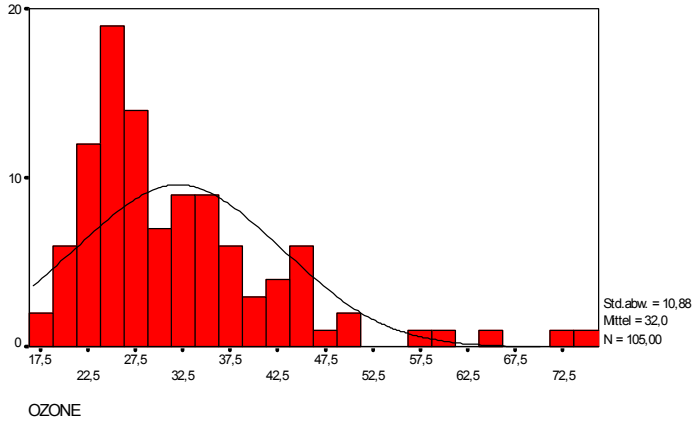
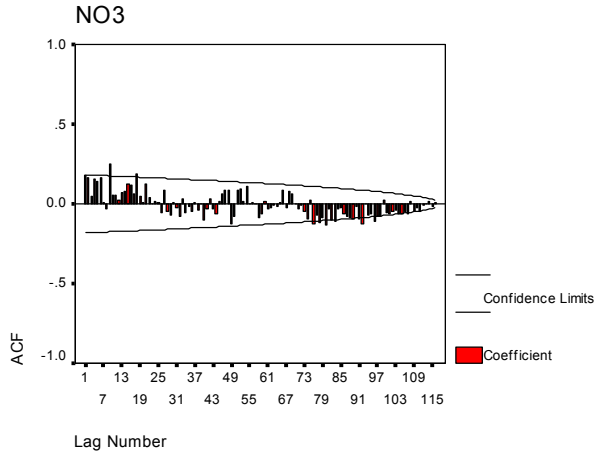


Freiberg Online Geology

FOG Vol.7 2002

FOG is an electronic journal registered under ISSN 1434-7512.



Integrierte Datenauswertung Hydrogeologie

Broder J. Merkel

Britta Planer-Friedrich

Institut für Geologie

Technische Universität Bergakademie Freiberg

Überarbeitet: 03/2003

Inhalt

1 Einleitung

Das Wort "*statista*" bedeutet im italienischen soviel wie *Staatsmann*. Gemeint ist damit im weitesten Sinne das Wissen des Staatsmannes über seine Ressourcen (Menschen, Tiere, Waffen, Ländereien etc.). Die Zählung der Ressourcen seitens der Herrschenden war somit die erste statistische Tätigkeit. Überliefert ist die älteste Volkszählung 3000 v. Chr. in Ägypten. Weitere Hinweise finden sich z.B. im Neuen Testament.

Wissenschaftliche Statistik wird in der Literatur erstmals um 1700 erwähnt. Sie wurde von Medizinern begründet und erst ab 1850 von anderen Disziplinen aufgegriffen.

Die kürzeste und prägnanteste Definition ist:

> *Determination of the probable from the possible*
>
> *(Bestimmung des Wahrscheinlichen des Möglichen)*

Eine moderne Definition des Leistungsspektrums der Statistik geht davon aus, dass alles, was mit der Gewinnung, Speicherung, Reduzierung, Verifizierung von Daten über die eigentlichen statistischen Verfahren bis hin zur Prognose (statistisches Modell) und Entscheidungsfindung zu tun hat, ebenfalls zur Statistik im weitesten Sinn gehört. Somit kann untergliedert werden in:

- Gewinnung von Daten
- Speicherung von Daten
- Reduzierung von Daten
- Beschreibung von Daten und Zusammenhängen
- Analyse
- Schätzung
- Hypothese
- Prognose
- Entscheidung

2 Datenerfassung und -speicherung

2.1 Erfassung von Daten

Bevor Daten gespeichert und/oder statistischen Prozeduren unterworfen werden können, müssen diese Daten vielfach erst gewonnen werden. Dabei ist das Design, also die Planung, in welcher Weise die Daten gewonnen werden, von entscheidender Bedeutung. Umgekehrt ist bei bestehendem Datenmaterial zu prüfen, ob die vorhandenen Daten überhaupt für die angestrebte Auswertung geeignet sind. Je nach wissenschaftlicher Disziplin und Fragestellung kann unterschieden werden in natürliche Messdaten und geplante Experimente im Bereich Naturwissenschaft und Medizin sowie systematische Erhebungen (Umfragen) im Bereich Volkswirtschaft und Sozialwissenschaft.

Unabhängig davon, ob die Daten selbst erhoben werden, oder dem Statistiker von Dritten zur Verfügung gestellt werden, gilt, dass Fehler, die bei der Gewinnung von Daten gemacht werden, nachträglich nur sehr schwer oder gar nicht mehr eliminiert werden können.

Als einfaches Beispiel dient die fehlende Motivation bei einer Meinungsumfrage:

Ziel sei die repräsentative Befragung der Gesamtbevölkerung über ein politisches Thema. Beim Befragen beliebiger Passanten vormittags in der Fußgängerzone macht der Untersucher zumindest zwei schwerwiegende Fehler:

- Der Kreis der Personen, der sich vormittags in einer Fußgängerzone einer Stadt aufhält, ist sicherlich nicht repräsentativ für die Gesamtbevölkerung, da z.B. Schüler und Berufstätige kaum angetroffen werden.

- Nicht alle angesprochenen Personen werden bereit sein, Auskunft zu geben; dadurch erfolgt eine nochmalige Selektion aus der Grundgesamtheit.

Als 2. Beispiel wird die Probenahme von Bodenproben angeführt: Wenn aus Zeitgründen oder aus Bequemlichkeit die Probenahme vom Auto aus erfolgt, so läuft der Probenehmer Gefahr, wenig repräsentative Proben zu bekommen, denn die Proben können einerseits durch die Straße bereits verfälscht sein (z.B. erhöhte Bleigehalte) und andererseits wird das vorhandene Straßennetz nicht unbedingt alle Bereiche (z.B. Bodentypen) eines zu untersuchenden Areals abdecken.

Dieser Effekt der persönlichen Bequemlichkeit bzw. auch dessen, was z.B. finanziell machbar ist, darf nicht unterschätzt werden. Viele Datensätze sind somit einseitig verfälscht (*biased sample*). Eine systematische Probenahme gibt "*unbiased samples*" aber auch nur bei ideal statistischer Anordnung.

Es besteht zudem ein enger Zusammenhang zwischen natürlicher Variabilität von Proben (*sampling error*) und Probenahmefehler (*non-sampling error*). Die Zunahme der Probenahmefehler mit der Größe eines Projektes liegt im Wesentlichen darin begründet, dass bei großen Projekten viele und wechselnde Probenehmer, Analysengeräte, Chemotechniker, etc. eingesetzt werden.

Es lassen sich verschiedene Konzepte zur Erfassung von Daten auf einer Fläche darstellen:

- Zufallsprobenahme (*random sample*)
- Sonderfall: *Monte-Carlo-Sampling*
- differenzierte Probe (*stratified sample*)
- aggregierte Probe (*clustered sample*)

Eine besondere Schwierigkeit besteht in der Festlegung der Anzahl der Probenahmepunkte. Dies gilt für Probennahmen auf einer Fläche, Probenahmen über ein vertikales Profil (z.B. Bohrung) oder für Messungen versus Zeit gleichermaßen. Wesentliches Kriterium ist die erwartete *Varianz*, also die Variation und die Art der Variation.

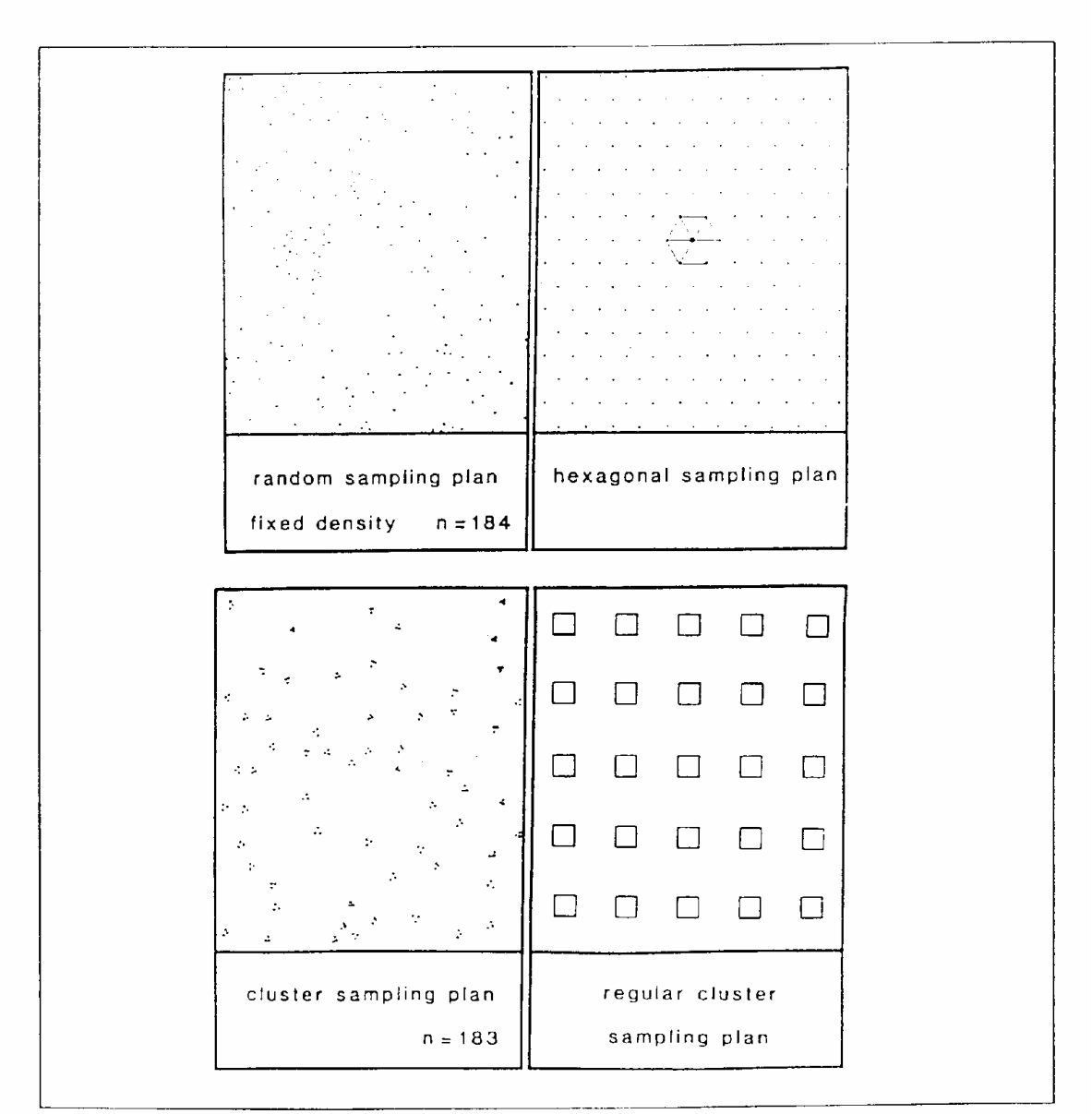


Abb. 1. Konzepte zur Erfassung von Daten auf einer Fläche

Die Problematik der Anzahl der notwendigen Messwerte läßt sich an einem Zeitreihenbeispiel aufzeigen. Gegeben ist ein Signal (z.B. die Temperatur des Grundwassers oder die Konzentration eines Wasserinhaltstoffes), das jahreszeitlichen Schwankungen unterworfen ist. Diese Zeitreihe kann entweder in einem festen Zeittakt (z.B. jede Woche) oder in einem variablen Takt aufgezeichnet sein.

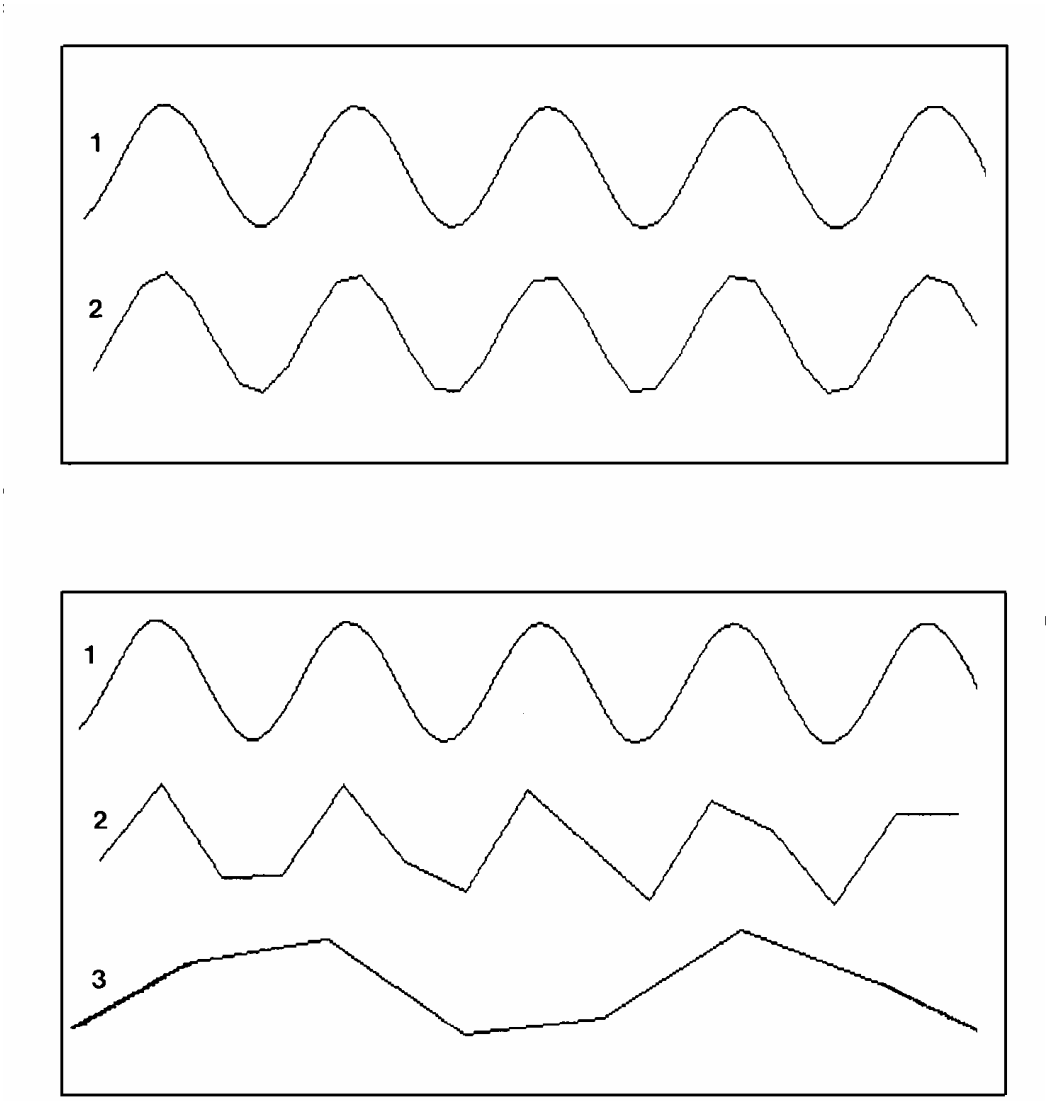


Abb. 2. Aliasing-Effekt

Werden 8 Messpunkte je Periode verwendet, so wird eine hinreichend gute Abbildung der Realität erreicht (Kurve 2 im oberen Teilbild). Reduziert sich die Zahl der Stützstellen pro Periode, wird der Verlauf der Ganglinie schon sehr "eckig"; insgesamt lässt sich der sinusförmige Verlauf aber noch erahnen (Kurve 2 im unteren Teilbild). Bei einer Reduzierung auf drei Messungen je Teilbild kommt es zum sogenannten *Aliasing-Effekt* (Kurve 3 im unteren Teilbild): Es wird eine andere Frequenz vorgetäuscht. Auch die Amplitude kann verfälscht werden und im Extremfall kann sich sogar ein nahezu monotones Signalverhalten einstellen. Bei unzureichender Messdichte enthalten die gewonnenen Daten in jedem Fall einen systematischen Fehler (*biased data*).

