8 Ähnlichkeiten, Skalierung, Clusteranalyse

8.1 Unähnlichkeiten

a **Ähnliche Muster.** Oft will man Beobachtungseinheiten auf ihre Ähnlichkeiten hin untersuchen. **Ähnlichkeit** oder Unähnlichkeit soll dabei auf Grund von mehreren oder vielen Variablen, also multivariat, bestimmt werden.
Beispiele sind:

- Vegetation: Probeflächen sind sich ähnlich und repräsentieren deshalb das gleiche Ökosystem, wenn die Artenzusammensetzung ähnlich ist.
- Pflanzenarten sind sich ähnlich, wenn sie häufig an gleichen Standorten vorkommen...
- ... oder wenn sie sich evolutionsgeschichtlich vor „kurzer“ Zeit aus der gleichen Art entwickelt haben.
- Gensequenzen sind sich ähnlich, wenn sie viele übereinstimmende Abschnitte haben.
- Wetterlagen sind sich ähnlich, wenn sich die Wetterkarten gleichen – in welchen Aspekten?
- Abstimmungvorlagen können auf Grund der Einzelresultate in den Kantonen oder Gemeinden auf Ähnlichkeit hin beurteilt werden.
- **Kunden-Management.** Welche Kunden zeigen ähnliches Kaufverhalten?
- Welche Produkte werden oft miteinander gekauft?
- **Pattern Recognition.** Buchstaben sind ähnlich, wenn sie oft verwechselt werden.
- Archäologische „Artefakte“, deren charakteristische Merkmale ähnlich sind, kommen von ähnlichen Kulturen und Zeitabschnitten.
- Mittelalterliche Handschriften oder andere Texte haben aufgrund der Verwendung gemeinsamer Quellen und durch Abschreiben eine Art Stammbaum, der durch Ähnlichkeiten erschlossen werden kann. Anonyme Texte können durch Stilmerkmale als ähnlich und damit vermutlich von der gleichen Urheberschaft erkannt werden.

Allgemein geht es darum, Muster *patterns* als ähnlich zu oder unähnlich zu charakterisieren.

b **Ziel.** Ein Ziel ist oft die Bildung von Gruppen ähnlicher Objekte, Orte oder Individuen. Dies ist das Thema der **Clusteranalyse**.
Es kann aber auch nützlich sein, eine grafische Darstellung durch Punkte zu suchen, in der starke Ähnlichkeiten der Beobachtungseinheiten durch kleine Distanzen der Punkte wiedergegeben werden und schwache durch grosse Abstände. Solche Darstellungen liefern die **Skalierungs-Methoden**.

c **Euklidische Distanz.** Die anschaulichste Art, Unähnlichkeit zu messen, stellt der gewöhnliche Abstand – die so genannte Euklidische Distanz – zwischen Punkten dar; im Fall von nur zwei Variablen zwischen den Punkten im Streudiagramm, im Falle von mehreren Variablen zwischen Punkten im mehrdimensionalen Raum. Besonders sinnvoll ist dieses Mass für quantitative Variable.

Die Distanz d_{hi} zwischen zwei Beobachtungseinheiten h und i mit den Variablenwerten $\underline{x}_h$ und $\underline{x}_i$ ist die Wurzel aus

$$d_{hi}^2 = \|\underline{x}_h - \underline{x}_i\|^2 = \sum_{j=1}^{m} \left(x_h^{(j)} - x_i^{(j)}\right)^2 .$$

Wenn Variable mit unterschiedlichen Einheiten und Streuungen miteinander verrechnet werden, dasn dominieren natürlich die Unterschiede für jene Variable, die einen grossen Wertebereich aufweisen. Es drängt sich deshalb meistens auf, die Daten zuerst zu standardisieren, so dass alle Variablen die gleiche Streuung haben.

d ▷ Zur Veranschaulichung wählen wir einen kleinen Ausschnitt aus den Daten der Vegetationsstudie, nämlich die Anzahlen der Arten Nardus stricta (Nardstri), Calluna vulgaris (Caluvulg), Festuca rubra (Festrubr), Deschampsia flexuosa (Descflex) und Agrostis capillaris (Agrotenu) für 9 Beobachtungen (diejenigen im östlichen Teil, auf dem Alpboden). Sie sind in Tabelle 8.1.d festgehalten.

	Nardstri	Caluvulg	Festrubr	Descflex	Agrotenu
54	40	0	10	6	0
58	33	0	5	8	4
59	35	0	5	3	0
60	30	25	0	3	0
64	0	0	13	0	38
66	0	0	20	0	40
67	0	0	10	0	12
68	3	0	13	6	0
70	2	0	0	0	2

Tabelle 8.1.d: Beispiel-Daten

Für die ersten beiden wird die Distanz gleich $\sqrt{((33-40)^2+0+(5-10)^2+(6-8)^2+4^2)} = 9.7$. Die ersten vier Zeilen führen zur folgenden Distanztabelle:

```
> dist(t.d)
      54    58    59
58  9.70
59  7.68  6.71
60 28.88 26.46 25.98
```

In dieser Unähnlichkeit kommen vor allem die Unterschiede zwischen den Anzahlen der Arten, die den grössten Zahlenbereich überdecken, am stärksten zum Ausdruck. Es ist deshalb sinnvoll, die Variablen zuerst zu standardisieren. Für Anzahlen kann man gleiche Mittelwerte anstreben. Man dividiert also jede Spalte durch ihren Mittelwert. Dann erhält man

```
> t.mn <- apply(t.d, 2, mean)
> t.dt <- sweep(t.d, 2, t.mn, "/")
> dist(t.dt[1:4,])
     54   58   59
58 1.08
59 1.24 1.78
60 9.16 9.19 9.02
```

Für kontinuierliche Variable ist es oft sinnvoll, sie zu transformieren (siehe 8.3.c) und dann auf Varianz (oder Median-Abweichung) 1 zu standardisieren. ◁

e **Manhattan-Distanz.** Wenn sich zwei Beobachtungen bezüglich einer einzigen Variablen stark unterscheiden, dann wird die Euklidische Distanz gross, auch wenn die anderen Variablen perfekt übereinstimmen. In diesem Sinne reagiert dieses Unähnlichkeitsmass nicht robust.

Diese Eigenschaft ist wesentlich abgeschwächt, wenn man die Differenzen für die einzelnen Variablen *nicht* quadriert und als Unähnlichkeit die Summe der Absolutbeträge der Differenzen

$$d^*_{hi} = \sum\nolimits_{j=1}^{m} \left| x_h^{(j)} - x_i^{(j)} \right|$$

benützt. Man kann sich überlegen, dass dies die Weglänge misst, die man in einer Stadt mit strenger Block-Struktur zurücklegen müsste, um von einem Ort zu einem anderen zu gelangen. Diese Distanz heisst deshalb **city block distance** oder **Manhattan distance**. Mathematiker nennen sie auch L_1**-Distanz**.

▷ Die Manhatten-Distanz der unstandardisierten Daten ergibt

```
> dist(t.d[1:4,],method="manhattan")
   54 58 59
58 18
59 13 11
60 48 42 35
```

Es empfiehlt sich, für die Bildung von Unähnlichkeiten wie für andere statistische Methoden, Anzahlen zunächst mit der Wurzel zu transformieren (8.3.c). Dann erhält man für alle 9 Beobachtungen die folgende Tabelle:

```
>   dist(sqrt(t.d),method="manhattan")
      54    58    59    60    64    66    67    68
58  3.89
59  2.05  3.27
60  9.73 10.60  7.67
64 15.38 14.11 15.18 21.98
66 16.41 15.13 16.21 23.01  1.03
67 12.24 10.96 12.04 18.84  3.14  4.17
68  5.04  7.76  6.27 13.07 10.35 11.37  8.09
70 11.94  9.98  9.88 12.21  9.77 10.80  6.63  7.79
```

◁

* Die Standardisierung der Variablen auf Varianz 1 reagiert ebenfalls stark auf Ausreisser. Es ist naheliegend, auch hier auf das Quadrieren zu verzichten und die Variablen durch ihre Mittlere Absolute Abweichung (mean deviation from the median)
$(1/n)\sum_i |x_i^{(j)} - \mathrm{med}\langle x_i^{(j)}\rangle|$ zu teilen.

f **Gemeinsame Elemente.** In speziellen Anwendungen gibt es andere naheliegende Masse für Ähnichkeiten oder Unähnlichkeiten. Man kann beispielsweise für Untersuchungsflächen eine Ähnichkeit auf Grund des Vorkommens von Pflanzenarten bestimmen. Die Variable $x^{(j)}$ sei also die zweiwertige Variable, die das Vorkommen der Art j anzeigt. Es ergibt sich für jedes Paar h, i von Probeflächen eine Vierfeldertafel, deren Anzahlen üblicherweise mit a für gemeinsames Vorkommen, d für gemeinsame Abwesenheit und b und c für die beiden nicht übereinstimmenden Fälle bezeichnet wird (Tabelle 8.1.f). (Beachten Sie, dass diese Tafel nicht für einen Test auf Unabhängigkeit geeignet ist: die Arten können kaum als zufällig aufgefasst werden, und schon die Hypothese der Unabhängigkeit macht wenig Sinn.)

Probefläche h	i vorhanden	i abwesend	total
vorhanden	a	b	$a+b$
abwesend	c	d	$c+d$
total	$a+c$	$b+d$	$m=a+b+c+d$

Tabelle 8.1.f: Vierfeldertafel des Vorkommens der m Arten auf zwei Probleflächen: Bezeichnungen

Mit diesen vier Zahlen werden verschiedene Ähnlichkeitsmasse definiert. Die einfachsten sind

$$s_{hi}^{(s)} = \frac{a+d}{a+b+c+d} \qquad \text{und} \qquad s_{hi}^{(J)} = \frac{a}{a+b+c}\,,$$

der **simple matching coefficient** und der **Jaccard-Koeffizient**. Der zweite geht davon aus, dass daraus, dass einige (viele) Arten auf beiden Flächen nicht vorkommen, nichts über die Änlichkeit der Flächen aussagt.

g **Direkt bestimmte Unähnlichkeiten.** Manchmal sind Unähnlichkeiten nicht auf Grund von Variablen $x^{(j)}$ gegeben, sondern direkt bestimmt. Man kann beispielsweise Verwechslungen von Buchstaben bei automatischer Erkennung zählen oder Ähnlichkeiten durch Experten oder Versuchspersonen direkt beurteilen lassen.

