Clusteranalyse Kapitel 2

Die Clusteranalyse ist eine multivariate statistische Methode. Sie dient der Datenexploration und zeigt in Baumdiagrammen (Dendrogrammen) (Abb. 2.1) mögliche Datenstrukturen auf. Ihre Anwendbarkeit auf Zooplanktondaten soll in dieser Untersuchung anhand der Zooplanktonergebnisse der ZISCH Fahrt vom Winter 1987 erprobt werden.

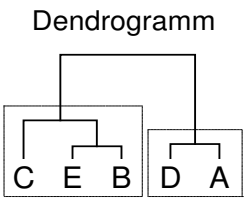


Abb.: 2.1

Stationscluster oder Artencluster 2.1

Grundsätzlich ist es möglich, sowohl Stationscluster als auch Artencluster zu erzeugen. Allerdings wären Artencluster für den vorliegenden Datensatz nicht sinnvoll. Bevor diese Aussage verdeutlicht wird, soll zuerst auf Stationscluster und ihre weitere Bearbeitung eingegangen werden.
Die Werte der Zooplanktonkonzentrationen werden in einer Datentabelle der folgenden Form (Tab. 2.1) erfasst:

	[Ind/m³]	**Variablen**					
		Station 1	Station 2	Station 3	Station 4	Station 5	→ **bis zu 136**
Fälle	Art 1	*21*	*0*	*0*	*71*	*3*	
	Art 2	*0*	*0*	*391*	*63*	*0*	
	Art 3	*1*	*5*	*880*	*0*	*0*	
	↑ **193**						

Tab.: 2.1 Datentabelle
(*Daten* nur zur Illustration)

Die Stationen (Variablen) werden mit Hilfe der Clusteranalyse in Gruppen (Cluster) ähnlicher Stationen zusammengefasst. Das bedeutet, dass die Stationen eines Clusters ein ähnliches Artenspektrum (Fälle) besitzen.

Nach der Ermittlung der Stationscluster wird pro Cluster für jede Art der Konzentrationsmittelwert **MW** über alle zum Cluster gehörenden Stationen berechnet (Tab. 2.2). Für diese Berechnung werden die Originalwerte (Ind./m^3) verwendet (d.h. nicht transformiert und nicht standardisiert).

[Ind/m^3]	**Cluster A MW**	Stat 11	Stat 2	Stat 103	Stat 44	**Cluster B MW**		**Cluster C MW**	
Art 1	***6***	*0*	*8*	*4*	*12*	***0***	*analog*	***90***	*analog*
Art 2	***0***	*0*	*0*	*0*	*0*	***4***	*"*	***42***	*"*
Art 3	***25***	*32*	*21*	*17*	*30*	***67***	*"*	***0***	*"*
usw.									

Tab.: 2.2 Zuordnung der Arten zu Clustern
(*Daten* zur Illustration)

Anhand der größten Konzentration (Tab. 2.2, grau) wird dann jede Art einem Cluster zugeordnet (Tab. 2.3, grau). Auf diese Weise werden Stationscluster indirekt auch zur Gruppenbildung der Arten verwendet.

Darüber hinaus wird berechnet in welchen Clustern (2.C, 3.C) die zweit- und drittgrößten Konzentrationen auftreten. So ergibt sich ein differenziertes Bild über die Verteilung der Arten im Seegebiet (vergleiche z. B. Tab. 3.1a,b).

Zooplankton	**1.Cluster**	**2.C**	**3.C**
Art 1	**C**	**A**	**B**
Art 2	**C**	**B**	**A**
Art 3	**B**	**A**	**C**
usw.			

Tab.: 2.3 Artengruppierung

Nun zu den Artenclustern. Im Vergleich zur Datentabelle 2.1 sind jetzt die Arten die Variablen (senkrecht) und die Stationen die Fälle (waagerecht).

Werden Arten direkt geclustert, dann tritt das folgende Problem auf: Es gibt Arten, die weit über die Nordsee verteilt vorkommen und andere, die nur in einem Gebiet zu finden sind. Weit verbreitete Arten müssten **daher in mehreren Clustern vorkommen**, aber mit der hier verwendeten Analyse kann jede Art nur einmal zugeteilt werden.

Wie verhält sich die Clusteranalyse unter diesen Bedingungen?
Dies kann mit Hilfe der Multidimensionalen Skalierung (MDS), einem weiteren multivariaten Verfahren, verdeutlicht werden. Ein zweidimensionaler MDS Graph (Abb. 2.2) projiziert die Variablen (jetzt Arten) auf eine Ebene, so dass sich ähnliche Variablen möglichst nah sind und unähnliche weit voneinander entfernt stehen.
Daher ist logisch betrachtet, in der MDS (Abb. 2.2) von den Arten folgende Anordnung zu erwarten: Weit verbreitete Arten werden sich in der Mitte ansammeln, da sie sich nur dort **mehreren Clustern gleichzeitig** zuordnen können und die restlichen Arten liegen strahlenförmig Drumherum. In jedem Strahl befinden sich einander ähnliche Arten. Die Strahlen berühren die häufigen Arten in der Mitte dort, wo sich jene häufigen Arten befinden, die jeweils zu den Strahlen (Artengemeinschaften) gehören.

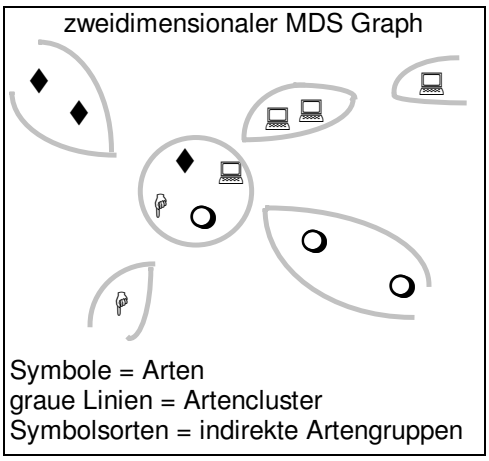


Abb.: 2.2

Artencluster

Wir führen eine Clusteranalyse durch, in der die Variablen die Arten sind und erhalten so Artencluster, die wir im MDS Graphen kennzeichnen. Logisch (was sich nah ist wird zusammengefasst) ist zu erwarten, dass diese Markierung den grauen Linien entspricht (Abb. 2.2):
Die häufigen Arten in der Mitte werden zu einem Cluster zusammengefasst. Die Strahlen (Artengemeinschaften) bilden je ein Cluster, manche Strahlen werden auch in zwei oder mehrere Cluster zerfallen.

Es ergibt sich eine Aufteilung ohne viel Aussagekraft. Das Cluster der häufigen Arten ist nutzlos, denn es enthält Arten aus dem gesamten Seegebiet. Auch dass die Strahlen in mehrere Cluster zerfallen können ist einer Interpretation eher hinderlich.

Indirekte Artengruppen

Zum Vergleich betrachten wir, wie die indirekt durch die Stationscluster entstandenen Artengruppen (siehe oben) sich im MDS Graph anordnen. Logisch (alle Arten haben einen Konzentrationsschwerpunkt und werden dementsprechend in Gruppen geteilt) ist ein Ergebnis zu erwarten, dass durch die Symbolsorten in Abb. 2.2 dargestellt wird.

Wichtig ist: Die Artengruppen sind im MDS (Abb. 2.2) gut voneinander getrennt, d.h. die Gruppen sind von einander verschieden und das bedeutet wiederum, dass eine zulässige Lösung erreicht wurde. Vorteilhaft ist, dass sich kein aussageloser Cluster von häufigen Arten bildet.

Fazit

Wenn wir die durch Symbolsorten markierten Artengruppen betrachten, dann stellen wir fest, dass die Abstände der Arten voneinander innerhalb der Artengruppen zum Teil sehr viel größer sind, als die Abstände zwischen den Arten verschiedener Gruppen (in der Mitte). Die Clusteranalyse aber macht das Gegenteil: Arten die nahe beieinander sind werden zusammengefasst.

