Maschinelles Lernen in der klinischen Bioinformatik: Clustering I+II

Dr. Meik Kunz

Lehrstuhl für Medizinische Informatik

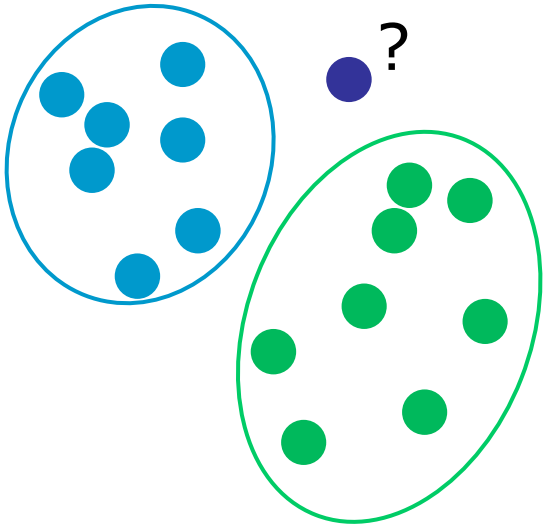
23.10.2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Machine Learning

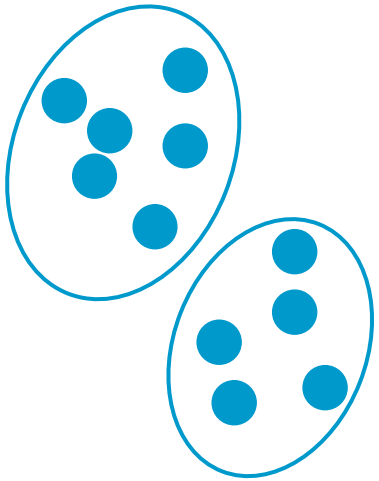
Machine Learning kann grob in drei Bereiche gegliedert werden:

Klassifikation



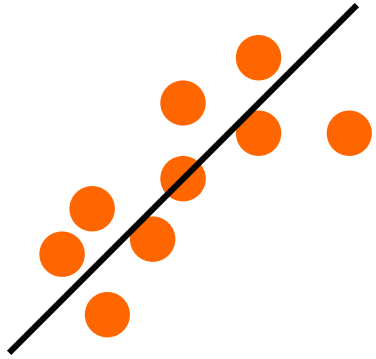
- Gruppen existieren

Clustering



- Keine Gruppen existieren

Regression



- Trends identifizieren

Multivariate Analysen

- Ausgangspunkt
- Große Datenmengen („Big Data“)
- Viele Einflussgrößen
- Komplexe Zusammenhänge
- Problemstellung
- (versteckte) Zusammenhänge herausfinden
- Hypothesen über Abhängigkeiten generieren
- Hypothesen über Zusammenhänge prüfen
- Neue Daten einordnen
- Vorhersagen treffen

Ansätze

- Hypothesengenerierend (Strukturen entdecken)
  - Clusteranalyse
    - Zahl der Objekte verringern
  - Faktoranalyse
    - Zahl der Variablen verringern
- Hypothesenprüfend
  - Regressionsanalyse
  - Varianzanalyse
  - Diskriminanzanalyse

Begriffe

- Objekt: individuelle Entität mit bestimmten Merkmalen
- Merkmal: Eigenschaft, Attribut
- Ausprägung: Messwert, Wert/Größe eines Merkmals
- Metrik: Abstandsmaß
- Cluster: Klasse von Objekten mit ähnlicher Merkmalsausprägung

Skalenniveaus: kategorial

- Nominalskala
- keine Rangfolge (gleich, ungleich)
    - Geschlecht, Farbe, Ort, PLZ
- Ordinalskala
- Rangfolge (kleiner, gleich, größer)
    - diskrete Werte: Noten, Einstufungen (z.B. Rangskalen)

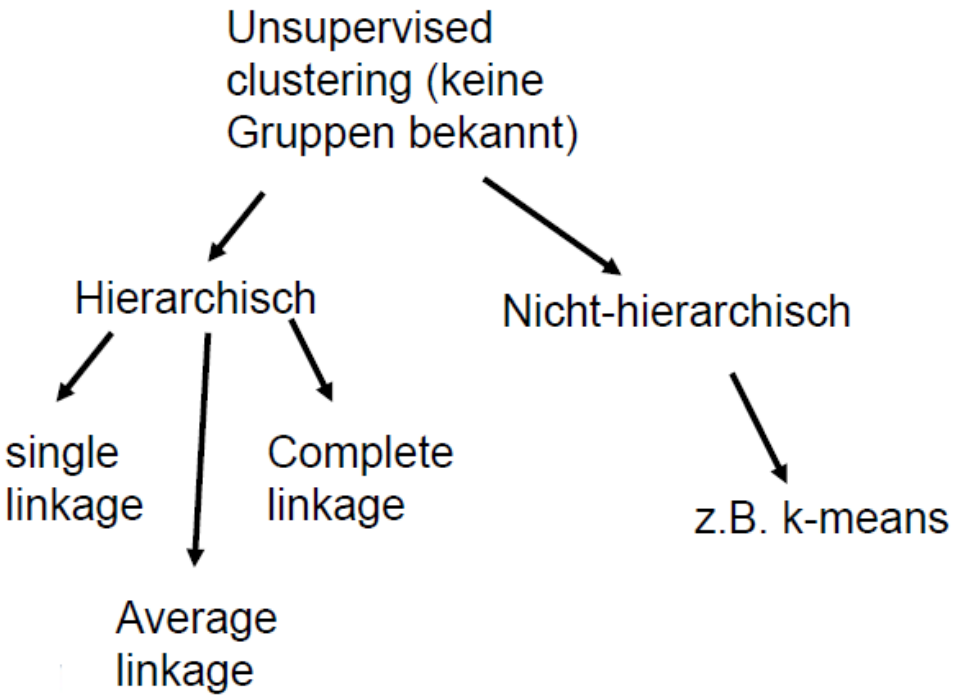
Skalenniveaus: metrisch

- Intervallskala
- Abstände zwischen Werten exakt (Bildung von Differenzen)
    - Temperatur (Celsius), Zeitpunkte
- Proportional-/Verhältnisskala
- Absolutskala (rationale Werte, Verhältnisse)
    - physikalische Größen, Temperatur (Kelvin), Anteile

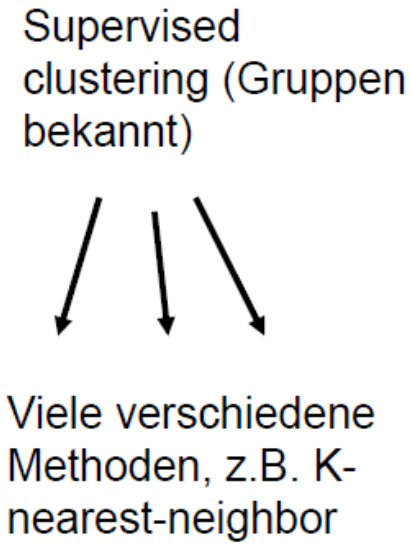
Wichtige Fragen für Interpretation der Ergebnisse

- Welche Annahmen wurden gemacht?
- Welche Methode wurde verwendet?

Clustermethoden



Gruppen finden, Einteilungsregeln erkennen



Einteilungsregeln erkennen

Welche Clusteranalyse?

- Zum Finden von bislang unbekannten Varianten einer Krankheit

→ unsupervised

- Zum Finden von Genen, mit denen man bekannte Krankheitsvarianten unterscheiden kann

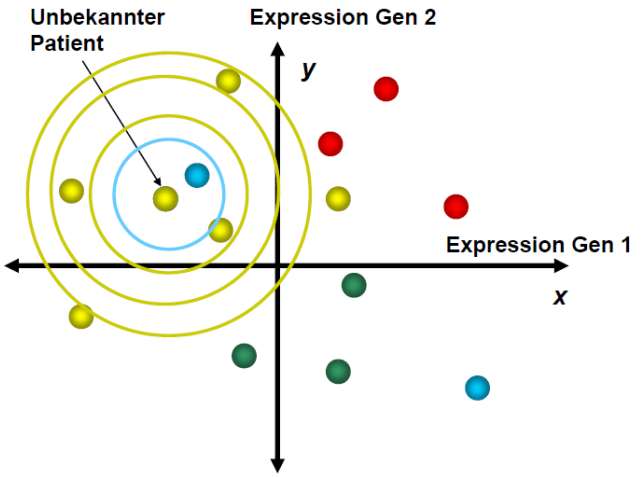
→ supervised

Supervised: K-nearest neighbour

- Eingabe bekannter Cluster, Klassifizierung neuer Objekte (z.B. Patienten)
- Algorithmus:

