S. 46, demonstriert einen Fall, in dem zwei Aufnahmen mit gleichem Artspektrum unterschiedlicher erscheinen als solche mit verschiedener floristischer Zusammensetzung (Tab. 4.1, A). Tatsächlich findet man

$$d(1,3) = [(0 - 0)^2 + (1 - 4)^2 + (1 - 4)^2]^{1/2} = 18^{1/2} = 4.24$$

Die floristisch total verschiedenen Aufnahmen 1 und 2 stehen sich wesentlich näher:

$$d(1,2) = [(0 - 1)^2 + (1 - 0)^2 + (1 - 0)^2]^{1/2} = 3^{1/2} = 1.73$$

Als Konsequenz kann man festhalten, dass die Euklidsche Distanz nur dann zu empfehlen ist, wenn die Artenvielfalt der Aufnahmen nicht allzu unterschiedlich ist. Für den Vergleich des gemeinsamen Auftretens von Arten eignet sie sich wohl in den seltensten Fällen, da sich Arten bezüglich ihrer Häufigkeit (absolute Stetigkeit) in der Regel zu sehr unterscheiden.

Tabelle 4.1

Ausgangsdaten als Beispiel zur Berechnung der Euklidschen- und der Sehnendistanz (nach ORLOCI 1978). Rohdaten (A), normalisierte Aufnahmen (B).

A

	Aufnahme 1	Aufnahme 2	Aufnahme 3
Art 1	0	1	0
Art 2	1	0	4
Art 3	1	0	4

B

	Aufnahme 1	Aufnahme 2	Aufnahme 3
Art 1	0	1	0
Art 2	0.707	0	0.707
Art 3	0.707	0	0.707

4.2 Die Sehnendistanz

Die Sehnendistanz wird von ORLOCI (1978) vorgeschlagen, um den eben erwähnten Nachteil der Euklidschen Distanz zu eliminieren. Implizit werden die Aufnahmen normalisiert, das heisst auf Einheitslänge (1.0) gebracht. Quantitative Unterschiede in den Artmächtigkeiten, welche zu den im vorigen Kapitel erwähnten, unerwünschten Aehnlichkeitsbeziehungen geführt haben, werden ausgeglichen. Die Einheitslängentransformation ist geometrisch darstellbar. In Abb. 4.2 sind die ursprünglichen Verhältnisse zwischen den untransformierten Arten (punktiert) und diejenigen nach der Transformation (strichpunktiert) dargestellt. Jede neue Artmächtigkeit x_i' ergibt sich aus der ursprüglichen durch Division mit der Länge des Aufnahmevektors La:

$$x_i' = x_i \; / \; La$$

Die Länge des Aufnahmevektors berechnet sich mit

$$La = (\Sigma x_i^2)^{1/2} \; .$$

Man kann diese Transformationsvorschrift in die Formel für die Euklidsche Distanz einsetzen und erhält dann nach einigen Umformungen eine neue Formel für die direkte Berechnung der Sehnendistanz:

$$ds(A,B) = [2(1 - \Sigma_i x_{Ai} \, x_{Bi} \; / \; (\Sigma_i x_{Ai} \; \Sigma_i x_{Bi})^{1/2})]^{1/2} \; .$$

Darin ist x_{Ai} die Artmächtigkeit der Art i in der Aufnahme A. Dem Rechengang ist bei schrittweisem Vorgehen leichter zu folgen. Transformiert man das Beispiel in Tab. 4.1, A, wie besprochen, so erhält man die Werte in 4.1, B. Die Anwendung der Formel für die Euklidsche Distanz ergibt

$$ds(1,3) = [(0-0)^2+(0.707-0.707)^2+(0.707-0.707)^2]^{1/2} = 0.0$$

$$ds(1,2) = [(0-1)^2+(0.707-0.0)^2+(0.707-0.0)^2]^{1/2} = 1.414$$

Die quantitativen Unterschiede zwischen den Aufnahmen 1 und

3 fallen damit nicht mehr ins Gewicht, dafür wird dem als wichtig empfundenen Artenspektrum Rechnung getragen. Im Gegensatz zur Euklidschen Distanz gibt es für die maximale Unterschiedlichkeit einen festen Höchstwert, nämlich $2^{1/2}$ oder 1.414. Das Beispiel zeigt auch gleich einen Nachteil der Sehnendistanz, indem artenreiche Aufnahmen (Nr. 1 und 3) im allgemeinen schlecht, artenarme (Nr. 2) jedoch eher gut differenziert werden. Je nach Zielsetzung der Analyse kann diese Eigenschaft erwünscht oder unerwünscht sein. Günstig wirkt sich die Sehnendistanz aus, wenn Aufnahmen ähnlicher Diversität verglichen werden sollen. Auch bei der Untersuchung von Aufnahmen schwer vergleichbarer Skalierung, z.B. bei gemischten Daten aus der Literatur und aus Originaluntersuchungen, kann die Sehnendistanz die nötige Anpassung bringen. Dagegen führt der Vergleich von Arten oft zu unbefriedigenden Ergebnissen: Die Sehnendistanz bewirkt eine Ueberkorrektur der Nachteile der Euklidschen Distanz. Die Folge ist, dass seltene Arten gut, abundante schlecht differenziert werden.

4.3 Skalarprodukt und Kovarianz

Diese beiden Masse für die Aehnlichkeit sollen hier gemeinsam behandelt werden. Im Gegensatz zu den Distanzmassen führt eine gute Uebereinstimmung von Arten oder Aufnahmen zu hohen positiven, Unterschiedlichkeit zu negativen Werten. Die auftretenden Höchst- und Tiefstwerte besitzen keine festen Grenzen, was die Interpretation erschwert. Beide Masse enthalten auch eine Transformation der Daten: Sie zentrieren die Vektoren (Das Skalarprodukt ist auch auf unzentrierte Daten anwendbar. Es besitzt dann dieselben Vor- und Nachteile wie die Euklidsche Distanz). Im Falle des Vergleiches von Aufnahmen sind alle Artmächtigkeiten um ihren Mittelwert in der betreffenden Aufnahme zu vermindern:

$$x'_i = x_i - \bar{x}, \quad \bar{x} = 1/p \; (\Sigma x_i) \; .$$

Dabei ist p die Zahl aller in der betreffenden Vegetationstabelle auftretenden Arten. Ein Beispiel zeigt Tabelle 4.2. Links enthält sie die rohen Ausgangsdaten, rechts die

Tabelle 4.2

Ausgangsdaten für die Berechnung von Skalarprodukt und Kovarianz. Rohdaten (A), zentrierte Daten (B).