Wir kommen auf mögliche Festlegungen von Ähnlichkeiten und Unähnlichkeiten zurück, sobald wir eine mögliche Verwendung eingeführt haben.

h **(Un-) Ähnlichkeitsmatrix.** Jede Definition von Ähnlichkeit s_{hi} oder Unähnlichkeit d_{hi} zwischen Objekten (Beobachtungen) führt zu symmetrischer $n \times n$ Matrix. Diese bildet den Ausgangspunkt für die folgende Anwendung und für die meisten Verfahren der Clusteranalyse, die wir in Abschnitt 8.4 besprechen werden.

8.2 Multidimensionale Skalierung.

a **Grundidee.** Man sucht eine Anordnung von Punkten $\underline{z}_i$ in der Ebene (oder eventuell im höher dimensionalen Raum), so dass die Euklidschen Distanzen zwischen Punkten die Unähnlichkeit zwischen Objekten möglichst genau wiedergeben.

Sobald man festgelegt hat, was „möglichst genau“ heissen soll, kann ein Optimierungsprogramm eine solche Anordnung bestimmen. Das ist zwar eine Aufgabe, die nicht wirklich lösbar ist, denn es ist eine nichtlineare Optimierung mit sehr vielen variablen Grössen (den $2n$ Koordinaten der gesuchten Punkte). Man kann aber mit geeigneten Programmen einen vernünftigen Lösungs-Vorschlag erhalten, von dem man einfach nicht garantieren kann, dass er das „globale“ Optimum darstellt.

Da nur für die Distanzen in der Darstellung eine Optimalität verlangt wird, kann eine erhaltene Lösung beliebig rotiert und gespiegelt werden, ohne dass sich etwas verschlechtert. Will man also zwei Darstellungen vergleichen, so muss man sie zuerst mit einem weiteren Verfahren möglichst gut zur Deckung bringen. Eine naheliegende Lösung ist es, jeweils die Hauptkomponenten-Transformation anzuwenden, um die Lösung eindeutig zu machen.

b **Stress.** Ein naheliegendes Mass für den Unterschied ist wieder eine Quadratsumme, nämlich

$$Q\langle \underline{z}_1, ..., \underline{z}_n \rangle = c \sum_{h,i} (g\langle d_{hi} \rangle - \|\underline{z}_h - \underline{z}_i\|)^2 \ .$$

Dabei ist g eine beliebige monotone Funktion, die ebenfalls so bestimmt wird, dass die Quadratsumme minimal wird. Der Sinn dieser Funktion besteht darin, zu berücksichtigen, dass der geanue Wert der Unähnlichkeit unwichtig ist und nur die Rangfolge zählen soll.

Im allgemeinen ist es nützlich, diese Summe noch für kleinere und grössere Unähnlchkeiten unterschiedlich zu gewichten. Man kann dann erreichen, dass eher die kleinen oder eher die grossen Unähnlchkeiten gut wiedergegeben werden. Für den zweiten Fall kann man für grössere Datensätze weniger auf Erfolg hoffen.

c ▷ Abbildung 8.2.c (i) zeigt das Resultat für das kleine Datenbeispiel der 9 Beobachtungen aus der Vegetationsstudie bei Verwendung der Manhattan-Distanz der wurzeltransformierten Daten (8.1.e). In Abbildung 8.2.c (ii) wird die Distanz und die darzustellende Unähnlichkeit verglichen. Für das Punktepaar $[68, 58]$ ist die Unähnlichkeit wesentlich grösser (7.76) als die Distanz in der Darstellung (5.40); für das Paar $[68, 54]$ ist sie wesentlich kleiner (5.04), während die Distanz grösser ist (7.10).

d ▷ Für den ganzen Datensatz der Vegetationsstudie erhält man die in Abbildung 8.2.d gezeigte Darstellung. Die 9 Beobachtungen, die auch im kleinen Beispiel auftauchen, sind hervorgehoben, um einen Vergleich ihrer Positionen mit ihrer vorhergehenden Darstellung zu erleichtern. ◁

e **Vergleich mit Hauptkomponenten.** Die Hauptkomponenten-Analyse haben wir eingeführt mit dem Ziel, hochdimensionale Daten in niedrig-dimensionale zu verwandeln – wie das jetzt auch durch die multidimensionale Skalierung erreicht wird. Welches Verfahren ist jetzt besser?

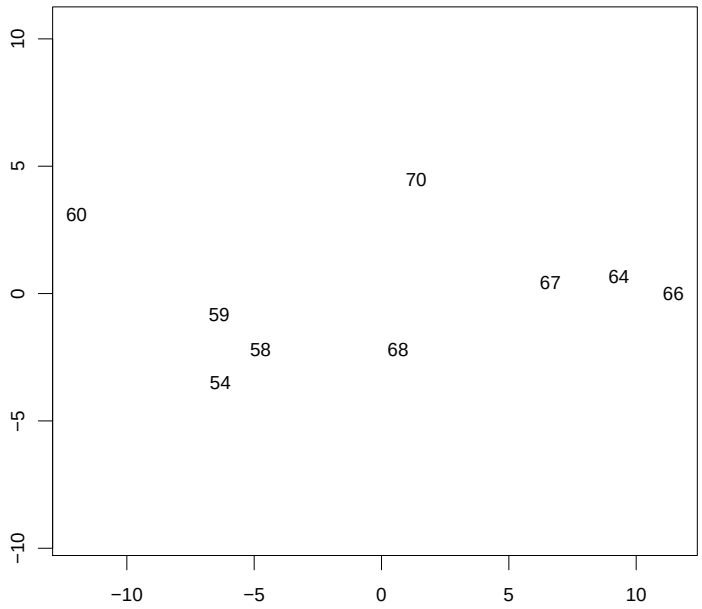

Abbildung 8.2.c (i): Multidimensionale Skalierung für den kleinen Beispiel-Datensatz

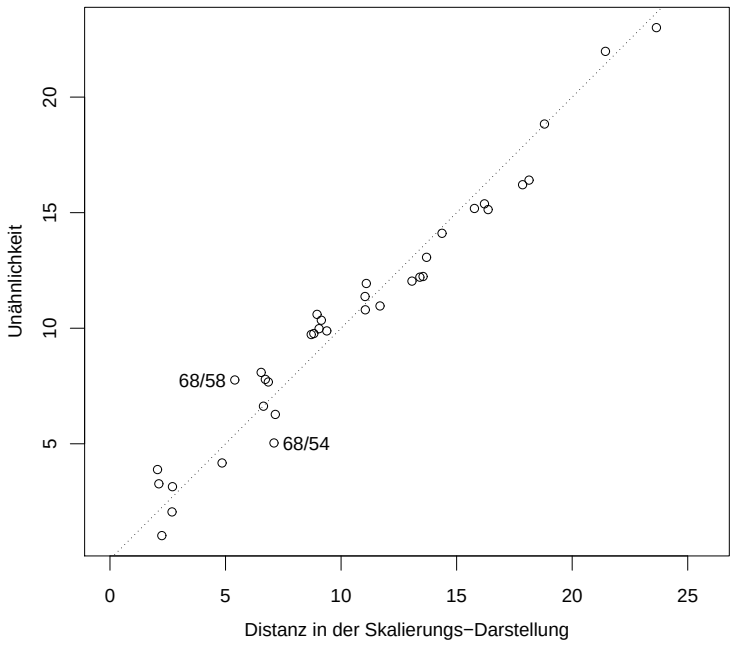

Abbildung 8.2.c (ii): Vergleich der Distanz in der Multidimensionalen Skalierung mit den vorgegebenen Unähnlichkeiten im kleinen Beispiel. Zwei Punktepaare, für die die Übereinstimmung nicht gut ist, sind markiert.

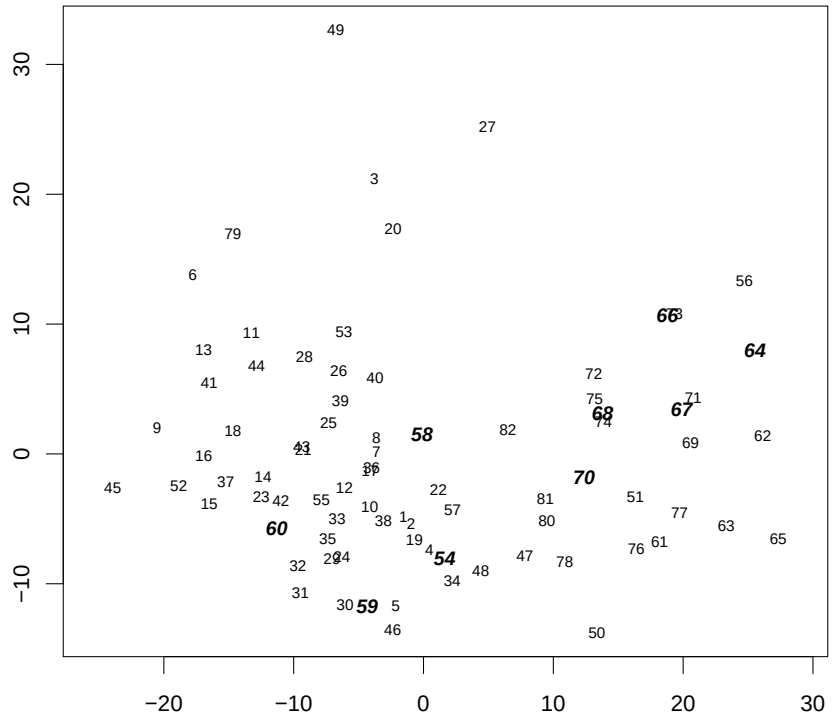

Abbildung 8.2.d: Multidimensional Scaling mit Euklidischer Distanz für die Art-Variablen des Beispiels der Vegetationsstudie

Die Hauptkomponenten sind Linearkombinationen von den ursprünglichen X-Variablen. Wir haben p solche Linearkombinationen gesucht mit dem Ziel, einen möglichst grossen Anteil der Variabilität in den Daten durch die Variabilität dieser Hauptkomponenten zu erfassen. Man kann zeigen, dass sich das auf die (Euklidischen) Distanzen überträgt: Die Distanzen im ursprünglichen Raum werden möglichst genau durch die Distanzen im Hauptkomponenten-Raum wiedergegeben.

Der Vorteil der multidimensionalen Skalierung besteht darin, dass wir weniger Einschränkungen machen.