Während für den Bereich der Wirtschaftswissenschaften relativ einfache Berechnungsgrundlagen entwickelt wurden und allgemein für das Design einer Umfrage oder die Anzahl der zu betrachtenden Objekte anerkannt sind, ist dies im Bereich der Naturwissenschaften und im Bereich Umweltschutz noch nicht der Fall. Insbesondere bei Fragestellungen, wie die optimale Einrichtung von Messnetzen (Luft, Gewässergüte, Grundwasserstand- und Güte) erfolgen muss, besteht nach wie vor Forschungsbedarf, so dass hier keine bestimmte Methode favorisiert wird.

Vor allem im Hinblick auf die Speicherung von Daten (Kapitel 2.2) ist wichtig, in welcher Form die Daten erfaßt werden:

- manuell (Laborbuch, Feldheft, Karten, Listen)
- maschinell auf Papier (z.B. konventionelle Pegelschreiber)
- elektrische/elektronische Sensoren mit digitaler Aufzeichnung
- hybride Methoden

Insbesondere die *hybriden Methoden* werden in Zukunft an Bedeutung gewinnen. So ist es durchaus möglich, direkt im Gelände Befunde in ein digitales Erfassungssystem einzugeben, das sich gleichzeitig den Raum- und Zeitbezug über Sensoren (GPS, Uhr) erfragt und mitspeichert.

2.2 Speicherung von Daten

Das Speichern von Daten muss nicht zwangsläufig in digitaler Form erfolgen. Vielfach sind analoge Aufzeichnungen nach wie vor sinnvoll. So ist beispielsweise der Aufbau und die Verwaltung eines Bohrarchivs, das alle relevanten Daten über die Bohrungen in einer Region enthalten soll, in digitaler Form nur mit erheblichem technischem und finanziellem Aufwand machbar. Dazu müssen z.B. alle verfügbaren technischen Zeichnungen und auch Karteiblätter und sonstige historische Unterlagen optisch gescannt und als Bilder im Computer verwaltet werden. Eine Vektorisierung eingescannter Vorlagen kann den Datenumfang reduzieren. Ein Verzicht auf diese Details im Sinne einer Reduzierung auf ein reines Zahlenwerk hat sich als wenig sinnvoll erwiesen. Ähnliches gilt für die Informationen, die in Karten vorliegen, auch wenn sich durch die Verbreitung von Geo-Informationssystemen (GIS) neue Aspekte ergeben.

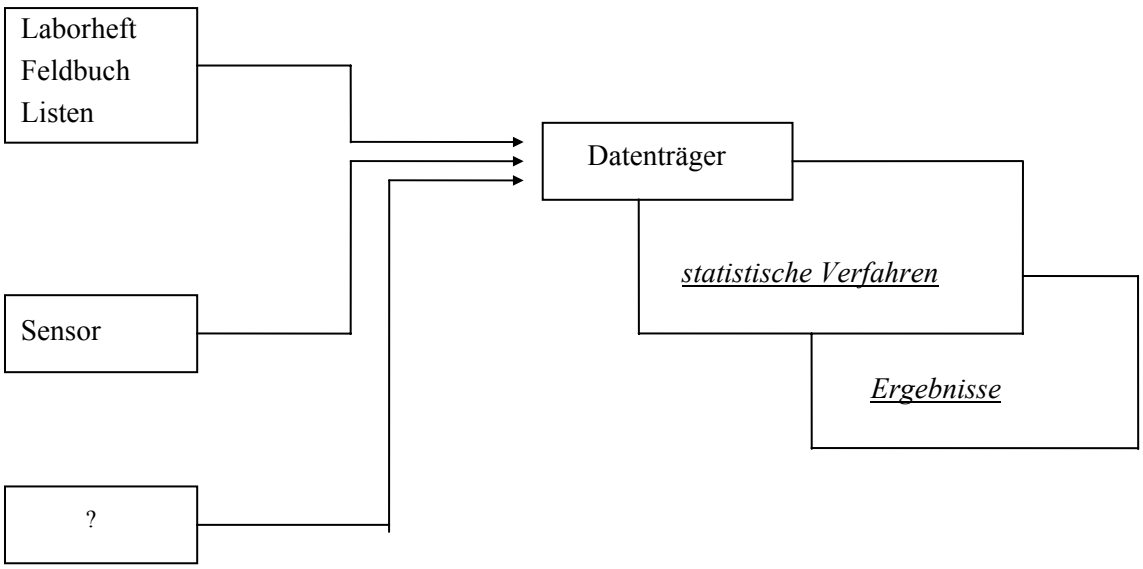


Abb. 3. Schema Datenerfassung

Um allerdings mit Daten Statistik "machen" zu können, ist die Überführung der Daten in einen Computer unabdingbar (von kleinen Datensätzen, die mittels Taschenrechner zu bearbeiten sind, abgesehen). lm einfachsten Fall werden die fraglichen Daten in eine Datei (*File*) im Spread Sheet

Format geschrieben, z.B. Excel. Solch ein Spread Sheet entspricht von der Struktur her einer Tabelle mit Reihen und Spalten, in denen die Messgrößen (Variablen) und die Fälle (Messereignisse) entsprechend angeordnet sind. Dabei bestehen generell zwei Möglichkeiten:

Variante 1

	Messung 1	Messung 2	Messung 3
Temperatur	12.5	13.6	10.4
Leitfähigkeit	503	490	567
pH-Wert	7.51	7.86	7.76

Variante 2

	Temperatur	Leitfähigkeit	pH-Wert
Messung 1	12.5	503	7.51
Messung 2	13.6	490	7.86
Messung 3	10.4	567	7.76

Während Spread Sheets keinerlei Struktur vorgeben und Files ohne jeglichen Bezug zueinander verwalten, werden Files in einer Datenbank innerhalb einer fest vorgegebenen Struktur mittels eines eindeutig definierten Idents miteinander verknüpft.

Folgenden Anforderungen werden an Datenbanksysteme gestellt:

- Redundanzfrei
- Vielfach verwendbar
- Unabhängig
- Funktionsintegriert
- Benutzerfreundlich
- Strukturflexibel
- Integer

Eine Datenbank ist gemäß Definition der Informationstheorie dann *redundanzfrei*, wenn jedes Datenelement in ihr nur genau einmal auftaucht. So soll z.B. das Datum einer Grundwasserprobe nur genau einmal gespeichert sein. Redundanzfreiheit bedeutet, dass der Speicher relativ effizient verwaltet wird (nicht in jedem Fall), es aber beim Suchen zu erheblichen Zeitproblemen kommen kann.

Eine Datenbank soll *vielfach verwendbar* sein in dem Sinn, dass mehr als ein Benutzer sinnvoll (auch teilweise) auf sie zurückgreifen kann, wobei die inhaltlichen Fragen an die Datenbank sehr unterschiedlich sein können. So interessiert sich Benutzer A für die chemischen Wasserinhaltsstoffe, während Benutzer B aus der Datenbank Informationen über die Auslastung des Labors herausfiltern möchte.

Die *Unabhängigkeit* einer Datenbank bezieht sich auf die Verbindung zwischen Datenverwaltung und Auswerteprogrammen. Eine unabhängige Datenbank soll somit von beliebigen Programmen aus erreichbar sein. Dies setzt voraus, dass die interne Struktur der Datenablage jedem Programmierer detailliert bekannt sein muss. Diese Forderung ist gegenwärtig nur sehr eingeschränkt realisiert, da nicht alle Softwarehersteller diesbezüglich ihre Strukturen offen darlegen.

Die Forderung nach einer *funktionsintegrierten Datenbank* hebt auf den schnellen und ordnungsbezogenen Zugriff auf verschiedene Datenelemente ab. Dies wird in der Regel durch die Einführung eines Identifikators erreicht, der als Ordnungskriterium in allen zusammengehörigen Datenelementen auftaucht und syntaktisch gesehen somit eine Redundanz bedeutet.

Die *Benutzerfreundlichkeit* einer Datenbank ist weniger durch die semantische und syntaktische Struktur einer Datenbank als durch die Benutzeroberfläche und die Rechnerleistung (schnelle Plattenzugriffe, schnelle CPU) bestimmt. So gesehen ist die Benutzerfreundlichkeit kein eigentliches Kriterium für die Datenbank.

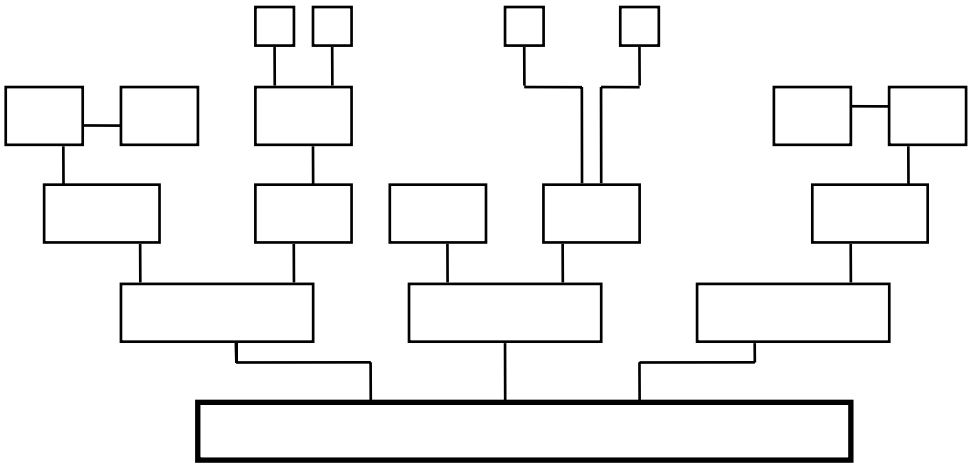
Strukturflexibel ist eine Datenbank, wenn jederzeit ohne Probleme neue Messgrößen oder Tabellen an eine bestehende Datenbank hinzugefügt werden können. Dies kann insbesondere aber bei Vielfachnutzungen zu ernsthaften Problemen führen. Eine Überbewertung führt leicht dazu, dass beim primären Datenbankdesign zu wenig nachgedacht wird.

Mit dem Begriff *integer* sind vor allem Aspekte der Datensicherheit angesprochen. Dies betrifft sowohl die Sicherheit der Daten einer Datenbank gegenüber Soft- und Hardwarefehlern (Vorbeugen von Datenverlusten durch geeignete Maßnahmen wie z.B. regelmäßiges Sichern) als auch der Schutz gegen unerlaubte Zugriffe (Datenschutz) und unerlaubte Veränderungen.

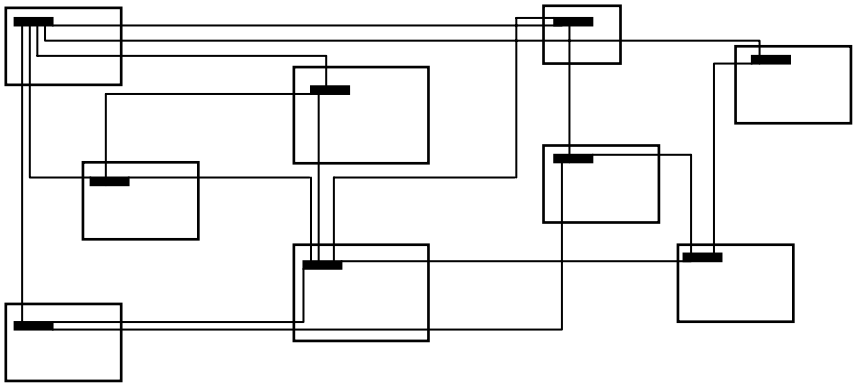
Wesentliche Prozeduren in Verbindung mit Datenbanken sind die folgenden Operationen:

- Definieren und Erfassen
- Ergänzen — (1) Aufbau + Pflege
- Korrigieren
- Wiederfinden
- Reduzieren & Kombinieren — (2) Bearbeitung & Interpretation
- Erstellen von Reporten

Wesentliche Begriffe aus der Informationstheorie sind die Worte *Satz*, *Segment* und *Raum*. Diese Begriffe sind eng verbunden mit Indizierungstechniken, die ein schnelles Suchen überhaupt erst ermöglichen. Während mit Beginn der Entwicklung von Datenbanken zunächst *hierarchische Baumstrukturen* bevorzugt wurden, wurden im letzten Jahrzehnt vor allem die relationalen Datenbankstrukturen weiterentwickelt. *Relationale Datenbanken* sind weitgehend redundanzfrei, vielfach verwendbar, funktionsintegriert und strukturflexibel.



hierarchische Datenbank



relationale Datenbank

(beliebige Ver-knüpfungen über Ident möglich)

Abb. 4. Grafische Darstellung einer hierarchischen und relationalen Datenstruktur

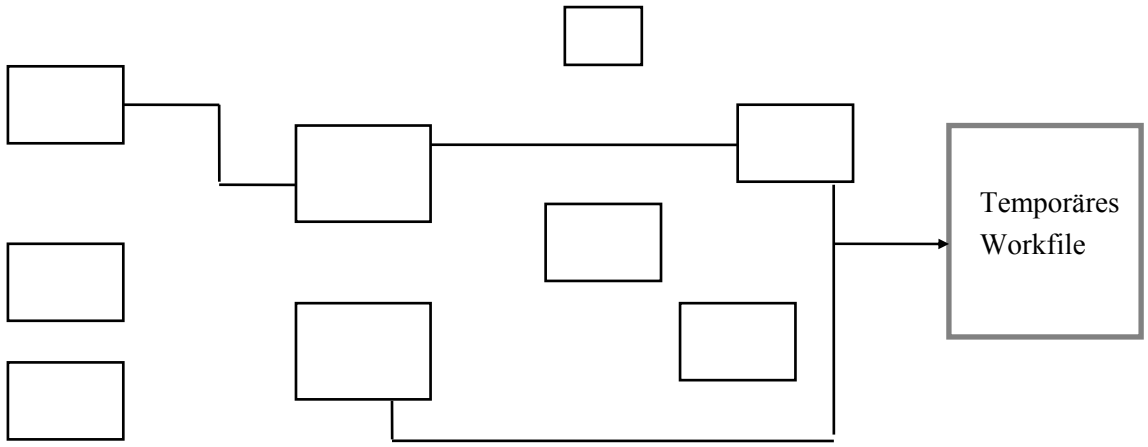
Die Variablen in Datenbanken müssen nach ihrem Typ definiert werden. Dies sind z.B. Datum, Text, Memo, sowie Zahlen. Bei den Zahlen wird unterschieden in Integer (Zahlen ohne Nachkommastellen) und Real (Zahlen mit Kommastellen). Integer- und Real-Zahlen werden noch bezüglich ihrer Bit-Tiefe unterschieden. Der Nutzer muss sich vor der Definition darüber im Klaren sein, welchen Umfang und welche Genauigkeit er für die jeweiligen Variablen benötigt. Die Bit-Kodierung von Zahlen ergibt sich aus der Formel 2^n (1 Byte = 8 Bit). Zahlen mit Dezimalstellen (Fließkommazahlen) werden durch interne Exponentialdarstellung ermöglicht. Daraus ergeben sich üblicherweise 7 signifikante Stellen für 32 Bit-Zahlen (single precision) und 14 signifikante Stellen für 64-Bit Zahlen (double precision).