A

	Aufn 1	Aufn 2
Art 1	0	0
Art 2	0	1
Art 3	1	0
Art 4	1	0
$\bar{x}$=	1/2	1/4

B

	Aufn 1	Aufn 2
Art 1	-1/2	-1/4
Art 2	-1/2	3/4
Art 3	1/2	-1/4
Art 4	1/2	-1/4
$\bar{x}$=	0	0

Tabelle 4.3

Wirkung der Zentrierung auf Arten unterschiedlicher Dominanz. Rohdaten (A), zentrierte Daten (B).

A

Aufn.	1	2	3	4	$\bar{x}$
Art 1	3	3	4	2	3
Art 2	1	1	2	0	1

B

Aufn.	1	2	3	4	$\bar{x}$
Art 1	0	0	1	-1	0
Art 2	0	0	1	-1	0

über die Aufnahmen zentrierten. Das Skalarprodukt entspricht sodann einfach der Summe der Produkte der Mächtigkeit jeder Art:

$$S(A,B) = \sum_i x'_{Ai}\, x'_{Bi} \; .$$

Für die Werte in Tabelle 4.2 erhalten wir

$$S(A,B) = (-1/2*-1/4)+(-1/2*3/4)+(1/2*-1/4)+(1/2*-1/4)$$
$$= 1/8 - 3/8 - 1/8 - 1/8 = -1/2$$

Um die Kovarianz zu erhalten, ist das Ergebnis noch durch p-1 zu dividieren:

$$C(A,B) = \frac{1}{p-1} S(A,B) = \frac{1}{p-1} \sum_i x'_{Ai}\, x'_{Bi}$$

Da sich die Zahl p innerhalb eines Datensatzes nicht ändert, sind S(A,B) und C(A,B) stets proportional. Verwendet man sie z.B. für eine Gruppierungsanalyse, so werden die Ergebnisse identisch ausfallen.

Wie schon bei den Distanzmassen, so ergeben sich auch bei Skalarprodukt und Kovarianz Schwierigkeiten mit dem in der Pflanzensoziologie typischen Nebeneinander von häufigen und seltenen Arten. Fast leere (d.h. mit vielen Nullen besetzte) Vektorpaare besitzen immer eine kleine Kovarianz. Gut besetzte Vektorpaare weisen dagegen eine grössere Streuungsbreite auf, sowohl in positiver wie in negativer Richtung. Sie werden daher durch nachfolgende Analysen besser differenziert.

Ansonsten sind Euklidsche Distanz und Kovarianz recht verschiedene Konzepte. Die Zentrierung, welche letztere impliziert, bringt unter Umständen grosse quantitative Verschiebungen mit sich. Tab. 4.3 verdeutlicht diesen Verhalt. Die beiden unterschiedlich gut vertretenen Arten werden nach der Transformation genau gleich bewertet. Tab. 4.4 zeigt ein Beispiel, in welchem der Unterschied zwischen Euklid-

Tabelle 4.4

Beispiel zur Unterschiedlichkeit von Euklidscher Distanz und Skalarprodukt der zentrierten Daten. Erstere beträgt für die Aufnahmen beider Tabellen 1.414. Das Skalarprodukt beträgt in Tabelle A -1/4, in Tabelle B 0.0.

A

Aufn.	1	2
Art 1	1	0
Art 2	0	0
Art 3	0	0
Art 4	0	1
$\bar{x}$=	1/4	1/4

B

Aufn.	1	2
Art 1	1	0
Art 2	0	0
Art 3	1	1
Art 4	0	1
$\bar{x}$=	1/2	1/2

Tabelle 4.5

Ausgangsdaten als Beispiel zur Berechnung des Korrelationskoeffizienten zwischen zwei Aufnahmen 1 und 2. A: Rohdaten. B: Zentrierte Daten. C: Standardisierte Daten.

A

Aufn.	1	2
Art 1	1	1
Art 2	2	1
Art 3	3	7
$\bar{x}$=	2	3

B

Aufn.	1	2
Art 1	-1	-2
Art 2	0	-2
Art 3	1	4
$(\Sigma x^2)^{1/2}$=	1.41	4.89

C

Aufn.	1	2
Art 1	-0.707	-0.408
Art 2	0.0	-0.408
Art 3	0.707	0.816
s=	1	1
$\bar{x}$=	0	0

scher Distanz einerseits und Skalarprodukt und Kovarianz andererseits voll zum Tragen kommt. In den Aufnahmen der Tabelle A fehlt Art 3, während sie in B vorkommt. Die Euklidsche Distanz beträgt in beiden Fällen 1.414. Für das Skalarprodukt erhält man in Tabelle A -1/4 (Gegenläufigkeit), in Tabelle B 0.0 (Unabhängigkeit). Es reagiert damit auf die unterschiedlichen Varianzverhältnisse.

4.4 Der Korrelationskoeffizient

Der Korrelationskoeffizient ist das bekannteste und in Verfahren der schliessenden Statistik wohl am häufigsten verwendete Aehnlichkeitsmass (vgl. z.B. GAENSSLEN und SCHUBÖ, 1973, S. 15ff.). Seine Anwendung in der Analyse von Vegetationsdaten kann jedoch zu überraschenden, meist enttäuschenden Ergebnissen führen. Die Untersuchung seiner Eigenschaften zeigt einige Gründe auf.

Zunächst liegt dem Korrelationskoeffizienten die Formel zur Berechnung der Kovarianz zugrunde. Er hat auch deren im vorigen Kapitel gezeigte Nachteile. Zusätzlich unterwirft er die Datenvektoren einer weiteren Transformation, der z-Transformation oder Standardisierung. Dabei werden alle Elemente um den Mittelwert ihres Vektors vermindert und durch die Standardabweichung dividiert:

$$z = (x - \bar{x}) / s ,$$

wobei z der neue, standardisierte Wert ist. In Tabelle 4.5 ist die Wirkung dieser Transformation anhand eines Beispiels dargestellt. Die Verhältnisse zwischen den Artmächtigkeiten ändern sich gegenüber den ursprünglichen Daten grundsätzlich. Insbesondere ist zu beachten, dass die Streuungen der Aufnahmen vereinheitlicht werden. Damit fällt ein für sie als wesentlich erachtetes Charakteristikum ausser Betracht. Insbesondere Tabelle 4.5, C, zeigt, dass die Gewichtung der Abundanzen nach einer Standardisierung wohl selten den ursprünglichen Absichten eines Vegetationskundlers entsprechen dürfte. Daher befriedigt auch eine Aussage über

die Struktur von Vegetationstabellen selten, sofern sie auf Grund des Korrelationskoeffizienten ermittelt wurde. Anders verhält es sich mit Standortsdaten. Sind diese durch Verwendung unterschiedlicher Messskalen gewonnen worden, z.B. wenn pH-Werte und Grundwasserstände vorliegen, so gewährleistet die Standardisierung - durch die Anwendung des Korrelationskoeffizienten - die gewünschte Vergleichbarkeit weitgehend.