Die Tatsache, dass der Datensatz sowohl aus sehr häufigen, als auch aus sehr seltenen Arten besteht macht die Clusterung von Arten sinnlos. Es wird daher nur mit Stationsclustern und indirekten Artengruppen gearbeitet.

Dateneigenschaften 2.2

Die vorliegenden Zooplankton Daten sind durch verschiedene Eigenschaften gekennzeichnet. Im letzten Abschnitt (Kapitel 2.2) wurde bereits abgehandelt:

- Die Arten weisen extrem unterschiedliche Häufigkeiten auf. Häufigkeit ist identisch mit der Eigenschaft Anzahl der Nullwerte (6 bis 98% der Werte einer Art).

Weiterhin gibt es zwei für Datensätze biologischer Gemeinschaften typische Eigenschaften:

- Ein sehr hohes Verhältnis von Variablen (136 Stationen) zu Fällen (193 Arten) und
- Ein insgesamt hohes Vorkommen von Nullwerten (bis zu 98% der Werte einer Art).

Multivariate Standardverfahren hingegen gehen davon aus, dass es fünf- bis zehnmal mehr Fälle als Variablen gibt, um zu verlässlichen Aussagen zu kommen (Held, pers. Kom.) und davon, dass die Variablen normalverteilt sind. (Clarke & Warwick, 1994) Die große Anzahl von Nullwerten macht eine Normalverteilung unmöglich. Diese Besonderheiten müssen in der Analyse berücksichtigt werden (z.B. Kapitel 2.6).

Zooplankton 2.2.1

In dieser Untersuchung gehen verschiedene taxonomische Kategorien (Fälle) zusammen in die Clusteranalysen ein.

In dieser Untersuchung wird nur mit Stationsclustern (siehe Kapitel 2.1) gearbeitet. Die Zooplanktonangaben beschreiben die Stationen. Es wäre auch möglich, Stationscluster auf der Basis von Umweltparametern wie z.B. Salzgehalt, Temperatur, Nährstoffe, usw. zu erstellen. Die unterschiedliche Einheiten sind kein Problem, wenn die Daten mittels geeigneter Transformation und/oder Standardisierung auf eine vergleichbare Skala gebracht werden.

Datengüte und -auswahl 2.3

Einige Zooplankter wurden nur an wenigen Stationen und in so geringer Anzahl gefunden, dass die Möglichkeit einer zufälligen Verteilung in Betrachtung gezogen werden muss. Das Problem besteht darin zu begründen, welche Daten verwendet werden und welche nicht. Zooplanktondaten können bis heute nur durch mühsames und zeitraubendes Auszählen von Hand gewonnen werden. Werden bei der Auswertung zu viele Daten verworfen, dann verringert sich der zur Verfügung stehende Informationsgehalt und das ist schlecht, denn es könnten entscheidende Daten sein, die leichtfertig weggelassen werden.

Im folgenden werden vier Ansätze für die Datenauswahl durchdacht:

2.3.1 Es werden alle Variablen verwendet.

2.3.2 Es werden solche Variablen verwendet, die an mindestens einer Station 4% des Zooplanktonbestandes ausmachen. (nach Clarke & Warwick, 1994)

2.3.3 Die Variablen werden nach der Anzahl ihrer Nullwerte geordnet. Alle Variablen, die mehr als eine bestimmte Anzahl von Nullwerten haben werden verworfen.

2.3.4 Für jede Art wird der Mittelwert MW↑0 (siehe **I**, unten) über all jene Stationen gebildet an denen überhaupt etwas gefunden wurde, wo also die Konzentration größer Null ist. Die Variablen werden entsprechend ihres MW↑0 geordnet. Variablen unterhalb eines Schwellenwertes von MW↑0 werden verworfen.

2.3.5 Kompletter Datensatz 2.3.1

Die Vorteile sind, eine leichte und eindeutige Entscheidung, sowie die Sicherheit, keine Daten zu verlieren. Der Nachteil ist, dass Arten, die zum Teil nur mit einem Organismus pro Unterprobe gezählt wurden berücksichtigt werden. Diese Organismen könnten aber rein zufällig und nicht repräsentativ für die untersuchte Station sein. Außerdem wirken sich Zählfehler und andere Fehlerquellen bei sehr niedrigen Individuenzahlen stärker aus.

Minimum 4% Datensatz 2.3.2

Hier werden nur die zahlenmäßig wichtigen Organismen verwendet. Es lässt sich aber weder für 4% noch für eine andere Prozentzahl eine klare Begründung finden. Problematisch ist weiterhin - da es im Zooplankton zum Teil Räuber-Beute Beziehungen gibt - dass der zahlenmäßig unterlegene Räuber möglicherweise unter die gewählte Prozentmarke fällt und verworfen wird und somit interspezifische Zusammenhänge nicht erkannt werden. Grundsätzlich ist zu sagen, dass die Erforschung der Verteilungen und Beziehungen der häufigsten Organismen auch ohne multivariate Verfahren seit vielen Jahrzehnten betrieben wird. Der Vorteil multivariate Verfahren besteht eben darin, eine Vielzahl von Variablen gleichzeitig zu betrachten, und so die Grenzen unserer Analysefähigkeit von den häufigen zu den seltenen Organismen auszudehnen. Dieser Vorteil wird bei diesem Ansatz zumindest teilweise wieder aufgegeben.

Minimaler Nullwert Datensatz 2.3.3

Es gibt einen guten Grund zumindest die Variablen, die nur ein oder zwei Werte größer Null besitzen zu verwerfen, wird das nicht getan, dann müssen die Auswirkungen solcher Fälle in der weiteren Analyse bedacht werden (siehe z.B. Kapitel 2.2, 2.4 und 2.6).
Allerdings: Unterteilt man das betrachtete Meeresgebiet in biologische Einheiten so ist es möglich, dass diese unterschiedliche Größenordnungen aufweisen. Ein Gebiet umfasst möglicherweise nur wenige Stationen, während in einer anderen Gegend viele Stationen eine biologische Einheit bilden. Organismen, die spezifisch für ein kleinräumiges Gebiet sind fallen weg, wenn Variablen mit sehr vielen Nullwerten verworfen werden. In der Analyse könnte das dazu führen, dass ein solches Gebiet nicht mehr als Einheit zu erkennen ist.

MW↑0 Datensatz 2.3.4

MW↑0 gibt an wie viele Individuen einer Art im Durchschnitt pro Station (kann umgerechnet werden in pro Probe) zum Zählen vorlagen (siehe **I**). Je kleiner MW↑0 ist, desto größer ist die Wahrscheinlichkeit von Zählfehlern und zufälligem Fang und desto stärker wirken sich Zählfehler und andere Fehlerquellen aus.

Diese Methode verwirft solche Organismen, die in den Proben mit nur wenigen Individuen repräsentiert waren. Es lässt sich anhand der Zählmethode (wurden Unterproben genommen, gibt es doppelte oder sogar dreifache Proben, usw.) abschätzen wie viele Organismen pro Probe und damit pro Station vorhanden sein müssen, um zu einem verlässlichen Ergebnis zu kommen. Dieser Wert kann dann als unterer Schwellenwert für MW↑0 verwendet werden.

$$\mathbf{I} \quad MW\uparrow 0_{Art} = \frac{\sum_{i=1}^{p\uparrow 0} \text{Konzentrationen}}{p\uparrow 0}$$

mit:
p↑0 = Anzahl der Stationen größer Null

Diese Methode kann allerdings nicht schichtweise angewendet werden, da die Probenschichten wie bereits erwähnt nicht mit den betrachteten interpolierten Schichten übereinstimmen. Das bedeutet, dass eine Art entweder in allen Schichten verwendet wird oder in keiner Schicht.