- Vor allem müssen die Koordinaten, die wir für die Darstellung brauchen, nicht linear von den Ausgangsgrössen $X^{(j)}$ abhängen.
- Im Stress-Mass wurde ausserdem nur gute Übereinstimmung der Distanzen im Darstellungsraum mit einer monoton transformierten Version der Unähnlichkeit übereinstimmt.
- Schliesslich brauchen wir gar keine X-Variablen für die multidimensionale Skalierung; eine Unähnlichkeitsmatrix, die auf irgendeine Weise zustande gekommen ist, genügt als Ausgangspunkt.

Die multidimensionale Skalierung hat auch einen bedeutsamen Nachteil: Sie bestimmt nur für die vorliegenden Beobachtungen eine Darstellung. Wenn neue Beobachutungen dazu kommen, ist nicht klar, wo sie in eine bestehende Abbildung eingezeichnet werden sollen. Bei Hauptkomponenten ist das klar: Man kann die Koordinaten auch für neue Beobachtungen ausrechnen, ohne die Analyse gesamthaft zu wiederholen. Damit kann man auch neue Beobachtungen zu visuell bestimmten Gruppen der vorliegenden zuordnen, was bei multidimensionaler Skalierung nur über eine nochmalige Analyse mit dem erweiterten Datensatz möglich ist und schief gehen kann, wenn die neue Darstellung allzu stark von der alten verschieden ist.

8.3 Weitere Überlegungen zu Unähnlichkeiten

a Es lohnt sich, auf die Bildung eines Unähnlichkeits-Masses viel Sorgfalt zu verwenden, da der Erfolg (was immer das heisst) meist entscheidend von diesem Mass abhängt. Es soll die zu den Daten und dem Auswertungsziel passende Intuition von (Un-) Ähnlichkeit möglichst gut wiedergeben.

b **Unähnlichkeit als Summe von Unähnlichkeiten für einzelne Variable.** Die meisten sinnvollen Unähnlichkeitsmasse entstehen als Summe von Beiträgen der einzelnen Variablen – allenfalls einer gewichteten Summe oder einem (gewichteten) Mittel der Beiträge,

$$d\langle \underline{x}_h, \underline{x}_i \rangle = \frac{\sum_j w_j \, d^{(j)} \langle x_h^{(j)}, x_i^{(j)} \rangle}{\sum_j w_j} \; .$$

Mittelwerte haben gegenüber Summen den Vorteil, dass fehlende Werte sinnvoll behandelt werden, indem die entsprechenden Terme (oben und unten) weggelassen werden.

Noch etwas flexibler ist die folgende Formel, die mit transformierten Beiträgen $d^{(j)}$ arbeitet:

$$\ell\langle d\langle \underline{x}_h, \underline{x}_i \rangle \rangle = \frac{\sum_j w_j \, \ell\Big\langle d^{(j)}\Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle \Big\rangle}{\sum_j w_j} \; .$$

Wenn $\ell\langle d \rangle = d^2$ und $d^{(j)}\Big\langle x_h^{(j)}, x_i^{(j)} \Big\rangle = |x_h^{(j)} - x_i^{(j)}|$ ist, erhält man bei gleichen Gewichten den gewöhnlichen Euklidischen Abstand, bis auf den Faktor $1/\sqrt{m}$.

c Grundlegend für die gesamthafte Unähnlichkeit ist also die Definition der Unähnlichkeiten $d^{(j)}\left\langle x_h^{(j)}, x_i^{(j)} \right\rangle$ für einzelne Variable.

Für **stetige Variable** wird man den Betrag der Differenz der Werte als Unähnlichkeit wählen. Damit aber gleiche Differenzen wirklich gleiche Unähnlichkeiten bedeuten, muss man die Variable oft **vorher transformieren**. Wenn beispielsweise eine Anzahl von Individuen auf einer Probefläche gezählt wird, bedeutet der Unterschied zwischen 0 und 2 etwas ganz anderes als der Unterschied zwischen 10 und 12.

Ein sinnvoller Vorschlag für eine Transformation besteht meistens aus der „first aid"-Transformation:

- die Logarithmus-Transformation für **Konzentrationen und Beträge** – also für stetige Variable, die nur positive Werte haben können –
- die Wurzeltransformation für **Zähldaten** und
- die so genannte Arcus-Sinus-Transformation $\ell\langle x \rangle = \arcsin \sqrt{x}$ für **Anteile** (Prozentzahlen/100).

Diese Transformationen haben von J. W. Tukey den Namen **first aid transformations** erhalten und **sollten für solche Daten immer angewendet werden**, wenn es keine Gegengründe gibt – es ist aber gerade im Zusammenhang mit Unähnlichkeiten nicht verboten, eine andere Transformation anzuwenden, wenn sie dazu führt, dass **gleiche Differenzen der transformierten Variablen gleiche Unähnlichkeiten bedeuten**.

Zusammenfassend können wir für die Unähnlichkeit einer stetigen Variablen also schreiben

$$d^{(j)}\left\langle x_h^{(j)}, x_i^{(j)} \right\rangle = \left| g^{(j)}\langle x_h^{(j)} \rangle - g^{(j)}\langle x_i^{(j)} \rangle \right| .$$

d Die naheliegendste **Unähnlichkeit für geordnete Variable** ist die absolute Differenz der Ränge. Diese Festlegung folgt dem gerade genannten Schema der Transformation und Differenzenbildung. Die Rang-Differenz der Beobachtungen h und i ist, wenn alle Werte verschieden sind, die Anzahl Beobachtungen mit Werten zwischen $x_h^{(j)}$ und $x_i^{(j)}$, vermindert um 1. Sie hängt damit von der Gesamtheit aller Beobachtungen ab. Man kann diesen Vorschlag auch für quantitativ interpretierbare Grössen (vorheriger Fall) verwenden.

e Für **binäre Variable** ist es naheliegend, nur zwischen Übereinstimmung ($x_h^{(j)}$ und $x_i^{(j)}$ beide 0 oder beide 1) und Verschiedenheit zu unterscheiden. Es gibt dann auch für die Unähnlichkeit nur die Werte 0 und 1, und man kann dies wie für quantitative Variable schreiben: $d^{(j)}\left\langle x_h^{(j)}, x_i^{(j)} \right\rangle = |x_h^{(j)} - x_i^{(j)}|$.

Oft bedeutet allerdings „positive Übereinstimmung" (Art an beiden Orten vorhanden) mehr Ähnlichkeit als negative Übereinstimmung. Dann soll man $d^{(j)}\langle 0, 0 \rangle \neq 0$ setzen. Man kann in diesem Fall auch das Gewicht w_j reduzieren – verletzt dadurch allerdings das oben anvisierte Schema (8.3.b), in dem w_j nicht von den $x^{(j)}$-Werten abhängt.

f Für nominale oder **kategorielle Variable** schliesslich kann man wie für binäre $d^{(j)}\langle x_h^{(j)}, x_i^{(j)}\rangle = 0$ setzen, falls $x_h^{(j)}$ und $x_i^{(j)}$ übereinstimmen, und =1 bei Nicht-Übereinstimmung. Will man differenzieren, so gibt es hier viele Möglichkeiten, die Häufigkeiten der möglichen Werte sowie ihre Ähnlichkeiten zu berücksichtigen.

g **Standardisierung und Gewichtung von Variablen.** Kontinuierliche Variable haben meistens verschiedene Einheiten. Damit sie in der Definition der Unähnlichkeit (8.3.b) in gewissem Sinne „gleiches Gewicht“ erhalten, standardisiert man sie auf (robuste) Varianz 1. Damit äquivalent ist es, in der Formel 8.3.b $w_j = 1/\widehat{\sigma}_j$ zu setzen.

h ▷ Im Beispiel führen alle Empfehlungen dazu, aus den Individuenzahlen der Arten zuerst die Wurzel zu ziehen, sie dann mit der Mittleren Absoluten Abweichung zu standardisieren und schliesslich die Manhattan-Distanz zu rechnen. So erhält man

```
      1     2     3     4
2  2.60
3 11.37 13.49
4  4.14  3.41 13.35
5  4.63  3.90 11.70  1.65
```

(Die Wurzel-Transformation macht in diesem kleinen Beispiel wenig aus.)

Für den Fall von Individuen-Zahlen kann man auf die Standardisierung auch verzichten. Man gibt damit den seltenen Arten wenig Einfluss auf die Unähnlichkeit – ihre Individuenzahl ist ja auch weniger verlässlich bestimmbar. ◁

i **Explizite Gewichtung.** Es kann sehr nützlich sein, zur Festlegung einer aussagekräftigen Unähnlichkeit eine Gewichtung explizit zu wählen. Eine starke Aussage zu diesem Punkt formuliert Anderberg (1973, p.13):

“Some investigators recommend reducing all variables to standard form (zero mean and unit variance) at the outset. Such suggestions simplify the mechanics of analysis but constitute a complete abdication of the analyst’s responsibilities and prerogatives to a mindless procedure.”

j **Hauptkomponenten.** Eine ungleiche Gewichtung von Variablen scheint sich vor allem dann aufzudrängen, wenn gewisse Variable stark korreliert sind. Wir haben die Hauptkomponenten-Analyse kennengelernt, die aus beliebigen Variablen unkorrelierte macht. Man ist deshalb versucht, die Euklidische Distanz auf Hauptkomponenten-Koordinaten anzuwenden. Das hat allerdings nur dann einen Effekt, wenn man sich auf wenige Hauptkomponenten beschränkt. Wenn man alle m verwendet, ändert sich die Distanz nicht; das war ja eine der Eigenschaften der Hauptkomponenten-Analyse.