Nach der Standardisierung der Daten ergibt sich der Korrelationskoeffizient durch die Anwendung der Formel für die Kovarianz. Für die Aufnahmen in Tab. 4.5 findet man:

$$r^2 = 1/3\ [(-0.707*-0.408)+(0.0*-0.408)+(0.707*0.816)]$$
$$= 0.228$$
$$r = 0.537$$

Das Resultat weist auf einen mässig positiven Zusammenhang der Aufnahmen 1 und 2 hin. Da r zwischen +1 und -1 liegen kann, treten bei Gegenläufigkeit (z.B. wenn zwei Aufnahmen keine gemeinsamen Arten aufweisen) natürlich auch negative Werte auf.

Die oben angewandte Art der Berechnung des Korrelationskoeffizienten ist in der Praxis recht umständlich. Der Vollständigkeit halber sei noch eine zweckmässigere Formel angeführt:

$$r^2 = \frac{1}{p} \cdot \frac{\Sigma(x_A - \bar{x}_A)(x_B - \bar{x}_B)}{[\Sigma(x_A - \bar{x}_A)^2]^{1/2} * [\Sigma(x_B - \bar{x}_B)^2]^{1/2}}$$

A und B sind die zu vergleichenden Aufnahmen, p die Anzahl Arten. Damit lässt sich nun r direkt aus den Rohdaten errechnen.

4.5 Kontingenzmasse

In dieser Gruppe finden sich die frühesten Formeln zur Berechnung der Aehnlichkeiten von Aufnahmen, nähmlich diejenigen von JACCARD (1901) und SOERENSEN (1948). Varian-

ten von diesen sind in grosser Zahl beschrieben worden und sie tauchen ihrer Beliebtheit wegen auch in neueren Lehrbüchern immer wieder auf (MUELLER-DOMBOIS und ELLENBERG 1974, ORLOCI 1978, GAUCH 1982 usw.). Ausgangspunkt zur Berechnung der Aehnlichkeiten sind ursprünglich reine Präsenz-Absenz Daten, welche in einer Kontingenztafel auszuwerten sind:

Aufn. A / Aufn. B	+	-	
+	a	b	a+b
-	c	d	c+d
	a+c	b+d	n

Im Feld a wird die Anzahl der Arten eingetragen, die sowohl in A wie auch in B vorkommen. In b werden die Arten gezählt, wenn sie in A fehlen, in B aber vorkommen. In c finden sich Arten, die in B fehlen und in A vorkommen. Feld d schliesslich enthält die Anzahl Arten, welche in beiden Aufnahmen fehlen. In der letzten Kolonne bzw. Spalte findet man die sogenannten Randsummen a+c, b+d, a+b, c+d sowie die Gesamtfrequenz n. Mit dieser Notation lassen sich sowohl metrische als auch nichtmetrische Masse formulieren. Welcher Fall vorliegt, ist auf formalem Wege abzuklären: Wird die Dreiecksungleichung nicht verletzt, so ist das Mass metrisch (Kap. 3.1.3). Für die meisten Funktionen sind diese Eigenschaften direkt bei LEGENDRE und LEGENDRE (1979, Bd II, S. 31f) nachzuschlagen. Die nicht metrischen unter ihnen haben dabei den Vorzug, dass sie gegenüber Leerstellen mehr oder weniger unempfindlich sind. Verzerrungen, wie sie bei der Euklidschen Distanz besprochen wurden, können dadurch vermieden werden. Für die Anwendung verschiedener Methoden bedeutet dies jedoch eine Einschränkung, wie später noch zu zeigen ist.

Die Berechnung der Felder a, b, c und d wird in Tabelle 4.6 anhand eines Beispiels gezeigt. Ausgehend von 4 Aufnahmen

Tabelle 4.6

Messung des Zusammenhanges zweier Arten einer Vegetationstabelle (A) mit Hilfe einer Kontingenztabelle (B). "+" bedeutet, dass die betreffende Art vorkommt, "-" dass sie fehlt.

A

	Aufn. A	Aufn. B	Aufn. C	Aufn. D
Art 1	1	1	1	0
Art 2	0	1	0	0

B

Art 1 / Art2	+	-	
+	a=1	b=0	a+b=1
-	c=2	d=1	c+d=3
	a+c=3	b+d=1	n=4

soll darin das Zusammenfallen von 2 Arten ermittelt werden können. Die Gesamtfrequenz n entspricht zwangsläufig der Summe der Aufnahmen, d.h. 4. Alle für die Bestimmung der Aehnlichkeit benötigten Parameter sind damit bekannt.

Da nun hier eine echte Kontingenztafel vorliegt, drängt sich als erstes ein aus der schliessenden Statistik bekanntes Mass zur Bestimmung des Zusammenhanges von A und B geradezu auf. Werden nämlich die Randsummen a+c, b+c, a+b und c+d (Tabelle 4.6) als invariante, naturgegebene Grössen angenommen (was nach PIELOU (1977) in Wirklichkeit durchaus nicht immer zutreffen muss), so lässt sich ein sogenanntes **Chiquadrat** χ^2 formulieren. Diese statistische Testgrösse gibt an, inwiefern die Inhalte der Felder a bis d von zufällig zu erwartenden Werten abweichen. Sie berechnet sich als Summe der quadrierten Differenzen zwischen Erwartungswerten E und beobachteten Werten O, dividiert durch die Erwartungswerte jedes Feldes i:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Als Erwartungswert für das Feld a nimmt man das Produkt der Randsummen der Art A, also (a+c) und (a+b) und misst dieses mit der Gesamtfrequenz:

$$E(a) = \frac{(a+c)(a+b)}{N}$$

Die beobachtete Frequenz ist natürlich der Wert a selbst dividiert durch N. Durch einfache Ableitung (vgl. z.B MUELLER-DOMBOIS und ELLENBERG 1974, S. 238) lässt sich die Formel für die ganze Tafel finden:

$$\chi^2 = \frac{N((ad-bc)^2)}{(a+b)(a+c)(c+d)(b+d)}$$

Diese Grösse ist unter bestimmten Voraussetzungen χ^2-verteilt (vgl. BARTEL 1972, S. 71). Bedienen wir uns ausnahmsweise der schliessenden Statistik, so kann einer χ^2-Tabelle (Anz. Freiheitsgrade = 1) entnommen werden, ob die Unterschiedlichkeit im Auftreten von Art 1 und 2 bis auf normale Abweichungen rein zufälliger Natur sind. Im Beispiel von Tabelle 4.6 erhält man

$$\chi^2 = \frac{4((1-0)^2)}{1 * 3 * 1 * 3} = \frac{4}{9} = 0.44$$

Bei einer Irrtumswahrscheinlichkeit von 5% findet man in entsprechenden Tabellen einen Zufallshöchstwert von 3.84, womit die Unterschiedlichkeit der beiden Arten erwartungsgemäss nicht gesichert werden kann. Selbstverständlich ist diese Aussage mit Vorbehalten zu interpretieren, wurden doch die Voraussetzungen für einen χ^2-Test nicht überprüft. Sie kann aber - nun wieder im Sinne der explorativen Statistik - als Interpretationshilfe ohne Bedenken verwendet werden.

