

DIN SPEC 38402-80



ICS 13.060.45

Einsprüche bis 2018-10-03

**Deutsche Einheitsverfahren zur Wasser-, Abwasser- und
Schlammuntersuchung –
Allgemeine Angaben (Gruppe A) –
Teil 80: Anwendung der Clusteranalyse für Wasseruntersuchungen (A 80)**

German standard methods for the examination of water, waste water and sludge –
General information (group A) –
Part 80: Application of cluster analysis for water analysis (A 80)

Méthodes normalisées allemandes pour l'analyse des eaux, des eaux résiduaires et des
boues –
Informations générales (groupe A) –
Partie 80: Application de l'analyse de cluster pour l'analyse de l'eau (A 80)

Dieser Entwurf mit Erscheinungsdatum 2018-08-03 wird der Öffentlichkeit zur Prüfung und Stellungnahme vorgelegt.

Weil die beabsichtigte Spezifikation von der vorliegenden Fassung abweichen kann, ist die Anwendung dieses Entwurfs besonders zu vereinbaren.

Stellungnahmen werden erbeten

- vorzugsweise online im Norm-Entwurfs-Portal von DIN unter www.din.de/go/entwuerfe bzw. für Norm-Entwürfe der DKE auch im Norm-Entwurfs-Portal der DKE unter www.entwuerfe.normenbibliothek.de, sofern dort wiedergegeben;
- oder als Datei per E-Mail an naw@din.de möglichst in Form einer Tabelle. Die Vorlage dieser Tabelle kann im Internet unter www.din.de/go/stellungnahmen-norm-entwuerfe oder für Stellungnahmen zu Norm-Entwürfen der DKE unter www.dke.de/stellungnahme abgerufen werden;
- oder in Papierform an den DIN-Normenausschuss Wasserwesen (NAW), 10772 Berlin, Burggrafenstr. 6, 10787 Berlin.

Die Empfänger dieses Entwurfs werden gebeten, mit ihren Kommentaren jegliche relevanten Patentrechte, die sie kennen, mitzuteilen und unterstützende Dokumentationen zur Verfügung zu stellen.

Zur Erstellung einer DIN SPEC können verschiedene Verfahrensweisen herangezogen werden:
Das vorliegende Dokument wurde nach den Verfahrensregeln einer Vornorm erstellt.

Gesamtumfang 45 Seiten

DIN-Normenausschuss Wasserwesen (NAW)

Inhalt

	Seite
Vorwort	3
Einleitung	5
1 Anwendungsbereich	6
2 Normative Verweisungen	7
3 Begriffe	7
4 Symbole und Abkürzungen	10
5 Einteilung von Clusterverfahren und Einsatzgebiet	11
5.1 Allgemeines	11
5.2 Datenvorbehandlung	14
5.2.1 Allgemeines	14
5.2.2 Häufige Verfahren der Datenvorbehandlung	15
5.3 Abstands- und Ähnlichkeitsmaße	16
5.3.1 Allgemeines	16
5.3.2 Minkowski-Distanz	16
5.3.3 Euklidische Distanz	17
5.3.4 Quadratische Euklidische Distanz	17
5.3.5 Chebychev-Distanz	17
5.3.6 Manhattan-Distanz	17
5.3.7 Power (Potenz)/Customized	18
5.3.8 Pearson-Distanz	18
5.3.9 Kosinus	18
5.3.10 Mahalanobis-Distanz	18
5.3.11 Einfluss der verschiedenen Distanzen	18
5.4 Hierarchische Clusterverfahren	19
5.4.1 Allgemeines	19
5.4.2 Divisive Verfahren	19
5.4.3 Agglomerative Verfahren	19
5.5 Nichthierarchische Verfahren	25
5.5.1 Allgemeines	25
5.5.2 Partitionierende (optimierende) Verfahren	25
5.5.3 Fuzzy-Clusterung	27
5.6 Dichtebasierte Verfahren	28
5.6.1 Allgemeines	28
5.6.2 CLUPOT	28
5.6.3 EM-Clusterung	28
5.7 Parametrische Clusterverfahren	29
5.7.1 Allgemeines	29
5.7.2 Unscharfe Clusterung nach Bocklisch [7]	29
6 Ausführungsbeispiel zur Clusteranalyse	32
6.1 Beispieldatensatz einer Fließgewässeruntersuchung	32
6.2 Beispiel - Ermittlung der Herkunft von Dioxinkontaminationen im Elbebereich	41
7 Diskussion der Ergebnisse und Empfehlungen	42
7.1 Diskussion der Ergebnisse	42
7.2 Empfehlungen	43
Literaturhinweise	45