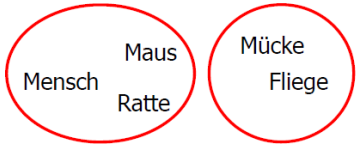
1. Finde die k-nächsten Nachbarn des Objektes

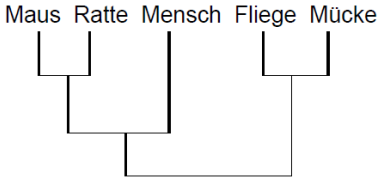
2. Objekt wird Cluster zugeordnet, dem Mehrheit der k-Nachbarn angehört



Einteilung unbekannt: Unsupervised Clustering

1. Nicht-hierarchisch: Die einzelnen Gene sollen nur in Gruppen eingeteilt werden, kein Baum





2. Hierarchisch: Baum wird erstellt

- divisiv: alle Punkte in einem Cluster, Cluster wird aufgeteilt bis es nur noch einen Punkt enthält
- agglomerativ: jeder Punkt ist ein Cluster, ähnlichsten Cluster werden zu einem kombiniert

unsupervised Clustering: k-Means

- nicht hierarchisch, partitionierendes Verfahren
- Anzahl Cluster (k) von Beginn festgelegt
- Jedes Cluster hat Clustermittelpunktes
- k-Means besonders bei *a priori*-Hypothese über die Anzahl der Cluster geeignet

unsupervised Clustering: k-Means

- Algorithmus:

1. Anzahl der Cluster festlegen

2. K-Cluster bilden (K-Punkte als Clustermittelpunkte wählen, K-Punkte nächstgelegenem Clustermittelpunkte zuordnen)

3. Mittelpunkt jedes Clusters berechnen

4. Teile Gene dem Cluster zu, dessen Mittelpunkt am Nächsten liegt

5. Wiederhole Schritte 2-4 bis sich nichts mehr ändert

unsupervised Clustering: PAM

- PAM=Partinioning Around Medoids
- Robustere Alternative zu k-Mean
- city block oder euklidische Distanz
- Algorithmus:

1. Anzahl der Cluster festlegen

2. K-Cluster bilden (K-Objekte als Clusterzentren (Medoids) wählen, K-Objekte nächstgelegenem Medoid zuordnen)

3. Berechne Summe der Distanzen der Objekte in allen Clustern

4. Wiederhole Schritte 2-3 bis sich nichts mehr ändert (Optimierung Summe Distanzen der Objekte zu den Medoids durch Austausch Medoids)

Divisiv hierarchisches Verfahren

- Top-Down-Verfahren
- Algorithmus:

1. Beginn: alle Objekte in einem Cluster
2. Bilde mit unähnlichstem Objekt neues Cluster
3. Wiederhole bis jedes Objekt eigenes Teilcluster bildet

Wie wählt man die Anzahl der Cluster?

- Silhouette:

Für jeden Objekt Pi wird die

• Mittlere Distanz $\hat{A}_i$ zu allen Objekten innerhalb des selben Clusters

• Mittlere Distanz $\hat{B}_i$ zu den Objekten des zweitbesten/nächsten Cluster

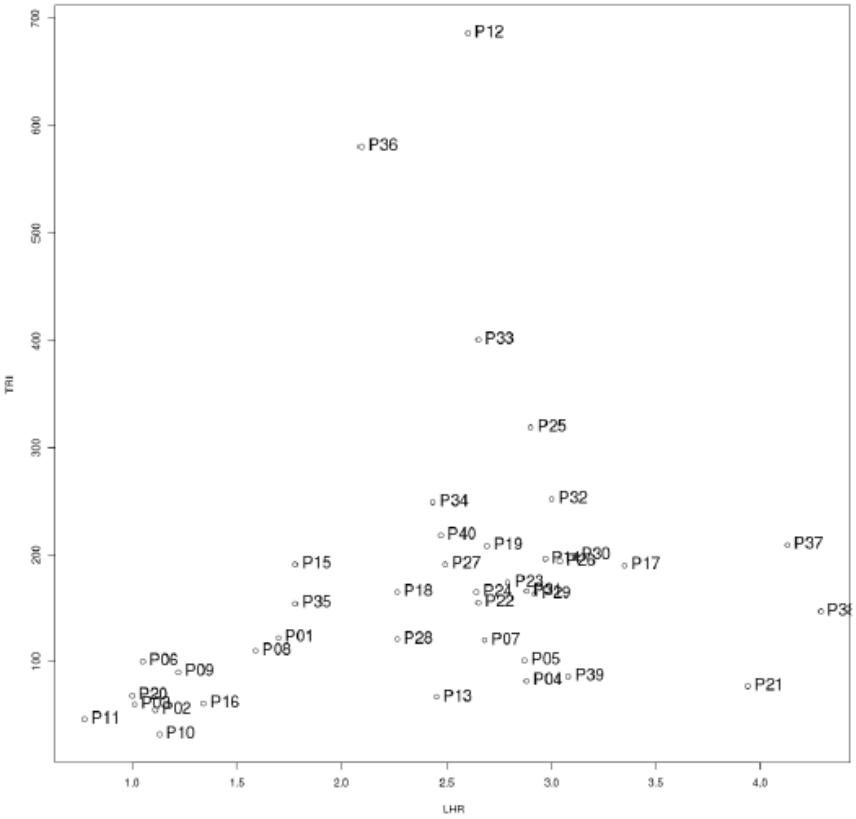
ergibt Silhouette: $$s(P_i) = \frac{\hat{B}_i - \hat{A}_i}{\max(\hat{A}_i, \hat{B}_i)}$$

und $-1 \leq s(Pi) \leq 1$

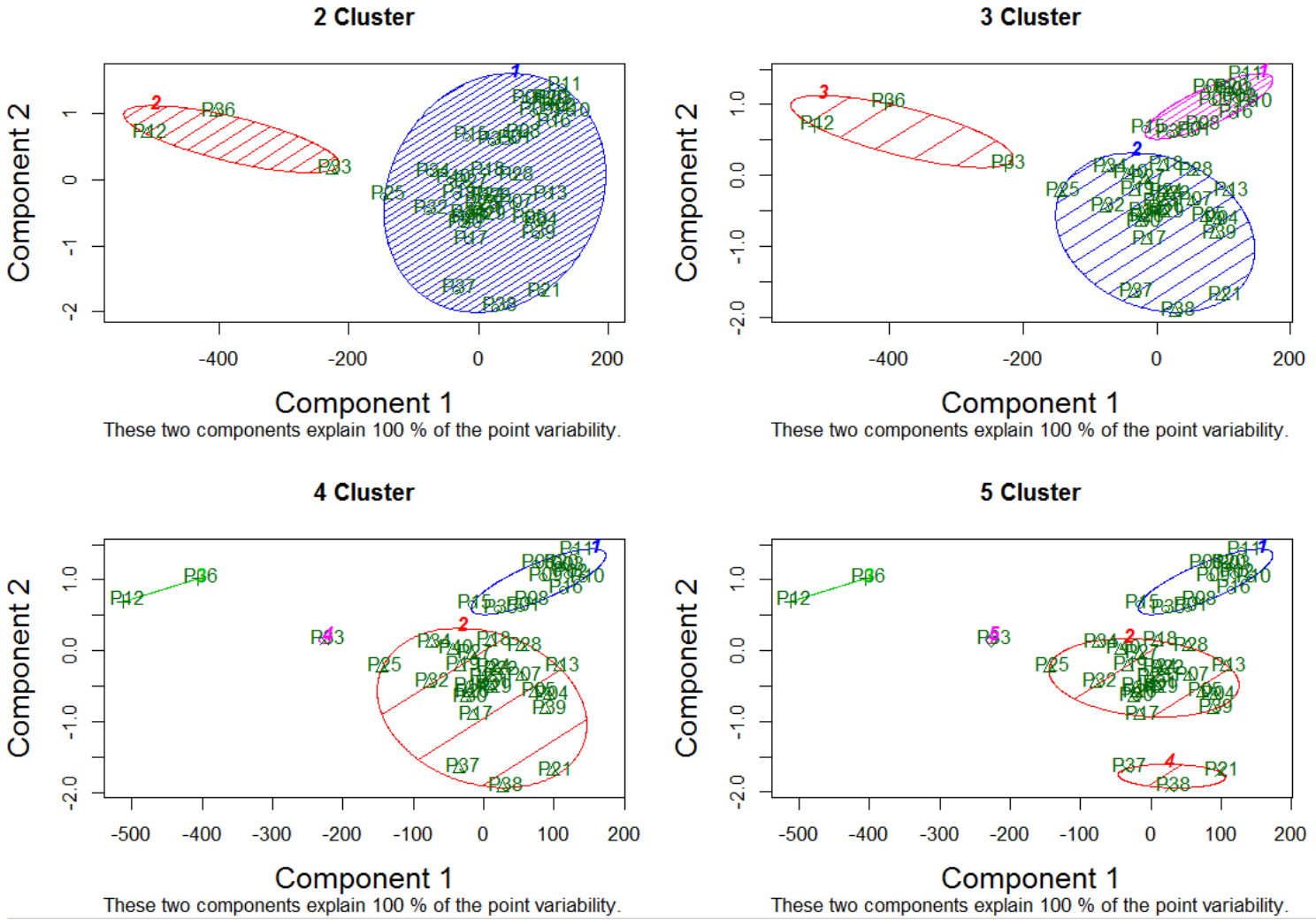
	Bewertung
0.71 - 1.0	optimale Clusterstruktur
0.51 - 0.70	gute Clusterstruktur
0.26 - 0.50	schwache Clusterstruktur evtl. artifiziell
< 0,25	unzureichende Clusterstruktur