Eine Vernachlässigung der letzten Hauptkomponenten führt allerdings dazu, dass „eigenständige“ Variable, die mit allen andern nur schwach korreliert sind, unterdrückt werden. Ob das erwünscht ist, muss sorgfältig überlegt werden. Vielleicht sind das ebenso wichtige Variable, die ihre Beiträge zur Unähnlichkeit haben sollten.

k **Mahalanobis-Distanz.** Ein Distanzmass, das Korrelationen zwischen Variablen berücksichtigt, ist die Mahalanobis-Distanz, die auf der Kovarianzmatrix aller Beobach-

tungen beruht,

$$d\langle \underline{x}_h, \underline{x}_i \rangle = (\underline{x}_h - \underline{x}_i)^T \widehat{\mathbf{\Sigma}}^{-1} (\underline{x}_h - \underline{x}_i) \; .$$

Sie erscheint auf den ersten Blick sehr geeignet zu sein. Die folgende Überlegung führt aber zu Bedenken: Führt man zunächst eine Hauptkomponenten-Analyse durch und standardisiert dann die erhaltenen Werte (scores) auf Varianz 1, dann liefert die gewöhnliche Euklidische Distanz das Gleiche wie die Mahalanobis-Distanz der ursprünglichen Beobachtungen. Hier werden die Hauptkomponenten mit kleinen Varianzen (kleinen Eigenwerten) also nicht vernachlässigt, sondern im Gegenteil auf gleiche Varianz wie die ersten Hauptkomponenten „aufgeblasen“. Das ist wohl meistens nicht sinnvoll.

l Ein sinnvoller Umgang mit **stark korrelierten Variablen** besteht wohl darin, dass man Gruppen von zwei oder mehreren solchen Variablen durch je eine einzige ersetzt, indem man eine aus der Gruppe auswählt oder eine neue definiert, die sie ersetzt – beispielsweise die Summe, den Mittelwert oder einen anderen „Index“.

m **Ähnlichkeiten.** Hier wurde die Festlegung von *Unähnlichkeiten* ausführlich diskutiert. Intuitiv spricht man eher von Ähnlichkeiten. Die vorhergehende Analyse ist für Unähnlichkeiten gedacht; sie auf Ähnlichkeiten zu übertragen, scheint schwierig.

Aus jeder Unähnlichkeit kann aber eine Ähnlichkeit erhalten werden, beispielsweise durch

$$s_{hi} = 1/(1 + d_{hi}) \; .$$

Das Resultat liegt dann zwischen 0 und 1 – was für Ähnlichkeitsmasse üblich ist. Umgekehrt kann man aus einer Ähnlichkeit s eine Unähnlichkeit erhalten über $d = 1/s - 1$ oder, einfacher, durch $d = 1 - s$.

n **Ähnlichkeiten von Variablen.** Man kann auch (Un-) Ähnlichkeiten von Variablen einführen, um diese dann grafisch darzustellen oder die Variablen in Cluster einzuteilen. Den Prototyp der Ähnlichkeit von Variablen bildet ihre Korrelation. Sie ist gleich

$$s\left\langle \underline{x}^{(j)}, \underline{x}^{(k)} \right\rangle = \frac{1}{n-1} \sum_i x_i^{(j)} x_i^{(k)} \; ,$$

falls die Variablen standardisiert wurden. Es ist dann

$$\begin{aligned} d\left\langle \underline{x}^{(j)}, \underline{x}^{(k)} \right\rangle &= \frac{1}{n-1} \sum_i (x_i^{(j)} - x_i^{(k)})^2 \\ &= \frac{1}{n-1} \sum_i (x_i^{(j)})^2 + \frac{1}{n-1} \sum_i (x_i^{(k)})^2 - \frac{2}{n-1} \sum_i x_i^{(j)} x_i^{(k)} \\ &= 2 - 2s\left\langle \underline{x}^{(j)}, \underline{x}^{(k)} \right\rangle \; . \end{aligned}$$

Für die Anwendungen sinnvoller kann der Betrag der Korrelation sein, da eine negative Korrelation auch „enge Beziehung“ bedeuten kann.

▷ Die Darstellung für die Arten der Vegetationsstudie zeigt sich in Abbildung 8.3.n. ◁

o Zusammenfassend: **In die Definition der Unähnlichkeit soll möglichst viel intuitives oder präzises Wissen über die Variablen eingehen.**

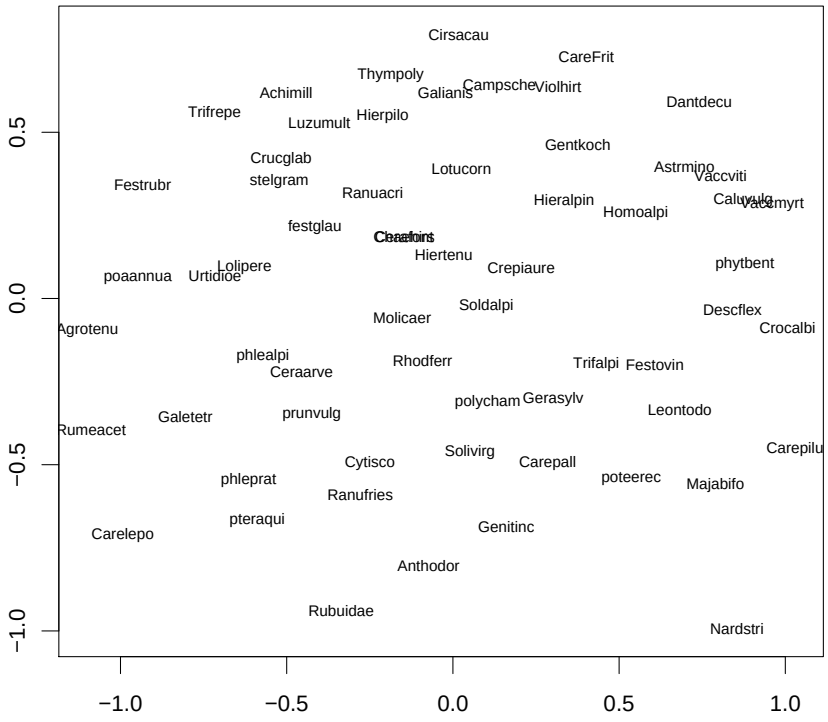

Abbildung 8.3.n: Multidimensionale Skalierung für die Arten in der Vegetationsstudie aufgrund ihrer Korrelationen

8.4 Clusteranalyse: Optimale Partitionen

a **Grundidee.** Gesucht ist eine Einteilung der Beobachtungen in homogene Gruppen, also Gruppen von möglichst ähnlichen Beobachtungseinheiten. Eine Gruppeneinteilung heisst mathematisch **Partition**. Die Gruppen $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_q$ sind Teilmengen der Gesamtmenge $\mathcal{S}$. Eine Partition ist dadurch charakterisiert, dass die Vereinigung aller $\mathcal{G}_k$ die Grundmenge ausschöpft, $\cup_{k=1}^{q} \mathcal{G}_k = \mathcal{S}$ und die $\mathcal{G}_k$ sich gegenseitig ausschliessen, $\mathcal{G}_k \cap \mathcal{G}_\ell = \emptyset$ für alle $k \neq \ell$.

Wenn nun eine Einteilung in *homogene* Gruppen gefragt ist, besteht die für „mathematisch sozialisierte“ Leute naheliegende Leitidee darin, ein **Gütemass** Q für die Homogenität einer Partition in q Gruppen zu definieren und dann die Partition mit dem besten Gütewert für gegebenes q zu suchen. Wenn sich das Gütemass auch eignet, Einteilungen mit verschiedener Gruppenzahl zu vergleichen, kann man auch über q optimieren.

b Wie erhält man ein Gütemass Q? Wie bei *Un*ähnlichkeiten ist es leichter, ein **Inhomogenitätsmass** (oder Hetereogenitätsmass) zu definieren. Die Gesamt-Hetereogenität wird jeweils als Summe der Heterogenitäten der Cluster festgelegt,

$$Q = \sum\nolimits_k h\langle \mathcal{G}_k \rangle \ ,$$

wobei die Inhomogenität h üblicherweise auf einer Unähnlichkeits-Definition für Beobachtungen beruht, wie wir sie im letzten Abschnitt eingeführt haben.

Dazu folgt ein gebräuchliches Beispiel.

c **K-means.** Für einen Cluster $\mathcal{G}$ bildet man den Mittelpunkt oder das „Zentroid“

$$\underline{x}_{\mathcal{G}} = \operatorname*{ave}_{i\in\mathcal{G}}\langle x_i \rangle$$

($\text{ave}_i\langle . \rangle$ bezeichnet den Mittelwert über die Argumente, wie $\sum_i$ die Summe bezeichnet.) Das Inhomogenitätsmass ist dann die Summe der quadrierten Abstände zwischen den Beobachtungen des Clusters und seinem Mittelpunkt,

$$h\langle \mathcal{G} \rangle = \sum\nolimits_{i\in\mathcal{G}} d\langle \underline{x}_i, \underline{x}_{\mathcal{G}} \rangle^{\,2} .$$

Die Suche eines Minimums des entsprechenden Gütemasses $Q = \sum_k h\langle \mathcal{G}_k \rangle$ wird als **K-means algorithm** bezeichnet.

Ersetzt man für die Bestimmung des Zentroids den Mittelwert mit dem „komponentenweisen Median“ $\underline{x}_{\mathcal{G}} = [\text{med}_i\left\langle x_i^{(1)} \right\rangle, ..., \text{med}_i\left\langle x_i^{(m)} \right\rangle]^T$ und die Manhattan-Distanz in der Definition von h, dann erhält man den so genannten **K-medians** Algorithmus.

d **Optimierung.** Die Optimierung eines solchen Kriteriums ist rechnerisch nur für sehr kleine Datensätze wirklich zu bewältigen. Für $n = 25$ Objekte und $q = 3$ Cluster wären 1.4×10^{11} Einteilungen durchzurechnen!

Wie immer in solchen Fällen begnügt man sich mit einer **„lokalen Optimierung“**: Man geht von einer Einteilung in k Gruppen aus und definiert eine Änderung, die zu einer Verbesserung des Kriteriums führt. Dieser Schritt wird dann so lange wiederholt, bis sich keine Verbesserung des Kriteriums mehr erreichen lässt. Je nach der anfänglichen Einteilung können sich verschiedene solche „lokale Lösungen“ ergeben – bei k-means und k-medians gibt es oft viele. Deshalb ist es empfehlenswert, von möglichst vielen anfänglichen Einteilungen aus jeweils das lokale Optimum zu bestimmen und dann die Lösung zu wählen, die zum besten Wert des Kriteriums führt. Aber auch dies garantiert nicht, dass es keine noch bessere Lösung gäbe. (Es bleibt: Was ich nicht weiss, macht mich nicht heiss...)

e ▷ Im **Beispiel der Vegetationsstudie** wurden in der Auswertung der Autoren vier Gruppen gebildet mit Hilfe der Vegetationsdaten. Der k-means Algorithmus mit wurzel-transformierten, unstandardisierten Individuenzahlen für alle 63 Arten und $k = 4$ liefert Cluster, die in Tabelle 8.4.e mit den im Artikel gefundenen Gruppen verglichen werden.

Eine Beurteilung der erhaltenen Gruppierung kann man erlangen, indem man die Beobachtungen in Hauptkomponenten (Abbildung 8.4.e (i)) oder mit multidimensionaler Skalierung (Abbildung 8.4.e (ii)) darstellt und die Gruppenzugehörigkeit symbolisiert. Die Figuren zeigen, dass die Distanzen in zwei Dimensionen recht gut dargestellt werden können – jedenfalls trennen sich die gefundenen Gruppen in diesen Dimensionen recht klar.