Vorwort

Eine DIN SPEC nach dem Vornorm-Verfahren ist das Ergebnis einer Normungsarbeit, das wegen bestimmter Vorbehalte zum Inhalt oder wegen des gegenüber einer Norm abweichenden Aufstellungsverfahrens vom DIN noch nicht als Norm herausgegeben wird.

Erfahrungen mit dieser DIN SPEC sind erbeten:

- vorzugsweise als Datei per E-Mail an naw@din.de in Form einer Tabelle. Die Vorlage dieser Tabelle kann im Internet unter <http://www.din.de/stellungnahme> abgerufen werden;
- oder in Papierform an den DIN-Normenausschuss Wasserwesen (NAW), 10772 Berlin, Burggrafenstr. 6, 10787 Berlin.

Dieses Dokument wurde vom Arbeitskreis NA 119-01-03-01-04 AK „Chemometrie“ des Arbeitsausschusses NA 119-01-03 AA „Wasseruntersuchung“ im DIN-Normenausschuss Wasserwesen (NAW) erarbeitet.

Es ist erforderlich, bei den Untersuchungen nach diesem Dokument Fachleute oder Facheinrichtungen einzuschalten und bestehende Sicherheitsvorschriften zu beachten.

Bei Anwendung dieses Dokuments ist im Einzelfall je nach Aufgabenstellung zu prüfen, ob und inwieweit die Festlegung von zusätzlichen Randbedingungen erforderlich ist.

Das vorliegende Dokument enthält das vom DIN-Normenausschuss Wasserwesen (NAW) und von der Wasserchemischen Gesellschaft - Fachgruppe in der Gesellschaft Deutscher Chemiker (GDCh) - gemeinsam erarbeitete Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung:

Anwendung der Clusteranalyse für Wasseruntersuchungen (A 80).

Die als DIN-Normen veröffentlichten Deutschen Einheitsverfahren sind bei der Beuth Verlag GmbH einzeln oder zusammengefasst erhältlich. Außerdem werden die genormten Deutschen Einheitsverfahren in der Loseblattsammlung „Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung“ gemeinsam von der Beuth Verlag GmbH und der Wiley-VCH Verlag GmbH & Co. KGaA publiziert.

Normen oder Norm-Entwürfe mit dem Gruppentitel „*Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung*“ sind in folgende Gebiete (Haupttitel) aufgeteilt:

Allgemeine Angaben (Gruppe A)

Sensorische Verfahren (Gruppe B)

Physikalische und physikalisch-chemische Kenngrößen (Gruppe C)

Anionen (Gruppe D)

Kationen (Gruppe E)

Gemeinsam erfassbare Stoffgruppen (Gruppe F)

Gasförmige Bestandteile (Gruppe G)

Summarische Wirkungs- und Stoffkenngrößen (Gruppe H)

Mikrobiologische Verfahren (Gruppe K)

Testverfahren mit Wasserorganismen (Gruppe L)

Biologisch-ökologische Gewässeruntersuchung (Gruppe M)

Einzelkomponenten (Gruppe P)

Schlamm und Sedimente (Gruppe S)

Suborganismische Testverfahren (Gruppe T)

Über die bisher erschienenen Teile dieser Normen gibt die Geschäftsstelle des Normenausschusses Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V., Telefon 030 2601-2448, oder die Beuth Verlag GmbH, 10772 Berlin (Hausanschrift: Am DIN-Platz, Burggrafenstr. 6, 10787 Berlin), Auskunft.