In dieser Arbeit wird ein fertiger Datensatz verwendet. Die Abschätzung des Schwellenwertes für MW↑0 fällt deshalb schwer.

<u>Fazit</u>

Aus den unter den einzelnen Punkten genannten Gründen und nach vielfachen Analyseversuchen wurde **Ansatz 2.3.1** gewählt - damit werden alle Daten verwendet. Dieser Ansatz macht es besonders notwendig während der weiteren Analyse die Auswirkungen von potentiell zufälligen oder stark fehlerbehafteten Arten, sowie die Auswirkungen von extrem vielen Nullstellen zu bedenken.

Transformation 2.4

Rohdaten biologischer Gemeinschaften erstrecken sich üblicherweise über eine logarithmische Skala, da sich Lebewesen bei ungünstigen Bedingungen kaum, bei günstigen Bedingungen dann aber exponentiell vermehren. Weiterhin unterscheiden sich Arten in einer Räuber-Beute Beziehung in ihren Individuenzahlen möglicherweise um mehrere Größenordnungen voneinander.

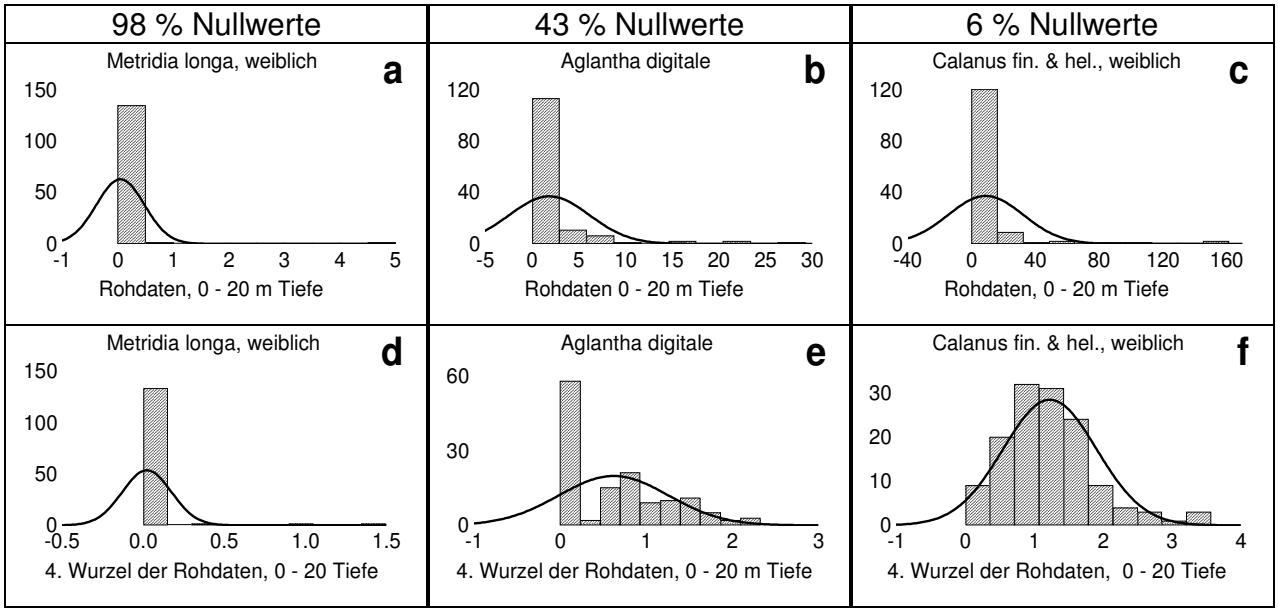


Abb.: 2.3 (Die x-Achse ist jeweils in 10 Säulen aufgeteilt.) Häufigkeiten mit Normalverteilungskurve

Eine Transformation (Abb. 2.3) mit der vierten Wurzel (Clarke & Warwick, 1994) normalisiert die Daten mit Ausnahme der vielen Nullwerte z.B. in Metridia longa und bringt die Daten (= x-Achse) auf eine ähnliche Skala (Größenordnung). Nicht normalverteilte Daten können bei Mittelwert Fusionsverfahren (Kapitel 2.7) das Ergebnis verfälschen.

Die Wurzeltransformation staucht die Daten. Das bedeutet das sich große Werte stärker verkleinern als kleine Werte (z.B. 16 wird zu 4, 4 zu 2; 16 ist viermal so gross wie 4 aber 4 nur noch zweimal so groß wie 2). Werte zwischen Null und Eins werden sogar größer (alle Werte gehen gegen Eins). Diese Stauchung tritt sowohl innerhalb der Arten auf, also in Richtung der Arten = waagerecht in der Datentabelle, als auch innerhalb der Stationen, also in Richtung der Stationen = senkrecht in der Datentabelle (Tab. 2.1).

Standardisierung 2.5

In der hier durchgeführten Analyse werden die Daten (nach der Transformation) in Richtung der Stationen (senkrecht, Tab. 2.1) standardisiert. Diese Standardisierung wird mit Wert ÷ ΣWerte durchgeführt.

Dies bedeutet anschaulich, dass jeder Wert (= 4.√ Konzentration einer Art) als Prozentzahl aller in der Station vorhandenen Arten dargestellt wird (Prozentzahl als Dezimalbruch von Eins, nicht multipliziert mit 100). Alle Stationen sind damit in der Summe gleich 1 und haben einen Mittelwert von 0.5.

Werden Stationen verglichen, dann geht es nicht mehr um die absoluten Konzentrationswerte sondern darum, ob das Verhältnis der Arten innerhalb einer Station dem Verhältnis der Arten in der anderen Station ähnlich ist. Weil die Prozentzahl eine Rangfolge ausdrückt, kann der Vergleich zweier Stationen auch als Vergleich der Rangfolge ihrer Arten interpretiert werden.

Es ist noch notwendig zu begründen, warum die Werte überhaupt standardisiert werden. Das ist eine Folge des gewählten Distanzmaßes und darauf wird in dem entsprechenden Absatz 2.6 (insbesondere Formel **VI**) eingegangen.

Distanzmaß 2.6

Die Clusteranalyse wird nicht von der Datentabelle (Tab. 2.1) aus durchgeführt. Statt dessen wird die (transformierte und standardisierte) Datentabelle in eine Unähnlichkeits- bzw. Ähnlichkeitsmatrix umgerechnet (Tab. 2.4).

	Stat 1	Stat 2	Stat 3	Stat 4	usw.
Stat 1		*5*	*23*	*1*	
Stat 2	***5***		*0*	*0*	
Stat 3	***23***	***0***		*70*	
Stat 4	***1***	***0***	***70***		
usw.					

Tab.: 2.4 Matrix

(*Daten* nur zur Illustration)

Es gibt verschiedene Methoden diese Berechnung durchzuführen. Was also ist diese Matrix? Jede Variable (Station) kann anschaulich als ein Punkt in einem vieldimensionalen Raum verstanden werden. Die Koordinaten eines Punktes werden durch die Fälle (Arten) beschrieben. Alle Variablen zusammen bilden also Punktewolken. Um über ihre Verteilung nachzudenken, dient hier die Vorstellung von Punktewolken in zweidimensionaler Ebene (Abb. 2.3) als Analogie zum vieldimensionalen Raum. Diese Vorstellung ist zulässig, so ist z.B. ein zweidimensionaler MDS-Graph (Abb. 2.2) eben das, die Projektion der vieldimensionalen Punktewolken auf eine zweidimensionale Ebene.

Anschaulich beschrieben enthält die Unähnlichkeitsmatrix die **Distanz** von jeder Variablen zu jeder anderen Variablen.