Beispiel R: unsupervised Clustering

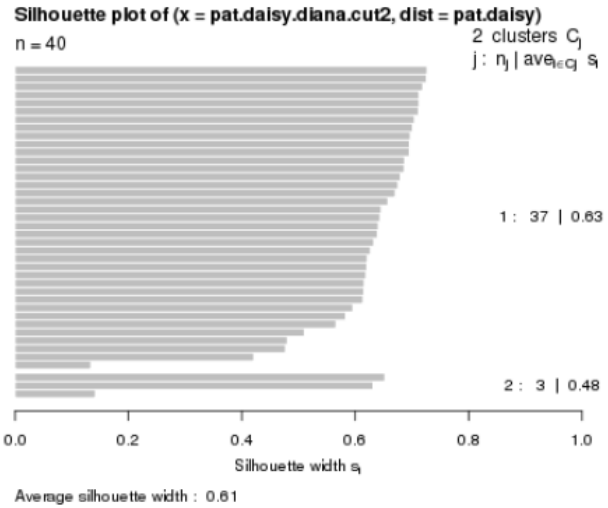
- Ansatz
  - Labordaten von Patienten, Risikopersonen, Gesunden
- Merkmale
  - 2 Analyte (Merkmale)
- Elemente
  - 40 zufällig ausgewählte Personen
- Daten
  - Stichproben

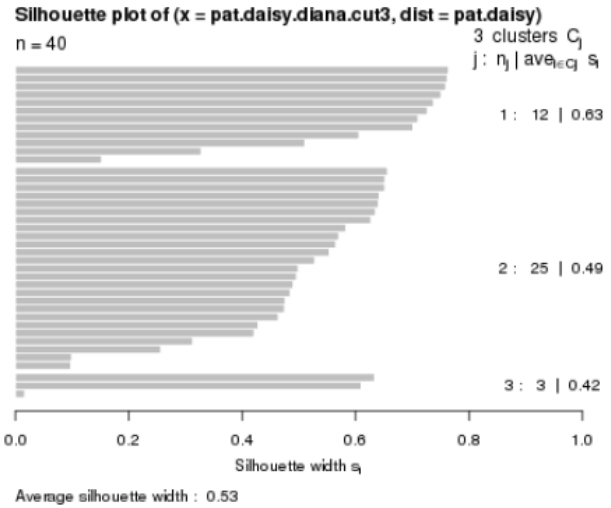


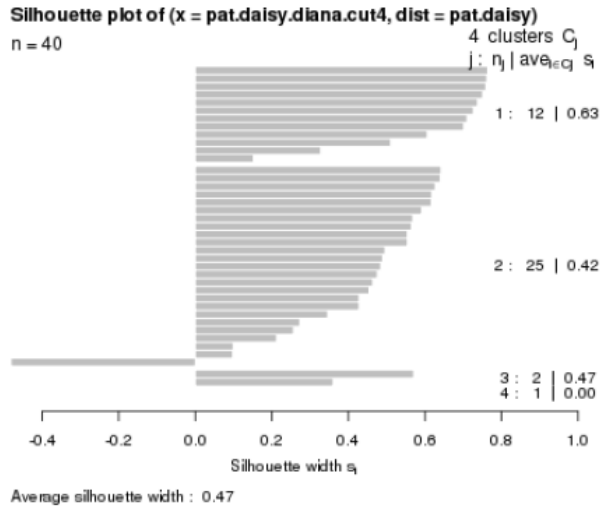
Divisiv hierarchisch: function cutree(pat.daisy.diana,k)

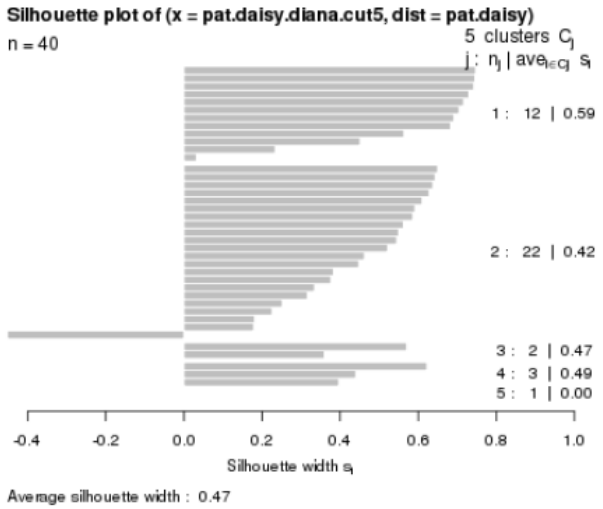


Silhouetteplot: function silhouette(pat.daisy.diana.cut…,pat.daisy)

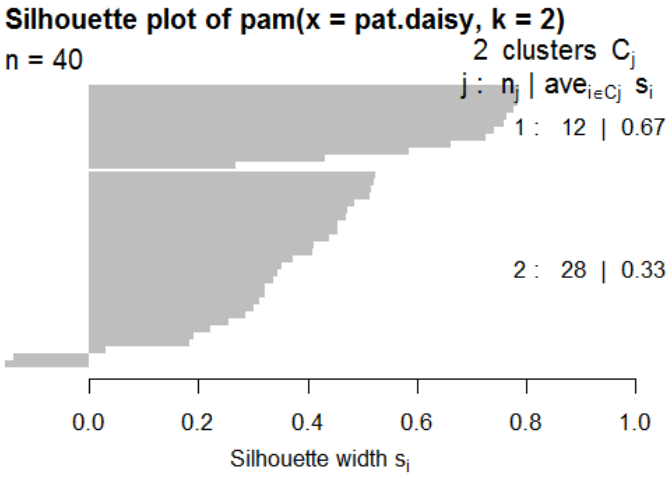


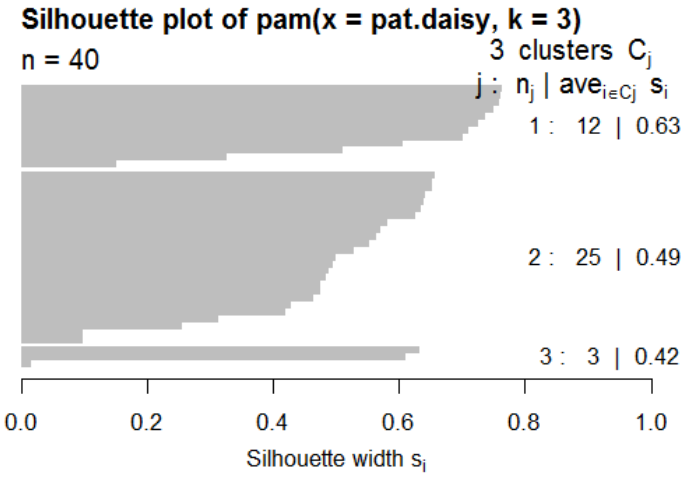


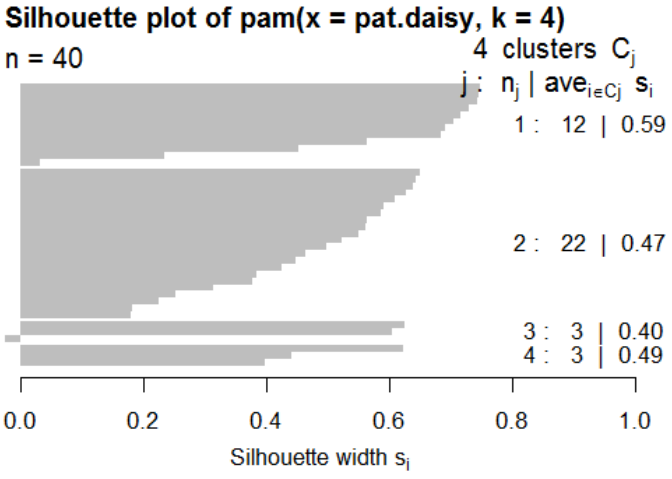


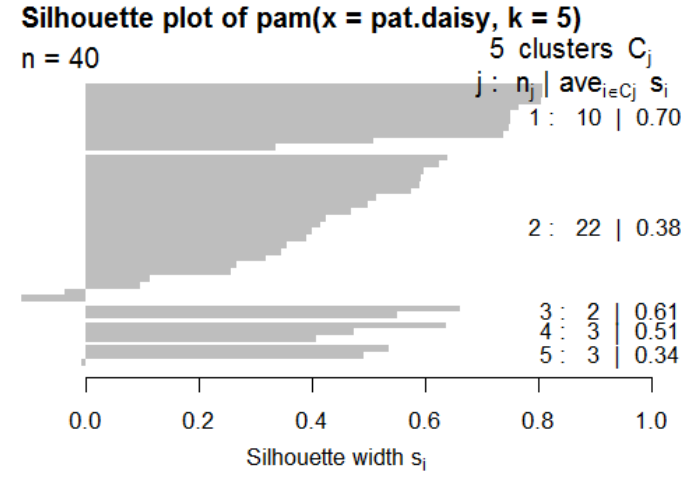


Silhouetteplot PAM: function silhouette(pat.daisy.diana.cut…,pat.daisy)