Einleitung

Zur Auswertung der vielen bei Wasseruntersuchungen gewonnenen Daten gibt es zahlreiche statistische Methoden, so auch die Clusteranalyse, von der verschiedene Verfahren bekannt sind. Verfügbare Software zur Clusteranalyse bietet deshalb mehrere Auswahlmöglichkeiten an. Ohne das nötige Hintergrundwissen werden häufig Fehler bei der Datenvorbehandlung oder bei der Wahl des problemangepassten Abstandsmaßes gemacht. Die Ausführungen sollen dem Anwender eine Hilfestellung geben, indem die gebräuchlichsten Verfahren der Clusteranalyse vorgestellt, Vor- und Nachteile dargelegt und Empfehlungen zur Anwendung gegeben werden.

1 Anwendungsbereich

Clusteranalysenverfahren sind mathematisch-statistische und heuristische Verfahren zum Auffinden und Visualisieren von Strukturen sowie inneren Zusammenhängen in mehrdimensionalen Datensätzen. Sie ist ein Teilgebiet der multivariaten Statistik/Datenanalyse und klassifiziert die Elemente der untersuchten Grundgesamtheit anhand ihrer mehrdimensionalen Ähnlichkeit. Umfangreiche Datenmengen mit beliebiger Anzahl von Messgrößen lassen sich in Bezug auf ihre Ähnlichkeit übersichtlich darstellen.

Die zu untersuchenden Objekte (hier: Wasserproben) werden durch einen Satz vorgegebener, problembezogener Merkmale (Synonym für Messgrößen, Variablen, Messparameter) und ihre individuelle Ausprägung charakterisiert. Die Gesamtheit der Merkmalsausprägungen für ein Objekt wird als sein Muster bezeichnet. Die Clusteranalyse nutzt Ähnlichkeitsmaße, um über die Muster der Objekte eine Klassifizierung durchzuführen. Dabei wird die Objektmenge eines Datensatzes in homogene Klassen (Cluster) unterteilt, innerhalb derer alle Objekte ähnliche Variablenmuster aufweisen, während sich die Muster von Objekten verschiedener Klassen deutlich unterscheiden. Neben der Clusterung von Objekten nach ihrer Ähnlichkeit ist auch eine Clusterung der Merkmale/Variablen möglich. Im Objektdendrogramm werden die Ähnlichkeiten von Objekten in dem durch die Merkmale aufgespannten Raum dargestellt, während im Merkmalsdendrogramm die Ähnlichkeiten der Merkmale in dem durch die Objekte (über die Korrelationsmatrix) charakterisierten Raum dargestellt werden.

Zur Clusterung von Daten wurden abhängig von der Fragestellung verschiedene Clusteralgorithmen entwickelt. Sie können bei gleichen Datensätzen zu unterschiedlichen Ergebnissen führen.

Viele Verfahren der Clusteranalyse benötigen eine geeignete der konkreten Fragestellung angepasste Datenvorbehandlung, ohne die es zu Fehlern bei der Interpretation der Ergebnisse und somit zu falschen Aussagen kommen kann.

Das Ergebnis der Auswertung ist daher immer im Zusammenhang mit dem verwendeten Clusterverfahren und seiner Datenvorbehandlung zu sehen, das problem- oder fragestellungsbezogen gewählt werden muss.

Die vorliegende Anleitung betrachtet ausschließlich Clusterungen metrisch skalierten Variablen (z. B. Konzentrationen). Auf die Clusterung nominaler (trüb, durchsichtig), ordinaler (sehr wenig, wenig, viel, sehr viel) oder binärer Variablen wird nicht eingegangen, da sie in der Wasseranalytik nur eine untergeordnete Rolle spielen. Aus Gründen der Übersichtlichkeit werden nur die gebräuchlichsten Verfahren vorgestellt und diskutiert.

Folgende Methoden der Clusteranalyse werden näher beschrieben:

- Single Linkage
- Complete Linkage
- Unweighted Average Linkage
- Weighted Average Linkage
- Centroid Linkage
- Median Linkage
- Ward
- Flexible Strategie
- Zwei-Wege-Clusterung

- k-means
- Forgy
- Kohonen-Netze
- Fuzzy-c-means
- FCV-Clusterung
- CLUPOT
- EM-Clusterung
- Bocklisch

Anhand eines Beispiels aus der Wasseranalytik wird das Vorgehen bei der hierarchischen Clusteranalyse erläutert und dabei auf die Unterschiede bei der Verwendung verschiedener Ähnlichkeitsmaße, Clusteralgorithmen oder fehlender Datenvorbehandlung eingegangen. Zur Wahrung der Übersichtlichkeit werden bewusst nur wenige Proben und Messgrößen in diesem Beispiel verwendet. Deshalb ist die Anwendung einiger Clusterverfahren, die umfangreichere Datenmengen benötigen, nicht möglich.