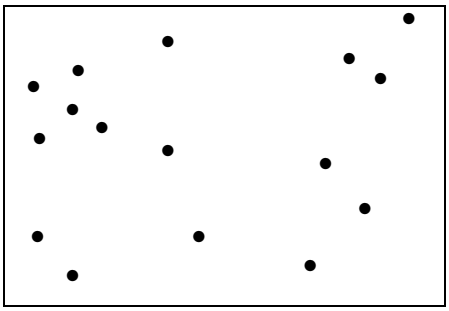


Abb.: 2.4 Variablen

Es werden im folgenden drei Möglichkeiten der Matrixberechnung, d.h. drei verschiedene Distanzmasse betrachtet:

2.6.1 Euklidische Distanz

2.6.2 Bray-Curtis Koeffizient

2.6.3 CityBlock Distanz

Dieses sind in der Literatur wohlbekannte Konzepte (z.B. Kaufman 1990; Clarke & Warwick 1994; Bacher 1996).

Euklidische Distanz 2.6.1

Die Euklidische Distanz ist die grundlegendste und anschaulichste aller Distanz Berechnungen. Sie ist unmittelbar verständlich als der Abstand zwischen zwei Punkten (Abb. 2.4) berechnet mit dem Pythagoras (siehe **II**).

Je größer der Abstand zwischen zwei Variablen, desto größer ist auch ihre Unähnlichkeit. Die mit der Euklidischen Distanz ED berechnete Matrix ist eine Unähnlichkeitsmatrix.

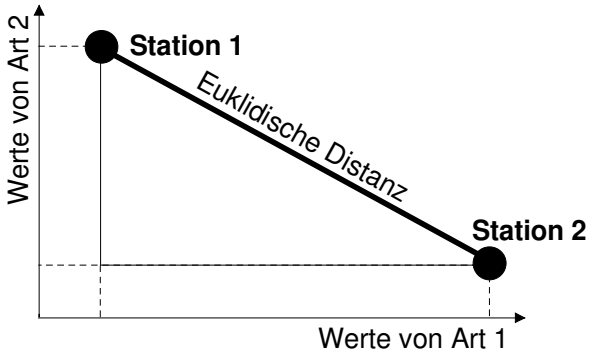


Abb.: 2.5 Euklidische Distanz

$$\mathbf{II} \quad ED_{Sta1zu2} = \sqrt[2]{\left(Art1_{Sta1} - Art1_{Sta2}\right)^2 + \left(Art2_{Sta1} - Art2_{Sta2}\right)^2}$$

Euklidische Distanz bezogen auf Abbildung 2.5

Beispiel: Haben zwei Variablen (Stationen) in allen Fällen (Arten) die gleichen Daten, dann liegen sie am gleichen Ort im Raum und der Abstand zwischen ihnen ist gleich Null. Es ist also ihre Unähnlichkeit = 0 und das ist natürlich sinnvoll, denn die Variablen sind identisch.

Da es im untersuchten Datensatz viele Nullwerte gibt (Kapitel 2.2), ist es wichtig zu beachten, was passiert, wenn zwei Variablen in einem Fall (bei einer Art) beide den Wert Null besitzen.

Beispiel (Tab. 2.5, Rech. 2.1): Die Euklidische Distanz hat im Fall A den Wert 2, im Fall B auch den Wert 2 und im Fall C den Wert 2.8.

$$ED_A = \sqrt[2]{(2-4)^2 + (2-2)^2} = 2$$

$$ED_B = \sqrt[2]{(2-4)^2 + (0-0)^2} = 2$$

$$ED_C = \sqrt[2]{(2-4)^2 + (2-4)^2} = 2.8$$

Rechnungen: 2.1 Tab. 2.5 in **II**

A	Station 1	Station 2
Art 1	2	4
Art 2	2	2

B	Station 1	Station 2
Art 1	2	4
Art 2	0	0

C	Station 3	Station 4
Art 1	2	4
Art 2	2	4

Tab.: 2.5 Beispiel

Im Vergleich von A mit B (siehe Art 2) wird deutlich, dass bei der Euklidischen Distanz gemeinsame identische Werte die gleiche Wirkung haben wie gemeinsame Nullwerte. Im Vergleich von B mit C (siehe wieder Art 2) wird deutlich, dass gemeinsame Nullwerte die Variablen ähnlicher machen, als unterschiedlich-große gemeinsame Werte.

Anschaulich erklärt: Die Euklidische Distanz ist ein absoluter (kein relativer) Wert, der sich für jede Variable aus der gleichen Anzahl von Fällen zusammensetzt. Besitzen zwei Variablen nun einen gemeinsamen Nullwert, dann ist das ein Fall weniger in dem Unterschiede auftreten können. Für den hier betrachteten Datensatz ist diese Eigenschaft „gemeinsame Nullwerte machen Variablen ähnlicher“ der Euklidischen Distanz nicht sinnvoll, weil z.B. eine Station an der Niederländischen Küste und eine Station in der Norwegischen Rinne sich nicht dadurch ähnlicher sind, dass ihnen beiden die gleichen Arten z.B. der Doggerbank fehlen. (nach Clarke & Warwick, 1994) Die Euklidische Distanz ist also für diese Analyse nicht geeignet.

Bray-Curtis Koeffizient 2.6.2

Der Bray-Curtis Koeffizient S_{jk} (siehe **III**) berechnet das Verhältnis von tatsächlich auftretender Ähnlichkeit $\sum_{i=1}^{p} \min(y_{ij}, y_{ik})$ zur maximal möglichen Ähnlichkeit $\sum_{i=1}^{p} \frac{1}{2}(y_{ij} + y_{ik})$.

Multipliziert mit 100 so ergibt sich daraus die Ähnlichkeit zweier Variablen als Prozentsatz (mit 100% = 100%ige Ähnlichkeit, also identisch). Die Berechnung nach Bray-Curtis **III** resultiert in einer Ähnlichkeitsmatrix.

$$\mathbf{III} \quad S_{jk} = 100\left(\frac{\sum_{i=1}^{p} \min(y_{ij}, y_{ik})}{\sum_{i=1}^{p} \frac{1}{2}(y_{ij} + y_{ik})}\right)$$

mit:

j, k = zwei Variablen
p = Anzahl Fälle
y = Wert aus Datentabelle

(nach Clarke & Warwick, 1994)

Die Formel **III** lässt sich für das einfache Beispiel aus Tab. 2.6 graphisch als Verhältnis zweier Strecken darstellen (Abb. 2.6). Der Mittelwert MW aus x und y (im Nenner) ist die maximal mögliche Ähnlichkeit, sie wird erreicht, wenn x und y gleich lang sind. Sind x und y verschieden, dann steht die kürzere Strecke (im Zähler) für die tatsächlich realisierte Ähnlichkeit.

	Station 1	Station 2
Art 1	x	y

Tab.:2.6 Beispiel

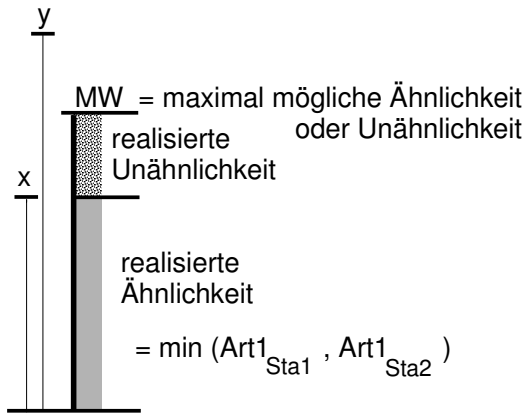


Abb.: 2.6 S_{jk} nach **III** und Tab. 2.6

Es lässt sich **III** in **IV** umformen (Clarke & Warwick, 1994).

Wird bei **IV** die 100 weggelassen, ergibt sich S_{jk} als Dezimalbruch (ohne Abbildung).

Weiterhin gilt: Ähnlichkeit + Unähnlichkeit = 1 oder

$S_{jk} + S^{|}_{jk} = 1$.