Die Inhalte der Kontingenztafel können nun zur Formulierung weiterer Aehlichkeitsmasse verwendet werden. Ausgehend vom Chiquadrat diskutiert PIELOU (1977, S. 208ff.) eingehend den **V- Wert** von YULE (1912):

$$V = \frac{ad - bc}{[(a+b)(a+c)(c+d)(b+d)]^{1/2}}$$

Es wird auch der Nachweis erbracht, dass es sich um einen Korrelationskoeffizienten handelt mit den Eigenschaften

V = 1 für vollständigen Zusammenhang
V = 0 für fehlenden Zusammenhang
V =-1 für Gegenläufigkeit

In unserem Beispiel (Tabelle 4.6) erhalten wir

$$V = (1-0)/9^{1/2} = 1/9^{1/2} = 0.33$$

Die eben besprochenen Aehnlichkeitsmasse haben für pflanzensoziologische Anwendungen den Vorteil, dass anstelle der Rohdaten mit Erwartungswerten operiert wird. Sie interpretieren eine Zunahme der Leerstellen (d.h. von Arten, die in beiden Aufnahmen fehlen) so, dass das gemeinsame Vorkommen zweier Arten weniger wahrscheinlich werde und bewerten es daher höher. Diese Art der Betrachtung ist in der meisten vegetationskundlichen Anwendungen sinnvoll - sofern der Datensatz eine gewisse Grösse besitzt und einigermassen homogen ist.

Wahrscheinlich das erste in der Pflanzensoziologie bekannt gewordene Aehnlichkeitsmass ist der **Koeffizient von Jaccard** (JACCARD 1901). Er berechnet sich nach der Formel

$$SJ = a / (a+b+c) .$$

Für das Beispiel in Tab. 4.6 erhält man

$$SJ = 1 / (1+0+2) = 0.33$$

Als Grenzwerte sind möglich:

max(SJ) = 1 (A und B identisch)
min(SJ) = 0 (A und B ohne gemeinsame Arten) .

Dem Koeffizient von Jaccard verwandt ist der **Koeffizient von Soerensen** (SOERENSEN 1948). Er birgt jedoch eine etwas stärkere Gewichtung gemeinsamer Arten in sich:

$$SS = 2a / (2a+b+c) .$$

Sein Wertebereich erstreckt sich ebenfalls von 0 bis 1. Für das Beispiel in Tab. 4.6 erhält man

$$SS = 2 / (2+0+2) = 0.5 .$$

Bei teilweise übereinstimmender Artengarnitur ergeben sich also etwas höhere Aehnlichkeitswerte als bei Jaccard.

Beide Koeffizienten sind unbeeinflusst von fehlenden Arten sowie vom Total n. Werden sie zur Analyse einer Vegetationstabelle eingesetzt, so ändern die Nenner von Aufnahmepaar zu Aufnahmepaar. Das gemeinsame Vorkommen der Arten wird mithin mit ständig änderndem Massstab gemessen, sodass die resultierende Aehnlichkeitsstruktur nicht-metrisch ist. Mit SJ und SS werden aber immer wieder gute Erfahrungen gemacht; wohl deshalb, weil sie der oft vorkommenden Heterogenität der Daten entgegenkommen.

Die bisher betrachteten Koeffizienten berücksichtigen nur die An- oder Abwesenheit einer Art. Es besteht jedoch öfters das Bedürfnis, auch die Artmächtigkeit in die Berechnungen mit einzubeziehen. Um dies bei Kontingenzmassen verwirklichen zu können, wurden verschiedenste Formeln vorgeschlagen (vgl. z.B. die Uebersicht bei MUELLER-DOMBOIS und ELLENBERG 1974). ELLENBERG (1956) verwendet in Anlehnung an Jaccard

$$SE = \frac{1/2\ \Sigma a_i}{\Sigma b_i + \Sigma c_i + 1/2(\Sigma a_i)} * 100$$

Darin ist Σa_i die Summe der Elemente der Zelle a in der Kontingenztafel (d.h. alle gemeinsamen Arten), Σb_i die Summe der Elemente in b und Σc_i diejenige in Zelle c.

Van der Maarels Koeffizient SM (VAN DER MAAREL et al. 1978) entspricht ebenfalls demjenigen von Jaccard, modifiziert für quantitative Daten:

$$SM = \frac{\sum_i x_{Ai}\ x_{Bi}}{\sum_i x_{Ai}^2 + \sum_i x_{Bi}^2 - \sum_i x_{Ai}\ x_{Bi}}$$

Das folgende Beispiel zeigt die Berechnung dieser beiden Koeffizienten:

Art	1	2	3	4
Aufn. A	1	2	3	0
Aufn. B	0	1	2	3

$$SE = \frac{1/2(2+1+3+2)}{1+3+1/2(2+1+3+2)} * 100 = \frac{4}{1+3+4} * 100 = 50$$

$$SM = \frac{(1*0)+(2*1)+(3*2)+(0*3)}{(1+4+9)+(1+4+9)-[(1*0)+(2*1)+(3*2)+(0*3)]}$$

$$= \frac{8}{14+14-8} = 0.4$$

Werden in den Ausgangsdaten alle von Null verschiedenen Werte durch 1 ersetzt, so ergibt sich für SE 50(%), für SM 0.5 und für SJ (Jaccard) ebenfalls 0.5. Konzeptionell erbringen die Koeffizienten Ellenbergs und van der Maarels keine neuen Aspekte.

Damit sind die Möglichkeiten der Aehnlichkeitsbestimmung auf Grund der Kontingenztafel bei weitem nicht erschöpft. Eine Uebersicht findet sich bei LEGENDRE und LEGENDRE (1979), wo in Band II, Seite 32/33 nicht weniger als 14 verschiedene Koeffizienten aufgeführt sind.