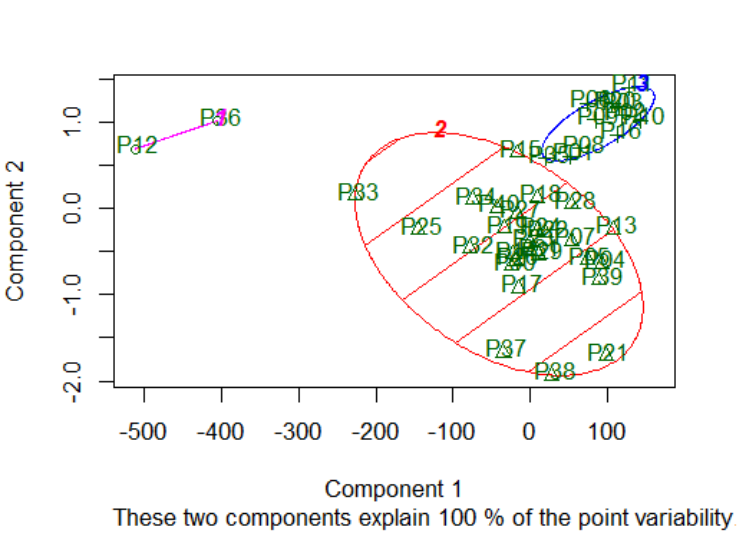




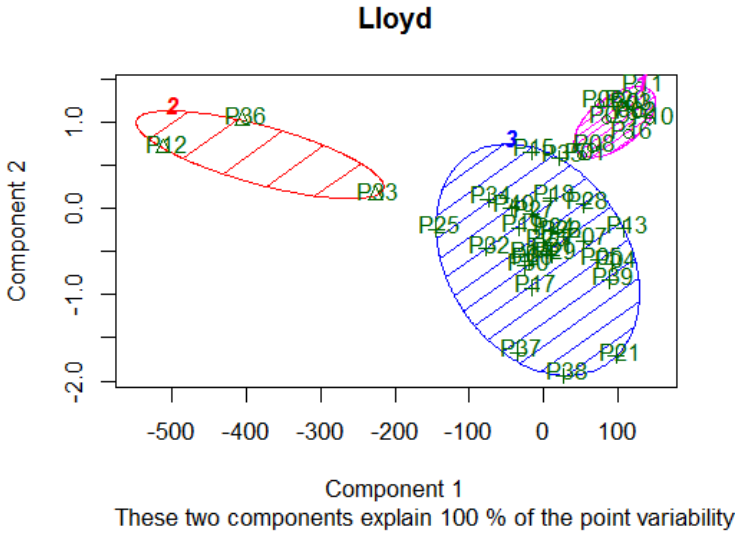




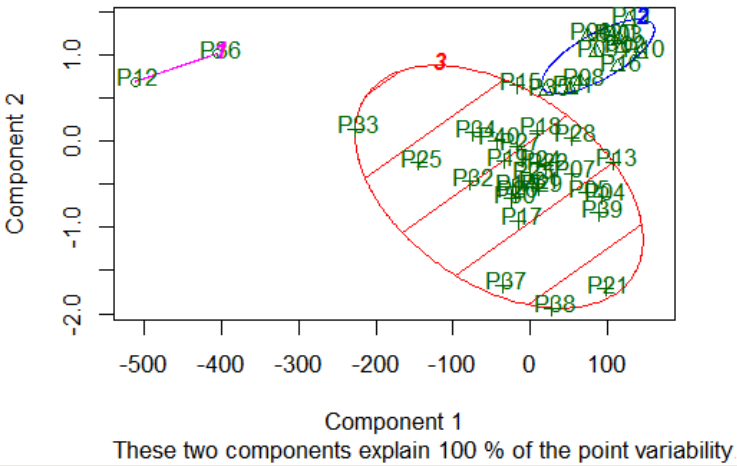
Silhouetteplot k-Means: function kmeans(pat.daisy,3,algorithm='...')



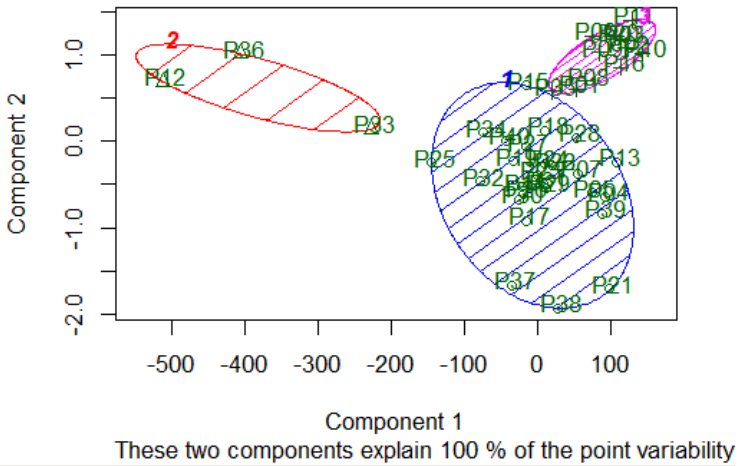
Lloyd



Hartigan-Wong



MacQueen



Unsupervised Hierarchisches Clustering

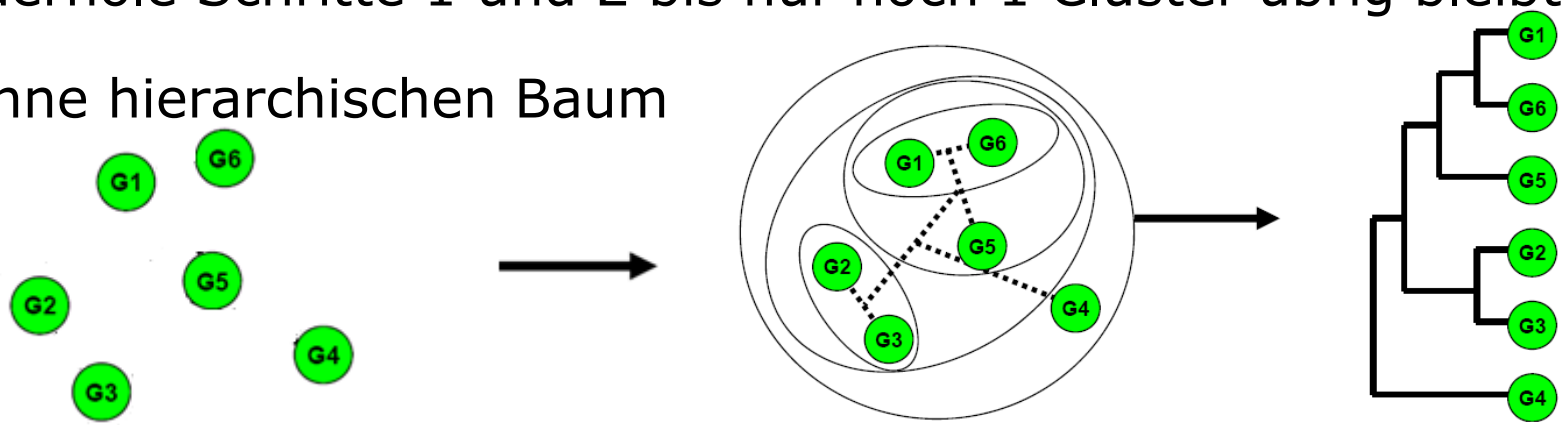
- Agglomeratives Verfahren (Bottom-Up-Verfahren)
- Algorithmus:

1. Jeder Punkt ist Cluster: Suche den kleinsten Abstand. Wenn mehrere Paare den gleichen Abstand (vorher festgelegte Regel)

2. Verbinde die 2 Cluster zu einem Neuen (größeren Cluster). Berechne die Abstände zwischen dem neuen Cluster und allen anderen.