Daraus folgt: Unähnlichkeit = 1 - Ähnlichkeit oder

$S^{|}_{jk} = 1 - S_{jk}$ und es entsteht dann **V**.

$$\mathbf{IV}\quad S_{jk} = 100\left(1 - \frac{\sum_{i=1}^{p}\left|y_{ij} - y_{ik}\right|}{\sum_{i=1}^{p}\left(y_{ij} + y_{ik}\right)}\right) \qquad \mathbf{V}\quad S^{|}_{jk} = \frac{\sum_{i=1}^{p}\left|y_{ij} - y_{ik}\right|}{\sum_{i=1}^{p}\left(y_{ij} + y_{ik}\right)}$$

(Clarke & Warwick, 1994)

$S^{|}_{jk}$ (**V**) ist ein Unähnlichkeitsmaß, daraus entsteht eine Unähnlichkeitsmatrix.

Wie verhält sich Bray-Curtis, wenn zwei Stationen bei der gleich Art einen Nullwert besitzen? (Bei der Euklidischen Distanz trugen gemeinsame Nullwerte zur Ähnlichkeit bei - siehe oben.)

Das Beispiel aus Tabelle 2.5 lässt sich auch für den Bray-Curtis Koeffizienten **IV** S_{jk} durchrechnen (Rech. 2.2).

Im Gegensatz zur (absoluten) Euklidischen Distanz gehen aber gemeinsame Nullwerte der Variablen nicht in die Berechnung ein, denn Bray-Curtis ist ein relationales Distanzmaß, d.h. es wird ein Verhältnis von tatsächlicher relativ zur maximalen Ähnlichkeit errechnet (Rech. 2.2, $S_{jkB} = S_{jkC}$).

$$S_{jkA} = 100\left(1 - \frac{|2-4| + |2-2|}{(2+4) + (2+2)}\right) = 80\%$$

$$S_{jkB} = 100\left(1 - \frac{|2-4| + |0-0|}{(2+4) + (0+0)}\right) = 66.\bar{6}\%$$

$$S_{jkC} = 100\left(1 - \frac{|2-4| + |2-4|}{(2+4) + (2+4)}\right) = 66.\bar{6}\%$$

Rechnungen: 2.2 Tab.2.5 in **III**

Da gemeinsame Nullwerte beim Bray-Curtis Koeffizienten nicht als Ähnlichkeit berechnet werden, bietet sich das Distanzmaß für diese Untersuchung an. Bray-Curtis ist aber nicht im Statistika Computer Programm enthalten!

Werden jedoch standardisierte Daten verwendet, dann geht der Bray-Curtis Koeffizient $S^{|}_{jk}$ in die CityBlock Distanz über und die CityBlock Distanz ist in Statistika enthalten.
Nach der Standardisierung (Kapitel 2.5) besitzt jede Variable die Summe Eins. Im Nenner von **V** haben wir die Summe von zwei Variablen und der Nenner ist gleich zwei (**VI**).

$$\mathbf{VI} \quad \sum_{i=1}^{p}\left(y_{ij} + y_{ik}\right) = 2$$

Aus **V** und **VI** folgt **VII**:

$$\mathbf{VII} \quad S^{*}_{jk} = \frac{\sum_{i=1}^{p}\left|y_{ij} - y_{ik}\right|}{2}$$

Weitere Erläuterungen zu **VII** im nächsten Absatz über die CityBlock Distanz.

CityBlock Distanz 2.6.3

Die CityBlock Distanz **VIII** (Abb. 2.7) ist der Euklidischen Distanz (Abb. 2.5) verwandt. Allgemein ist die CityBlock Distanz definiert:

$$\mathbf{VIII} \quad CB_{jk} = \sum_{i=1}^{p}\left|y_{ij} - y_{ik}\right|$$

mit: j, k = zwei Variablen
p = Anzahl Fälle
y = Wert aus Datentabelle

(Clarke & Warwick, 1994)

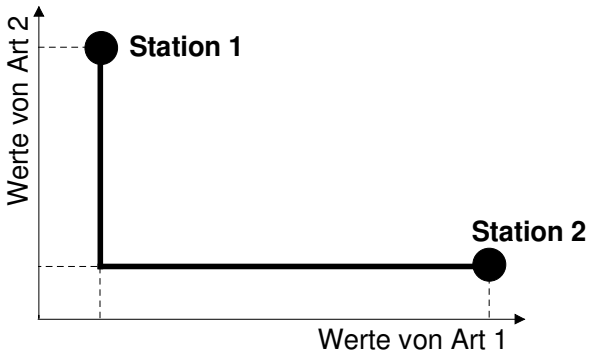


Abb.:2.7 CityBlock Distanz

S^{*}_{jk} aus **VII** unterscheidet sich von CB_{jk} **VIII** nur um den Faktor 0.5. Dieser Faktor betrifft jeden Wert in der Unähnlichkeitsmatrix gleichermaßen und verändert das Ergebnis der Clusteranalyse folglich nicht. Zusammenfassend (**IV**, **VI** und **VIII**) können wir (nur für standardisierte Werte) schreiben:

$$\mathbf{IX} \quad S_{jk} = 100\left(1 - \frac{CB_{jk}}{2}\right) \qquad \text{oder} \qquad \mathbf{X} \quad CB_{jk} = 2\left(1 - \frac{S_{jk}}{100}\right)$$

Fusionsverfahren 2.7

Die Unähnlichkeitsmatrix (vorheriger Abschnitt) gibt uns einen Wert für die Unähnlichkeit von jeder Variablen (Station) zu jeder anderen Variablen (Station). In der nun folgenden Fusion der Stationen werden zuerst die zwei Stationen fusioniert, die den kleinsten Wert (Abstand, Unähnlichkeit, Fusionsdistanz) in der Unähnlichkeitsmatrix besitzen. Für dieses Cluster aus zwei Stationen muss nun wiederum der Abstand zu allen noch verbliebenen Stationen berechnet werden. Wie diese Berechnung durchgeführt wird hängt von dem gewählten Fusionsverfahren ab.

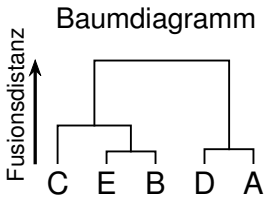


Abb.: 2.8

Wurde die Berechnung der Abstände durchgeführt, dann kann die nächste Fusion stattfinden. Danach werden wieder die Abstände neu berechnet. Dieser Prozess schreitet solange fort, bis alle Stationen zu einem einzigen Cluster verbunden sind.

Werden zwei Cluster fusioniert, so ist die Distanz zwischen ihnen die Fusionsdistanz des neu entstehenden Clusters. Die Verbindung der Cluster wird durch einen Querstrich (Abb. 2.8) in Höhe der Fusionsdistanz eingetragen. So ergibt sich ein Baumdiagramm (Dendrogramm), dass die Reihenfolge der Fusionen abbildet.

Beschreibung der Fusionsverfahren 2.7.1

Es wurden für diese Arbeit vier in der Literatur (z.B. Kaufman 1990; Clarke & Warwick 1994; Bacher 1996) wohlbekannte Fusionsverfahren auf ihre Nützlichlichkeit für diese Arbeit hin durchdacht. Zuerst werden die vier Verfahren kurz vorgestellt.