Im Artikel wird auch eine Vegetationskarte angegeben, die auf einer von der Studie unabhängigen Kartierung beruht. Abbildung 8.4.e (iii) zeigt, dass die im Artikel aufgrund

	k-means A	B	C	D	total
1	0	3	0	4	7
Arti- 2	4	17	0	0	21
kel 3	6	0	22	2	30
4	2	0	0	22	24
total	12	20	22	28	82

Tabelle 8.4.e: Vergleich der Vegetationsgruppen gemäss Artikel mit den Ergebnissen des k-means Algorithmus

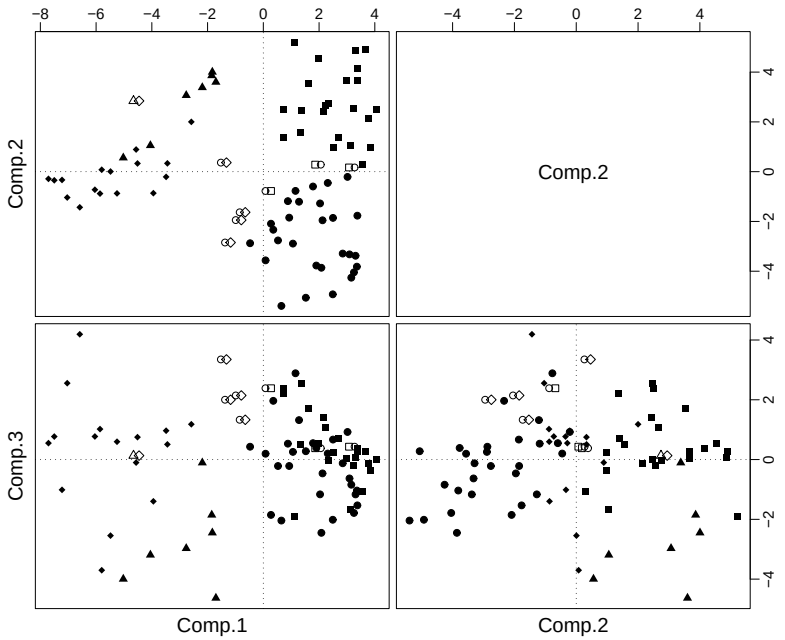

Abbildung 8.4.e (i): Gruppierung gemäss k-means und gemäss Artikel, dargestellt in Hauptkomponenten

der Daten hergeleitete Einteilung nicht überall mit der Kartierung übereinstimmt. Vor allem die Gruppen 1 und 2 sind sich wohl sehr ähnlich und ihre Unterscheidung deshalb schwierig – und wohl auch weniger wichtig. Abbildung 8.4.e (iv) stellt die von k-means gefundene Gruppierung dar. ◁

f **PAM.** Benützt man eine Beobachtung als Zentroid, dann ist die Gruppeneinteilung durch die Auswahl der Zentroide aus den Beobachtungen gegeben. Man muss also die Auswahl suchen, die das Gütekriterium für die entsprechende Gruppeneinteilung optimiert. Diese Methode wird **k-medoid** method genannt. In R heisst sie entsprechend den Promotoren Kaufman and Rousseeuw (1990) PAM. Da keine Zentren berechnet werden müssen, ist sie einerseits auch anwendbar, wenn nur Distanzen gegeben sind, und andererseits ist sie dank einfacherer Rechnung für grössere Datenmengen geeignet.

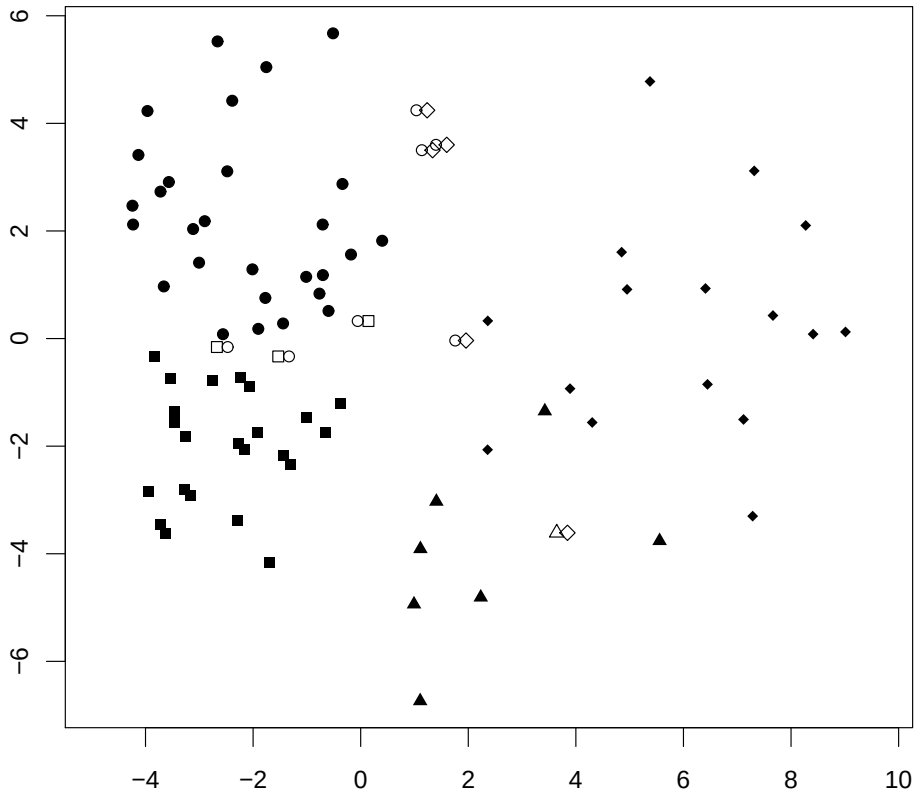

Abbildung 8.4.e (ii): Gruppierung gemäss k-means und gemäss Artikel, dargestellt mit multidimensionaler Skalierung

Der in der S-Funktion `pam` implementierte Algorithmus enthält zwei Teile. Im ersten wird gemäss einer ad-hoc-Methode die k Beobachtungen bestimmt, die als Anfangslösung dienen. Diese werden im iterativen Teil gemäss dem Kriterium verbessert. Man kann, wie bei k-means und k-median, studieren, welche Lösungen man erhält, wenn man mehrere Male jeweils eine Zufallsauswahl von k Beobachtungen als Anfangslösung wählt und die entsprechende „lokale Lösung“ bestimmt.

▷ Im Beispiel gibt die Verwendung von Manhattan-Distanzen und PAM die Gruppierung, die in der Karte Abbildung 8.4.f gezeigt wird. Sie scheint sich hier weniger zu bewähren als k-means mit Euklidischen Distanzen. ◁

g* Man kann einfach ausrechnen, dass auch für die k-means-Methode keine X-Werte bekannt sein müssen. Wir berechnen

$$\begin{aligned}
\sum_{h,i\in\mathcal{G}} d_{hi}^2 &= \sum_{h,i\in\mathcal{G}} (\underline{x}_i - \underline{x}_h)^T(\underline{x}_i - \underline{x}_h) \\
&= \sum_{h,i\in\mathcal{G}} \big((\underline{x}_i - \bar{\underline{x}}_\mathcal{G}) - (\underline{x}_h - \bar{\underline{x}}_\mathcal{G})\big)^T(...) \\
&= \sum_{h,i\in\mathcal{G}} \big((\underline{x}_i - \bar{\underline{x}}_\mathcal{G})^T(\underline{x}_i - \bar{\underline{x}}_\mathcal{G}) - 2(\underline{x}_i - \bar{\underline{x}}_\mathcal{G})^T(\underline{x}_h - \bar{\underline{x}}_\mathcal{G}) + (\underline{x}_h - \bar{\underline{x}}_\mathcal{G})^T(\underline{x}_h - \bar{\underline{x}}_\mathcal{G})\big) \\
&= 2n_\mathcal{G} \sum_{i\in\mathcal{G}} \big((\underline{x}_i - \bar{\underline{x}}_\mathcal{G})^T(\underline{x}_i - \bar{\underline{x}}_\mathcal{G}) - 2\sum_{i\in\mathcal{G}} (\underline{x}_i - \bar{\underline{x}}_\mathcal{G})^T \sum_{h\in\mathcal{G}} (\underline{x}_h - \bar{\underline{x}}_\mathcal{G})\big) \\
&= 2n_\mathcal{G} \sum_{i\in\mathcal{G}} d(\underline{x}_i, \bar{\underline{x}}_\mathcal{G})^2
\end{aligned}$$

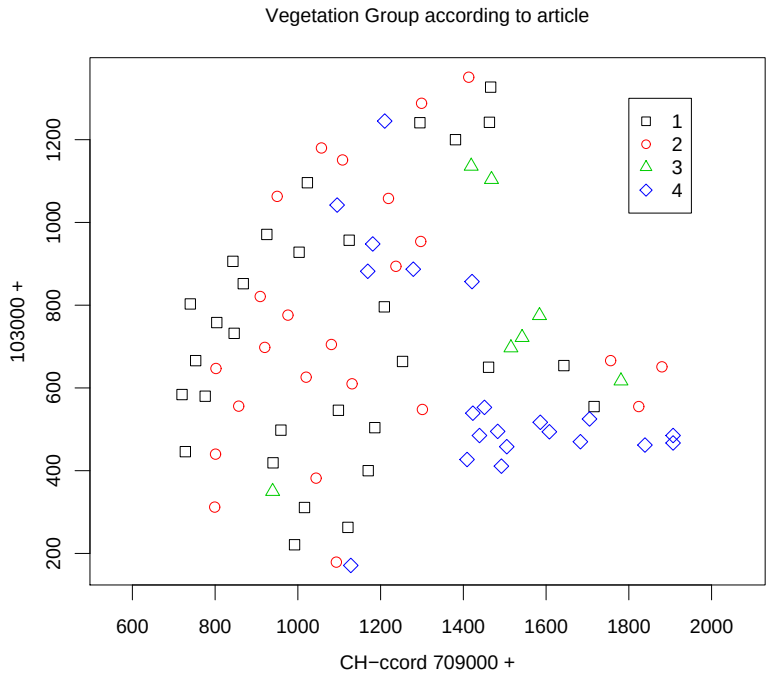

Abbildung 8.4.e (iii): Geografische Orte der Probeflächen mit Gruppierung gemäss Artikel