3. Wiederhole Schritte 1 und 2 bis nur noch 1 Cluster übrig bleibt

4. Zeichne hierarchischen Baum

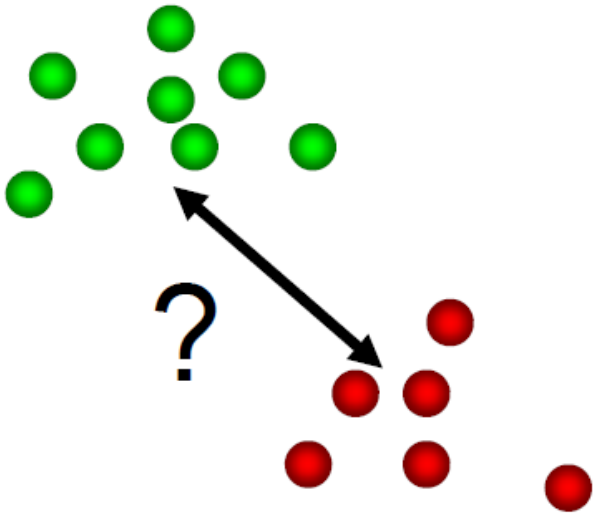


Fragen für Interpretation der Ergebnisse

- es kommt immer ein Baum raus, aber ist dieser biologisch sinnvoll?

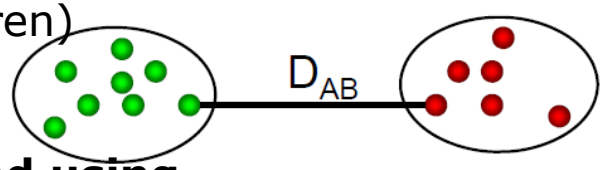
Wie verbinden wir die Cluster?

- Linkage-Methoden sind Regeln, nach denen der Abstand von einem Cluster zum nächsten gemessen wird, also wie Cluster miteinander verbunden werden.
- definieren unterschiedliche Abstände zwischen Clustern

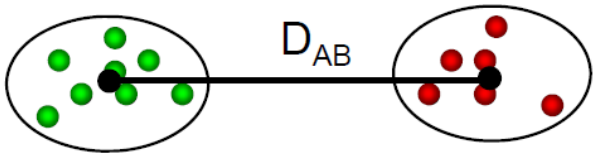


Linkage-Methoden: 3 weit verbreitete Methoden

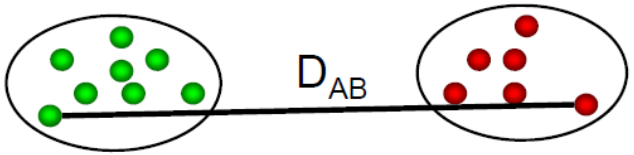
Single linkage (Nearest-Neighbour-Methode)
- Minimale Distanz zwischen Clustern
- tendieren wenige große Gruppen zu bilden, denen viele kleine Gruppen gegenüberstehen (gut um Ausreißer zu identifizieren)
- kategorial und metrische Skalen



Average linkage (WPGMA „weighted pair-grop method using arithmetic averages)
- Mittlere Distanz zwischen Clustern
- Nur metrische Skalen



Complete linkage (Furthest-Neighbour-Methode)
- Maximale Distanz zwischen Clustern
- tendieren verstärkt einzelne, etwa gleich große Gruppen zusammenzufassen
- kategorial und metrische Skalen



Weitere Methoden:
• Zentroid: Abstand der Clusterschwerpunkte (UPGMC, „unweighted pair-grop method using centroids“)
• Ward: minimierte Varianz (hohe Trennleistung, für metrische Daten optimal; euklidische Distanz, nur für unkorrelierte Merkmale)

Wie ist die Distanz zur Bestimmung des Abstandes zwischen zwei Objekten?

→ Distanzmaße

Minkowski-Metrik

- Distanz der Objekte A und B für die Ausprägungen x aller Merkmale p

$$d_{AB} = \left(\sum_{i=1}^{i=p} |x_{A,i} - x_{B,i}|^r \right)^{\frac{1}{r}}$$

r: Minkowski Exponent

Minkowski-Metrik: r=1 (Manhattan-Distanz)

- city-block Metrik: Distanz, wenn nur parallel zu den Koordinatenachsen gelaufen werden darf (wie in den Straßen von Manhattan)
- Betrag der Abstände: negative Distanzen vermeiden
- Ausreißer fallen weniger ins Gewicht

$$D = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Manhattan: Weg vom Central-Park zum UN-Gebäude

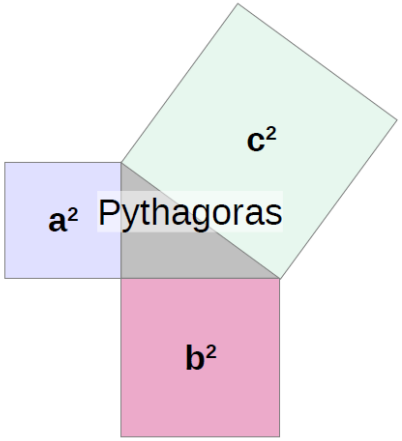
Minkowski-Metrik: r=2 (Euklidische Distanz)

- quadratischer Abstand: “Luftlinie” zwischen zwei Punkten

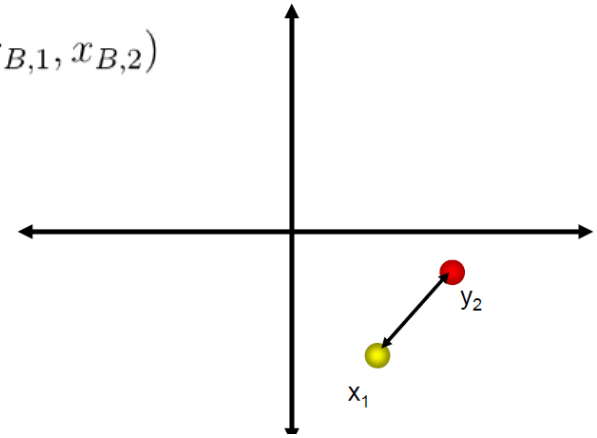
$$A = (x_{A,1}, x_{A,2})$$

$$B = (x_{B,1}, x_{B,2})$$

$$d_{AB} = \sqrt{(x_{A,1} - x_{B,1})^2 + (x_{A,2} - x_{B,2})^2}$$



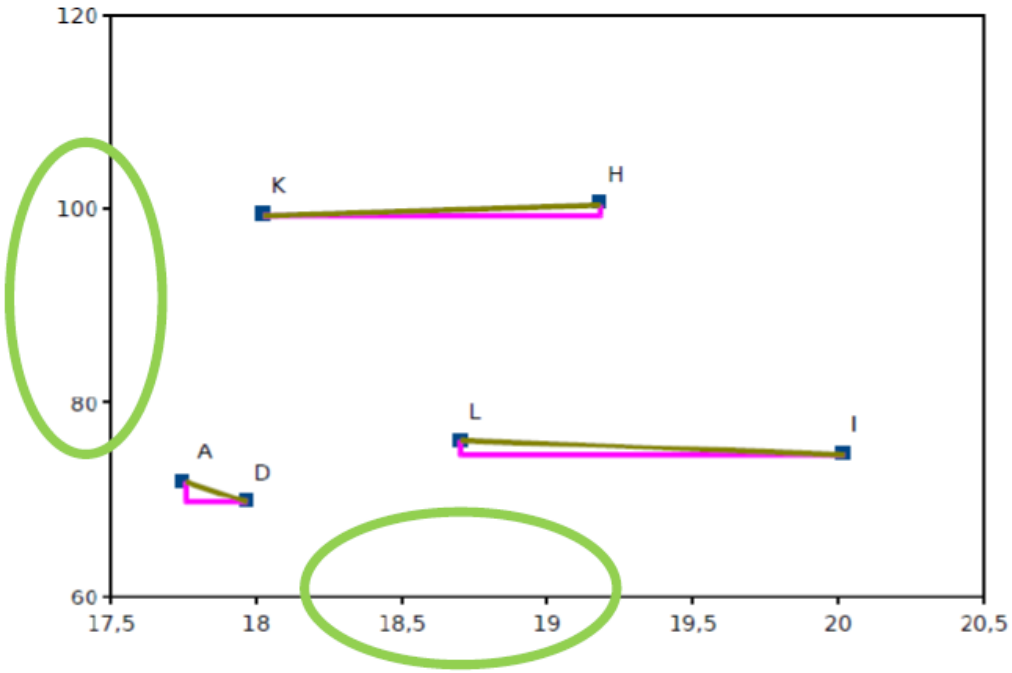
$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