Bit-Kodierung	Zahl der Ausprägungen	Integer Werte-Bereiche	Real Nachkommazahlen
1 Bit (binär)	$2^1 = 2$	0 oder 1	
8 Bit (1 Byte)	$2^8 = 256$	unsigned Integer: 0 bis 255 signed Integer: -128 bis +127	
16-Bit (2 Byte)	$2^{16} = 65\,536$	unsigned Integer: 0 bis 65535 signed Integer: -32768 bis 32767	

32-Bit (8 Byte)	2^{32} = 4 294 967 296	unsigned Integer: 0 bis 4.294.967.295 signed Integer: -2.147.483.648 bis + 2 147 483 647	Single precision 7 signifikante Stellen
64-Bit (16 Byte)	$2^{64} = 1.8447 \times 10^{19}$		Double precision 14 signifikante Stellen

2.3 Abfrage von Daten

Für den benutzungsorientierten Anwender sind nach der Speicherung von Daten in Datenbanken vor allem die Abfragemöglichkeiten und die dabei verwendete *Abfragesprache* (*query language*) von größter Bedeutung, um die Vielzahl der in der Datenbank gespeicherten Daten schnell und gezielt weiter bearbeiten zu können. Dabei werden nur aus einzelnen Files bestimmte Daten abgefragt und in ein temporäres Workfile geschrieben. Dieses Workfile kann nach Beendigung der Arbeit gelöscht werden, da es redundante Daten enthält.



Grundkenntnisse in einer beliebigen Programmiersprache sind zum Verständnis von Query - Sprachen hilfreich, aber nicht zwingend.

Als Beispiel ist eine Datenbank mit folgender Struktur gegeben:
Relation BRUNNEN mit Name, Rechts- und Hochwert, Betreiber etc.
Relation AUSBAU mit Angaben zu Brunnenausbau (Sperrohre, Filterrohre etc.)
Relation CHEMIE1 mit Angaben zu Temperatur, pH-Wert, Wasserstand etc.
Relation CHEMIE2 mit Angaben zu Ca, Mg, N, K, $N0_3$, Cl etc.

Eine mögliche Query ist im Folgenden gegeben, wobei bewusst keine spezielle Query-Syntax (z.B. SQL) verwendet wird:

```
FOR EACH Ident
        IF BRUNNEN:RW > 4100000 AND IF BRUNNEN:RW < 4200000 _
        AND IF BRUNNEN:HW > 5250000 AND IF BRUNNEN:HW < 5300000 _
        AND IF AUSBAU:TIEFE > 100
LIST CHEMIE1:TEMP, PH CHEMIE2: Ca, Mg, N03, S04
```

- es werden alle Brunnen (alle Idents) betrachtet
- mittels "IF" wird ein Fenster bezüglich Rechtswert ("RW") und Hochwert ("HW") in der Relation "BRUNNEN" selektiert, zusätzlich werden nur Brunnen mit einer Tiefe größer 100 m selektiert (Diese Information steht in der Relation "AUSBAU")
- aus den Relationen "CHEMIE1" und "CHEMIE2" werden bestimmte Parameter in eine Liste geschrieben

Beim Design dieses Datenbank Beispiels wurden diese vier Relationen angelegt, um möglichst Redundanzfreiheit und Strukturflexibilität zu erreichen, um beliebige Auswertungen machen zu können. Eine Trennung in mehr als eine "CHEMIE"- Relation empfiehlt sich, da manche Parameter sehr häufig gemessen werden (Wasserstand, Leitfähigkeit), andere Parameter (z.B. Schwermetalle und Pestizide) dagegen nur sehr selten. Eine mögliche Strategie ist es, die verschiedenen Wasserinhaltsstoffe (in der Realität kommen leicht über 100 relevante Parameter in Frage) in sinnvolle Gruppen zusammenzufassen. Werden z.B. die Pestizide in einer Relation verwaltet, wird mit großer Wahrscheinlichkeit nicht immer der gleiche Parameterumfang bestimmt werden. Spätestens wenn ein neues Pestizid oder Metabolit in die Liste des Untersuchungsumfanges aufgenommen wird, müssen die bereits in der Datenbank befindlichen Datensätze bezüglich dieses neuen Parameters mittels eines "*Missing Values*" verwaltet werden. Um „Missing Values“ weitgehend in Sätzen zu vermeiden, kann auch die Strategie vertreten werden, pro Relation nur einen einzigen Parameter aufzunehmen. So könnte eine Datenbank, die wasserchemische Daten aus Brunnen verwaltet, neben den Orts beschreibenden Relationen über je eine Relation für jeden Chemie-Parameter verfügen. In der NITRAT-Relation würden sich dann genau drei Messgrößen befinden:

- IDENT
- DATUM (incl.Uhrzeit) und
- WERT.

Diese "NITRAT"-Relation wäre frei von „Missing Values“, da ein Satz nur geschrieben wird, wenn ein Nitrat-Wert für einen bestimmten Brunnen zu einer bestimmten Zeit vorliegt. Es ist einsichtig, dass über eine Query eine Liste (oder eine neue Relation) erstellt werden kann, die die Nitrat-Werte eines Ortsfensters für ein Zeitfenster oder einen Stichtag enthält (Grundlage für eine Isolinienkarte der Nitrat-Werte). Ebenso kann mittels einer Query, die Zeitreihe für einen oder mehrere frei wählbare Brunnen über ein Zeitfenster als Liste oder Relation gebildet werden.

3 Datencheck

3.1 Generelles Vorgehen

Nach der Datenerfassung und der Erstellung einer Tabelle (File) mit den Daten, erfolgt zunächst ein Datencheck. Bei hydrochemischen Daten müssen Bestimmungs- und Nachweisgrenze festgelegt werden, Werte kleiner der Nachweisgrenze ersetzt und bei allen Daten das Problem fehlender (Meß-)Werte berücksichtigt werden. Zudem kann der Datensatz auf sog. „Ausreißer“ getestet werden.

Für die eigentliche statistische Analyse geht man in folgenden Schritten vor:

1. Signifikanzniveau festlegen (für die gesamte Studie einheitlich)
2. Skalierungsniveau feststellen (jede Variable)
3. auf Normalverteilung prüfen (jede Variable)
4. an Hand der Ergebnisse von 2 und 3 das statistische Verfahren wählen
5. Verfahren durchführen
6. an Hand des Signifikanzniveaus prüfen, ob Ziel erreicht.

Diese Vorgehensweise gilt für die meisten statistischen Verfahren. Lediglich die Faktorenanalyse und die Clusteranalyse bieten direkt keine Prüfmöglichkeit an Hand eines Signifikanzniveaus.

Beispiel: Dem Auftraggeber einer hydrogeochemischen Studie reicht eine Irrtumswahrscheinlichkeit von 1% bzw. 99% statistische Sicherheit. Somit wird an Hand eines P-Wertes von 0,01 geprüft. Die Daten der Variablen "Nitrat" und "Atrazin" haben Verhältnisskalenniveau. Ein P-P-Plot zeigt aber, dass die Daten der Variable "Atrazin" im Gegensatz zu "Nitrat" nicht normal verteilt sind. Somit ist die PEARSON-Korrelation nicht sinnvoll anzuwenden. Statt dessen wird das Rang-Verfahren KENDALL oder SPEARMAN angewendet. Der berechnete Korrelationskoeffizient ist positiv, der P-Wert wird mit = 0,0054 ausgegeben. Somit ist die Aussage des statistischen Test: Die Variable "Nitrat" korreliert positiv mit der Variablen "Atrazin" auf dem vorgegeben Signifikanzniveau von P=0,01.

3.2 Bestimmungs- und Nachweisgrenze

Im Bereich naturwissenschaftlicher Messungen werden vielfach Sensoren eingesetzt, die ein elektrisches Signal erzeugen, welches proportional zu der zu messenden physikalischen Größe ist. Wenn die Beziehung zwischen dem Sensorsignal und der physikalischen Messgröße linearer Natur ist, wird mit Hilfe der linearen Regression aus dem Sensorsignal der gesuchte Wert berechnet. Dazu wird eine Eichkurve erstellt.

Im Bereich sehr niedriger Konzentration, also nahe dem Nullpunkt, taucht das Problem auf, inwieweit gemessene Sensor-Effekte noch als physikalische Messwerte interpretiert werden dürfen. Definitionsgemäß ist die *Bestimmungsgrenze* eines Verfahrens die kleinste quantitativ bestimmbare Menge (Konzentration), die signifikant von Null verschieden ist. Der Begriff *Nachweisgrenze* ist definiert durch den Wert, der signifikant aus dem *Grundrauschen* (*Störpegel*) des *Blindwertes* herausragt. Die Nachweisgrenze liegt deshalb immer unter der Bestimmungsgrenze. Beide Grenzwerte können rechnerisch oder graphisch ermittelt werden. Rechnerisch erfolgt dies mit Hilfe folgender Formel

$$XN = 2 \cdot \frac{sdv(x) \cdot t}{a_1} \cdot \sqrt{\frac{1}{N} + 1 + \frac{(y_c - \overline{y})^2}{a_1^2 \, \Sigma(x_i - \overline{x})^2}}$$

$$XB = \frac{y_h - a_0}{a_1} + \frac{sdv(x) \cdot t}{a_1} \cdot \sqrt{\frac{1}{N} + 1 + \frac{(y_h - \overline{y})^2}{a_1^2 \, \Sigma(x_i - \overline{x})^2}}$$

Graphisch gewinnt man die Grenzwerte sehr einfach aus der Eichregressionsgerade und den 95 %-Vertrauenskurven der Schätzung gewonnen werden. Sollen die Grenzwerte auf 1 %-Niveau abgesichert sein, muss sinngemäß die 99%-Vertrauensgrenze herangezogen werden.

Zur grafischen Bestimmung der Nachweisgrenze X_N und der Bestimmungsgrenze X_B wird die Regressionsgerade und der 95 %-Vertrauensbereich so geplottet, dass die Schnittpunkte von Regressionsgerade und Vertrauenskurven mit der y-Achse im Fenster liegen.

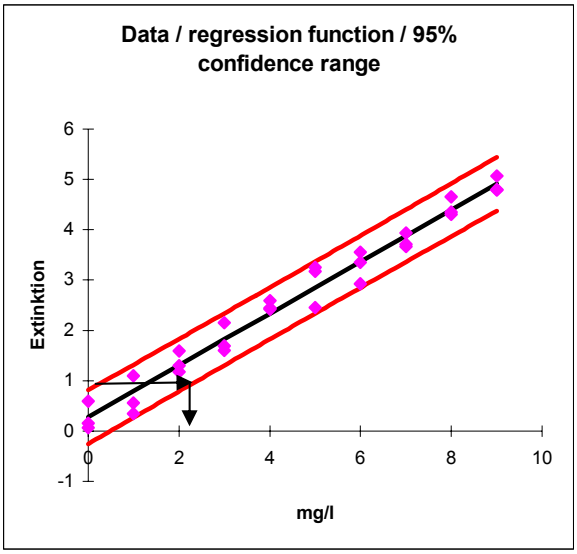
Für die Ermittlung der Nachweisgrenze wird vom Schnittpunkt der oberen Vertrausenskurve mit der y-Achse eine Parallele zur x-Achse eingezeichnet. Wo diese die untere Vertrauenskurve schneidet, wird der x-Wert als Nachweisgrenze abgelesen.

Für die Ermittlung der Bestimmungsgrenze wird vom Schnittpunkt der Regressionsgerade mit der y-Achse eine Parallele zur x-Achse eingezeichnet. Wo diese die untere Vertrauenskurve schneidet, wird eine Parallele zur y-Achse eingezeichnet. Wo diese die obere Vertrauenskurve schneidet, wird eine zweite Parallele zur x-Achse eingezeichnet. Wo diese die untere Vertrausenskurve schneidet, wird der x-Wert als Bestimmungsgrenze abgelesen.

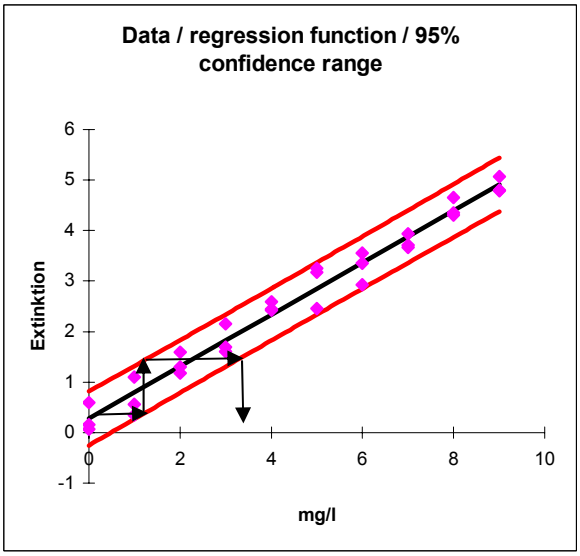
Die nachfolgende Tabelle zeigt ein einfaches Beispiel von 3fach-Bestimmungen einiger Messwerte in mg/L und deren Extinktionswerten. In Abb. 5 wurde daraus graphisch die Nachweis- und die Bestimmungsgrenze ermittelt.

Tab 1. Messwerte (Extinktion) und zugehörige Konzentrationen (mg/L) für jeweils drei Wiederholungsmessungen

mg/l	Extinktion
0	0.60
0	0.16
0	0.07
1	0.35
1	0.56
1	1.10
2	1.30
2	1.59
2	1.18
.....	.....



Nachweisgrenze



Bestimmungsgrenze

Abb. 5. Graphische Ermittlung von Nachweis- und Bestimmungsgrenze

3.3 Werte „kleiner als“ und „Missing values“

Im Bereich der wasserchemischen Analytik aber auch bei anderen Messverfahren taucht ständig das Problem auf, dass wir Werte kleiner als die Nachweisgrenze (nwg) erhalten. Das Problem mit solchen Werten beginnt bereits bei der Datenverwaltung. In einer Datenbank können Werte in der Form <0,5 nur so gespeichert werden, wenn das Feld als Text-Variable vereinbart wird. Statistik-Programme können mit solchen Angaben oft gar nicht umgehen. Oft behilft man sich, in dem man die Werte als in der Form „-0.5“ abspeichert. Für Konzentrationen ist dies ein machbarer Weg, da negative Konzentrationen definitionsgemäß nicht auftreten können. Bei pe-Wert oder Temperatur, die negative Werte aufweisen können, tritt das Problem „kleiner als Nachweisgrenze“ nicht auf.

Ersetzt man die <Wert-Information für Statistiken mit dem „Missing Value“, also einem fehlendem Wert, so verwirft man a) eine wichtige Information und schränkt b) den Datenumfang für bi- und multivariate Verfahren ein.

Eine häufig praktizierte Variante ist es daher, den „kleiner als“ Wert durch 0,5*Nachweisgrenze oder 0,33*Nachweisgrenze zu ersetzten. Man nimmt somit einfach an, dass der Messwert 50% bzw. 33% der Nachweisgrenze entsprechen würde. Ein deutliches besseres Verfahren ist aber, sich die Häufigkeitsverteilung des Messwertes anzusehen und die Häufigkeitsverteilung im Bereich zwischen Nachweisgrenze und Null entsprechend dem Kurvenverlauf anzupassen. Vielfach macht man auch keinen großen Fehler, wenn man diesen Verlauf linear approximiert. Mit Hilfe eines Zufallszahlengenerators ersetzt man nun die "<Werte" durch Zufallszahlen zwischen Null und der Nachweisgrenze.