4.6 Absolutwertfunktionen

Aus dem Bestreben heraus, den Rechenaufwand für die Aehnlichkeitsbestimmung möglichst klein zu halten, wird gelegentlich die folgende, einfache Absolutwertfunktion verwendet:

$$A(A,B) = \sum_i ABS(x_{Ai} - x_{Bi}) .$$

Meistens wird diese Funktion als Alternative zur Euklidschen Distanz gesehen. Für ihre geometrische Interpretation sei auf Abb. 4.1 verwiesen. Die Euklidsche Distanz zwischen A und B entspricht dort der direkten Verbindung zwischen den beiden Punkten. Die Absolutwertfunktion misst dagegen ausschliesslich entlang der horizontalen und der vertikalen Achse und addiert die Beträge (A(A,B)=2), weshalb auch von City-Block- oder Manhattandistanz gesprochen wird. Gegenüber der Euklidschen Distanz ergeben sich kaum Vorteile. Es treten auch die bei ORLOCI (1978) erwähnten Fälle auf, bei welchen Aufnahmen mit identischem Artenspektrum als sehr unterschiedlich bewertet werden (Kap. 4.1). WHITTAKER (1952) skaliert deshalb die Aufnahmen innerhalb seines Koeffizienten W:

$$W(A,B) = \sum_i \text{ABS}\,(x_{Ai}/Q_A - x_{Bi}/Q_B) \quad \text{, wobei}$$

$$Q_A = \sum_i x_{Ai} \quad \text{und} \quad Q_B = \sum_i x_{Bi} \; .$$

Die Werte für W(A,B) sind dadurch auf den Bereich 0 bis 2 begrenzt. Betrachten wir unser aus dem letzten Kapitel bekanntes Beispiel:

Art	1	2	3	4
Aufn. A	1	2	3	0
Aufn. B	0	1	2	3

Für Whittakers Koeffizient ist $Q_A = Q_B = 6$ und

$$W(A,B) = \text{ABS}((1/6)-(0/6)) + \text{ABS}((2/6)-(1/6)) + \text{ABS}((3/6)-(2/6)) + \text{ABS}((0/6)-(3/6)) = 1.$$

ORLOCI (1978) erwähnt, dass W(A,B) auf die Darstellung einfacher Vegetationsgradienten linearisierende Wirkung hat, was natürlich auf die Bereichsanpassung zurückzuführen ist (und auch von der Sehnendistanz gesagt werden könnte).

4.7 Die Mahalanobis Distanz

Alle bisher besprochenen Aehnlichkeitsmasse sind von der Korrelation der Arten beeinflusst. Für viele Anwendungen ist diese Eigenschaft durchaus erwünscht. Pflanzenarten, die auf einen Standortsgradienten fast gleich reagieren, gehen dabei in die Berechnung der Aufnahmeähnlichkeit mit gleichem Gewicht ein wie solche, die eine ganz andere Eigenschaft der Standortes wiederspiegeln. Dem kann bei Bedarf abgeholfen werden. An einem einfachen Beispiel aus ORLOCI (1978) soll dies erläutert werden (Tabelle 4.7, A). Darin ist sofort ersichtlich, dass Art 1 und 2 hoch korrelieren (weil sie möglicherweise ähnliche Standortsansprüche besitzen). Art 3 ist dagegen annähernd unabhängig. Die Korrelationsmatrix bestätigt diese Befunde (Tabelle 4.7, B). Die Aufnahmen 1 bis 5 entstammen offensichtlich zwei verschiedenen Vegetationsgradienten, wobei der eine mit den Arten 1 und 2 besser vertreten ist als der andere mit Art 3. Dieses Ungleichgewicht soll die Mahalanobisdistanz ausgleichen. In Abbildung 4.3 wird das Vorgehen anhand eines zweidimensionalen Falles erläutert. Man erinnert sich, dass zur Berechnung der Euklidschen Distanz jede Art mit gleichem Gewicht in die Berechnung eingeht. In Abb. 4.3, A, erhält man

$$D^2(A,B) = (2-1)^2 + (1-2)^2 = 2.$$

Darin wird stillschweigend angenommen, dass die Arten 1 und 2 voneinander unabhängig sind und das Koordinatensystem deshalb rechtwinklig sei. Dies trifft hier eigentlich nicht zu. Den allgemeinen Fall zeigt Abb. 4.3 B. Darin korrelieren Art 1 und 2 positiv, was geometrisch durch schiefwinklige Koordinaten dargestellt wird. Die Distanz DM(A,B) vermindert sich gegenüber der Euklidschen Distanz. Die Berechnung ergibt sich direkt aus den Cosinussatz:

$$DM^2(A,B) = [(2-1)^2 + (1-2)^2] - 2(2-1)(1-2)\cos\alpha$$

Der Winkel α ist Ausdruck der Korrelation der Arten 1 und 2. Den benötigten Zusammenhang bringt die Formel

Tabelle 4.7

Beispiel zur Berechnung der Mahalanobisdistanz (A) mit der Korrelationsmatrix der Arten (B).

A

	Aufn.1	Aufn.2	Aufn.3	Aufn.4	Aufn.5
Art 1	2	5	2	1	0
Art 2	3	4	1	0	0
Art 3	0	1	4	3	1

B

	Art 1	Art 2	Art 3
Art 1	1.000	0.883	-0.163
Art 2		1.000	-0.536
Art 3			1.000

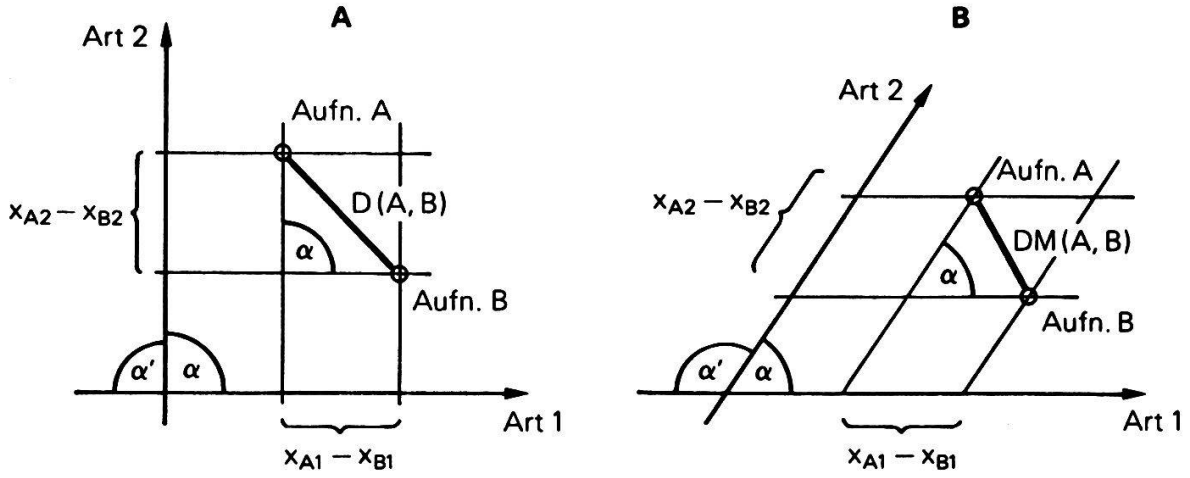

Abb. 4.3 Berechnung der Euklidschen Distanz (A) und der Mahalanobisdistanz (B).