2.7.1.1 Nächster-Nachbar Verfahren (single linkage)

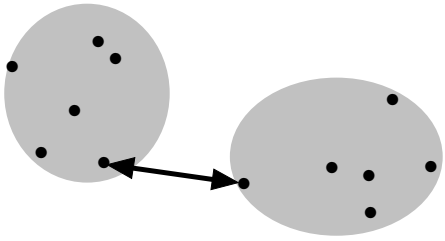
2.7.1.2 Weitester-Nachbar Verfahren (complete linkage)

2.7.1.3 Verfahren ungewichteter paarweiser Ähnlichkeiten

2.7.1.4 Verfahren gewichteter paarweiser Ähnlichkeiten

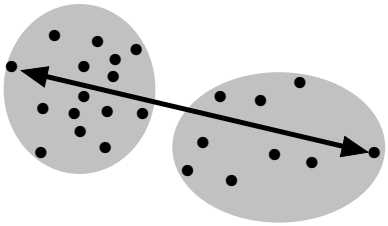
Nächster-Nachbar Verfahren 2.7.1.1

Bei dem Nächster-Nachbar Verfahren (Abb 2.9) ist der Abstand zwischen zwei Clustern gleich dem kleinsten Abstand zwischen ihren Variablen.
Diese kleinste Unähnlichkeit zwischen zwei Clustern ist gleich der Fusionsdistanz des daraus entstehenden neuen Clusters.



Pfeil = die nächsten Nachbarn

Abb.: 2.9 Nächster-Nachbar Verfahren



Pfeil = die weitesten Nachbarn

Abb.: 2.10 Weitester-Nachbar-Verfahren

Weitester-Nachbar Verfahren 2.7.1.2

Bei dem Weitester-Nachbar-Verfahren (Abb 2.10) ergibt sich die Distanz zwischen zwei Clustern aus der größten Unähnlichkeit zwischen ihren Variablen. Diese Fusionsdistanz des neuen Clusters (= „Durchmesser“) sagt etwas über die Homogenität des neuen Clusters aus, denn wir wissen, dass sich alle darin enthaltenen Variablen mindestens so ähnlich sind wie die weitesten Nachbarn sich ähnlich sind.

Verfahren der ungewichteten paarweisen Ähnlichkeiten 2.7.1.3

Im Verfahren der ungewichteten paarweisen Ähnlichkeiten werden im neuen Fusionscluster für alle Variablenpaare die Ähnlichkeiten berechnet und der Mittelwert daraus gebildet.

Verfahren der gewichteten paarweisen Ähnlichkeiten 2.7.1.4

Die Berechnung der gewichteten paarweisen Ähnlichkeit basiert auf den Fusionsähnlichkeiten der beiden beteiligten Cluster. Jede Fusionsähnlichkeit, wird mit der Anzahl der im zugehörigen Cluster vorhandenen Variablen multipliziert, beide Ergebnisse werden addiert und dann alles durch die Gesamtanzahl der Variablen im neu entstehenden Cluster geteilt.

Ziel der Fusionsverfahren 2.7.2

Allgemein lassen sich für Cluster zwei Arbeitsvorschriften oder Ziele definieren:

A **Die Variablen im Cluster sind untereinander ähnlich.**
(IntraCluster-Heterogenität = minimal)

B **Die Variablen im Clusters sind den Variablen außerhalb des Clusters unähnlich.** (InterCluster-Heterogenität = maximal)

Beide Ziele sollen durch die Fusionsverfahren optimiert werden. Der Erfolg der Clusteranalyse muss sich an diesen beiden Bedingungen **A** und **B** messen lassen.

Eigenschaften der Fusionsverfahren 2.7.3

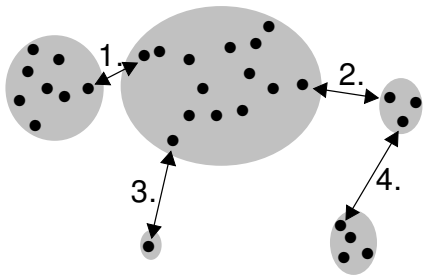
Jedes Fusionsverfahren hat bestimmte Vor- und Nachteile. Auch diese Eigenschaften werden in der Literatur (z.B. Kaufman 1990; Clarke & Warwick 1994; Bacher 1996) beschrieben. Für diese Arbeit musste eine Entscheidung getroffen werden, welches Fusionsverfahren ausgewählt werden soll. Deshalb wurden die wichtigsten Eigenschaften der einzelnen Verfahren durchdacht und werden im folgenden dargestellt.

Nächster-Nachbar Verfahren 2.7.3.1

Im Nächster-Nachbar Verfahren treten z.B. die Probleme der Verkettung, Brückenbildung und Formvariabilität auf.

Verkettung: Es entstehen im Dendrogramm wenige Clusterkerne, die dann in einer langen Kette alle einzeln nach und nach zusammen mit noch unfusionierten Variablen an den größten Clusterkern angehängt werden (Abb. 2.12).

- **Überlegung:** Cluster, die bereits groß sind haben eine große Ausdehnung und es ist wahrscheinlich, dass sie anderen Clustern oder Variablen am nächsten sind und mit ihnen verbunden werden. Dem gegenüber ist die Wahrscheinlichkeit von kleinen Clustern geringer sich mit weiteren kleinen Clustern oder Variablen zu verbinden, bevor sie mit dem größten Clusterkern fusioniert werden.



Zahlen = Fusionsreihenfolge
Abb.: 2.12 Verkettung

Brückenbildung: (Abb. 2.13): Es kann auch passieren das zwei eigentlich gut getrennte Cluster durch einige Variablen (= Brücke) zwischen ihnen zu „früh“ zu einem Cluster verbunden werden.

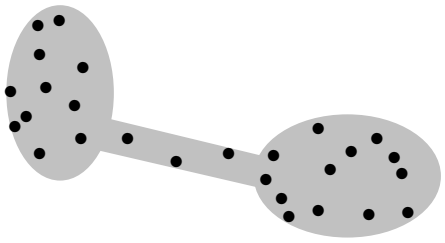


Abb.:2.13 Brückenbildung

Formbeliebigkeit: Dieses Verfahren stellt keine Ansprüche an die Form der Cluster. Sie können kugelig oder lang und dünn sein, sich sogar umeinander winden.

Das kann z.B. folgende Konsequenz haben (Abb. 2.14):
Die zu verschiedenen Clustern gehörenden Variablen {a, a^{I}, a^{II}} sowie {b, b^{I}, b^{II}} sind sich jeweils näher (ähnlicher), als die zu demselben Cluster gehörenden Variablen {a, b}, {a^{I}, b^{I}} und {a^{II}, b^{II}}.

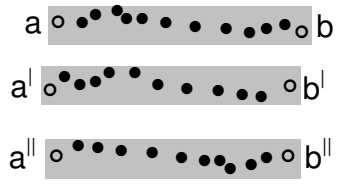


Abb.: 2.14 Intra- & Inter-heterogenität

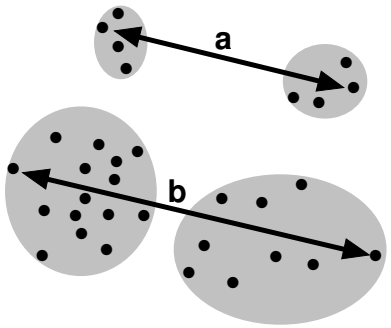
Das ist nach den im Kapitel 2.7.2 formulierten Zielen keine gute Lösung

Weitester-Nachbar Verfahren 2.7.3.2

In diesem Verfahren sind z.B. Größenabhängigkeit und Kugelform problematisch.

Größenabhängigkeit: Kleine Cluster (Abb. 2.15) haben eine höhere Wahrscheinlichkeit sich miteinander zu verbinden als große Cluster. Unter Umständen selbst dann, wenn die kleinen Cluster weiter voneinander entfernt sind als die großen Cluster.

- **Grund:** Die Distanz hängt ausschließlich von den zwei extremsten Variablen ab.



Größenabhängigkeit: **Distanz a < b**

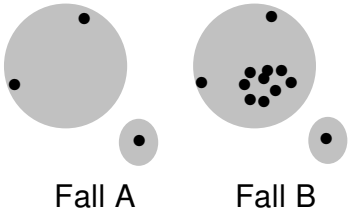
Abb.: 2.15 Weitester-Nachbar Verfahren

Formstarrheit: Es entstehen streng kugelige Cluster, da der „Durchmesser“ minimiert wird. Es ist fraglich, ob der Datensatz einer solch starren Clusterbedingung entspricht.