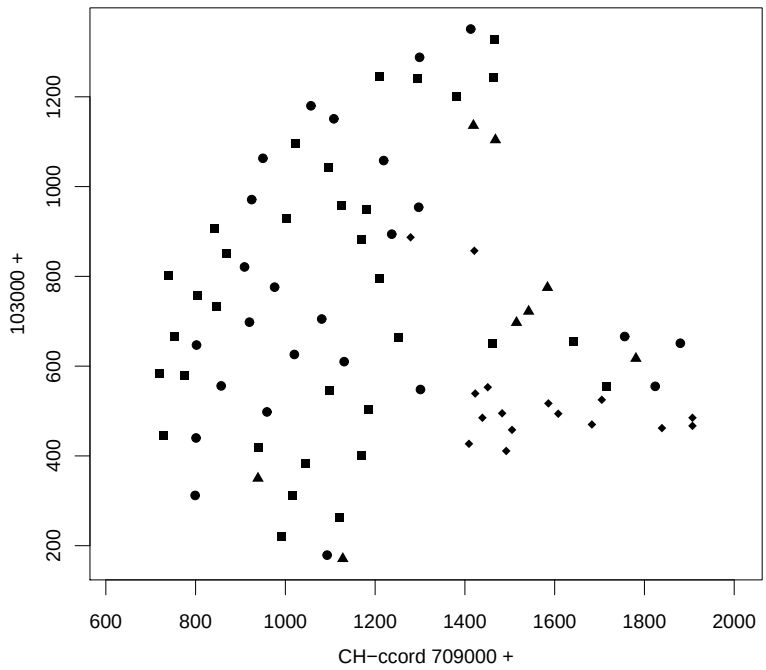

Abbildung 8.4.e (iv): Geografische Orte der Probeflächen mit Gruppierung gemäss k-means

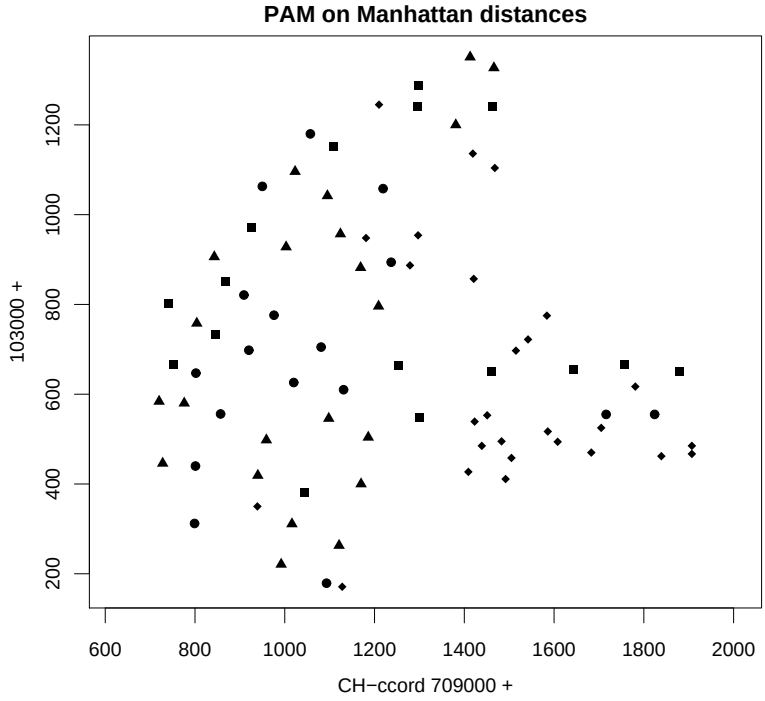

Abbildung 8.4.f (iv): Geografische Orte der Probeflächen mit Gruppierung gemäss PAM mit Manhattan-Distanzen

Also ist

$$h\langle \mathcal{G} \rangle = \sum_{i \in \mathcal{G}} d\langle \underline{x}_i, \underline{x}_{\mathcal{G}} \rangle^{\,2} = \frac{1}{n_{\mathcal{G}}} \sum_{h,i \in \mathcal{G}} d_{hi}^2 \,.$$

Diese Formel lässt sich anwenden, wenn nur die Distanzen bekannt sind. Für grosse Datensätze ist dies allerdings kein Vorteil, da die Distanzmatrix dann viel grösser wird als die Datenmatrix. Man kann aber andererseits diese Formel auch für die Festlegung weiterer möglicher Inhomogenitätsmasse nehmen, indem man statt d_{hi}^2 irgendeine Unähnlichkeit verwendet.

h **Silhouetten.** Wie gut sind die Cluster getrennt? Das Verfahren liefert ja immer eine Lösung, und die Frage stellt sich, ob es sich um natürliche, gut abgegrenzte Cluster handelt oder ob in einer gleichmässig verteilten Gesamtheit durch die Einteilung mehr oder weniger willkürliche Grenzen gezogen werden.

Zur Beantwortung der Frage haben Kaufman and Rousseeuw (1990) Grössen eingeführt, die für jede Beobachtung feststellen, wie eindeutig sie zu „ihrem“ Cluster gehören, Sei $\widetilde{d}\langle i, \mathcal{G} \rangle$ due Distanz von Objekt i zu Cluster $\mathcal{G}$ – in k-means und PAM die Distanz der Beobachtung i zum Zentrum oder Zentroid des Clusters $\mathcal{G}$. Zudem bezeichne $\mathcal{G}\langle i \rangle$ den Cluster, zu dem i gehört ($i \in \mathcal{G}$), und $\mathcal{B}\langle i \rangle = \arg\min_{\mathcal{G} | i \notin \mathcal{G}} \left\langle \widetilde{d}\langle i, \mathcal{G} \rangle \right\rangle$ den Nachbar-Cluster von i. Der Silhouetten-Wert für i ist dann 1− das Verhältnis der Unähnlichkeiten von i von diesen beiden Clustern

$$\widetilde{s}\langle i \rangle = 1 - \frac{\widetilde{d}\langle i, \mathcal{G}\langle i \rangle \rangle}{\widetilde{d}\langle i, \mathcal{B}\langle i \rangle \rangle}$$

(Die Original-Definition wurde hier für negative Werte vereinfacht.) Der Wert wird 0, wenn die Beobachtung auf der Grenze zwischen zwei Gruppen liegt.

Die Silhouetten-Werte werden in einer grafischen Darstellung gegen eine geeignete Anordnung der i aufgetragen: Die Mitglieder jedes Clusters werden zusammengezogen und absteigend sortiert.

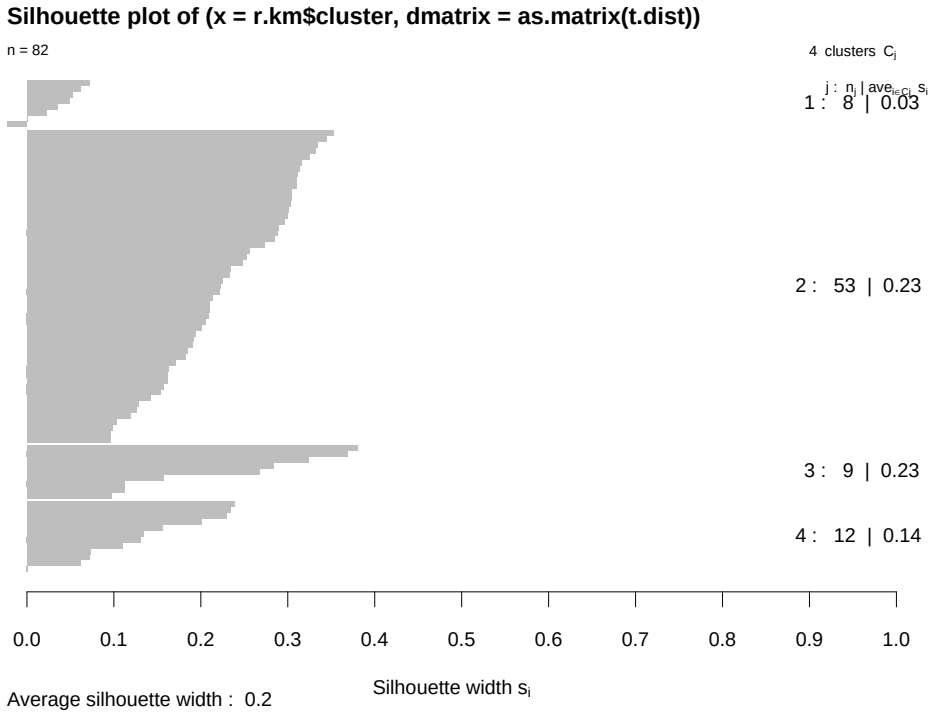

Abbildung 8.4.h: Silhouette für die k-means-Gruppen im Beispiel

Abbildung 8.4.h zeigt eine solche Silhouette für unser Beitspiel. Die Cluster sind recht gut definiert; Nur wenige Beobachtungen zeigen kleine Silhouetten-Werte.

i Für die **Wahl der Anzahl Cluster** werden sinnvoller Weise ad-hoc-Ideen benützt, die mit der Problemstellung in der einzelnen Anwendung zusammenhängen. Silhouetten sind auch hier hilfreich. Vergleicht man Silhouetten für verschiedene q, so kann sich zeigen, dass die Erhöhung von q um 1 jeweils zu einer Unterteilung eines Clusters in zwei Teile führt. Das deutet auf gut abgegrenzte Cluster hin.

j **Grosse Datensätze.** Heute werden oft Gruppierungen in grossen Datensätzen (n gross) gesucht. Die Berechnung und Speicherung aller Unähnlichkeiten ist dann aufwändig und sollte vermieden werden.

Ein Vorschlag für einen Algorithmus heisst CLARA und stammt ebenfalls von Kaufman and Rousseeuw (1990): Begonnen wird mit einer zufällig ausgewählten Stichprobe, mit der PAM durchgeführt werden kann und klassiere alle Objekte mit den resultierenden Zentroiden. Das wird 5 mal wiederholt, und schliesslich wird die beste der 5 Lösungen als Ergebnis angegeben.