$$r\ (1,2) = -\cos\ \alpha' = -\cos\ (180^{o} - \alpha)$$

Es ist leicht zu prüfen, dass sich DM(A,B) auf D(A,B) reduziert, wenn r (1,2) null wird, die Arten 1 und 2 also voneinander unabhängige Information tragen. In der allgemeinen Form, d.h. bei mehr als zwei Arten, berechnet sich die Mahalanobis Distanz wie folgt:

$$DM^2(A,B) = \sum_i (x_{Ai} - x_{Bi})^2 - 2 * \sum_j \sum_k (x_{Aj} - x_{Bj})(x_{Ak} - x_{Bk})$$

Ist p die Anzahl der Arten, so laufen die Summationen wie folgt:

$i=1,...,p$
$j=1,...,p-1$
$k=1,...,p$

Für die beiden ersten Aufnahmen in Tab. 4.7 erhält man demzufolge:

$$\begin{aligned} DM^2\ (1,2) = {} & (2-5)^2 + (3-4)^2 + (0-1)^2 \\ & -2[(2-5)(3-4)(-0.883) \\ & +(2-5)(0-1)(0.163) \\ & +(3-4)(0-1)(0.536)] = 3.775^2 \end{aligned}$$

Das Beispiel zeigt, dass bei der Berechnung der Mahalanobisdistanz laufend auf die Korrelationsmatrix zugegriffen werden muss, was bei der Verwendung von Rechenautomaten einen hohen Speicherbedarf erfordert und lange Rechenzeiten verursacht. Für die Praxis empfiehlt sich ein anderes Vorgehen. Anstelle von Artvektoren können die Faktorenwerte einer Hauptkomponentenanalyse verwendet werden, die immer unkorreliert sind (Kap. 7). Eine nachfolgend gerechnete Distanz ist dann ebenfalls frei vom Einfluss der Korrelation der Arten.
Die Mahalanobisdistanz hat nur einen Sinn im Kontext einer grösseren Vegetationstabelle. Ihr Vorteil gegenüber der Euklidschen Distanz ist meist unbedeutend, wird doch der nichtlinearen Struktur der Daten nicht Rechnung getragen.

4.8 Informationsmasse

Die Mehrheit bisher besprochener Aehnlichkeitsmasse wurde für metrische Daten entwickelt. Ihre Anwendung auf Artmächtigkeitsskalen ist nur mit Kunstgriffen möglich (Abschnitt 3.2). Echte Nominaldaten sind jedoch typisch für vegetationskundliche Untersuchungen. So wird bei Waldaufnahmen häufig unterschieden zwischen Krautschicht, Strauchschicht und Baumschicht. Solche Unterschiede lassen sich metrisch nicht treffend beschreiben. Hingegen kann das Spektrum der Merkmale durch die Entropiefunktion von SHANNON (1948) erfasst werden. Damit kommen die Methoden der Informationstheorie zur Anwendung. Ueber deren Grundzüge orientiert z. B. ein umfassender Anhang bei RENYI (1962). Die nachfolgenden Ausführungen halten sich an die Nomenklatur von PIELOU (1977) und ORLOCI (1978).

Gegeben seien also Stichproben (z.B. von einzelnen Arten), die in Klassen (z.B Artmächtigkeitsklassen) mit den relativen Häufigkeiten p1,p2,...,pn aufgeteilt werden. Die mittlere Entropie für die Aufnahme h ergibt sich dann als

$$H(P_h) = - \sum_j p_{hj} \ln p_{hj} \quad , \; j=1,\ldots,n.$$

Wie PIELOU (1977) ausführt, kann $H(P_h)$ als Informationsgehalt der Aufnahme h aufgefasst werden. Die Entropie eignet sich z.B. als Mass für die Diversität derselben. Genauer gefasst ist $H(P_h)$ ein Ausdruck für die Ungewissheit über die Zusammensetzung von h. Anhand von Tabelle 4.8 kann dies gezeigt werden. Zu diesem Zwecke ist die Formel für die Entropie so umzuschreiben, dass statt der relativen Häufigkeiten p_j die absoluten innerhalb der Klassen j, f_j verwendet werden können. Wir setzen

$$p_{hj} = f_{hj} / f_{h.}$$

p_{hj} ist die relative Häufigkeit für die Klasse j in Aufnahme h. $f_{h.}$ steht für die Summe der Häufigkeiten in allen Klassen, also für die Gesamtzahl der Individuen der Aufnahme h. In Tabelle 4.8 ist zum Beispiel $p_{12} = f_{12} / f_{2.} = 9/9 = 1.$

Tabelle 4.8

Beispiel zur Berechnung der Entropie. "0" bedeutet, dass die betreffende Art fehlt, "-" bedeutet, dass sie nicht untersucht wurde.

Aufnahme h / Art j	1	2	3
1	3	9	3
2	3	0	3
3	3	0	3
4	-	-	0
5	-	-	0
$f_{h.}$ =	9	9	9

Man erhält für die Entropie

$$H(P_h) = -\Sigma f_{hj}/f_{h.} \ln f_{hj}/f_{h.}$$

$$= \ln f_{h.} - 1/f_{h.} \Sigma f_{hj} \ln f_{hj}$$

$$= 1/f_{h.} (f_{h.} \ln f_{h.} - \Sigma f_{hj} \ln f_{hj})$$

Die Entropie ist ein relatives Mass (mit der Kolonnensumme $f_{h.}$ als Massstab). Für die weiteren Betrachtungen verwenden wir nun aber deren $f_{h.}$-faches:

$$I(F_h) = H(P_h)f_{h.} = f_{h.} \ln f_{h.} - \Sigma f_{hj} \ln f_{hj}$$

$I(F_h)$ bezeichnen wir als Informationsgehalt von h. Er eignet sich ebenfalls als Mass für die Diversität der Aufnahme h. Dessen Eigenschaften sollen anhand der drei Aufnahmen in Tab. 4.8 gezeigt werden:

1. Die Diversität wird maximal, wenn die Individuen der Aufnahme gleichmässig auf alle Arten verteilt sind. Aufnahme 1 zeigt diesen Fall: $I(F_1) = 9 \ln 9 - (3\ln 3 + 3\ln 3 + 3\ln 3) = 9.89$. Das Minimum wird erreicht, wenn alle Individuen einer einzigen Art angehören, nämlich $I(F_2) = 9 \ln 9 - (9\ln 9 + 0\ln 0 + 0\ln 0) = 0$. Erinnern wir uns an die Definition von $I(F_h)$ als Mass für die Ungewissheit über die Zusammensetzung einer Aufnahme: Wird bei Aufnahme 2 ein beliebiges Individuum aus der Gesamtstichprobe herausgegriffen, so kann mit absoluter Gewissheit gesagt werden, dass es der Art 1 angehört. Die Unsicherheit der Aussage ist gleich null. Im Falle der Aufnahme 1 jedoch ist die Wahrscheinlichkeit für eine Angehörigkeit zu jeder der Arten genau gleich, nämlich 1/3. Die Ungewissheit über die Artzugehörigkeit ist also maximal.