Zusammenfassend stehen also folgende Verfahren zum Ersetzen von Werten „kleiner als“ zur Verfügung:

- Ersetzen durch 0.5 * nwg

- Ersetzen durch 0.3 * nwg
- Ersetzen durch Werte aus linearer / multipler linearer Regression (ggf. auch nichtlineare Regression)
- Ersetzen durch Zufallswerte (RND(1)* nwg)

„Missing Values“ werden in Datenbanken üblicherweise als Wert -999 verwaltet. Ersetzt werden können Sie durch:

- arithmetische Mittel
- Median entweder der Grundgesamtheit oder
- Median bezogen auf Gruppen.
- Approximation aus einfacher / multipler Regression

3.4 Plausibilitätsprüfung ("Ausreißertests")

Die Verwendung so genannter *Ausreißertests* ist sehr problematisch, da eine Beobachtung, die aus der Reihe fällt, keineswegs eine Fehlmessung sein muss. Ein tausendjähriges Hochwasser ist nun mal ein seltenes Ereignis und passt sicherlich nicht in die Verteilungskurve einer 20 jährigen Messperiode. Jede Art von Automatismus ist daher abzulehnen. Als warnendes Beispiel kann auch die Entdeckung des "Ozon-Loches" dienen: Über einen Zeitraum von 10 Jahren wurden von der NASA niedrige Ozon-Gehalte in der Atmosphäre zu bestimmten Zeiten gemessen, die aber auf Grund eines automatischen Ausreißertests aus dem Datensatz eliminiert wurden. (Glücklicherweise wurden die Original-Messwerte aber abgespeichert.) Ausreißertester sind deshalb nur zu verwenden, um Hinweise auf mögliche Daten- oder Erhebungsfehler zu geben. Erst eine individuelle, inhaltliche Plausibilitätsprüfung darf dazu führen, dass der betreffende Wert aus der weiteren statistischen und sonstigen Auswertung ausgeschlossen wird.

Ein einfacher Test auf Ausreißer ist der nach *DAVID*, *HARTLEY* und *PEARSON*:

Es wird die Prüfgröße z aus Spannweite und Standardabweichung berechnet und mit dem Tabellenwert Z verglichen (z = Spannweite / Standardabweichung).

Wenn z < Z ist, ist die Nullhypothese, dass die Extremwerte keine Ausreißer sind, nicht widerlegt. Die Daten können somit als ausreißerfrei angesehen werden.

Tab 2. Ausreißertest nach DAVID - HARTLEY - PEARSON

Probengröße	(1 - α) in Prozent			
n	90 %	95 %	99 %	99,5 %
3	1,997	1,999	2,000	2,000
4	2,409	2,429	2,445	2,447
5	2,712	2,753	2,803	2,813
6	2,949	3,012	3,095	3,115
7	3,143	3,222	3,338	3,369
8	3,308	3,399	3,543	3,585
9	3,449	3,552	3,720	3,772
10	3,57	3,685	3,875	3,935
11	3,68	3,80	4,012	4,079
12	3,78	3,91	4,134	4,208
13	3,87	4,00	4,244	4,325
14	3,95	4,09	4,34	4,431
15	4,02	4,17	4,44	4,53
16	4,09	4,24	4,52	4,62
17	4,15	4,31	4,60	4,70
18	4,21	4,37	4,67	4,78
19	4,27	4,43	4,74	4,85
20	4,32	4,49	4,80	4,91
25	4,53	4,71	5,06	5,19
30	4,70	4,89	5,26	5,40
35	4,84	5,04	5,42	5,57
40	4,96	5,16	5,56	5,71
45	5,06	5,26	5,67	5,83
50	5,14	5,35	5,77	5,93
55	5,22	5,43	5,86	6,02
60	5,29	5,51	5,94	6,10
65	5,35	5,57	6,01	6,17
70	5,41	5,63	6,07	6,24
75	5,46	5,68	6,13	6,30
80	5,51	5,73	6,18	6,35
85	5,56	5,78	6,23	6,40
90	5,60	5,82	6,27	6,45
95	5,64	5,86	6,32	6,49
100	5,68	5,90	6,36	6,53
150	5,96	6,18	6,64	6,82
200	6,15	6,39	6,84	7,01
500	6,72	6,94	7,42	7,60
1 000	7,11	7,33	7,80	7,99

3.5 Plausibilitätskontrollen

Wasserchemische Analysen enthalten eine Vielzahl von Parametern, die durch unterschiedlichste Verfahren gewonnen wurden. Die Möglichkeiten, dass sich Fehler einschleichen sind vielfältig: es beginnt bei der Probennahme, setzt sich fort über Art der Probenlagerung (Licht, Temperatur) und Dauer der Lagerung bis hin zu dem eigentlichen Analysenverfahren. Daher ist es unerlässlich. Eine Analyse auf Plausibilität zu überprüfen. Diesbezüglich gibt es verschiedene Verfahren:

- Bilanzfehler
- an Hand der elektrischen Leitfähigkeit
- unplausible Spezies

Bilanzfehler

Definitionsgemäß müssen in einem Wasser die Anzahl der positiv geladen Ionen und Komplexen unter Berücksichtigung ihrer Wertigkeit denen der negativ geladenen Ionen und Komplexen entsprechen. Dazu müssen, wenn die Angaben in mg/l vorliegen, alle Konzentrationen in mmol(eq)/l umgerechnet werden. Der Bilanzfehler ergibt sich dann wie folgt:

$$F1 = \frac{\sum \text{Kationen} - \sum \text{Anionen}}{\sum \text{Kationen} + \sum \text{Anionen}} \cdot 100$$

In Deutschland empfiehlt unter anderem der DVWK (1992) die Formel:

$$F2 = \frac{\sum Kationen - \sum Anionen}{\left(\sum Kationen + \sum Anionen\right) \cdot 0.5} \cdot 100$$

Durch den Faktor 0.5 im Nenner wird Fehler F2 doppelt so groß wie der Fehler F1. Die DVWK-Regeln 128 (1992) empfehlen z.B. Analysen mit einem F2-Fehler größer als 2% (Analysen mit über 2 meq/l Kationen- und Anionensumme) respektive 5% (Analysen mit kleiner 2 meq/l kationen- und Anionensumme) als fehlerhaft zu verwerfen.

In diese beiden Formeln (F1 und F2) werden üblicherweise nur die 4 Haupt-Kationen und Anionen in meq/L berücksichtigt. Bei sehr niedrigen pH-Werten müssen die Protonen (H^+) in die Rechnung einbezogen werden und ggf. auch weitere Inhaltsstoffe (z.B. 2-wertiges Eisen). Bei hohen pH-Werten, wenn das Carbonat-Ion eine Rolle spielt, muss dieses zusätzlich zum HCO_3^- berücksichtigt werden. Unberücksichtigt bleibt die Bildung von Komplexen, die nur einbezogen werden, wenn man zur Berechnung ein Programm wie PHREEQC verwendet, das den Bilanzfehler aus der Summe aller negativ und positiv geladener Komplexe berechnet und damit zu anderen Werten für den Bilanzfehler kommt. Unberücksichtigt bleiben bei beiden Formeln die nullwertigen Komplexe

Eine andere Möglichkeit der Normierung ist die Ionenstärke

$$F3 = \frac{(Elektroneutralität(mol/kg))}{Ionenstärke(mol/kg)} \cdot 100$$

Dieser Fehler kann ebenfalls sehr einfach berechnet werden, wenn die Analyse in PHREEQC analysiert wird. Mittels des Schlüsselwortes SOLUTION_SPREAD ist es möglich EXCEL-Files mehr oder wenig direkt in PHREEQC einzulesen. Die Kopfzeile muss die Master-Spezies in der Form enthalten, wie sie in der thermodynamischen Datensammlung innerhalb von PHREEQC definiert sind. In einer zweiten Kopfzeile können zusätzliche Angaben zur Spezies (z.B. as NO3) gemacht werden respektive zur verwendeten Einheit (z.B. ug/l).

PHREEQC - What do you want to model today?

File Edit View Calculations Help

Input | Output | Database | Grid | Chart

SOLUTION_SPREAD
units mg/l

pH	pe	temp	Mg	Na	Ca	Fe	K	Alkalinity	Cl	N(+5)	S(6)
								meq/l		as NO3	as SO4
6.9	12	8.5	0.5	3.6	3.3	0.028	2.1	0.08	3.3	2.3	12
6.9	11.9	9.5	4.7	2.1	12.5	0.008	2.6	0.28	5.8	4.5347654	23.5
7.3	11.4	8.8	1.8	7.3	9.5	0.013	7.8	0.29	4.9	4.8	25
7.1	3.8	8.7	6.8	0.5	23.3	0.047	5.7	0.58	5.2	3.9	53.9
7.3	0.1	8.5	7.5	0.5	27.8	0.049	5.6	0.95	4.5455653	12.6	45.8
7.2	-0.2	7.4	7.1	9.5	24.5	0.14	4.2	1.1	24.1	34	19.2
7.1	-0.15	8.5	7	0.5	21	0.017	12.9	0.12	11.2	59.5	20.8
6.9	5	9.4	12.3	4.6	36.9	0.014	7.5	2.8	6.6	4.1	14.6
6.9	0.14	9.9	12.7	2.9	39	0.017	9.4	3.07	3.3	19.4	7.2
7.1	5.6	7.9	7.2	8.6	25.3	0.02	14.2	0.15	18.2	59.4	39.7
7.4	0.1	9.3	13.3	4.6	42.7	0.023	5.6	2.9	9	34.6	25.7
7.2	0.23	9.8	3.1	2.5	40	0.012	42.5	2.7	5	7.2	16.3
6.9	0.34	9	3.4	7.7	38.5	0.019	37.8	2.6	10.6	4.7	21.8
7.3	-0.23	9.1	13.2	3.8	39.4	0.29	5.5	2.2	15.4	2.2	36.9
7.2	-0.13	9.7	13.5	3.3	40	0.009	6.5	2.3	10.9	14.2	27.4
7.2	0.34	8.4	14.1	2.5	64.9	0.08	6.2	4.2	6.8	32.8	21.4
6.9	0.24	8.8	14.3	3.7	59.1	0.18	7.3	2.1	14.5	29.5	65.8
7.3	0.12	9.1	8.8	32.1	42.7	0.07	11.1	1.2	71.7	3.4	49.6
7.2	-0.05	7.2	9.8	32.4	32.2	0.01	8.8	0.12	116	7.5	27.4
8.4	-3.5	8.4	65.9	99.4	224.7	38	24.1	13.7	55	0.23	282.5
8.8	-2.3	9.9	126	1193	174	23	32.9	61.9	304	0.03	13.6

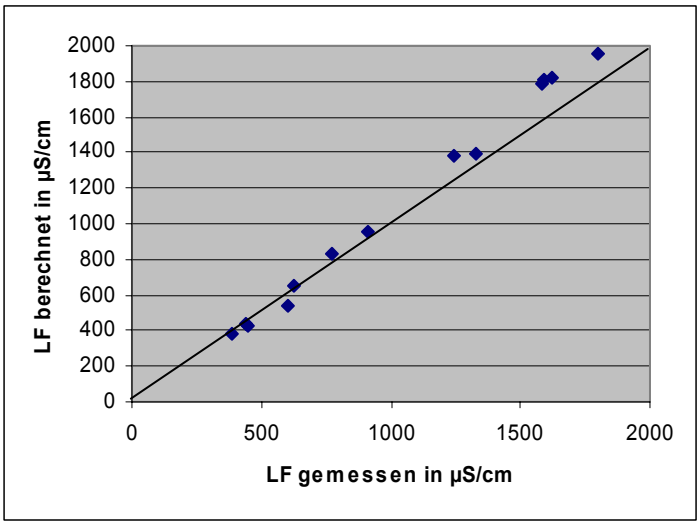
Vergleich mit der elektrischen Leitfähigkeit

Eine Faustregel besagt, dass die elektrische Leitfähigkeit in µS/cm multipliziert mit 0.7 den Abdampfrückstand in mg/L ergibt. Der Abdampfrückstand ist die Summe aller Wasserinhaltsstoffe, wobei die Neben- und Spurenstoffe in der Regel nur eine untergeordnete Rolle spielen. Eine genauere Approximation erhält man für Wässer im Bereich 200 bis 700 µS/cm aus der Formel gültig für 25°C nach Maier & Grohmann (1977):

$$LF = 63.0 \cdot 0.5 \Sigma c \cdot z^2$$

wobei c für die Ionenkonzentrationen in mmol/l steht. Letztlich wird also die Ionenstärke in mol/kg mit dem Faktor 63000 multipliziert. Die folgende Abbildung zeigt, dass im Bereich bis 1000 µS/cm die Übereinstimmung recht gut ist und dass oberhalb von 1000 µS/cm die berechneten Werte zu gross

sind. Berechnungen nach Laxen (1977) und Rossum (1975) berücksichtigen die Aktivitäten und liefern deshalb auch bei höheren Leitfähigkeiten eine bessere Übereinstimmung.



Unplausible Spezies

Bestimmte Parameter-Paare schließen sich gegenseitig aus, da es unter bestimmten Redox-Zuständen nicht möglich ist beide Spezies in bestimmten Konzentrationsbereichen zu finden. Die folgende Tabelle ist den DVWK-Regeln 128 (1992) entnommen:

Parameter	Ausschluss von
O_2 > 5 mg/L	Fe^{2+} > 0.05 mg/L Mn^{2+} > 0.05 mg/L NO_2^- > 0.05 mg/L NH_4^+ > 0.1 mg/L H_2S > 0.01 mg/L
Fe^{2+} > 0.2 mg/L Fe^{2+} > 1.0 mg/l	NO_3^- > 2.0 mg/L H_2S > 0.1 mg/L
Mn^{2+} > 0.2 mg/L	NO_3^- > 2.0 mg/L H_2S > 0.1 mg/L
H_2S > 0.1 mg/L 8.0 > pH < 5.5	NO_3^- > 1mg/L $Ca^{2+} + Mg^{2+}$ > 1mmol/L
Spektraler Absorptions-K. bei 254 nm > 10 E/m	DOC < 3 mg/L

4 Grundlagen der Statistik

Alle statistischen Verfahren basieren auf der *Wahrscheinlichkeit* (*probability*). Ein kurzer Ausflug in die Wahrscheinlichkeitstheorie ist daher angeraten. Als Beispiel dient das Wetter in Deutschland:

Die Aussage, es regnet morgen, wenn es heute regnet, ist mit einer statistischen Wahrscheinlichkeit von ca. 0,7 (70%) richtig. Dies ist ein Erfahrungswert, der beschreibt, dass Wetterlagen in Deutschland relativ stabil sind.

Die Vorhersage des amtlichen Wetterberichts hat eine Wahrscheinlichkeit des Eintretens von ≈ 0,8 (80 %); d.h. die Wissenschaft der Meteorologie samt Satelliten und weltumspannender Messnetze schafft "lediglich" eine Verbesserung von ca. 10%.

Ein Beispiel für eine diskrete Wahrscheinlichkeit ist das Werfen einer Münze. Diskret ist die Wahrscheinlichkeit, da nur eine definierte Anzahl von Ergebnissen eintreten kann (Kopf oder Zahl). Der theoretisch mögliche Fall, dass die Münze auf der Seite zu stehen kommt, ist für die folgenden Betrachtungen per Definition ausgeschlossen:

es gilt: jeder Wurf ist unabhängig (independent); dann ist die Wahrscheinlichkeit

für 2 Mal Kopf in 2 Würfen: ½ * ½ = ¼
und analog für 3 Mal Kopf in 3 Würfen: ½ * ½ * ½ = $^1/_8$

Lautet die Frage, wie groß die Wahrscheinlichkeit ist, bei 3 Würfeln nur einmal "Kopf" zu würfeln, so kann man sich dies anhand der überhaupt möglichen *Permutationen* ausrechnen:

KKK ZZZ ZKZ
KZK ZKK KZZ
KKZ ZZK

Da bei 8 möglichen Permutationen KOPF dreimal vorkommt, ist die gesuchte Wahrscheinlichkeit $^3/_8$ (37.5 %).