2 Normative Verweisungen

Es gibt keine normativen Verweisungen in diesem Dokument.

3 Begriffe

Für die Anwendung dieses Dokumentes gelten die folgenden Begriffe:

3.1

Abstandsmaß

Algorithmus zur Bestimmung des Abstandes zwischen zwei Punkten im mehrdimensionalen Raum

3.2

agglomerativ

beginnend mit einzelnen Objekten werden immer größere Einheiten erzeugt

3.3

Ähnlichkeit

Maß für die Übereinstimmung von Merkmalsausprägungen von Objekten

Anmerkung 1 zum Begriff: Als Maß für die Ähnlichkeit mehrdimensionaler Objekte dient oft ihr Abstand im m-dimensionalen Raum (kleiner Abstand = große Ähnlichkeit).

3.4

binäre Variable

charakterisiert durch Belegung 0 oder 1

3.5

Cluster

Anhäufung von Objekten im mehrdimensionalen Raum, bei denen sich die Objekte innerhalb eines Clusters ähnlicher sind als zwischen verschiedenen Clustern

3.6

Datenmatrix

Datensatz

betrachtete Gesamtdatenmenge, bestehend aus n Objekten und m Merkmalen

3.7

Dimension eines Datensatzes

Zahl, identisch mit der Anzahl der Merkmale

3.8

divisiv

systematische Aufspaltung einer Datenmenge, beginnend von der Gesamtmenge in Teilgruppen, bis hin zu einzelnen Objekten

3.9

fuzzy

unscharf, Zuordnung nicht alternativ zu einzelnen Gruppen/Klassen, Zuordnung zu mehreren Gruppen/Klassen und mit unterschiedlichen Zugehörigkeiten möglich

3.10

Gruppe

<Clusteranalyse> Synonym für Klasse, zumeist Teilklasse

3.11

hierarchische Clusterung

einmal getroffene Zuordnung eines Objektes zu einem Cluster, die im weiteren Verlauf der Clusterung nicht mehr rückgängig gemacht wird

3.12

Kovarianzmatrix

Matrix, die die paarweisen Kovarianzen der m Merkmale enthält

3.13

Merkmal

<Clusteranalyse> Synonym für Messgröße, Variable, Parameter (in den Naturwissenschaften abweichend vom Sprachgebrauch der Statistik verwendet)

3.14

metrische Variable

Größe, dargestellt durch reelle Zahlen

3.15

Muster

Merkmalsvektor, Ausprägung der registrierten Merkmale/Messgrößen an einem Objekt

3.16

natürliche Klasse

Klasse, wie sie sich als mathematisches Ergebnis eines Klassifizierungsverfahrens (z. B. der Clusterung) ergibt

3.17

nicht hierarchische Clusterung

Clusterverfahren, bei denen die Zuordnungen von einzelnen Objekten zu Clustern im Verlaufe der Clusterung auch wieder rückgängig gemacht werden können

3.18

nominale Variable

Vertreter der kategorialen Variable zur Bezeichnung der Kategorie

BEISPIEL Einteilung von Wasser in Reinstwasser, Abwasser, Trinkwasser

3.19

Objekt

Fall

Element aus der Menge der Untersuchungsgesamtheit

Anmerkung 1 zum Begriff: Die genaue Beschreibung/Abbildung eines Objektes ist theoretisch nur durch unendlich viele Merkmale möglich; praktisch wird das Objekt durch eine Anzahl festgelegter Merkmale beschrieben.