2. Zusätzliche Arten, welche keine Individuen enthalten, tragen nicht zur Diversität bei. In Tabelle 4.8 ist also $I(F_1) = I(F_3)$.

3. Welcher Natur die Merkmale sind, ist unerheblich. Artmächtigkeit, Lebensform oder Schichtzugehörigkeit sind

gleichermassen zulässig und erhalten gleiches Gewicht. Eine Transformation, wie sie für metrische Aehnlichkeitsmasse notwendig ist, entfällt.

4. Werden die Merkmale nach zwei verschiedenen Kriterien K und L in Klassen eingeteilt (z.B. in K, Arten und L, Schichtzugehörigkeit), so ist die gesamte Diversität gleich der Summe der Diversitäten K und L. In Tabelle 4.9, A, ist ein einschlägiger Fall konstruiert. Wir finden folgende Informationswerte:

$$I(F_{hK}) = 9\ln 9 - (3\ln 3 + 3\ln 3 + 3\ln 3)$$
$$= 19.77 - (3.3 + 3.3 + 3.3) = 9.88$$

$$I(F_{hL}) = 9\ln 9 - (3\ln 3 + 3\ln 3 + 3\ln 3) = 9.88$$

$$I(F_{hKL}) = 9\ln 9 - 9(1\ln 1) = 9\ln 9 = 19.77$$

Innerhalb von Rundungsfehlern bestätigt sich, dass die Diversitäten addiert werden dürfen:

$$I(F_{hKL}) = I(F_{hK}) + I(F_{hL}).$$

Im Falle, dass K und L teilweise abhängig sind, ist die Gesamtdiversität kleiner als die Diversitäten K plus L:

$$I(F_{hKL}) < I(F_{hK}) + I(F_{hL}).$$

Einen Fall vollständiger Abhängigkeit zeigt Tabelle 4.9, B. Darin gilt natürlich, dass die Diversität sowohl durch das Artspektrum wie auch durch die Schichtzugehörigkeit vollständig beschrieben werden kann:

$$I(F_{hKL}) = I(F_{hK}) = I(F_{hL}) = 9.88$$

Die bisher besprochenen Funktionen der Infomationstheorie können nun leicht verwendet werden, um Masse für den Zusammenhang von Vegetationsaufnahmen zu formulieren. Anhand des Beispieles in Tabelle 4.10 sollen verschiedene Möglichkeiten gezeigt werden. Ausgegangen wird von den beiden Aufnahmen

Tabelle 4.9

Diversitätsberechnungen. Beispiel vollständiger Unabhängigkeit von Merkmalen (A) sowie von vollständiger Abhängigkeit (B).

A

K= \ L=	Kraut-schicht	Strauch-schicht	Baum-schicht	f_{hK}
Art 1	1	1	1	3
Art 2	1	1	1	3
Art 3	1	1	1	3
f_{hL}	3	3	3	$f_{h.}=9$

B

K= \ L=	Kraut-schicht	Strauch-schicht	Baum-schicht	f_{hK}
Art 1	3	0	0	3
Art 2	0	3	0	3
Art 3	0	0	3	3
f_{hL}	3	3	3	$f_{h.}=9$

Tabelle 4.10

Berechnung des Zusammenhanges zwischen zwei Aufnahmen (A), Häufigkeit der Symbole (B) und Kontingenztafel (C).

A

Art	Aufn. 1	Aufn. 2
1	.	.
2	+	+
3	.	+
4	1	.
5	.	1
6	2	1
7	.	.
8	1	1
9	1	2
10	+	+

B

Aufn.	Klassen .	+	1	2	$f_{h.}$
1	4	2	3	1	10
2	3	3	3	1	10

C

Klassen in Aufn. 2	Klassen in Aufn. 1: .	+	1	2	F_2
.	2	0	1	0	$f_{21}=3$
+	1	2	0	0	$f_{22}=3$
1	1	0	1	1	$f_{23}=3$
2	0	0	1	0	$f_{24}=1$
F_1	$f_{11}=4$	$f_{12}=2$	$f_{13}=3$	$f_{14}=1$	$f_{1.}=f_{2.}=10$

in Tabelle 4.10, A.Darin sind die Artmächtigkeiten mit Hilfe von Symbolen notiert. Es liegen also Nominaldaten vor, die auch als solche behandelt werden sollen. Ohne auf Zusammenhänge zwischen den Aufnahmen einzugehen, lässt sich eine Statistik der auftretenden Klassen (Symbole) erstellen (Tabelle 4.10, B). Um nun Beziehungen darzustellen, ist eine Kontingenztafel zu konstruieren mit den Zelleninhalten $f_{1s,2t}$. Jede Zelle addiert die Anzahl Fälle, in welchen Aufnahme 1 bezüglich einer bestimmten Art das Symbol s, Aufnahme 2 das Symbol t enthält. Das leicht nachzuvollziehende Ergebnis zeigt Tabelle 4.10, C. Es ist offensichtlich, dass Aufnahme 1 und 2 teilweise abhängig sind. Schematisch ist dies in Abb. 4.4 dargestellt. Jeder Kreis entspricht dem Informationsgehalt einer Aufnahme. Es überlappen die gemeinsamen Informationsanteile, welche die Abhängigkeiten darstellen. Anhand von Tabelle 4.10, C und Abb. 4.4 lässt sich diese Abhängigkeit in folgender Weise fassen:

1. Information der Randverteilungen:

$$I(F_1) = f_{1.} \ln f_{1.} - \sum_t f_{1t} \ln f_{1t}$$

$$I(F_2) = f_{2.} \ln f_{2.} - \sum_j f_{2j} \ln f_{2j}$$

Die Summation erfolgt darin über alle Elemente der jeweiligen Randwerte. Die beiden Randverteilungen haben mit dem Zusammenhang der Aufnahmen nichts zu tun, sondern beschreiben deren Informationsgehalt (bezüglich der Häufigkeiten vorkommenden Symbole) separat. Dies ist in Abb. 4.4, A, dargestellt.