Verfahren der ungewichteten paarweisen Unähnlichkeit 2.7.3.3

In diesem Verfahren sind die Verteilungs- und Größenabhängigkeit problematisch, die gute Interpretierbarkeit der Fusionsdistanz hingegen ist positiv (Kapitel 2.7.2., Bed. **A**).

Verteilungsabhängigkeit: Bei Fall A und B (Abb. 2.16) gibt es je zwei paarweise Unähnlichkeiten die in beiden Fällen identisch sind. Zusätzlich gibt es im Fall B weitere paarweise Unähnlichkeiten, die alle kleiner sind als die eben genannten. Damit besitzt der Cluster im Fall B eine kleinere paarweise Unähnlichkeit als der Cluster im Fall A. Der Cluster im Fall B wird sich also im Fusionsablauf früher bilden als der Cluster im Fall A.

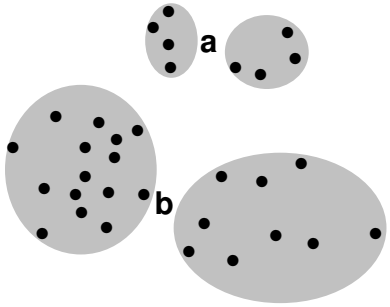


(kleine Distanz = kleine Unähnlichkeit)

Abb.: 2.16 Verteilungsabhängigkeit

Dieses Beispiel illustriert, dass eine ungleiche Verteilungsdichte (im Vergleich von einem Cluster zum anderen) zu Verzerrungen bei der Fusionierung führen kann. Eine solche ungleiche Verteilungsdichte ist bei den Eigenschaften des vorliegenden Datensatzes (Kapitel 2.2) zu erwarten.

Größenabhängigkeit: Bei ähnlicher Verteilungsdichte haben kleine Cluster (Abb. 2.17) eine höhere Wahrscheinlichkeit sich miteinander zu verbinden als große Cluster. Allerdings ist diese Eigenschaft nicht so stark ausgeprägt wie im Nächster-Nachbar Verfahren. Vergleiche dazu Abbildung 2.15:



(kleine Distanz = kleine Unähnlichkeit)

Abb.: 2.17 Größenabhängigkeit

Würden wir dort das Verfahren der ungewichteten paarweisen Unähnlichkeit anwenden, dann könnten wir nicht mehr mit Sicherheit folgern, dass das Clusterpaar **b** sich unähnlicher ist als Clusterpaar **a**. In der Abbildung 2.17 können wir diese Folgerung nur ziehen, weil das Clusterpaar **b** genauso nahe beieinander liegt, wie das Clusterpaar **a**.

Die **Fusionsdistanz** gibt bei diesem Verfahren den mittleren Abstand zwischen den zu fusionierenden Clustern an und damit die durchschnittliche IntraCluster-Homogenität .

Verfahren der gewichteten paarweisen Unähnlichkeit 2.7.3.4

Anstatt die einzelnen Variablen Schritt für Schritt zu fusionieren, können wir die Clusteranalyse auch in umgekehrter Reihenfolge als schrittweise Aufsplittung betrachten. Zur Gewichtung in Schritt 2., Aufsplittung oder Fusion der Cluster CEB und DA (siehe Kapitel 2.7.1.4; Abb. 2.11 und 2.18) wurde die folgende Überlegung gemacht:

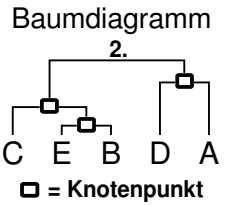


Abb.: 2.18

Bei jedem Aufteilungsschritt wird das Gewicht gleich verteilt. Es ergibt sich:

DAzuCEB = ((DzuC + (DzuE + DzuB)/2)/2 + (AzuC + (AzuE + AzuB)/2)/2)/2

Diese Gewichtung ergibt sich auch rechnerisch:

DAzuCEB = ((DzuC + 0.5 DzuE + 0.5 DzuB)/2 + (AzuC + 0.5 AzuE + 0.5 AzuB)/2)/2
= (0.5 DzuC + 0.25 DzuE + 0.25 DzuB + 0.5 AzuC + 0.25 AzuE + 0.25 AzuB)/2
= 0.25 DzuC + 0.125 DzuE + 0.125 DzuB + 0.25 AzuC + 0.125 AzuE + 0.125 AzuB

Rechnungen 2.3

Wie bei der ungewichteten paarweisen Unähnlichkeit so werden auch bei der gewichteten paarweisen Unähnlichkeit alle Variablenpaare zwischen den beiden beteiligten Clustern zur Berechnung herangezogen. Die Unähnlichkeiten können direkt aus der Unähnlichkeitsmatrix abgelesen werden. Ihr Gewicht ergibt sich nach der Formel **XI**:

$$\text{XI} \quad \text{Gewicht} = \frac{1}{2^k}$$

mit k = Anzahl der Knotenpunkte zwischen den Variablen (siehe Abb. 2.18)

Je größer k desto kleiner das Gewicht. Das bedeutet auch: je weiter eine Variable vom betrachteten Fusionsschritt entfernt liegt, desto geringer ist ihr Einfluss.

Welche Auswirkung kann die Gewichtung haben? Dazu im Folgenden ein Beispiel (siehe Abb. 2.16, 2.19 für Beispiel und Kapitel 2.7.1.3 und 2.7.1.4 für Methoden).

Für die Fusionsdistanz der **ungewichteten** paarweisen Unähnlichkeit gilt (Abb. 2.19):

Fall A $ab_{zu}N = (a_{zu}N + b_{zu}N)/2 = 0.5\ a_{zu}N + 0.5\ b_{zu}N$

Fall B $ab9V_{zu}N = (a_{zu}N + b_{zu}N + \sum_{i=1}^{9} V_{i\,zu}N)/11 = 0.09\ a_{zu}N + 0.09\ b_{zu}N + \underline{\mathbf{0.82}} \sum_{i=1}^{9} V_{i\,zu}N$

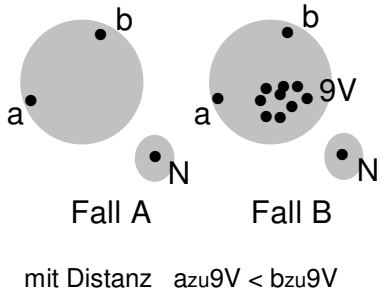
Rechnungen 2.4

Für die Fusionsdistanz der **gewichteten** paarweisen Unähnlichkeit gilt (Abb. 2.19 & 2.20):

Fall A $ab_{zu}N = 0.5\ a_{zu}N + 0.5\ b_{zu}N$

Fall B $ab_{zu}N = \frac{1}{2^2} a_{zu}N + \frac{1}{2^1} b_{zu}N + \frac{1}{2^2}\{9V\} = 0.25\ a_{zu}N + 0.5\ b_{zu}N + \underline{\mathbf{0.25}}\ \{9V\}$

Rechnungen 2.5



mit Distanz $a_{zu}9V < b_{zu}9V$

Abb.: 2.19 Verteilungsabhängigkeit

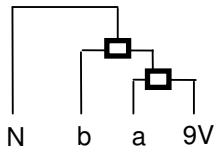


Abb.: 2.20 Schema Fall B

Positiv:

Im Vergleich zur ungewichteten paarweisen Unähnlichkeit wird auf diese Weise der verzerrende Einfluss der **Verteilungsabhängigkeit** verringert. Das Gewicht der 9 Variablen beträgt bei der gewichteten paarweisen Unähnlichkeit nur 25 statt 82%. (Rech. 2.4 & 2.5)

Negativ:

Die **Fusionsdistanz** ist bei diesem Verfahren der gewichteten paarweisen Unähnlichkeit (im Gegensatz zur ungewichteten paarweisen Unähnlichkeit) in Bezug auf die in Kapitel 2.7.2 formulierten Ziele nicht interpretierbar.