3.20

ordinale Variable

Vertreter der kategorialen Variable zur Charakterisierung der Rangfolge

BEISPIEL süß, sehr süß

3.21

Parameter

<Clusteranalyse> Synonym für Merkmal, Analysenparameter, Messgröße

3.22

semantische Klasse

Klasse, die sich durch „sprachliche“ Einteilung ergibt

BEISPIEL Wasser vom Fluss A oder Wasser vom Fluss B

3.23

skaleninvariant

Eigenschaft eines Merkmals, unabhängig von a) der Skaleneinteilung der Merkmale (Wahl der Maßeinheit z. B. mg oder µg) und/oder b) von der Art der zu vergleichenden Merkmale (unterschiedlichste physikalische Dimensionen z. B. Konzentrationen, Ionenbeweglichkeiten)

3.24

Standardisierung/Autoskalierung

Datenvorbehandlung, bei der alle Merkmale durch Bildung der Differenzen zu ihren Mittelwerten zuerst zentriert und anschließend durch die Standardabweichung dividiert, skaliert werden

3.25

translationsinvariant

Eigenschaft eines Merkmals, unabhängig von der Lage des Koordinatenursprungs

3.26

univariat

Betrachtungsweise, bei der der Einfluss einer unabhängigen Messgröße auf die Zielgröße untersucht wird und bei der zur Vereinfachung alle anderen Messgrößen als konstant angenommen werden

3.27

Varianz

mittlere quadratische Abweichung aller Messwerte vom Gesamtmittelwert eines Merkmals

3.28

Zentrierung

Datenvorbehandlung bei der von allen Messwerten ihr Mittelwert subtrahiert wird

3.29

optimierende Clusterung

nicht hierarchisches Clusterverfahren, die Clusterbildung erfolgt durch Optimierung einer Zielgröße im Verlauf der Clusterung

3.30

parametrische Clusterung

Berechnung der Cluster aus Parametern, die aus Merkmalsverteilungen ermittelt werden

4 Symbole und Abkürzungen

Symbole	Bedeutung
A, B	Cluster A , Cluster B
a und q	frei wählbare Parameter
a_H	Parameter der unscharfen Clusterung, beschreibt Maximalwert der Zugehörigkeit
α, β, γ	Parameter für die flexible Strategie
b	Parameter der unscharfen Clusterung, beschreibt die am Rand ($x_k \pm c_{L,R}$) vorhandene Zugehörigkeit, als b_L und b_R
C	Kovarianzmatrix
c	Parameter der unscharfen Clusterung, mit dem die Ausdehnung des scharf betrachteten Gebietes von c_L bis c_R beschrieben werden kann
D	Abstandsmatrix
d_{ij}	Abstand zwischen Objekt i und j
$d_{\max}$	größter vorkommender Abstand zwischen zwei Objekten
g	Glättungsparameter (Halbwertsbreite)
i	Index für Objekt i
j	Index für Objekt j
k	Index für Merkmal
m	Dimension, Anzahl der ursprünglichen Messgrößen/Merkmale
n	Gesamtanzahl von Objekten in der Datenmatrix
n_A	Anzahl der Objekte im Cluster A
n_B	Anzahl der Objekte im Cluster B
n_i	Anzahl der Objekte im Cluster i
P	Wahrscheinlichkeits-/Verteilungsdichte
p	vorzugebende Potenz
r	vorzugebende Potenz der Wurzelfunktion

Symbole	Bedeutung
S	Ähnlichkeitsmatrix
s	Standardabweichung
s_{ij}	Ähnlichkeitsmaß für Objekt i und j
x_{ik}	Merkmalsausprägung k für Objekt i
$x_{k,A}^A$	Wert des Analysenvektors bezüglich Merkmal x_k und des Clusters A
$\bar{x}_{k,A}$	Mittelwert des Merkmals x_k im Cluster A
x^A	Analysenvektor (z. B. für ein zu betrachtendes Objekt)
$\bar{x}_k$	Mittelwert des Merkmals k
$x_k^{\max}$	Maximum des Merkmals k
$x_k^{\min}$	Minimum des Merkmals k
z	Parameter der unscharfen Clusterung, der den Abfall der Zugehörigkeit zum Rand als z_L und z_R hin beschreibt
L, R	Index für links/rechts bei parametrischer Clusterung (Bocklisch)
μ	Zugehörigkeit
μ_G	Gesamtzugehörigkeit
$\mu(x_k)$	Zugehörigkeitsfunktion des Merkmals x_k /Randverteilung bezüglich Merkmal x_k

5 Einteilung von Clusterverfahren und Einsatzgebiet

5.1 Allgemeines

Um vorhandene innere Strukturen bei mehrdimensionalen Daten aufzufinden, werden Ähnlichkeitsbetrachtungen genutzt, die alle Merkmale gleichermaßen berücksichtigen. Ziel ist es, eine Gruppierung/Aufteilung in Cluster zu finden, bei der sich günstigenfalls semantische Klassen (z. B. Elbe- oder Saalewasser) mit den natürlichen Klassen (mathematisches Ergebnis der Clusterung) in Übereinstimmung bringen lassen.