2. Gesamtinformation der Kontingenztafel:

$$I(F_1,F_2) = f_{1.,2.} \ln f_{1.,2.} - \sum_s \sum_t f_{1s,2t} \ln f_{1s,2t}$$

Hier erfolgt die Summation über alle Elemente innerhalb der Kontingenztafel. Grafisch ist der Fall in Abb. 4.4, B, dargestellt. Sind nun unsere beiden Aufnahmen unabhängig, so kann die Gesamtinformation auch aus den Randverteilungen berechnet werden:

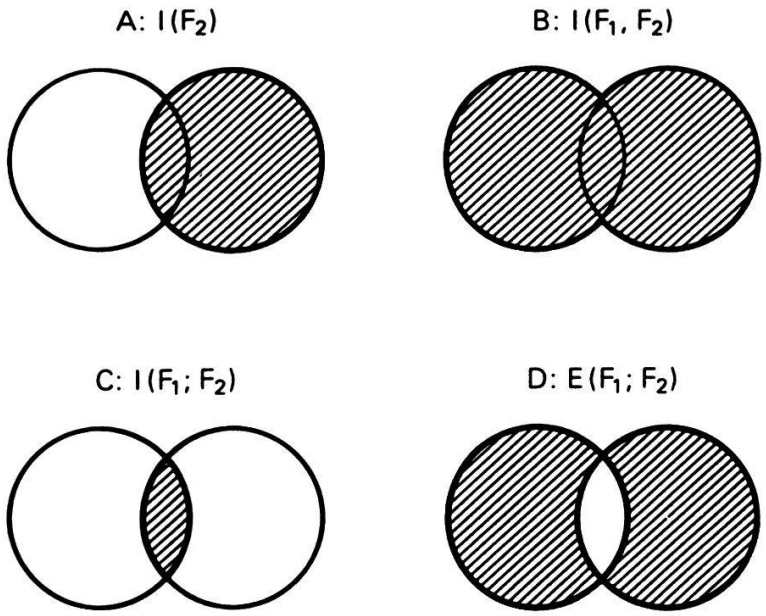

Abb. 4.4 Grafische Darstellung verschiedener Informationsmasse: Information einer Randverteilung (A), Gesamtinformation (B), Wechselseitige Information (C) und Spezifische Information (D).

$$I(F_1,F_2) = I(F_1) + I(F_2)$$

Grafisch dargestellt dürften sich dann die beiden Kreise in Abb. 4.4 nicht überschneiden. Sind hingegen beide Aufnahmen genau gleich, so müssen sich die Kreise decken. Die Information einer Randverteilung (einer Aufnahme) entspricht dann zugleich der Gesamtinformation:

$$I(F_1,F_2) = I(F_1) = I(F_2) \ .$$

Die Gesamtinformation ist also bereits ein Mass für die Uebereinstimmung der beiden Aufnahmen.

3. Wechselseitige Information:

$$I(F_1;F_2) = I(F_1) + I(F_2) - I(F_1,F_2)$$

Anhand der Abb. 4.4, C, ist leicht einzusehen, dass es sich hier um den beiden Aufnahmen gemeinsamen Teil der Information handelt. Bei vollständiger Unabhängigkeit ergibt sich natürlich der Grenzwert 0:

$$I(F_1;F_2)_{min} = I(F_1) + I(F_2) - I(F_1) - I(F_2) = 0$$

Bei Identität der Aufnahmen ist die wechselseitige Information gleich derjenigen einer einzelnen Aufnahme:

$$I(F_1;F_2)_{max} = I(F_1) + I(F_2) - I(F_1) = I(F_2) = I(F_1).$$

Damit qualifiziert sie sich als Aehnlichkeitsmass mit dem Bereich

$$0 \leqslant I(F_1;F_2) \leqslant I(F_1),\ I(F_2).$$

4. Spezifische Information

$$E(F_1;F_2) = I(F_1,F_2) - I(F_1;F_2)$$

Es handelt sich um die Information, welche jede Aufnahme für

sich allein trägt (Abb. 4.4, D). Bei Unabhängigkeit gilt, dass sie der Summe der Informationen beider Aufnahmen entspricht:

$$E(F_1;F_2)_{max} = I(F_1) + I(F_2).$$

Bei völliger Identität wird sie dagegen 0:

$$E(F_1;F_2)_{min} = I(F_1) - I(F_2) = 0.$$

Mit der spezifischen Information verfügt man somit über ein Unähnlichkeitsmass mit den Grenzwerten

$$I(F_1) + I(F_2) \geqslant E(F_1;F_2) \geqslant 0.$$

5. Rajskis Divergenzkoeffizient:

Um ein relatives Mass zu erhalten, schlug RAJSKI (1961) einen Koeffizienten vor, der den Anteil der spezifischen Information an der Gesamtinformation misst:

$$d(F_1;F_2) = \frac{E(F_1;F_2)}{I(F_1,F_2)} = 1 - \frac{I(F_1;F_2)}{I(F_1,F_2)}$$

Die Grenzwerte von 0 (Identität) und 1 (Unabhängigkeit) sind somit von der Gesamtinformation unbeeinflusst.

Auf das Beispiel in Tabelle 4.10 angewandt ergeben die hier besprochenen Masse fogende Werte:

1. Information der Randverteilungen:

$$I(F_1) = 10\ln10 - (4\ln4 + 2\ln2 + 3\ln3 + 1\ln1) = 12.798$$
$$I(F_2) = 10\ln10 - (3(3\ln3) + 1\ln1) = 13.138$$

2. Gesamtinformation:

$$I(F_1,F_2) = 10\ln10 - (2(2\ln2) + 6(1\ln1)) = 20.253$$

3. Wechselseitige Information:

$$I(F_1;F_2) = 12.798 + 13.138 - 20.253 = 5.683$$

4. Spezifische Information:

$$E(F_1;F_2) = 20.253 - 5.683 = 14.57$$

5. Rajskis Koeffizient:

$$d(F_1;F_2) = 14.57/20.253 = 0.719$$

Die hier gezeigten Informationsmasse behandeln die Daten so, als wären sie nominal. Beurteilt wird die Zufälligkeit des Zusammentreffens zweier Symbole wie "+" oder "1" der Skala Braun-Blanquet. Das ist ein ausserordentlich strenges Vergleichskriterium. Wie sich am Beispiel der Tabelle 4.10 leicht nachprüfen lässt, können Vegetationsaufnahmen als völlig zusammenhangslos eingestuft werden, wenn sie zwar lauter gemeinsame Arten, jedoch unterschiedliche Artmächtigkeiten aufweisen. Informationsmasse sollten eher angewandt werden, wenn mit echt qualitativen Kriterien gearbeitet wird, also Schichtzugehörigkeit, Wuchsform oder Phänologie, aber auch bei Präsenz- Absenz Daten.

Die Informationsmasse bieten eine beachtliche Auswahl von Möglichkeiten für die Beschreibung der Aehnlichkeit von Aufnahmen. Dabei fällt sofort ihre Verwandtschaft zu den Koeffizienten der Kontingenztafel auf (Abschnitt 4.7), wo die Felder a und d Gemeinsamkeiten, b und c Divergenzen repräsentieren. FEOLI et al. (1984) weisen denn auch nach, dass sich die meisten Informationsmasse im Falle von Präsenz-Absenz Daten auf das Chiquadrat oder ein Derivat davon reduzieren. Für Datensätze mit komplizierter Merkmalsstruktur (z.B. Arten und Lebensformen gemischt) bieten sich die Informationsmasse als ideale Lösung an. Beispiele für Anwendungen in grösserem Rahmen gibt es kaum und es ist in dieser Hinsicht noch einige Pionierarbeit zu leisten.