- Problem Ausreißer: fallen durch Quadrierung stark ins Gewicht

Größenordnung wichtig für Gruppierung

- Problem: Sortierung abhängig von der Größenordnung und Metrik (linera/quadratisch)



	manhattan		euclidean	
	distance	rank	distance	rank
A↔D	2.23	3	2.02	5
H↔K	2.39	4	1.69	3
I↔L	2.53	5	1.79	4

Ansatz 1: Canberra-Distanz

- gewichtete Manhattan-Distanz
- Distanz d der Objekte *A* und *B* für die Ausprägungen *x* aller Merkmale *p*:

- Manhattan-Distanz

$$d_{AB} = \sum_{i=1}^{p} |x_{A,i} - x_{B,i}|$$

- wird gewichtet

$$d_{AB} = \sum_{i=1}^{p} \frac{|x_{A,i} - x_{B,i}|}{|x_{A,i}| + |x_{B,i}|}$$

Ansatz 2a: Standardisierung/Studentisierung

- z-Wert-Normalisierung (nach W.S. Gosset, „student“)

- Bezug auf Mittelwert und Standardabweichung

- Mittelwert normalisierter Daten ist 0, Varianz=1:

$$z = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{k}^{k=n}\left(x_k - \bar{x}\right)^2}}$$

$$\bar{z} = 0$$

$$s^2 = 1$$

Ansatz 2b: Standardisierung

- Standardisierung auf ein Werteintervall:

- -1 .. +1
- 0 .. +1

Intervall -1 .. +1

$$x_{s[-1..+1]} = \frac{2 \cdot (x - x_{min})}{x_{max} - x_{min}} - 1$$

$$x_{s[-1..+1]}(x_{min}) = -1$$

$$x_{s[-1..+1]}(x_{max}) = 1$$

- mit Bezug auf:

- Minimum und Spannweite
- Mittelwert
- Standardabweichung
- Spannweite
- Maximal-/Minimalwert

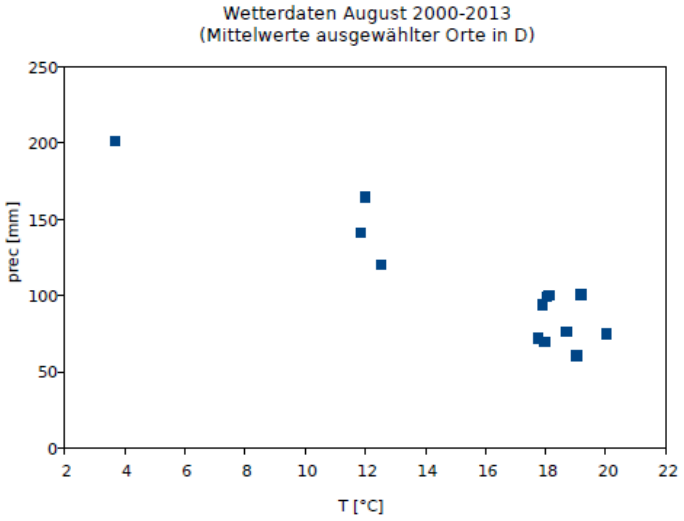
Intervall 0 .. +1

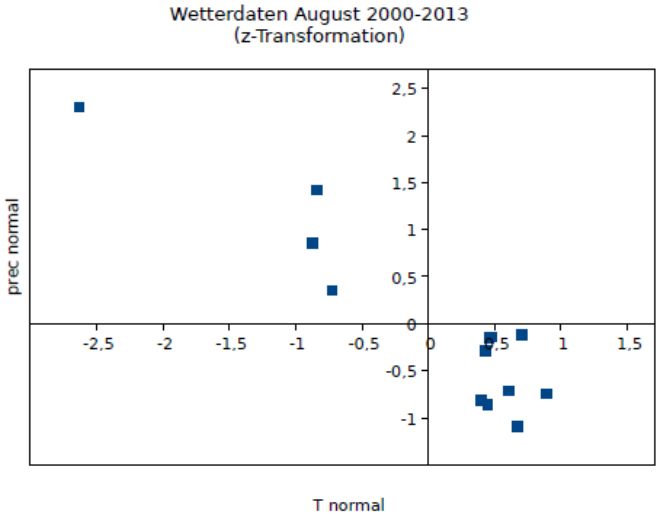
$$x_{s[0..+1]} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

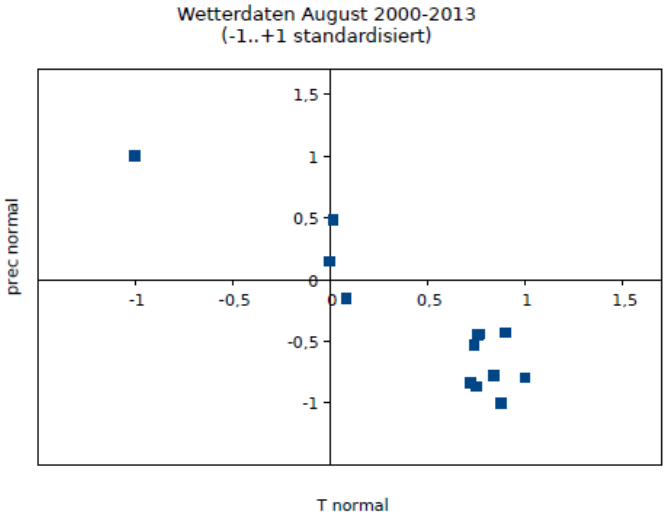
$$x_{s[0..+1]}(x_{min}) = 0$$

$$x_{s[0..+1]}(x_{max}) = 1$$

Standardisierung







Beispiel: Distanzmaße berechnen

	A	B
X	5	7
Y	6	1
Z	8	5

- Manhattan-Distanz $$d_{AB}=|A_X-B_X|+|A_Y-B_Y|+|A_Z-B_Z|=10$$

- Euklidische-Distanz $$d_{AB}=\sqrt{|A_X-B_X|^2+|A_Y-B_Y|^2+|A_Z-B_Z|^2}=\sqrt{38}=6,1644$$

- Canberra-Distanz $$d_{AB}=\frac{|A_X-B_X|}{|A_X|+|B_X|}+\frac{|A_Y-B_Y|}{|A_Y|+|B_Y|}+\frac{|A_Z-B_Z|}{|A_Z|+|B_Z|}=1.111722$$

Beispiel R: Distanzmaße, Standardisierung, Linkage

- Ansatz
  - Wetterdaten und geographische Lage
- Merkmale
  - mittlere Temperatur, Niederschlagsmenge, Sonnenstunden
- Elemente
  - 13 ausgewählte Orte in Deutschland
- Daten
  - Monatsmittel im August, gemittelt für 2000-2013

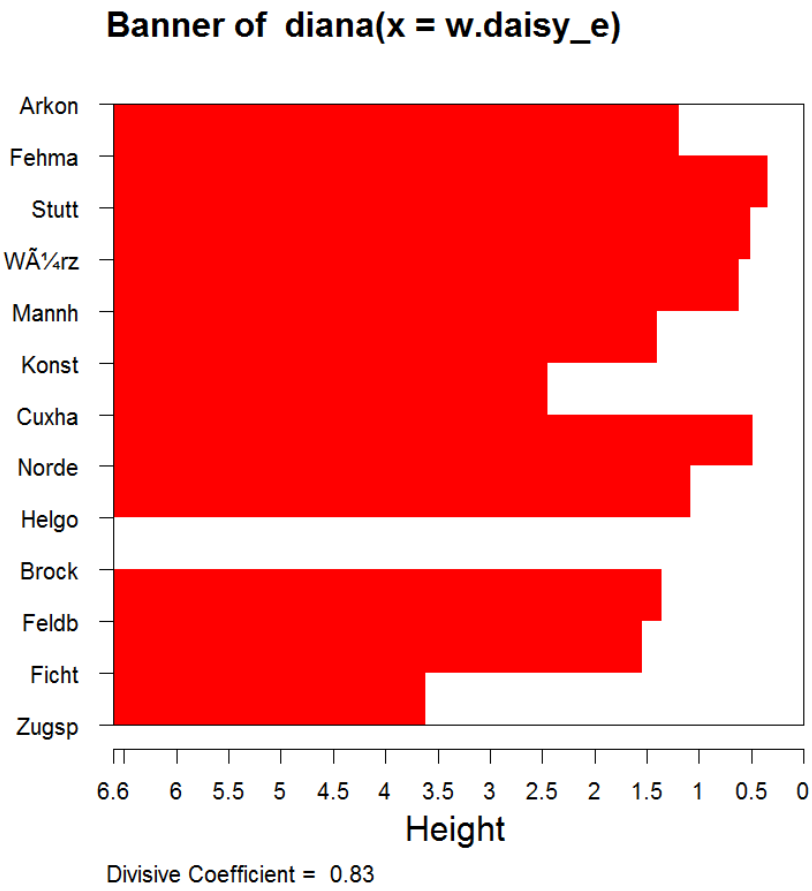
	Temperatur [°C]	Niederschlag [mm/m²]	Sonnenstunden [h]
A	17.74	71.99	232.46
B	11.85	141.24	166.19
C	18.11	99.96	192.34
D	17.96	69.99	218.19
E	11.99	164.76	186.07
F	12.53	120.34	190.64
G	17.89	93.86	211.08
H	19.18	100.81	230.30
I	20.01	74.93	214.76
K	18.02	99.57	200.95
L	18.70	76.14	221.62
M	19.02	60.66	220.39
N	3.69	201.24	178.34