Es existieren zahlreiche Methoden zur Zusammenfassung von Objekten in Cluster. Eine Einteilung ist in Bild 1 gegeben. Man unterscheidet zwischen hierarchischen und nicht hierarchischen Verfahren, je nachdem, ob eine einmal getroffene Zuordnung eines Objektes zu einem Cluster wieder rückgängig gemacht werden kann. Die nichthierarchischen Verfahren werden auch als partitionierende (optimierende) Verfahren bezeichnet, da sie die Clusterbildung durch die Optimierung einer Zielfunktion erreichen und bei denen man eine Anzahl der zu bildenden Cluster vorgeben kann. Neben „scharfen“ Methoden, bei denen ein Objekt ausschließlich zu einem Cluster gehört, existieren auch Fuzzy-Methoden. Hier kann ein Objekt gleichzeitig zu mehreren Clustern gehören, aber nicht zwingenderweise mit gleichen Zugehörigkeitswerten. So kann anhand der Maximalzugehörigkeit zu einem Cluster die Einordnung vorgenommen werden. Fuzzy-Clusterungen liefern auch bei komplizierten Datenstrukturen, z. B. wenn sich Objekte zwischen den Clustern befinden, noch gute Klassifikationsergebnisse.

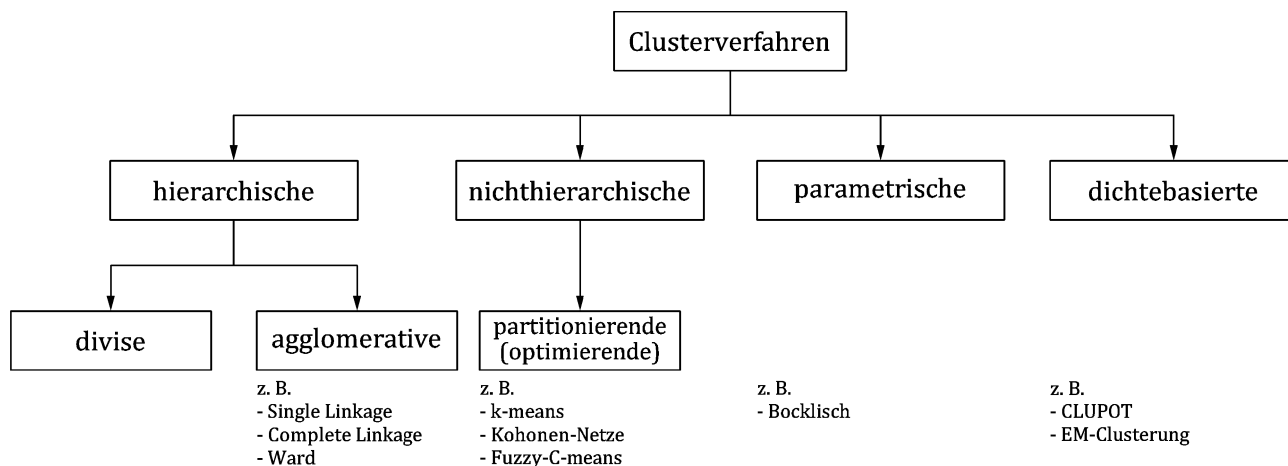


Bild 1 — Überblick ausgewählter Clusterverfahren

Die Ergebnisse hierarchischer Clusterverfahren (Merkmals- oder Objektclusterungen) werden in sogenannten Dendrogrammen dargestellt (siehe 5.4.3). Zur Veranschaulichung der nachfolgenden Ausführungen dient ein Datensatz der Untersuchungsergebnisse von Wasserproben.