Die klassische Definition der *LAPLACE-Wahrscheinlichkeit* ergibt sich aus dem Quotienten

$$P = \frac{Anzahl\ positiver\ Elementarereignisse}{Anzahl\ möglicher\ Elementarereignisse}$$

Die Wahrscheinlichkeit, dass ein bestimmtes Ereignis, für das bei einmaligem Versuch die Wahrscheinlichkeit P besteht, bei mehreren Versuchen Xmal auftritt, ergibt sich aus der *BERNOULLI-Formel*:

$$P(n, X) = \binom{n}{X} * P^x * (1 - P)^{n-X}$$

mit n = Anzahl der Versuche, X = Häufigkeit des bestimmten Ereignisses.

Der *Binomialkoeffizient* "n über X" errechnet sich aus der *Binominalverteilung*:

$$\binom{n}{X} = \frac{n!}{(n-X)!*X!}$$

Die Wahrscheinlichkeit für genau drei Sechser bei sechsmaligem Würfeln ist somit beispielsweise:

$$P(6,3) = \binom{6}{3} * \left(1/6\right)^3 * \left(5/6\right)^3$$

Der Binomialkoeffizient "**6** über **3**" ist gleich 20. Somit errechnet sich die Wahrscheinlichkeit zu:

$$P(6,3) = 20 * 1/216 * 125/216 = 0.0536$$

Die Wahrscheinlichkeit für genau drei Sechser bei sechsmaligem Würfeln beläuft sich also auf 5.36 %. Die Binomialverteilung ist die wichtigste diskrete Verteilung. Sie ist nur definiert für nichtnegative und ganzzahlige X-Werte.

Viele statistische Verfahren verwenden als Prüfgrüße die sog. „Normalverteilung“ und liefern nur dann vertrauenswürdige Werte, wenn die untersuchten Daten *normal verteilt* und auf einem bestimmten Skalenniveau (meist Intervallskalenniveau) sind. Deshalb sollen im Folgenden Skalierungsniveaus numerischer Daten und die Normalverteilung näher erläutert werden.

4.1.1 Skalierungsniveau numerischer Daten

In einer Grundgesamtheit von Daten werden die gemeinsamen Merkmale (z.B. die wasserchemischen Inhaltsstoffe) als Variable bezeichnet. Die jeweils für eine Variable beschriebene Größe bzw. Eigenschaft wird als *Ausprägung der Messgrößen* bezeichnet.

Grundsätzlich wird zwischen qualitativen und quantitativen Messgrößen unterschieden. Qualitative Messgrößen werden meist als alphanumerische, quantitative als numerische Messgrößen gespeichert. Alphanumerische Messgrößen (z.B. die organoleptischen Befunde einer Wasserprobe: Färbung, Geruch, Geschmack etc.) können jedoch auch durch eine geeignete Kodierung (z.B. Geschmack) in numerische Daten überführt werden. Einen eigenen Datentyp stellen Datum und Zeit dar, für die je nach beabsichtigter Auswertung eine gesonderte Behandlung erforderlich sein kann. Numerische Daten werden in 4 Skalierungsniveaus unterschieden.

Tab 3. Skalierungsniveaus numerischer Daten

Skalierungsniveau	Charakterisierung	Beispiel
Verhältnis	absoluter Nullpunkt, Verhältnis von Werten ist bedeutsam	Grad Kelvin, Konzentrationen
Intervall	Differenz zwischen Wertepaaren ist bedeutsam	Grad Celsius, Entfernungen
Ordinal	Aussage über Differenzen nicht sinnvoll, aber Ordnung vorhanden	Schulnoten, Mineralhärten
Nominal	Kodierung numerisch bedeutungslos	geol. Schichten

Je nach statistischem Verfahren ist darauf zu achten, welches Skalierungsniveau mindestens verlangt ist. Für eine *Pearson-Korrelation* z.B. müssen die Daten für beide Messgrößen auf Intervallskalenniveau vorliegen; die Korrelation einer beliebigen Messgröße mit der Messgröße "geologische Schicht" ist also, auch wenn sie numerisch kodiert ist, nicht zulässig. Es empfiehlt sich in solchen Fällen, den Zusammenhang durch eine *einfaktorielle Varianzanalyse* darzustellen.

Unabhängig vom Skalierungsniveau muss unterschieden werden zwischen *stetigen* und *diskreten Daten*. Daten sind stetiger Natur, wenn sie durch reelle Zahlen zu beschreiben sind. Nahezu alle physikalisch meßbaren Größen haben stetige Natur (Temperaturen, Massen, Konzentrationen etc.). Diskrete Daten sind gekennzeichnet durch eine Reihe i.d.R. ganzzahliger Ausprägungen innerhalb eines definierten Intervalls. Somit ist bei diskreten Daten die Anzahl der möglichen Ausprägungen a priori eingeschränkt, während sie bei stetigen Daten unendlich groß ist.

Daten, die auf *Nominalskalenniveau* gemessen sind, sind ausnahmslos diskreter Natur (Geschlecht, alle Ja-Nein-Fragen, Bundesland). Es gibt aber auch Übergangsbereiche zwischen diskreten und stetigen Daten: Die Angabe des Alters in Jahren ist zunächst diskreter Natur; durch Einführung eines Dezimalanteiles (z.B. 10.3 Jahre) können die Daten aber stetigen Charakter erhalten. Umgekehrt erhalten physikalische Messwerte möglicherweise durch die Art der Messwerterfassung diskrete Natur: z. B. Messung von Daten mit einem 8-Bit-Analog-Digital-Wandler (ADC); ein fiktiver Messbereich von 0...100 °C wird dabei auf genau 256 Ausprägungen abgebildet, d.h. in Klassen zu jeweils ca. 0.4 °C.

4.1.2 Normalverteilung

4.1.2.1 Theorie

Die GAUß'sche Normalverteilung (*Glockenkurve*) hat für die Statistik eine besondere Bedeutung; sie ist dadurch charakterisiert, dass *Maximum* und *Erwartungswert* zusammenfallen (Symmetrie) und das sie Wendepunkte an den Stellen "$\mu - \sigma$" bzw. "$\mu + \sigma$" aufweist.

$$Y = \frac{1}{\sigma * \sqrt{2\pi}} * e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

mit X als beliebiger Abszisse und Y als zugehöriger Ordinate. Der *Erwartungswert* μ bestimmt die Lage und die *Standardabweichung* σ die Form der Verteilung. Eine große Standardabweichung bewirkt eine breite Verteilung, eine kleine Standardabweichung dementsprechend eine schlanke Glockenkurve. Die Fläche unter der Normalverteilung ist immer gleich 1, entspricht also der Wahrscheinlichkeit von 100 %.

Die so genannte Standardnormalverteilung ergibt sich bei Annahme eines Erwartungswertes μ von 0 und einer Standardabweichung σ von 1.

$$Y_s = \frac{1}{\sqrt{2\pi}} * e^{-(1/2)\ X^2}$$

Die Wahrscheinlichkeit für das Auftreten von Werten in Intervallen ergibt sich aus dem Integral der Verteilungsfunktion, der Fläche unter der Verteilungskurve:

$$F(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-(1/2)\ X^2} dx$$

Tab 4. Standardnormalverteilung

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4324	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01381
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00400	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139

3.0	0.001350	3.2	0.000687	3.4	0.000337	3.6	0.000159	3.8	0.000072	4.0 0.000032
3.1	0.00967	3.3	0.000483	3.5	0.000233	3.7	0.000108	3.9	0.000048	4.1 0.000021

Beispiel: Wie groß ist die Wahrscheinlichkeit, dass die standardnormal verteilte Variable Z einen Wert von größer 2.35 annimmt?

Ein Blick in die Tabelle der Standardnormalverteilung ergibt eine Wahrscheinlichkeit von 0.00939, also 0.939 %.

Viele statistische Verfahren verwenden die Normalverteilung als Prüfgröße. Diese Verfahren liefern nur dann vertrauenswürdige Werte, wenn die untersuchten Daten auch *normal verteilt* sind. Daher müssen Daten vorab auf Normalverteilung geprüft werden.

4.1.2.2 Prüfung auf Normalverteilung

Die Prüfung auf Normalverteilung geschieht am einfachsten durch Darstellung der Daten in Form eines *Histogramms.*

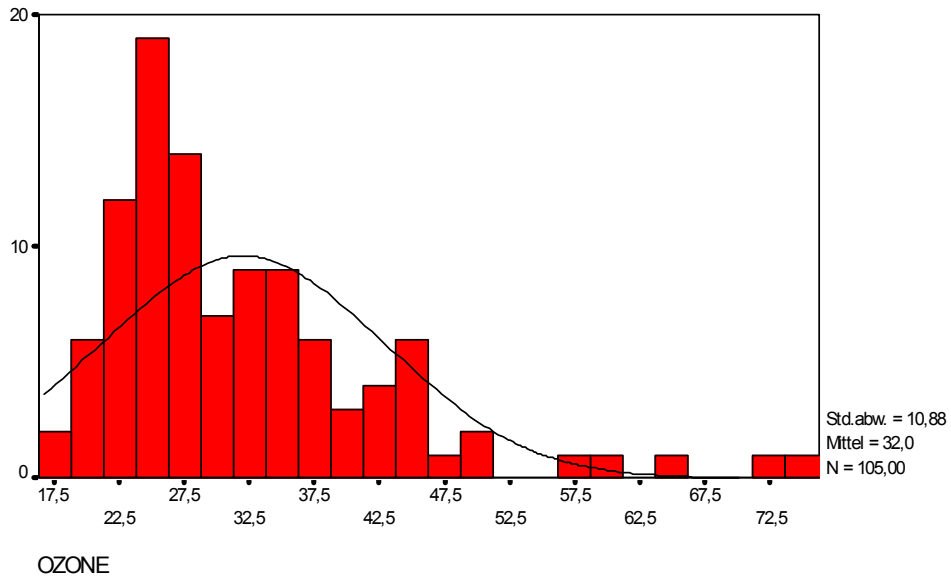


Abb. 6. Histogramm

Eine weitere grafische Möglichkeit, Daten hinsichtlich ihrer Normalverteilung zu prüfen, bietet der sog. P-P-Plot.

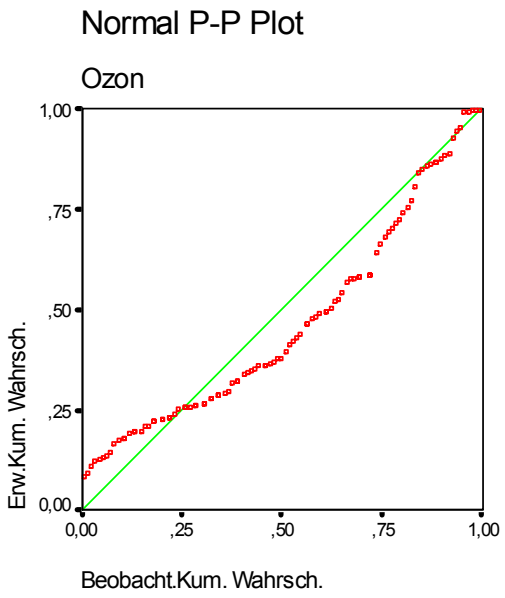


Abb. 7. Summenkurve einer Wahrscheinlichkeitsverteilung

Beide Grafiken zeigen, dass keine Normalverteilung vorliegt.

Überschlagsmäßig ist eine Prüfung auch mit Hilfe des *Variationskoeffizienten* (s/μ [s = Varianz, μ = Mittelwert]). Ist der Variationskoeffizient kleiner als 0.5 kann erfahrungsgemäß davon ausgegangen werden, dass die zugehörigen Daten vermutlich normal verteilt sind.

Beispiel: Der Mittelwert der Leitfähigkeit beträgt 562 µS/cm, bei einer Standardabweichung von 40.5, so ergibt sich:

$$\frac{40.5}{562} = 0.072$$ (Normalverteilung kann angenommen werden)

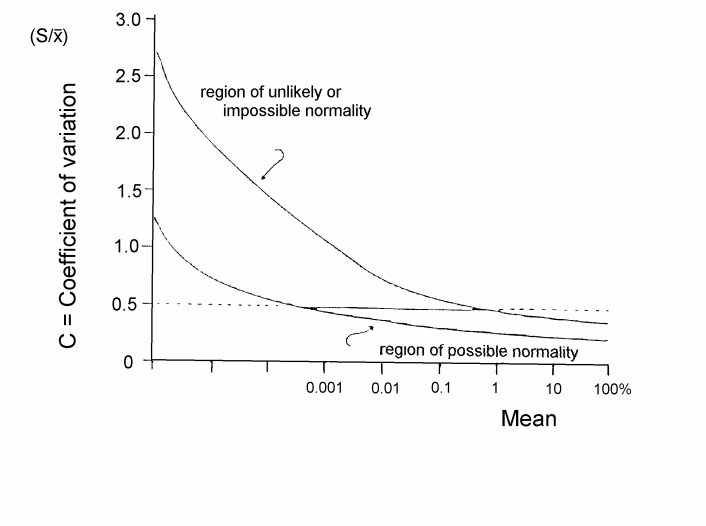


Abb. 8. Prüfung auf Normalverteilung in Abbhängigkeit des Variations-koeffizienten.

4.1.2.3 Transformationen

Nun sind aber Daten in der Regel nicht standardnormal verteilt, sondern allenfalls normal verteilt. Normal verteilte Daten (X) können aber durch eine einfache Transformation in standardnormal verteilte (Z) überführt werden:

$$Z = \frac{X - \mu}{s}$$

Allgemein können Zahlenwerte, die bestimmten Bedingungen statistischer Verfahren nicht genügen, durch Transformationen in geeignete Werte überführt werden. Gängige Operationen sind:

1. $X' = \sqrt{\overline{X}}$
2. $X' = LOG(X)$
3. $X' = LN(X)$
4. $X' = X - \overline{X}$
5. $X' = X / s$
6. $X' = \frac{X - \overline{X}}{s}$
7. $X' = \frac{X}{X_{max}}$
8. $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$

Nur die Transformationen 1, 2 und 3 sind nicht-linear und verändern somit die Form der Verteilung und sind u.U. geeignet, nicht normal verteilte Daten in normal verteilte zu überführen. Die Transformationen 4 bis 8 dienen der Standardisierung der Datenwerte und sind insbesondere für die Faktorenanalyse und die Clusteranalyse von Bedeutung.

Beispiel: Eine normal verteilte Zufallsvariable (Wasserstände in einem Bachbett) hat einen Erwartungswert von 112.5 cm und eine Standardabweichung von 37.34 cm. Gesucht ist die Wahrscheinlichkeit, mit der der Wert 200.0 cm überschritten wird.

Ergebnis:

$$\frac{200 - 112.5}{37.35} = 2.342$$

Unter Verwendung der Tabelle der Standardnormalverteilung ergibt sich aus einem Z-Wert von 2.342 eine Wahrscheinlichkeit von 0.00964 bzw. 0.964 % für die Überschreitung des Werts 200 cm.

Beispiel: Der Z-Wert für die Wahrscheinlichkeit von 0.025 (2.5 %) beträgt -1.96. Aufgrund der Symmetrie der Verteilungsfunktion ergibt sich für 1.96 eine gleichgroße Wahrscheinlichkeit. Hieraus folgt, dass alle Werte größer als 1.96 und kleiner als -1.96 eine Wahrscheinlichkeit von 5 % aufweisen oder, umgekehrt formuliert, der Wertebereich von -1.96 bis + 1.96 eine Wahrscheinlichkeit von 95 % hat.

Die Standardnormalverteilung wird als Prüfverteilung zur Verifizierung einer statistischen Hypothese genutzt:

- Liegt der jeweils berechnete Z-Wert im Bereich von -1.96 bis +1.96, so wird die Hypothese angenommen.
- Liegt der Z-Wert dagegen außerhalb dieses Bereichs, so wird die Hypothese auf dem 5 %- Niveau abgelehnt.