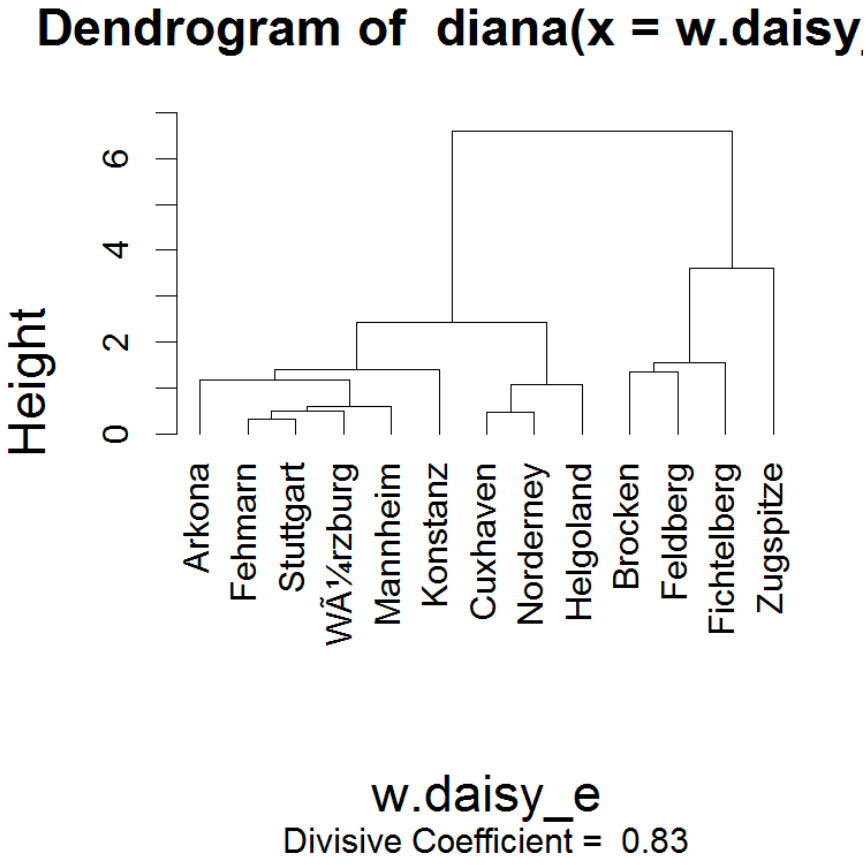
Erstellung der Proximitätsmatrix (Entfernungstabelle)

- Distanzen aller Objekte für jede Merkmalskombination ermitteln

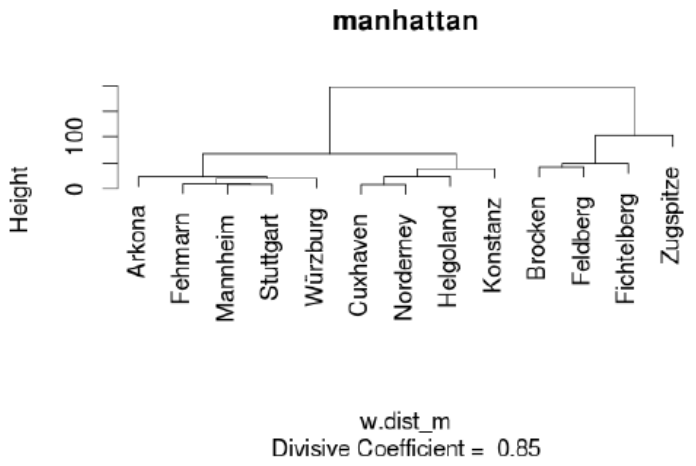
	Aachen	Augsburg	Bayreuth	Berlin	Bremen	Cottbus	Dresden	Erfurt	Essen	Frankfurt/Main	Frankfurt/Oder	Freiburg	Fulda	Garmisch-Part.	Hamburg	Hannover	Karlsruhe	Kassel	Kiel
Aachen		570	532	637	369	739	651	446	123	240	721	466	330	740	475	354	345	307	5
Augsburg	570		239	593	715	574	472	422	601	365	649	340	335	117	720	600	221	432	8
Bayreuth	532	239		352	572	339	237	187	494	264	414	457	1						
Berlin	637	593	352		375	125	214	288	480	564	91	800	4						
Bremen	369	715	572	375		496	478	351	249	450	467	722	3						
Cottbus	739	574	339	125	496		138	320	608	585	119	800	4						
Dresden	651	472	237	214	478	138		220	581	485	177	700	3						
Erfurt	446	422	187	288	351	320	220		367	268	366	533	1						
Essen	123	601	494	480	249	608	581	367		256	600	524	2						
Frankfurt/Main	240	365	264	564	450	585	485	268	256		661	262							
Frankfurt/Oder	721	649	414	91	467	119	177	366	600	661		873	5						
Freiburg	466	340	457	800	722	800	700	533	524	262	873		3						
Fulda	330	335	187	474	388	490	390	180	297	95	547	357							
Garmisch-Part.	740	117	334	686	856	686	586	515	735	502	759	490	4						
Hamburg	475	720	596	279	110	430	492	376	350	509	382	759	4						
Hannover	354	600	460	258	118	378	385	289	258	362	331	624	2						
Karlsruhe	345	221	337	670	595	670	570	403	397	135	743	130	2						
Kassel	307	432	278	367	288	502	402	135	188	190	440	457	1						
Kiel	556	820	693	343	205	511	573	459	454	599	446	852	5						