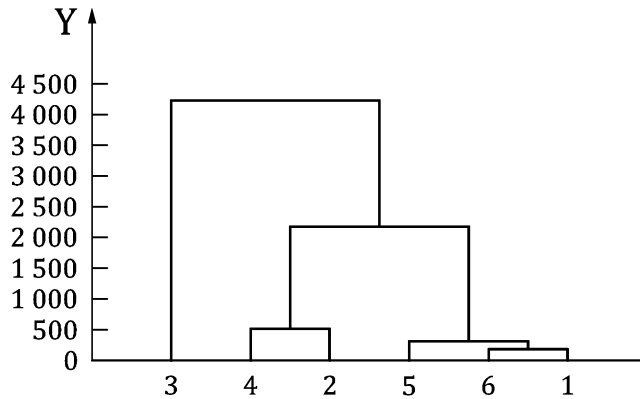
Bei einer Oberflächenwasserbeprobung verschiedener Gewässertypen wurden die in Tabelle 1 dargestellten Analysenergebnisse erhalten. Die Ähnlichkeit bei Berücksichtigung aller sieben Merkmale ist für die einzelnen Wasserproben im Objektdendrogramm (Bild 2a) und für die Messgrößen im Merkmalsdendrogramm (Bild 2b) dargestellt.

Tabelle 1 — Beispieldatensatz 6 Wasserproben (Originalmesswerte)

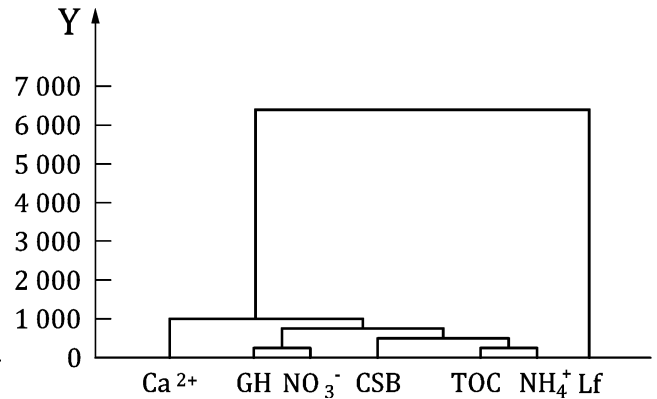
Merkmals- Objekt	Leit- fähigkeit (Lf) μS/cm	Gesamt- härte (GH) °dH	Ca²⁺ mg/l	CSB mg/l	TOC mg/l	NO₃⁻ mg/l	NH₄⁺ mg/l
Probe 1	420	10,3	60,4	77,8	29,3	13,9	0,24
Probe 2	1 610	51,3	307	19,2	6,8	32,5	0,31
Probe 3	3 430	82,9	431	30,7	7,3	42,4	2,14
Probe 4	1 100	29,3	192	12,0	2,6	7,7	0,06
Probe 5	96	1,8	7,6	10,9	4,3	5,6	0,02
Probe 6	280	7,0	35,9	21,8	7,2	1,9	0,05

Besonders ähnlich sind Merkmale, die schon bei geringen Distanzen zusammengefasst werden (siehe Bild 2b). Erwartungsgemäß sind sich die Merkmale CSB und TOC sehr ähnlich. In diesem Beispieldatensatz nimmt das Merkmal Leitfähigkeit aufgrund seiner hohen Zahlenwerte eine besondere Stellung ein. Die Ursache dafür liegt in der unsachgemäßen Verwendung unbehandelter Originaldaten der Tabelle 1 für die Clusteranalyse, in denen die Zahlenwerte der Leitfähigkeit bei der gewählten Einheit sehr hoch sind. Würde statt μS/cm die Einheit mS/cm verwendet, ergäben die tausendfach kleineren numerischen Werte andere Zusammenhänge, die Veränderungen im Dendrogramm zur Folge haben.