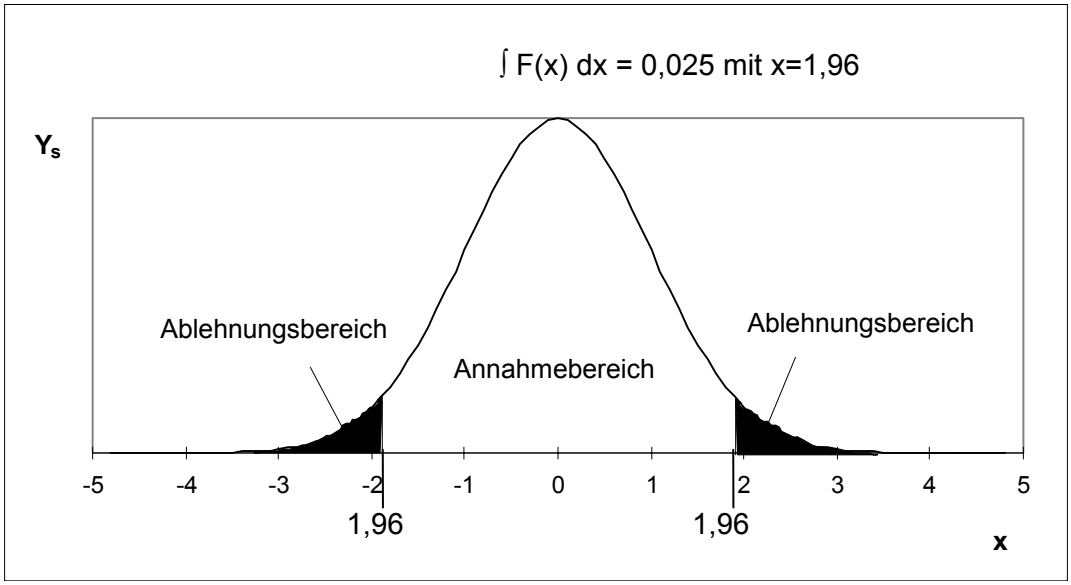


Abb. 9. Wahrscheinlichkeitsbereiche von 95 % bzw. 5 %

4.1.2.4 Tschebyscheff'

Sind die zu untersuchenden Daten nachweislich nicht normal verteilt, liefert die *TSCHEBYSCHEFF'sche Ungleichung* eine grobe Abschätzung der Wahrscheinlichkeit für eine bestimmte Abweichung *c* vom Erwartungswert.

$$P\left(\left|X-\mu\right| \geq c\right) \leq \frac{s^2}{c^2}$$

mit *s* als Standardabweichung (positiver Wert) und dem Erwartungswert μ.

Beispiel: Eine beliebig verteilte Zufallsvariable hat einen Erwartungswert μ von 200 und eine Standardabweichung von 15. Gesucht ist die Wahrscheinlichkeit, dass ein Wert von 170 unterschritten oder ein Wert von 230 überschritten wird. Die vorgegebene Abweichung vom Erwartungswert μ betragt 200 -170 = 30. Somit ergibt sich durch

Einsetzen: $15^2 / 30^2 = 225 / 900 = 0.25$

Die gesuchte Wahrscheinlichkeit ist somit höchstens 25 %.

4.1.2.5 Student-t-Verteilung

Die *t-Verteilung nach STUDENT* (Pseudonym des engl. Statistikers GOSSET) ist der Standardnormalverteilung sehr ähnlich. Sie ist stetig, glockenförmig und symmetrisch. Im Gegensatz zur Standardnormalverteilung ist sie jedoch abhängig von der Anzahl der Freiheitsgrade (*FG*). Die Anzahl der Freiheitsgrade ergibt sich aus der Differenz des Stichprobenumfangs *n* minus Anzahl der geschätzten Parameter *k*.

$$FG = n - k$$

Im Normalfall ist $k = 1$. Für große Freiheitsgrade ist die t-Verteilung identisch mit der Standard-Normalverteilung. Aber auch bei Freiheitsgraden von 100 ist der Unterschied zwischen t-Verteilung und Standardnormalverteilung nicht mehr sehr groß. Werte für verschiedene Freiheitsgrade sind in der Tabelle "T-Verteilung" aufgelistet. Verwendung findet die t-Verteilung insbesondere als Prüfgröße bei den t-Tests.

4.1.2.6 Fischer F-Verteilung

Die *F-Verteilung nach FISHER* (1925) ist eine *stetige, unsymmetrische linkssteile Verteilung*. Der Wertebereich erstreckt sich von 0 bis unendlich. Die Verteilung ist abhängig von zwei Freiheitsgraden n-1 und m-1, also dem Umfang von Beobachtungen zweier unabhängiger Stichproben n und m. Die F-Verteilung wird als Prüfverteilung herangezogen, wenn zwei oder mehr Stichproben verglichen werden (*F-Tests, Varianzanalyse*).

4.1.2.7 Chi-Quadrat-Verteilung

Die *CHI - Quadrat - Verteilung* ist ebenso wie die F-Verteilung *stetig und unsymmetrisch* (linkssteil) mit einem Wertebereich von 0 bis unendlich. Die CHI - Quadrat - Verteilung hängt ebenso wie die t-Verteilung nur vom Freiheitsgrad $FG = n - 1$ ab. Mit wachsendem Freiheitsgrad nähert sich die CHI - Quadrat - Verteilung der Normalverteilung an.

Die CHI - Quadrat - Verteilung wird u.a. als Prüfgröße beim *CHI - Quadrat - Test* sowie beim H - Test und *FRIEDMANN - Test* verwendet.

4.1.2.8 Log-Normalverteilung

Während die F-Verteilung und die CHI-Quadrat-Verteilung nur für relativ kleine Fallzahlen linkssteile, asymmetrische Prüfverteilungen ergeben, sind viele Verteilungen in der Natur unabhängig von der Fallzahl positiv schief, also linkssteil mit einem sehr flachen Auslaufen auf der rechten Seite. Dies ist häufig dadurch bedingt, dass die linke Seite durch einen nicht unterschreitbaren Wert (Konzentration = 0, Startzeit eines Experimentes etc.) gekennzeichnet ist. Solche Daten sind oft mit der *logarithmischen Normalverteilung* (*Log Normalverteilung*) darstellbar:

$$Y = \frac{1}{s * (2\pi)^{1/2}} * \frac{1}{X} * \quad e^{-1/2 \left(\frac{(\ln(X)-\mu)}{s}\right)^2}$$

Während die Standardnormalverteilung ebenso wie die t-Verteilung, die F-Verteilung und die CHI-Quadrat-Verteilung als Prüfverteilungen benutzt werden, dient die Log-Normalverteilung lediglich der Beschreibung eines Datenkollektivs.

Log normal verteilte Daten können durch Logarithmieren zu annähernd normal verteilten Daten transformiert werden.

Tab 5. STUDENT-Verteilung

Irrtumswahrscheinlichkeiten α für den zweiseitigen Test									
FG \ α	0.50	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001
1	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619	6366.198
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.598	99.992
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214	12.924	28.000
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610	15.554
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869	11.178
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959	9.082
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.885
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041	7.12
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781	6.594
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587	6.211
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437	5.921
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318	5.694
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221	5.513
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140	5.363
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073	5.239
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015	5.134
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965	5.044
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922	4.966
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.822	4.897
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850	4.837
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819	4.784
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792	4.736
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.767	4.693
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745	4.654
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725	4.619
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707	4.587
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690	4.558
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674	4.530
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659	4.506
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646	4.482
32	0.682	1.309	1.694	2.037	2.449	2.738	3.365	3.622	4.441
34	0.682	1.307	1.691	2.032	2.441	2.728	3.348	3.601	4.405
35	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591	4.389
36	0.681	1.306	1.688	2.028	2.434	2.719	3.333	3.582	4.374
38	0.681	1.304	1.686	2.024	2.429	2.712	3.319	3.566	4.346
40	0.680	1.303	1.684	2.021	2.423	2.704	3.307	3.551	4.321
42	0.680	1.302	1.682	2.018	2.418	2.698	3.296	3.538	4.298
45	0.680	1.301	1.679	2.014	2.412	2.690	3.281	3.520	4.269
47	0.679	1.300	1.678	2.012	2.408	2.685	3.273	3.510	4.251
50	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496	4.228
55	0.679	1.297	1.673	2.004	2.396	2.668	3.245	3.476	4.196
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460	4.169
70	0.678	1.294	1.667	1.994	2.381	2.648	3.211	3.435	4.127
80	0.678	1.292	1.664	1.990	2.374	2.639	3.195	3.416	4.096
90	0.677	1.291	1.662	1.987	2.358	2.632	3.183	3.402	4.072
100	0.677	1.290	1.660	1.984	2.364	2.626	3.174	3.390	4.053
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373	4.025
200	0.676	1.286	1.653	1.972	2.345	2.601	3.131	3.340	3.970
500	0.675	1.283	1.648	1.965	2.334	2.586	3.107	3.310	3.922
1000	0.675	1.282	1.646	1.962	2.330	2.581	3.098	3.300	3.906
∞	0.675	1.282	1.645	1.960	2.326	2.576	3.090	3.290	3.891
FG / α	0.50	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001

5 Statistische Kennzahlen

Statistische Kennzahlen oder *Maßzahlen* charakterisieren eine Stichprobe oder Grundgesamtheit in kürzester Form. Sie beschreiben grundsätzlich nur ein Datenkollektiv und beziehen sich auf eine Variable. Es kann unterschieden werden in Lage- und Streuungskennzahlen. *Lokalisationsmaße* (*Mittelwerte*) beschreiben die Lage einer Datenverteilung, während *Dispersionsmaße* (*Varianz*, *Standardabweichung*) ihre Breite charakterisieren.

5.1 Minimum, Maximum, Median, arithmetisches Mittel, Standardabweichung

Minimum und *Maximum* sind die Extremwerte einer Messgröße. Der Differenzbetrag zwischen Maximum und Minimum wird als *Spannweite* (*range*) bezeichnet.

Der Zentralwert eines Datenkollektivs wird als *Median* bezeichnet. Bei geraden Fallzahlen wird der Median aus den beiden mittleren Datenwerten gebildet.

$$Median = \frac{X_k + X_{k+1}}{2}$$

Als *Modus* (*mode*) bzw. *Gipfel-* oder *Modalwert* wird derjenige Wert benannt, der in einem Datenkollektiv mit der größten Häufigkeit auftritt. Der Modus ist undefiniert, wenn mehrere Ausprägungen der Messgröße die gleiche Häufigkeit aufweisen. Die am häufigsten verwendete Lagemaßzahl ist das *arithmetische Mittel* (*arithmetic mean*):

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Das Symbol $\overline{X}$ wird für den Mittelwert von Stichproben verwendet, für den Mittelwert von Grundgesamtheiten wird das Symbol μ benutzt. Die Zahl n bezeichnet die Anzahl der Fälle in dem Datenkollektiv. Neben dem arithmetischen Mittel wird auch das *getrimmte Mittel* (*trimmed mean*) verwendet. Dabei werden die größten und kleinsten Werte (z.B. 10% von beiden Extrema) nicht mitberücksichtigt.

Ein wichtiges Maß für die Streuung eines Datenkollektivs um den Mittelwert ist die *Varianz* s^2. Die Varianz der Grundgesamtheit wird mit d^2 bezeichnet. Die Varianz ergibt sich aus dem Quadrat der Abweichungen vom Mittelwert:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Beim Vergleich der Varianzen von zwei Messgrößen ist zu berücksichtigen, dass der Betrag der Varianz von der Dimension der Daten abhängig ist.

Aus der Varianz ergibt sich die *Standardabweichung* (*standard deviation*) zur Beschreibung der Streuung um den Mittelwert:

$$s = \sqrt{s^2}$$

Durch die Einführung des Variationskoeffizienten V_k

$$V_k = \frac{s}{\overline{X}} \qquad \text{für alle } X > 0$$

wird die Schwierigkeit bei der Interpretation der Varianzen verschiedener Messgrößen mit unterschiedlichen Wertebereichen umgangen. Der Variationskoeffizient ist ein dimensionsloses Streuungsmaß mit dem Mittelwert als Einheit.

Der relative *Variationskoeffzient* $V_{\%}$ bildet auf Werte zwischen 0 und 100 % ab:

$$V_{\%} = \frac{1}{\sqrt{n}} * \left(\frac{s}{\overline{X}} \right) * 100 \qquad \text{für alle } X > 0$$

5.2 Harmonisches und geometrisches Mittel

Die Anwendung des *harmonischen Mittels* ist dann sinnvoll, wenn in den Daten Reziprozität enthalten ist. Das harmonische Mittel ist definitionsgemäß nur für Daten ungleich 0 gültig.

$$\overline{X}_h = n * \frac{1}{\sum \frac{1}{X_i}}$$

Beispiel: Eine Strecke von A nach B (60 km) wird mit einer Durchschnittsgeschwindigkeit von 30 km/h zurückgelegt, auf dem Rückweg erreicht der Fahrer eine Geschwindigkeit von 60 km/h. Die Durchschnittsgeschwindigkeit errechnet sich wie folgt:

Für den Hinweg werden 2 Std., für den Rückweg 1 Std. benötigt, insgesamt werden in 3 Std. 120 km zurückgelegt. Daraus ergibt sich eine mittlere Geschwindigkeit von 40 km/h. Bildet man dagegen aus den beiden Durchschnittgeschwindigkeiten das arithmetische Mittel, erhält man fälschlicherweise eine durchschnittliche Geschwindigkeit von 45 km/h. Das harmonische Mittel liefert das richtige Ergebnis, ohne Rückrechnung über gefahrene Zeiten und zurückgelegte Kilometer:

$$\overline{X}_h = \frac{2}{1/30 + 1/60} = 40$$

Das harmonische Mittel liefert für das obige Beispiel aber nur dann das richtige Ergebnis, wenn die Teilstrecken gleich groß sind!

Was folgt daraus z.B. für die Hydrogeologie?

Die Mittelung von K_f-Werten sollte ebenfalls mit Hilfe des harmonischen Mittelwertes erfolgen, da es sich bei K_f-Werten auch um Geschwindigkeiten handelt.

Das *geometrische Mittel* wird verwendet, wenn Wachstumserscheinungen (Bakterienwachstum, Kristallisationsprozesse etc.) vorliegen.

Das geometrische Mittel ist nur für Werte größer als 0 definiert:

$$\overline{X}_g = (X_1 * X_2 * X_3 \quad \ldots \quad X_n)^{1/n}$$

5.3 Kreuztabellen

Kreuztabellen sind zweidimensionale Häufigkeitsverteilungen, wobei eine Stichprobe nach zwei Messgrößen (einer Zeilenvariablen und einer Spaltenvariablen) in Untergruppen gesplittet wird. Es werden die Fallzahlen innerhalb der Subgruppe, die sich durch eine solche Unterteilung ergeben, sowie ihre Prozentanteile von Zeilen-, Spalten- oder Gesamtsummen betrachtet.

Die beiden Messgrößen, die eine Kreuztabelle definieren, besitzen oft Nominalskalenniveau, jedoch ist dieses Verfahren auch für Messgrößen mit Ordinalskalenniveau gebräuchlich, sofern die Zahl der Stufen nicht allzu groß ist. Für Messgrößen mit Intervallskalenniveau und für stetige Zufallsgrößen ist eine Kreuztabellierung nur sinnvoll, wenn der Wertebereich vorher in Klassen gerastert wird.