8.5 Hierarchische Verfahren, Dendrogramme

a Die so genannten **agglomerativen** Verfahren der Clusteranalyse sind populär, weil sie einerseits auf sehr einfachen Rechenschritten beruhen und deshalb schon ein halbes Jahrhundert alt sind, und andererseits, weil sie ein interessante grafische Darstellung in Form eines „Dendrogramms“ ergeben.

b ▷ Die Grundlage für ein hierarchisches Verfahren bildet ein Unähnlichkeitsmass (oder Ähnlichkeitsmass) für die Beobachtungen, das als Ausgangspunkt eine entsprechende Unähnlichkeits-Matrix liefert. Um das Verfahren zu illustrieren, gehen wir von der in 8.1.e angegebenen Unähnlichkeitstabelle aus.

Die Einheiten, die sich ähnlich sind, sollen sich wenn möglich in einem Cluster wiederfinden. Also beginnen wir damit, die beiden Einheiten h_1 und h_2 mit der kleinsten Unähnlichkeit zu einem „Mini-Cluster“ zu vereinigen. Es handelt sich im Beispiel um die Beobachtungen 64 und 66 mit einer Unähnlichkeit von 1.03.

Wir wollen fortfahren, indem wir den nächstkleineren Wert der Unähnlichkeit suchen. Damit wir dies sinnvoll tun können, müssen wir je einen Wert für die Unähnlichkeit zwischen dem Mini-Cluster $\{h_1, h_2\}$ und jeder übrigen Beobachtungseinheit i festlegen. Dazu gibt es verschiedene Varianten, die weiter unten diskutiert werden. Ein naheliegender Vorschlag besteht darin, die Unähnlichkeit der Beobachtung i zu der ähnlicheren der beiden „vereinigten Beobachtungen“ h_1 und h_2 zu wählen. Im Beispiel hat für die Beobachtung Nummer 70 die Unähnlichkeit zu 64 den Wert 9.77 und zu 66 den Wert 10.80. Also wird ihre Unähnlichkeit zum Minicluster gleich 9.77.

In der Unähnlichkeitsmatrix ersetzen wir die beiden Zeilen und Spalten mit Nummern h_1 und h_2 je durch eine einzige mit den Werten $d\langle\{h_1, h_2\}, i\rangle$. Im Beispiel ist die Beobachtung 64 zu jeder anderen Beobachtung ähnlicher als 66. Die neue Matrix ist

```
>   dist(sqrt(t.d),method="manhattan")
      54    58    59    60    C1    67    68
58  3.89
59  2.05  3.27
60  9.73 10.60  7.67
C1 15.38 14.11 15.18 21.98
67 12.24 10.96 12.04 18.84  3.14
68  5.04  7.76  6.27 13.07 10.35  8.09
70 11.94  9.98  9.88 12.21  9.77  6.63  7.79
```

Nun können wir in der neuen Matrix wieder den kleinsten Unähnlichkeitswert bestimmen und die entsprechenden Elemente zu einem Cluster vereinigen. Im Beispiel zeigen die Beobachtungen 54 und 59 die Unähnlichkeit 2.05 und werden zu einem neuen Minicluster.

In den weiteren Schritten werden nicht nur Beobachtungseinheiten zu Mini-Clustern vereinigt, sondern auch Einheiten mit bereits geformten Clustern und Cluster mit einander. Als nächstes werden im Beispiel der erste Minicluster (64 und 66) mit der Beobachtung 67 vereinigt. ◁

c **Schema.** Dieses Vorgehen lässt sich in einen Algorithmus zusammenfassen:

0 Beginne mit der „Partition“, in der jede Beobachtungseinheit ein Cluster ist.

1 Vereinige die beiden Cluster mit der kleinsten Unähnlichkeit zu einem Cluster.

2 Berechne die Unähnlichkeit des neuen Clusters mit jedem verbleibenden Cluster. Verschiedene Formeln für diesen Schritt, genannt **Update-**Formeln – siehe 8.5.g – führen zu den verschiedenen Methoden der agglomerativen Clusteranalyse.

(It) Wiederhole 1 und 2, bis alle Objekte in einem Cluster vereinigt sind.

Es entsteht eine **Hierarchie**, d.h. eine Folge von „geschachtelten“ Partitionen. Zu jedem Vereinigungsschritt ℓ gehört ein **Index** d_ℓ – der Wert der „kleinsten Unähnlichkeit“.

d **Dendrogramm.** Alle Schritte dieses Verfahrens lassen sich grafisch in einem Dendrogramm festhalten. Zuunterst in Abbildung 8.5.d (i) stehen die beiden Beobachtungen, die im Beispiel zuerst vereinigt wurden, nämlich 64 und 66, als „Wurzeln“, die sich auf der Höhe von 1.03 zu einer einzigen Wurzel – zum ersten Minicluster – vereinigen. Später, auf der Höhe von 3.14, stösst Beobachtung 67 dazu. 3.14 war die Unähnlichkeit zwischen Beobachtung 67 und dem Cluster $\{64, 66\}$.

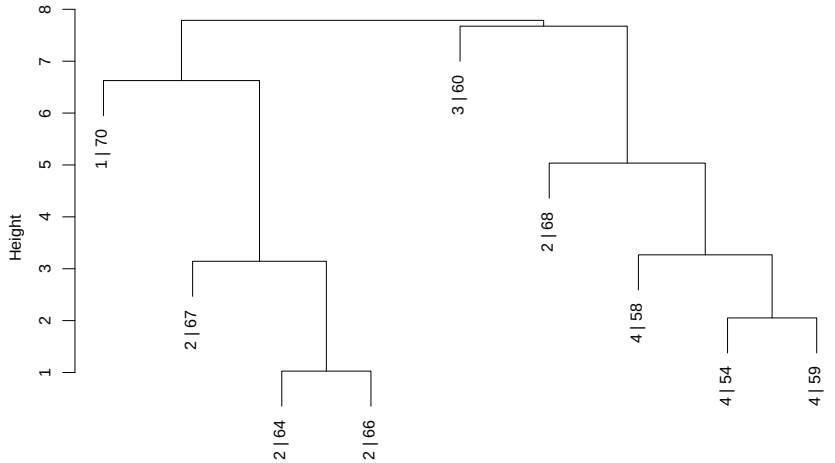

Abbildung 8.5.d (i): Dendrogramm für das Beispiel der 9 Beobachtungen

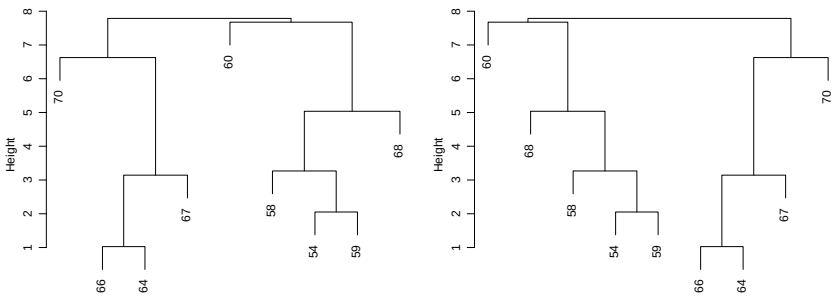

Abbildung 8.5.d (ii): Zwei weitere, gleichbedeutende Dendrogramme für das Beispiel

Die Anordnung der Objekte ist nicht eindeutig. Die gleiche Hierarchie kann durch recht verschieden aussehende Dendrogramme dargestellt werden, wie Abbildung 8.5.d (ii) zeigt. Die Freiheiten entsprechen denen eines „Mobiles“, an dem leichte Gegenstände zur Dekoration aufgehängt werden.

e **Gruppen-Einteilung.** Cluster-Analyse hat ja zum Ziel, Gruppen von ähnlichen Beobachtungen zu bilden. Ein Dendrogramm zeigt viele mögliche Einteilungen. Wenn man sich vergegenwärtigt, wie das dahinter stehende Verfahren abläuft, so wird klar, dass jede beliebige mögliche Anzahl Cluster in einem bestimmten Schritt gerade erreicht wird. Am Anfang sind ja n „Cluster“ vorhanden, am Schluss noch einer, und wenn man nach k Schritten abbricht, hat man $n - k$ Cluster. Das Dendrogramm kann man sich dem entsprechend auf einer gewissen Höhe horizontal durchgeschnitten denken. Die dann noch zusammenhängenden Wurzeln bilden je einen Cluster. Schneidet man die abgebildeten Dendrogramme auf der Höhe 4, dann gibt das die Gruppen $\{64, 66, 67\}$ und $\{54, 59, 58\}$; die übrigen drei Beobachtungen bilden je einen weiteren Cluster für sich selbst, zusammen also $q = 5$ Cluster.

f **Unähnlichkeitsmasse zwischen Clustern.** Um die Unähnlichkeit zwischen Clustern (und zwischen Einzelbeobachtungen und Clustern) zu bestimmen, haben wir im Beispiel eine einfache Regel benützt, die dem ersten der folgenden Unähnlichkeitsmasse zwischen Clustern entspricht. Es gibt aber noch zwei weitere gebräuchliche solche Masse. Die Definitionen sind:

- single linkage (nearest neighbor): $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \min\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- complete linkage (farthest neighbor): $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \max\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- average linkage: $d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle = \text{ave}\langle d\langle i_1, i_2 \rangle \mid i_1 \in \mathcal{G}_1, i_2 \in \mathcal{G}_2 \rangle$
- * Naheliegend als Mass für die Unähnlichkeit zwischen Clustern wäre auch der Abstand ihrer Mittelpunkte. Es stellt sich für die Bildung von Dendrogrammen als ungeeignet heraus, siehe 8.5.h.

g **Updating.** In Schritt 2 des Grundschemas 8.5.c müssen jeweils Unähnlichkeiten zwischen einem neu entstandenen Cluster und den bereits vorhandenen berechnet werden. Für die genannten Unähnlichkeitsmasse lassen sich diese aus den vorhergehenden berechnen:

- single linkage:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \min\langle d\langle \mathcal{G}_1, \mathcal{H} \rangle , d\langle \mathcal{G}_2, \mathcal{H} \rangle \rangle$$
- complete linkage: Analog
- average linkage:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \frac{n_1 d\langle \mathcal{G}_1, \mathcal{H} \rangle + n_2 d\langle \mathcal{G}_2, \mathcal{H} \rangle}{n_1 + n_2}$$
- * Allgemein:
$$d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle = \text{Funktion}\langle d\langle \mathcal{G}_1, \mathcal{H} \rangle , d\langle \mathcal{G}_2, \mathcal{H} \rangle , d\langle \mathcal{G}_1, \mathcal{G}_2 \rangle , n_1, n_2, n_H \rangle$$
Mit dieser Formel kann man sogar den Mittelpunkts-Abstand durch „updating“ erhalten.