Nach dem Objektdendrogramm in Bild 2a wären sich die Proben 1 und 6 am ähnlichsten, dann die Probe 5. Der Probe 4 ist die Probe 2 am ähnlichsten. Probe 3 fällt aus dem Rahmen. Aber auch diese Betrachtung erfolgt anhand eines aus unbehandelten Daten erstellten Dendrogramms und kann deshalb bezüglich der mehrdimensionalen Ähnlichkeit falsch sein! Um die unterschiedlichen Messgrößen und Wertebereiche vergleichbar zu machen, ist unbedingt eine Datenvorbehandlung erforderlich.



a) Objektdendrogramm
(Ähnlichkeit der Wasserproben)



b) Merkmalsdendrogramm
(Ähnlichkeit der Messgrößen)

Legende

Y Linkage Distance
1 Probe 1
2 Probe 2
3 Probe 3

X Merkmal
4 Probe 4
5 Probe 5
6 Probe 6

**Bild 2 — Dendrogramme als Ergebnis der Clusteranalyse
(Ward, Euklidische Distanz, unbehandelte Daten)**

5.2 Datenvorbehandlung

5.2.1 Allgemeines

Unterschiedliche Merkmale (Messgrößen) decken im allgemeinen unterschiedliche Bereiche (Konzentrationsniveaus) und Streuungen (Werteschwankungen) ab. Dadurch werden die Ergebnisse bestimmter statistischer Auswertungen beeinflusst. Die Datenvorbehandlung hat das Ziel, durch geeignete Transformationen das Datenmaterial in eine Form zu bringen, in der derartige Verzerrungen nicht mehr auftreten. So darf es nicht sein, dass das Ergebnis der multivariaten Auswertung z. B. von der Wahl der Maßeinheit eines Merkmals abhängt [1] (siehe nachfolgendes Beispiel) oder Merkmale nur aufgrund hoher Absolutwerte im Vergleich zu anderen Merkmalen dominierenden Einfluss auf das Ergebnis haben. Die Wahl der Datenvorbehandlung ist bestimmt durch die Problem- oder Fragestellung. Es gibt keine Form der Datenvorbehandlung, die für alle Fragestellungen gleichsam zu bevorzugen ist.

Die Auswirkung der Wahl verschiedener Maßeinheiten auf die mehrdimensionale Ähnlichkeit werden am folgenden Beispiel erläutert:

Gegeben sind drei Objekte A, B, C mit ihren Messwerten für zwei Merkmale. Welche Objekte sind sich am ähnlichsten? (Die Varianten a) und b) unterscheiden sich bei gleichen Messwerten nur durch unterschiedliche Maßeinheiten/Wahl der Skaleneinteilung).

a) A = (1 µg/l; 4000 ng/l), B = (3 µg/l; 3000 ng/l), C = (4 µg/l; 1000 ng/l)

b) A = (1000 ng/l; 4 µg/l), B = (3000 ng/l; 3 µg/l), C = (4000 ng/l; 1 µg/l)

Tabelle 2 — Ergebnisse der Ähnlichkeitsberechnung

Euklidische Distanz	Variante	
	a)	b)
d_{AB}	1000	2000
d_{AC}	3000	3000
d_{BC}	2000	1000

Die Ähnlichkeitsbetrachtung (siehe Tabelle 2) über die Euklidische Distanz führt bei Variante a) zur größten Ähnlichkeit (kleinste Distanz) zwischen den Objekten A und B, in Variante b) hingegen zwischen B und C. Die willkürliche Wahl der Maßeinheit der Merkmale kann bei unbehandelten Daten einen dominanten Einfluss ausüben und damit zu fehlerhaften Ergebnissen führen. Deshalb ist unbedingt darauf zu achten, dass die Datenvorbehandlung merkmalsbezogen vor Beginn jeder Auswertung durchgeführt wird.

Bei Datensätzen mit Untergrundrauschen wie Spektren und Zeitreihen können die in 5.2.2. dargestellten Verfahren der Datenvorbehandlung nicht angewendet werden. Sie würden die Bereiche ohne Informationsgehalt, z. B. das Rauschen, gleichermaßen stark bewerten wie die informationsenthaltenden Teile des Spektrums. Hier haben sich weitere, dafür geeignete Verfahren der Datenvorbehandlung wie Spektrennormierung, Glättung, Multiplicative Signal Correction (MSC), Extended Multiplicative Signal Correction (EMSC), Standardisierung der Spektren (SVN) etabliert, die ausführlich in [2] beschrieben sind.