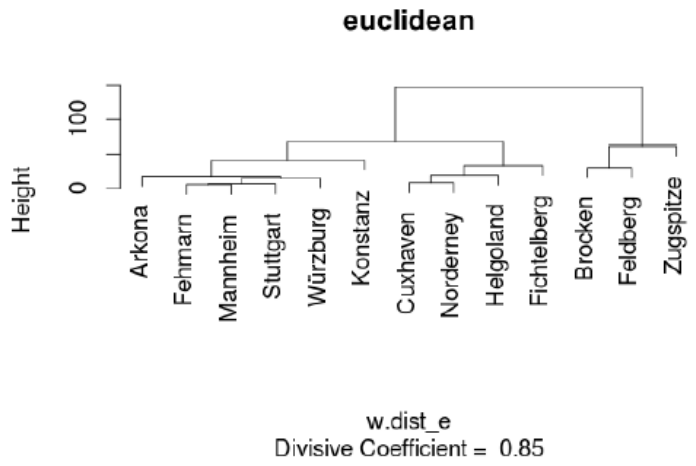
Divisives Cluster: euklidische Distanz

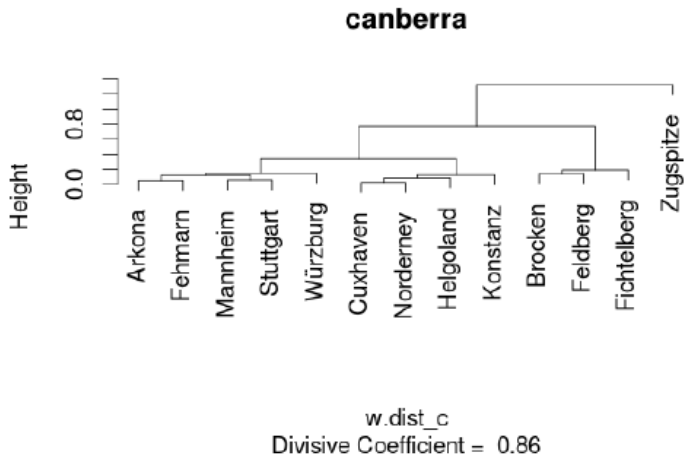


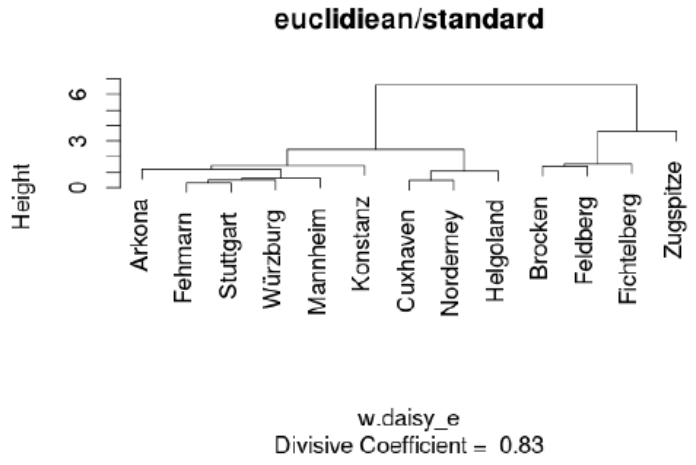


Divisive Cluster: Distanzmaße

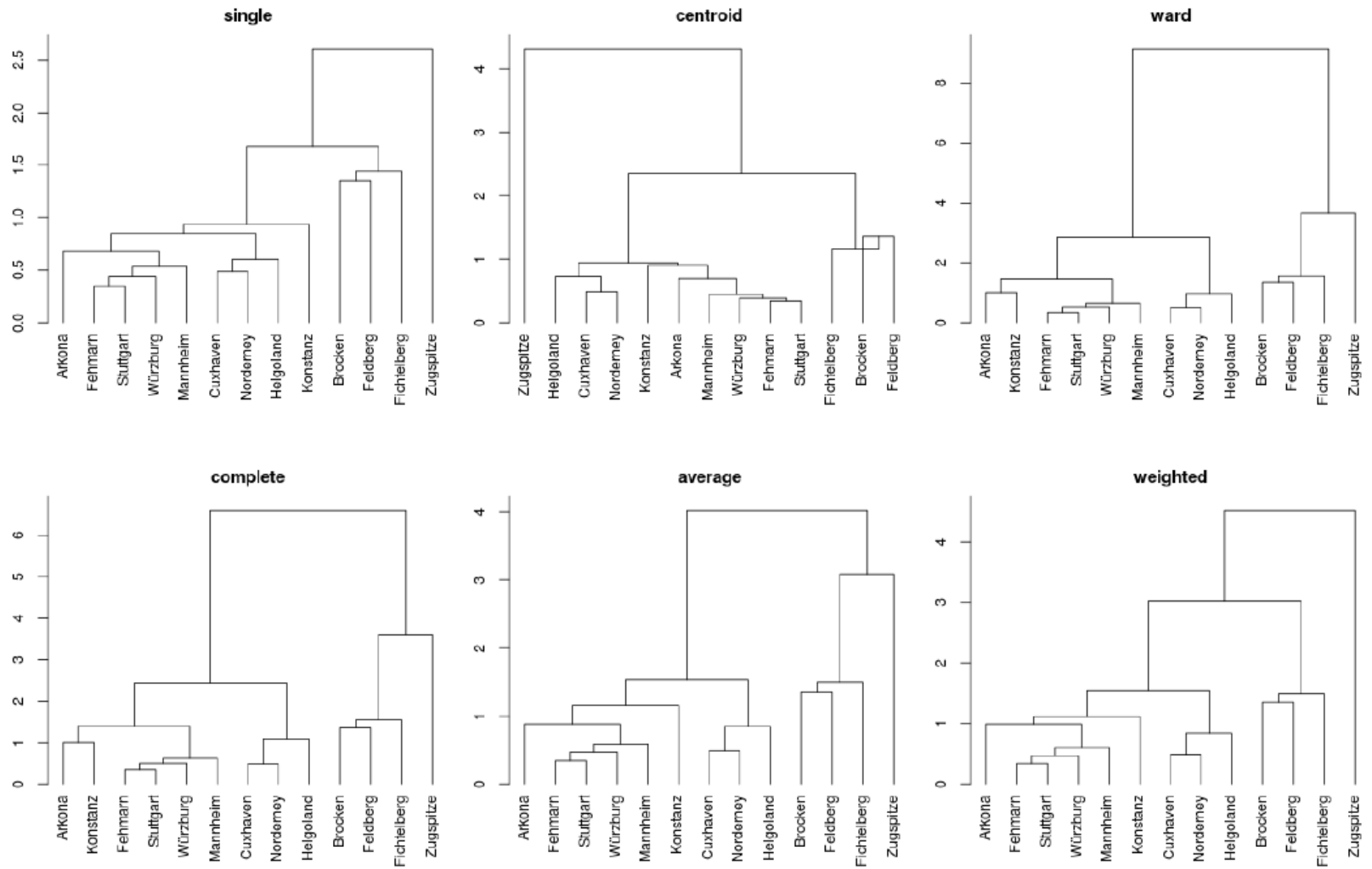




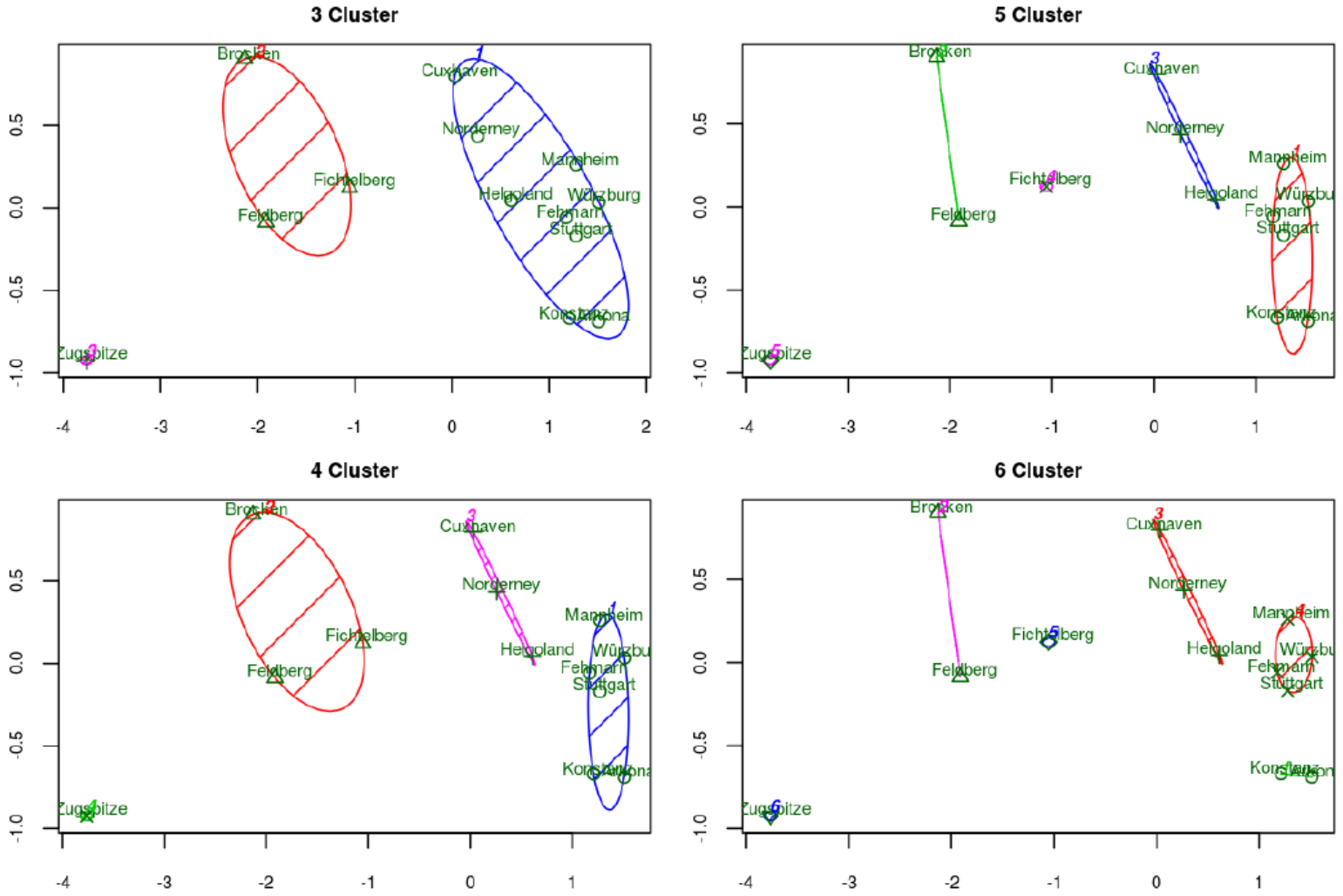




Agglomerative Cluster: Linkage-Verfahren



Agglomerative optimierte Cluster: Ward-Verfahren



Zusammenfassung: Methode bestimmt das Resultat