Aus zweidimensionalen Kreuztabellen lassen sich *interferenzstatistische Maße* errechnen; z.B. der *CHI-Quadrat-Koeffizient*, anhand dessen geprüft wird, inwieweit eine Kreuztabelle auffällige Verteilungsunterschiede zwischen den einzelnen Zeilen bzw. Spalten aufweist. Der CHI-Quadrat-Test nimmt als Hypothese die Unabhängigkeit der beiden Messgrößen an (*Null-Hypothese*) und verwirft diese, wenn die Abweichungen zu groß sind. Die Berechnung beruht auf dem Vergleich der unter der Null-Hypothese erwarteten Häufigkeit E_{ij} und der tatsächlichen Häufigkeit N_{ij}:

$$CHI^2 = \sum_{i=1}^{k} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

mit den Freiheitsgraden $FG = (k-1)(I-1)$, den aus den relativen Randhäufigkeiten berechneten, erwarteten Häufigkeiten $E_{ij} = n * f_i * f_j$ und der empirisch beobachteten Häufigkeit N_{ij}.

Innerhalb der Zellen können weitere Informationen nützlich sein:

- erwartete Häufigkeit
- Zell-Chi-Quadrat

Der Zell-Chi-Quadrat-Wert stellt den Beitrag der jeweiligen Zelle zum Chi-Quadrat-Koeffizienten für die gesamte Tabelle dar. Als recht grobes Testverfahren kann dieser Chi-Quadrat-Beitrag jeder Zelle mit $FG = 1$ auf Signifikanz geprüft werden. Als Ergebnis dieser *Signifikanzprüfung* wird die *Zell-Signifikanz* berechnet und mit Hilfe einer Sternsymbolik ausgegeben. Üblicherweise werden drei Signifikanzniveaus unterschieden:

*	P < 0.05	signifikant
**	P < 0.01	"sehr" signifikant
***	P < 0.001	"hoch" signifikant

Diese Sprechweise ist nicht unkritisch, weil sie dazu führen kann, das Signifikanzniveau nicht vor der Durchführung des statistischen Tests festzulegen.

```
Zeilen -Var: As         KAT =   4
Spalten-Var: Temp.°C    KAT =   4

         Zeilen -KAT   1 :      0.0008 <= As        <       0.0289
         Zeilen -KAT   2 :      0.0289 <= As        <       0.0569
         Zeilen -KAT   3 :      0.0569 <= As        <       0.0850
         Zeilen -KAT   4 :      0.0850 <= As        <=      0.1130

         Spalten-KAT   1 :         7.2 <= Temp.°C   <         9.9
         Spalten-KAT   2 :         9.9 <= Temp.°C   <        12.6
         Spalten-KAT   3 :        12.6 <= Temp.°C   <        15.2
         Spalten-KAT   4 :        15.2 <= Temp.°C   <=       17.9

        Temp.°C

          KAT  1 KAT  2 KAT  3 KAT  4
As       ┌──────┬──────┬──────┬──────┐
         |   11 |   38 |    8 |    7 |     64
KAT   1  |      |      |      |      |   59.3
         | 17.2 | 59.4 | 12.5 | 10.9 |
         | 84.6 | 73.1 | 36.4 | 33.3 |
         | 10.2 | 35.2 |  7.4 |  6.5 |
         |    - |    - |    - |    - |
         ├──────┼──────┼──────┼──────┤
         |      |   11 |    1 |    6 |     18
KAT   2  |      |      |      |      |   16.7
         |      | 61.1 |  5.6 | 33.3 |
         |      | 21.2 |  4.5 | 28.6 |
         |      | 10.2 |  0.9 |  5.6 |
         |    - |    - |    - |    - |
         ├──────┼──────┼──────┼──────┤
         |    2 |    2 |   12 |    8 |     24
KAT   3  |      |      |      |      |   22.2
         |  8.3 |  8.3 | 50.0 | 33.3 |
         | 15.4 |  3.8 | 54.5 | 38.1 |
         |  1.9 |  1.9 | 11.1 |  7.4 |
         |    - |   ** |   ** |    - |
         ├──────┼──────┼──────┼──────┤
         |      |    1 |    1 |      |      2
KAT   4  |      |      |      |      |    1.9
         |      | 50.0 | 50.0 |      |
         |      |  1.9 |  4.5 |      |
         |      |  0.9 |  0.9 |      |
         |    - |    - |    - |    - |
         └──────┴──────┴──────┴──────┘
           13     52     22     21      108
         12.0   48.1   20.4   19.4    100.0

Die Zell-Einträge bedeuten von oben nach unten:
                                                 Zell-Häufigkeit
                                                 Zeilen-Prozent
                                                 Spalten-Prozent
                                                 Total-Prozent
                                                 Zell-Signifikanz

Anzahl der Fälle mit Missing-Values  = 97

        Chi-Quadrat          =   36.326  mit DF =   9  P = .000
```

6 Gruppenvergleiche

Zwei Stichprobenmittelwerte aus derselben Grundgesamtheit sind in der Regel unterschiedlich. Dieser Unterschied kann als zufällig oder signifikant interpretiert werden. Zu diesem Zweck wird üblicherweise die so genannte *Null-Hypothese* aufgestellt ($\mu_1 = \mu_2$). Nun wird getestet, ob diese Null-Hypothese abzulehnen ist.

Dabei können grundsätzlich zwei Fehler auftreten:

Fehler 1. Art: Die Hypothese wird verworfen, obwohl sie richtig ist. Die Wahrscheinlichkeit für den Fehler 1. Art nennt man die *Irrtumswahrscheinlichkeit*.

Fehler 2. Art: Die Hypothese wird nicht verworfen, obwohl sie falsch ist. Das Risiko für einen Fehler 2. Art wird desto größer, je kleiner die geforderte Irrtumswahrscheinlichkeit ist.

Verfahren, die bezüglich der Skalierung lediglich Nominal- oder Ordinalskalenniveau voraussetzen und keine Anforderung an die Verteilung der Daten stellen, werden als verteilungsfreie bzw. **nicht-parametrische Tests** bezeichnet. Alle Verfahren, die eine Normalverteilung sowie zumindest Intervallskalenniveau voraussetzen, werden als **parametrische Tests** bezeichnet.

6.1 Parametrische Tests

6.1.1 T-Test

Mit dem t-Test für abhängige Stichproben wird untersucht, ob sich die Mittelwerte einer Messgröße bzw. die Differenzen von **zwei** Messgrößen signifikant von Null unterscheiden (Vorher-Nachher-Vergleiche):

$$t = \left|\overline{X}\right| * \frac{1}{\sqrt{\frac{s^2}{n}}}$$

Der t-Test für unabhängige Stichproben prüft, ob sich **zwei** Subgruppen einer Stichprobe hinsichtlich der Mittelwerte einer Messgröße unterscheiden. Für die Berechnung des t-Wertes werden zwei Ansätze unterschieden, je nachdem, ob die Varianzen innerhalb der beiden Subgruppen gleich groß (homogen) oder sehr unterschiedlich (heterogen) sind.

$$t = \left|\overline{X}_1 - \overline{X}_2\right| * \frac{1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Wurden mehr als zwei unabhängige Stichproben erhoben, bzw. mehr als zwei Untergruppen durch eine weitere Variable definiert, können die Mittelwerte der Stichproben bzw. Untergruppen mit Hilfe der Varianzanalyse untersucht werden.

Die Subgruppen müssen durch eine "unabhängige" Variable (*Klassifizierungsvariable*) definiert sein. Weisen die beiden Subgruppen die gleiche Fallzahl auf, kann die Prüfung auch mittels des t-Tests für abhängige Stichproben durchgeführt werden. Dies erfordert allerdings eine andere Datenstruktur.

6.1.2 Varianzanalyse

Die Varianzanalyse wird anstelle des T-Tests verwendet, wenn **mehr als zwei** Subgruppen untersucht werden sollen. Die einfache Varianzanalyse verlangt normal verteilte Grundgesamtheiten auf Intervallskalenniveau und unabhängige Stichproben, die die gleiche Varianz haben. Wenn die Anzahl der Fälle in den Subgruppen sehr unterschiedlich ist und dies den realen Verhältnissen der Stichprobe entspricht, ist es sinnvoll, die Subgruppenmittelwerte mit dem arithmetischen Mittel der Gesamtstichprobe zu vergleichen. Unter der Annahme, dass die unterschiedlichen Subgruppengrößen das Ergebnis von Stichprobenfehlern sind, wird das harmonische Mittel zum Vergleich herangezogen werden. Das Prinzip der Varianzanalyse besteht in einer Stufenweisen Zerlegung der Summe der Abweichungsquadrate (Sum of Squares) QS_{gesamt}:

$$QS_{gesamt} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2$$

mit k Anzahl der Subgruppen

. n_i Anzahl der Fälle je Stichprobe (Subgruppe)

X_{ij} Messwerte in der i-ten Stichprobe

$\overline{X}$ Mittel aller n-Werte (arithmetisches oder harmonisches Mittel)

Die Summe der Abweichungsquadrate vom jeweiligen Gruppenmittelwert wird als $QS_{innerhalb}$ und die Differenz QS_{gesamt} - $QS_{innerhalb}$ wird als $QS_{zwischen}$ bezeichnet.

$QS_{gesamt} = QS_{innerhalb} + QS_{zwischen}$

Für diese Summen ergeben sich bei k Subgruppen die folgenden Freiheitsgrade (FG):

$$FG_{gesamt} = n - 1$$

$$FG_{innerhalb} = n - k$$

$$FG_{zwischen} = k - 1$$

Die sogenannten *mittleren Quadrate* (*Mean square sums, MQS*) ergeben sich aus den Quotienten:

$$MQS_{zwischen} = \frac{QS_{zwischen}}{FG_{zwischen}}$$

$$MQS_{innerhalb} = \frac{MQS_{innerhalb}}{FG_{innerhalb}}$$

Wenn zwischen den Gruppen keine signifikanten Unterschiede bestehen, unterscheiden sich $MQS_{zwischen}$ und $MQS_{innerhalb}$ auch nur zufällig. Dies kann durch die F-verteilte Prüfgröße getestet werden.

$$F = \frac{MQS_{zwischen}}{MQS_{innerhalb}}$$

Die Signifikanz des F-Wertes ergibt sich aus k -1, n - k Freiheitsgraden und führt letztlich zu der Aussage, ob signifikante Unterschiede bestehen. Die Varianzanalyse macht keine Aussage, zwischen welchen Gruppen im Detail Unterschiede bestehen.

6.2 Nicht-parametrische Tests

Sind die Differenzen von Stichproben oder Subgruppen nicht normal verteilt und nicht auf Intervallskalenniveau (dies kann immer dann angenommen werden, wenn die Daten selber diese Kriterien nicht erfüllen), so haben parametrische Tests keine Aussagekraft. In diesem Fall sind nicht-parametrische Tests, so genannte Rangverfahren zu verwenden. Durch alleinige Betrachtung der Ränge, nicht mehr der Rohdaten, wird nur noch ein Ordinalskalenniveau betrachtet, welches keine Implikationen mehr über die Zahlenwerte gibt. Im Rahmen dieser Rangverfahren werden die Daten zunächst aufsteigend sortiert und anschließend in Ränge unter Berücksichtigung von "ties" (Werte mit gleichem Rang) transformiert.

6.2.1 Mann-Whitney-Test

Anstelle des T-Tests kann für **zwei** Subgruppen der WILCOXON-Test oder der MANN-WHITNEY-Test (auch als U-Test bekannt)) angewandt werden.

Beispiel:

Gruppe A: 4.6, 4.7, 4.9, 5.1, 5.2, 5.5, 5.8, 6.1, 6.5, 6.5, 7.2 → X = 5.6

Gruppe B: 5.2, 5.3, 5.4, 5.6, 6.2, 6.3, 6.8, 7.7, 8.0, 8.1 → X = 6.5

4.6	1	A
4.7	2	A
4.9	3	A
5.1	4	A
5.2	5.5	A
5.2	5.5	B
5.3	7	B
5.4	8	B
5.5	9	A
5.6	10	B
5.8	11	A
6.1	12	A
6.2	13	B
6.3	14	B
6.5	15.5	A
6.5	15.5	A
6.8	17	B
7.2	18	A
7.7	19	B
8.0	20	B
8.1	21	B

Rohdaten Gruppe A	Rohdaten Gruppe B	Rangdaten Gruppe A	Rangdaten Gruppe B
4.6	5.2	1	5.5
4.7	5.3	2	7
4.9	5.4	3	8
5.1	5.6	4	10
5.2	6.2	5.5	13
5.5	6.3	9	14
5.8	6.8	11	17
6.1	7.7	12	19
6.5	8.0	15.5	20
6.5	8.1	15.5	21
7.2		18	
Summe der Ränge T		96.5	134.5
Mittelwert der Ränge Tmean		8.8	13.5

Die weitere Berechnung kann anhand eines Verhältnisses z oder eines Wertes U erfolgen.

1.) Berechnung eines Verhältnis z (normalverteilt):

$$z = \frac{(T - Tmean) \pm 0.5}{\sigma_T}$$

mit T = Summe der Ränge jeder Gruppe

Tmean = Mittelwert der Probenverteilung

Der Wert z wird verglichen mit kritischen z-Werten aus Tabellenwerken in Abhängigkeit des Signifikanzniveaus.

2.) Berechnung von des Wertes U („U-Test“):

$U = T[max] - T$ mit T[max] = Max. Summe der Ränge bei Anzahl n an Werten in jeder Gruppe (21+20+19+18+17+16+15+14+13+12+11 = 176 für Gruppe A, n = 11)

T = tatsächlich ermittelte Summe der Ränge jeder Gruppe

	Tmax	T	ΔT
Gruppe A	176	96.5	79.5
Gruppe B	165	134.5	30.5

ΔTmean = 55, Vgl. mit $U_{krit.}$ (Tabellen) mit p = 90%, U zwischen 31 - 79

6.2.2 Kruskal-Wallis-Test

Bei **mehr als zwei** Subgruppen kann anstelle der Varianzanalyse der *Kruskal - Wallis -Test* durchgeführt werden. Bei diesem nicht-parametrischen Verfahren wird untersucht, ob sich die mehrere Subgruppen einer Stichprobe hinsichtlich des mittleren Ranges einer Messgröße signifikant unterscheiden.

6.3 Korrelation

Mit statistischen Verfahren wie der Korrelation lässt sich bestimmen, wie hoch der statistische Zusammenhang zwischen zwei Parametern, d.h. unterschiedlichen Messgrößen wie z.B. Wasserleitfähigkeit und Grundwasserflurabständen, ist. Dabei ist zu beachten, dass nicht prozessorientierte, sondern nur statistische Zusammenhänge betrachtet werden! Es kann also nicht berücksichtigt werden, ob zwei zu betrachtende Parameter überhaupt in irgendeiner (kausalen) Beziehung zueinander stehen. Ein signifikanter statistischer Zusammenhang kann somit durchaus zu unsinnigen Schlussfolgerungen verleiten, z.B. aus der Korrelation zwischen einem Rückgang der Störche und einem Geburtenrückgang darauf schließen zu wollen, dass der Storch die Kinder bringt.

Ein Maß für den statistischen Zusammenhang zwischen zwei Messgrößen ist *die Co-Varianz COV*:

$$COV = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

Je stärker der Zusammenhang ist, desto größer ist die Co-Varianz. Die Co-Varianz ist abhängig vom Wertebereich der untersuchten Messgrößen. Aus diesem Grund wird die Co-Varianz mit Hilfe der Varianzen der zwei zu untersuchenden Messgrößen S_x und S_y normiert, d.h. es wird das Verhältnis zwischen der Co-Varianz und dem Produkt der Standardabweichungen der Messgrößen X und Y betrachtet. Diese normierte Größe wird als Korrelationskoeffizient r bezeichnet. Er ist dimensionslos, auch als **PEARSON-Korrelationskoeffizient** bzw. Produktmomentkorrelation bekannt:

$$r = \frac{COV}{S_x * S_y}$$

Mit Hilfe des Korrelationskoeffizienten wird geprüft, ob es einen Zusammenhang zwischen zwei Messgrößen einer Stichprobe gibt. Definitionsgemäß kann r Werte von -1 bis +1 annehmen. Ist r = 0, sind die beiden Messgrößen unkorreliert. Bei positivem r verhalten sich die beiden Messgrößen gleichsinnig (Y nimmt zu, wenn X zunimmt), bei negativem r ist das Verhalten gegensinnig (inverse Korrelation).