Auswahl eines Fusionsverfahrens 2.7.4

Die beiden Nächster-Nachbar und Weitester-Nachbar Verfahren sind mit vielen Unwägbarkeiten und Problemen (siehe Kapitel 2.7.3.1 und 2.7.3.2) behaftet.

Es bleiben also die beiden Mittelwertverfahren. Es konnte aber keine schlüssige Begründung gefunden werden, die eine Entscheidung zwischen dem Verfahren der ungewichteten- oder gewichteten paarweisen Unähnlichkeit erlaubt hätte. Statt dessen ist aus den Überlegungen (Kapitel 2.7.3.3 und 2.7.3.4) hervorgegangen, das sich beide Verfahren in ihren Eigenschaften gut ergänzen.
Ungelöst ist auch der folgende Sachverhalt: In all den verschiedenen Überlegungen zu den Eigenschaften der Fusionsverfahren konnte keine Information über die Erfüllung des Ziels **B**: Die InterCluster-Heterogenität soll maximal sein, gefunden werden. (Kap. 2.7.2)

Aus den eben genannten Gründen heraus wurde ein anderer, ein neuer Ansatz gesucht. Im Rahmen dieser Diplomarbeit wurde die im folgenden beschriebene Lösung erarbeitet.

Ansatz für mehrere Fusionsverfahren 2.7.5

Doch zuerst zum Vergegenwärtigen hier die beiden in Kapitel 2.7.2 genannten Ziele:

A **Die Variablen im Cluster sind untereinander ähnlich.**
(IntraCluster-Heterogenität = minimal)

B **Die Variablen eines Clusters sind den Variablen außerhalb des Clusters unähnlich.** (InterCluster-Heterogenität = maximal)

Daraus ergibt sich folgender Ansatz:

> **Je besser ein Cluster beide Bedingungen A und B erfüllt, desto klarer grenzt sich das Cluster von der Umgebung ab.**

→ Ein Cluster das sich sehr gut von der Umgebung abgrenzt ist für die jeweiligen Nachteile der verschiedenen Fusionsverfahren unempfindlich und wird daher mit jedem Verfahren zu erkennen sein.
Es kann gesagt werden: Die Beständigkeit des Clusters ist groß.

→ Je schlechter sich ein Cluster von der Umgebung abgrenzt, desto anfälliger ist das Cluster für die Nachteile der Fusionsverfahren und desto weniger Verfahren werden das Cluster darstellen können (zum Ergebnis haben).
Es kann gesagt werden: Je weniger Verfahren ein Cluster darstellen können, desto kleiner ist die Beständigkeit eines solchen Clusters.

Weiterhin kann folgender Ansatz 2 formuliert werden:

> **Die Beständigkeit eines Clusters beruht auf dem Verhältnis von IntraCluster-Homogenität zu InterCluster-Homogenität.**

Unterschiede der Lösungsansätze 2.7.6

Ein ganz praktischer Unterschied ergibt sich in der Behandlung der Baumdiagramme.

Wird nur ein Fusionsverfahren ausgewählt, dann ist es notwendig (Christiansen, pers. Kom.) eine Fusionsdistanz auszuwählen und das Baumdiagramm auf der entsprechenden Höhe querzuschneiden (Abb. 2.21, z.B. bei).

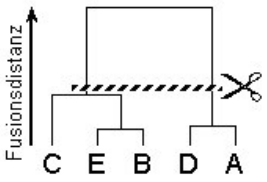


Abb.: 2.21

Bei der hier verwendeten Methode mit mehreren Fusionsverfahren (z.B. Abb. 3.2, 3.3) können die gefundenen identischen Cluster unterschiedliche Fusionsdistanzen aufweisen, also unterschiedliche IntraCluster-Homogenitäten (s.o.) besitzen.

In diesem Fall ist theoretisch zu erwarten, dass sich diese Cluster auch in ihrer InterCluster-Heterogenität (z.B. definiert als mittlerer Abstand zu den nächsten zwei Clustern) unterscheiden, so dass die Beständigkeit der Cluster gleich oder zumindest sehr ähnlich ist. Abbildung 2.22 zeigt ein Modell dieses postulierten Sachverhaltes.

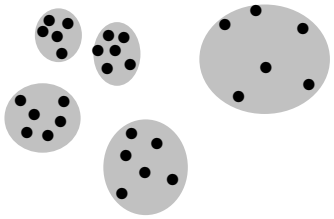


Abb.: 2.22 Beständigkeit

Das bedeutet das Intra- und InterCluster-Homogenität miteinander gekoppelt sind!

Eine Analyse der Daten zeigt verschiedene mögliche Gründe für diesen Umstand auf:

1) Die Anzahl der sehr seltenen Arten (mit nur ein oder zwei Werten größer Null) variiert von Stationscluster zu Stationscluster.

Beispiel: Die Stationscluster mit Atlantischem Einfluss besitzen viele in der Nordsee sehr seltene Arten. Das erhöht die InterCluster-Heterogenität dieser Cluster.

Diese Arten sind jedoch sogar so selten, dass sie häufig nur an einer oder zwei Stationen im Cluster vorkommen und so verkleinern sie gleichzeitig die IntraCluster-Homogenität. (Abb. 3.2, 3.3, 3.4)

2) Die Wassertiefe ist für die Stationscluster verschieden. Es wurden aber maximal fünf Proben pro Station genommen (Kapitel 1.3.4). Für je 20 m Wassertiefe ungefähr eine Probe. Die Proben wurden gleichmäßig über die Wassersäule verteilt. Das bedeutet, dass ab einer Wassertiefe größer als 100 m die Auflösung der Daten abnimmt. Das macht die Daten heterogener.

Beispiel: Die Stationscluster mit Atlantischem Einfluss befinden sich in einer viel tieferen Wassersäule (Abb. 3.1, 3.4) als die restlichen Cluster, speziell die Cluster der Doggerbank. (Abb. 3.2, 3.3, 3.4)

3) Die Cluster bestehen aus einer unterschiedlichen Anzahl von Stationen. Je mehr Stationen einem Cluster angehören, desto wahrscheinlicher ist es, dass die so zusammengefassten Daten heterogener sind als in kleineren Clustern.

Beispiel: Cluster 6 in 20 bis 40 m Tiefe (Abb. 3.6, 3.7 und 3.8)

4) Es ist biologisch betrachtet nicht sinnvoll zu erwarten, dass alle Gebiete vergleichbar sind, z.B. können an der Küste durch lokale Einflüsse Cluster von wenigen Stationen entstehen, die sehr homogen sind (Intra-) und sich durch Strömungen auch auf den angrenzenden Cluster auswirken, also auch eine große InterCluster-Homogenität aufweisen.

Fazit

Aufgrund der geschilderten Datenstruktur ist es nicht möglich durch querschneiden eines Baumdiagramms (Abb. 2.21, z.B. Abb. 3.10, 3.11) zu einer inhaltlich sinnvollen Lösung zu kommen.

Durchführung der Clusterung 2.7.7

Bevor die Zuordung zu Clustern erfolgen kann, ist es notwendig eine Arbeitsdefinition für Cluster aufzustellen:

a) Es werden die beiden Mittelwertverfahren (Kapitel 2.7.1) verwendet.

b) Ein Cluster muss in den Baumdiagrammen beider Verfahren identisch sein.

c) Ein Cluster ist in beiden Verfahren identisch, wenn die Variablen identisch sind. (Implizit heißt das, das die Fusionsreihenfolge der Variablen in den identischen Clustern nicht identisch sein muss.)

d) Ein Stationscluster muss aus mindestens vier Stationen bestehen, denn wenn es zu viele Cluster gibt oder die Cluster zu klein sind, dann ist die Lösung unübersichtlich und die Interpretation schwierig.