h Sinnvollerweise achtet man darauf, dass die neue Unähnlichkeit $d\langle \mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{H} \rangle$ mindestens so gross ist wie der kleinere der beiden Ausgangswerte $d\langle \mathcal{G}_1, \mathcal{H} \rangle$ und $d\langle \mathcal{G}_2, \mathcal{H} \rangle$ – sonst ergeben sich „verschlungene“ Dendrogramme. Für die ersten drei erwähnten Masse ist das gewährleistet, für den Mittelpunkts-Abstand dagegen nicht!

i ▷ Im Beispiel der Vegetationsstudie erhält man, ausgehend von der Euklidischen Distanz, mit „average linkage“ das Dendrogramm in Abbildung 8.5.i. Die Gruppen gemäss Originalartikel sind als erste Ziffern der „Wurzeln“ angegeben. Man sieht, dass die Gruppierung sehr gut mit einer übereinstimmt, die man mit Schneiden bestimmter Äste erhalten kann – allerdings mit unterschiedlichen Schnitthöhen. Die Gruppen 3 und 4 scheinen recht homogen zu sein, während die zweite Gruppe eher eine Sammlung schlecht passender Einheiten darstellt. ◁

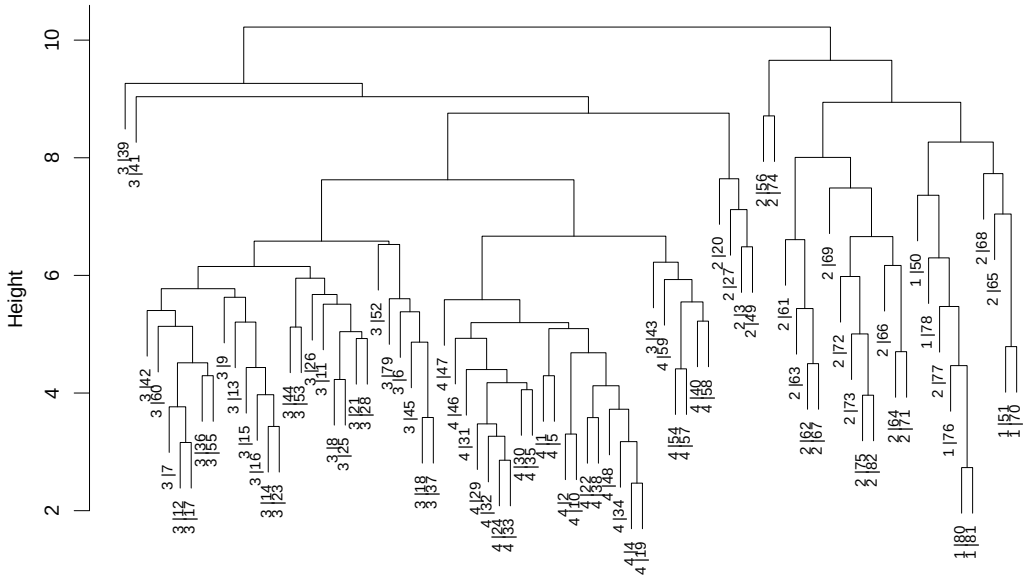

Abbildung 8.5.i: Dendrogramm der hierarchischen Clusterbildung mit „average linkage“, ausgehend von der Euklidischen Distanz

j **Eigenschaften.** Welches dieser Masse ist am sinnvollsten? Um diese Frage zu beantworten, stellen wir uns eine Situation mit zwei Gruppen unterschiedlicher Grösse vor und eine Beobachtung, die der einen oder anderen Gruppe angeschlossen werden soll. In Abbildung 8.2.c (i) liegt Beobachtung 68 näher bei der Gruppe $\{54, 58, 50\}$ als bei $\{64, 66, 67\}$, und zwar unabhängig davon, welches der drei Masse verwendet wird. Wäre es aber etwas weiter rechts, so dass die Distanz zum Punkt 67 kleiner würde als zu 58, dann würde es nach der single linkage-Regel zur anderen Gruppe geschlagen, während es nach dem complete linkage-Mass noch immer links angeschlossen würde – bis seine Distanz zu 66 kleiner würde als zu 54. Weitere solche Betrachtungen führen zu folgenden Beurteillungen:

- Mit single linkage können kettenförmige Cluster entstehen, deren „Enden“ sich sehr unähnlich werden.
- Dem gegenüber garantiert complete linkage, dass sich alle Mitglieder eines Clusters „nahestehen“. Es entstehen kompakte, tndenziell „kugelförmige“ Cluster mit ähnlichem „Durchmesser“.
- Average linkage stellt einen guten Kompromiss zwischen diesen Extremen dar.

k **Fazit.** Dendrogramme sind grafische Darstellungen, die recht viel Information enthalten und auch ästhetisch ansprechen. Sie sind deshalb recht beliebt. Wenn man aber das Ziel konsequent verfolgt, die Beobachtungseinheiten in Gruppen einzuteilen, muss man eine optimale Partition bestimmen. Gruppeneinteilungen durch Schneiden eines Dendrogramms sind nicht optimal – wie viel schlechter sie sind, hängt vom Datensatz und von der verwendeten Agglomerations-Methode ab.

Wie immer, wenn mehrere Methoden zur Verfügung stehen, ist die Versuchung gross, einige oder alle durchzuführen und nach gemeinsamen Schlüssen, hier also gemeinsamen Gruppen zu suchen. Das kann sinnvoll sein, wenn man bewusst Methoden wählt, die oft recht verschiedene Ergebnisse liefern, hier beispielsweise single linkage und average linkage (complete linkage hält der Autor nicht für empfehlenswert). Wenn dann beide Verfahren gleiche Gruppen ergeben, sind sie wohl recht klar gegeneinander abgegrenzt.

Dendrogramme graphisch zu vergleichen, ist allerdings schwierig; man müsste jeweils die Freiheiten, die man für die Dendrogramm-Darstellung hat (8.5.d) ausnützen, um die Anordnung der Beobachtungen möglichst ähnlich zu machen. Dafür kennt der Autor kein Programm.

Einfacher ist es, ein numerisches Mass für die Übereinstimmung von zwei Dendrogrammen zu definieren. Wir wollen das aber hier nicht ausführen.

l **Divisive Verfahren.** Statt von unten kann man auch versuchen, ein Dendrogramm von oben her zu entwickeln. Man startet mit einem einzigen Cluster, der alle Beobachtungen umfasst, und teilt diesen nach einer geeigneten Regel in zwei Cluster. In jedem weiteren Schritt wird einer der existierenden Cluster in zwei aufgespalten. So entsteht ebenfalls eine Hierarchie von möglichen Gruppen-Einteilungen, die durch ein Dendrogramm dargestellt werden können. Die Aufspaltung wird jeweils so vorgenommen, dass ein Unähnlichkeitsmass zwischen Gruppen möglichst klein wird.

- **Polythetische Verfahren.** Wenn dieses Mass, wie die bisher diskutierten, auf mehreren Variablen beruht, gerät man – ausser bei sehr kleinen Beobachtungszahlen – in rechnerische Schwierigkeiten bei der Optimierung, da es zu viele Aufteilungen eines Clusters in zwei Gruppen gibt. Man kann sich Auswege zurechtlegen, wie beispielsweise die Anwendung von k-means oder k-medians mit $k = 2$ Clustern in jedem Schritt.
- **Monothetische Verfahren.** Aufspaltung in jedem Schritt kann jeweils auf Grund einer einzigen Variablen erfolgen. Das hat den Vorteil, dass das Optimierungsproblem rechnerisch zu bewältigen ist und dass sich für die Anwendung sehr einfache Regeln ergeben, wo eine Beobachtung zuzuordnen ist – eine aus dem zugrunde liegenden Datensatz oder auch eine neue. Das Verfahren liefert einen **Entscheidungsbaum**, der einem klassischen Pflanzen-Bestimmungsschlüssel gleicht.

L Literatur zur Cluster-Analyse

a Kaufman and Rousseeuw (1990): Konzentriert sich auf 5 Programme, die es auch im R (library(cluster)) gibt. Benützerorientiert, einfach. Enthält aber auch gute Hinweise auf andere Methoden.
Weitere Methoden in R: Ripley, 1996, Kap. 9

b Deutsch-sprachige Bücher: Bock, 1974; Steinhausen and Langer, 1977; Späth, 1977; Späth, 1983; Deichsel and Trampisch, 1985

Weitere englische Bücher: Sokal and Sneath, 1963; Hartigan, 1975; Everitt, 1980; Gordon, 1981

8.S S-Funktionen

a **Unähnlichkeit.** Die Funktion `dist` liefert Euklidische und Manhattan-Distanzen (und noch ein paar andere).
`> t.dist <- dist(scale(sqrt(d.vegenv[,19:82]))), method="manhattan"`

Die Funktion `daisy` aus dem package `library(cluster)` eignet sich für Daten von gemischtem Typ.

b **Multidimensionale Skalierung.** Die Funktion
`> t.mds <- isoMDS(t.dist)` aus dem package `library(MASS)` lässt leider keine Gewichtung des Stress-Masses zu. Es minimiert

$$Q = \sum_{h,i} (g\langle d_{hi}\rangle - \|\underline{z}_h - \underline{z}_i\|)^2 \Big/ \sum_{h,i} \|\underline{z}_h - \underline{z}_i\|^2$$

Die Funktion `sammon` optimiert

$$Q = \sum_{h,i} \frac{d_{hi} - \|\underline{z}_h - \underline{z}_i\|)^2}{d_{hi}} \Big/ \sum_{h,i} d_{hi} \ .$$

Sie lässt also keine monotone Transformation g der Unähnlichkeiten zu.

c **Clusteranalyse: Optimale Partition**
`> kmeans(t.d, k=4)`
`R> t.cl <- pam(t.d, k=4, metric="manhattan"); plot(t.cl)`

Silhouetten werden von `pam` automatisch mitgeliefert und mit `plot(t.cl)` nach einer Darstellung der Hauptkomponenten gezeigt. Für `kmeans` muss man sie zuerst berechnen durch
`R> t.dist <- dist(t.d,method="manhattan")`
`R> t.sh <- silhouette(r.km$cluster,dmatrix=as.matrix(t.dist))`
Anschliessend erhält man die Grafik durch `plot(t.sh)`

d **Agglomerative Verfahren.**
`> t.cl <- hclust(t.dist, method="average")`
Dendrogramm:
`> plot(t.cl)`

e **Divisive Verfahren.**
`> t.cl <- diana(t.dist)`