Der Korrelationskoeffizient *r* ist nicht unabhängig von der Zahl der Fälle (N) der Stichprobe. Um einen Korrelationskoeffizienten zu beurteilen, muss man neben der Zahl der zugrunde liegenden Messungen die Art der Messgrößen berücksichtigen. Die Irrtumswahrscheinlichkeit P wird über eine Formel aus der F-Verteilung geschätzt oder ist aus entsprechenden Tabellen zu entnehmen:

$$F = \frac{r^2\left(n-2\right)}{1-r^2}$$

$$P = \left(0.5 * \log\frac{1+r}{1-r}\right)\sqrt{n-1} * \frac{1}{\sqrt{2}} \qquad \text{mit } r \neq 1$$

Beispiel: Für eine Stichprobe mit 12 Fällen ergibt sich für ein Messgrößenpaar X - Y ein Korrelationskoeffizient von r = 0.78 und ein Signifikanzniveau von P = 0.014. Das bedeutet, dass der Korrelationskoeffizient von r = 0.78 mit einer statistischen Sicherheit von 98.6 % von Null verschieden ist bzw. mit einer Wahrscheinlichkeit von 98.6 % ein Zusammenhang zwischen den beiden Messgrößen besteht.

6.3.1 Rangkorrelation

Der PEARSON-Korrelationskoeffizient darf nur dann zur Beurteilung herangezogen werden, wenn beide Messgrößen **normal verteilt** sind und auf Intervallskalenniveau interpretierbar sind. Ferner muss der Zusammenhang zwischen den beiden Messgrößen linearer Natur sein. Ist eine oder beide Messgrößen nicht normal verteilt bzw. nicht auf Intervallskalenniveau interpretierbar, müssen parameterfreie Verfahren wie die Rangkorrelation nach **SPEARMAN** oder **KENDALL** verwendet werden. Dabei werden die Originaldaten zunächst auf Ränge, d.h. eine Rangfolge abgebildet. Die Bewertung des Korrelationskoeffizienten erfolgt wie beim PEARSON-Korrelationskoeffizienten. Zwischen SPEARMAN und KENDALL - Verfahren besteht in der Regel kein großer Unterschied. Durch die Abbildung auf die Rangfolge und die Korrektur von "ties" (Fälle mit gleichen Werten) sind Rangkorrelationsverfahren wesentlich rechenzeitintensiver als das PEARSON-Verfahren.

Wenn der Zusammenhang zwischen den beiden Messgrößen nicht linearer Natur ist, muss der nicht-lineare Korrelationskoeffizient zur Überprüfung des Zusammenhanges benutzt werden. Bei einem kleinen Stichprobenumfang (n < 30) empfiehlt es sich nach OLKIN & PRATT (1958) den korrigierten Korrelationskoeffizienten r(korr) zu verwenden:

$$r(korr) = r\left(1 + \frac{1 - r^2}{2(n-3)}\right)$$

6.3.2 Partielle Korrelation

Mit Hilfe der *partiellen Korrelation* werden partielle Korrelationskoeffizienten zwischen Messgrößenpaaren (= Variablenpaaren) berechnet und dabei lineare Effekte anderer Messgrößen korrigiert. Es werden also Wechselwirkungen von zwei korrelierenden Messgrößen mit einer oder mehr zusätzlichen Messgrößen berücksichtigt. Diese Korrelationskoeffizienten messen die Stärke der linearen Assoziation zwischen zwei Messgrößen, außer derjenigen, die aus einer gemeinsamen Assoziation mit der bzw. den Kontroll-Messgrößen stammt. Auf diese Weise kann z.B. eine Beziehung zwischen Quellschüttung und Grundwasserneubildung bei Prüfung der Größe des Einzugsgebietes und der hydraulischen Parameter untersucht werden.

6.4 Regression

6.4.1 Lineare Regression

Zeichnet man in ein kartesisches Koordinatensystem die Wertepaare X - Y von zwei zu untersuchenden Messgrößen ein, erhält man ein Streudiagramm (Scatterplot). Ziel der linearen Regressionsanalyse ist es, beobachteten Datenverteilungen eine Regressionsgerade in der Form Y = B*X + A anzupassen. Dabei ist Y die Zielgröße (ahhängige Variable) und X die Einflußgröße (unabhängige Variable).

A ist der Achsenabschnitt; er gibt an, wie groß Y an der Stelle X = 0 ist. B ist die Steigung der Regressionsgeraden und beschreibt, um wie viel sich die Zielgröße Y im Mittel verändert, wenn X um eine Maßeinheit erhöht bzw. erniedrigt wird. Die Steigung (B) wird vielfach auch als Regressionskoeffizient bezeichnet. Ist er negativ, so bedeutet dies, dass Y abnimmt, wenn X zunimmt. Bei positivem Regressionskoeffizienten verhalten sich X und Y gleichsinnig. Die Konstanten A und B werden nach der Methode der kleinsten Fehlerquadrate berechnet.

Unter der Annahme, dass beide Messgrößen Zufalls-Messgrößen sind, kann sowohl Y aus X als auch X aus Y geschätzt werden. Hieraus ergeben sich zwei Regressionsgeraden, die ihren Schnittpunkt im Schwerpunkt (Mittelwerte von X und Y) haben. Nur wenn ein vollständiger linearer Zusammenhang zwischen X und Y besteht (r = 1, P = 0.00000), sind die beiden Regressionsgeraden deckungsgleich. Je kleiner der Absolutbetrag der Korrelation zwischen X und Y ist, desto größer ist die Schere zwischen den beiden Geraden:

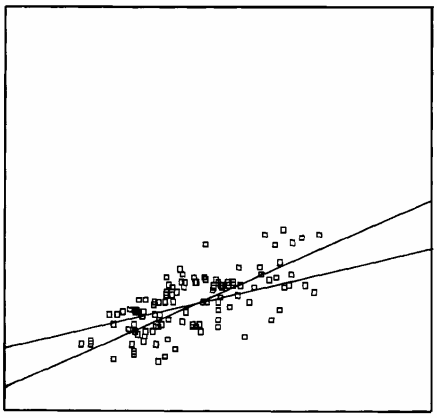


Abb. 10. Scatterplot der Messgrößen *X* und *Y* mit beiden möglichen Regressionsgeraden.

Der Korrelationskoeffizient *r* kann aus beiden Regressionskoeffizienten *B* berechnet werden:

$$r = \sqrt{B_{yx} * B_{xy}}$$

Die Regressionsgerade ist mit einem Streubereich behaftet, der durch den Standardfehler der Schätzung beschrieben wird:

$$S_{yx} = \sqrt{\frac{Q_y - Q_{xy}^2 / Q_x}{n-2}}$$

$$Q_x = \sum X^2 - \frac{\left(\sum X\right)^2}{N}$$

$$Q_y = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N}$$

$$Q_{xy} = \sum XY - \frac{\sum X * \sum Y}{N}$$

Aus dem Standardfehler der Schätzung kann für jedes X eine Abweichung von der geschätzten Regressionsgeraden berechnet werden und z.B. in Form eines 95 %-Vertrauensbereiches grafisch dargestellt werden:

$$Delta\ X = S_{yx} * \sqrt{\frac{1/N + \left(X - \overline{X}\right)^2}{Q_x}} * 2F$$

Eine beliebte Fehlinterpretation ist, zu sagen, alle Punkte außerhalb des Vertrauensbereichs seien unzuverlässig, fehlerhaft, Ausreißer etc. Das ist völliger Unsinn: Der Vertrauensbereich gibt an, in welchem Maße die Regressionsgerade um den Schwerpunkt gedreht werden kann unter Annahme einer statistischen Wahrscheinlichkeit von z.B. 95 %. Er macht damit eine Aussage möglich, ob man von einem signifikanten Trend sprechen kann.

In der folgenden Abbildung ist das Ergebnis somit, dass sich aus der Regression zwischen Nitratgehalt im Grundwasser und der Zeit keine auf dem 95% Signifikanz-Niveau abgesicherte Aussage ableiten lässt: Zwar zeigt die Regressionsgerade einen moderaten Anstieg; der 95% - Vertrauensbereich zeigt aber deutlich, dass die Regressionsgerade auch eine abnehmende Tendenz annehmen könnte.

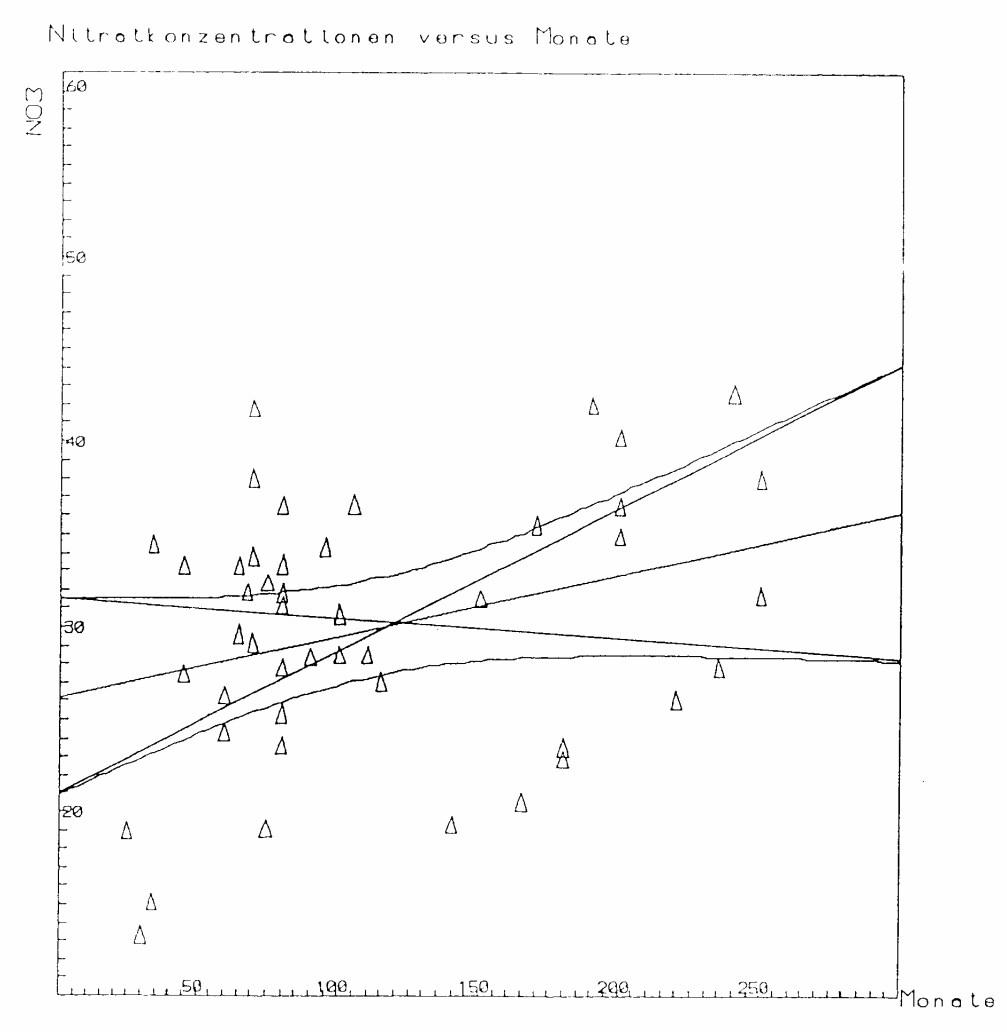


Abb. 11. Regressionsgrade im 95 % Vertrauensbereich

6.4.2 Nichtlineare Regression

Mit Hilfe der nichtlinearen Regression wird versucht, den Daten eine beliebige Funktion möglichst optimal anzupassen, z.B.:

Exponentialfunktion: $Y = A * e^{B*X}$

Potenzfunktion: $Y = A * X^{B}$

Hyperbolische Funktion $Y = A + B/X$:

Einen anderen häufig verwendeten nichtlinearen Funktionstyp stellen polynominale Regressionsansätze dar:

$$Y = A + B_1 * X + B_2 * X^2 + B_3 + X^3 + \ldots + B_n * X^n$$

Das ausgeprägte Schwingungsverhalten von Polynomen führt dazu, dass sie zur Interpolation von Datenwerten ungeeignet sein können. Das Extrapolieren (Trendberechnung) von Werten mit Hilfe der

polynominalen Regression sollte in jedem Fall unterbleiben. In allen Fällen, bei denen Werte mit Hilfe einer Regressionsfunktion außerhalb des Messintervalls geschätzt werden sollen, muss mit äußerster Vorsicht gearbeitet werden.

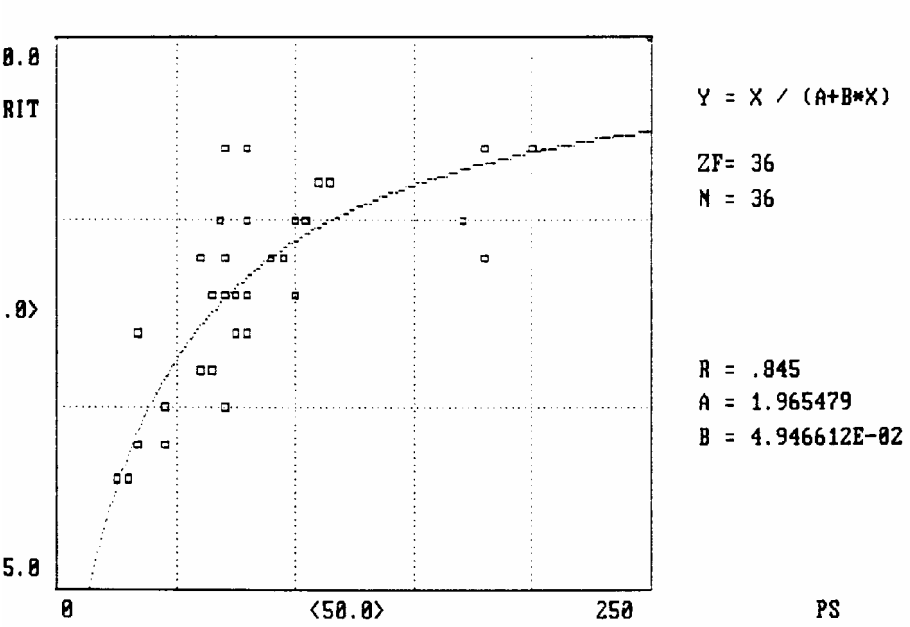


Abb. 12. Nichtlineare Regression

6.4.3 Multiple lineare Regression

Mit Hilfe der multiplen linearen Regression wird der Zusammenhang zwischen einer Messgröße Y und mehreren Messgrößen X_1, X_2 ... X_n untersucht. Der multiple lineare Korrelationskoeffizient wird durch die lineare Regressionsfunktion

$$Y = A + B_1 * X_1 + B_2 * X_2 + \ldots B_n * X_n$$

berechnet. Das Gleichungssystem kann nicht gelöst werden, wenn die Zahl der Fälle kleiner oder gleich der Zahl der Messgrößen ist oder falls lineare Zusammenhänge zwischen den Messgrößen bestehen.

Die sog. Residuen sind die Abweichungen zwischen dem berechneten Y und dem tatsächlich beobachteten Y. Neben dem Korrelationskoeffizienten und dem zugehörigen Signifikanzniveau ist eine Analyse der Residuen ein wichtiges Kriterium für die Güte der multiplen linearen Regression.

Durch entsprechende Transformationen der Einfluß-Messgrößen (z.B. Logarithmieren) ist es sehr einfach möglich, aus einer linearen multiplen Regression eine nichtlineare multiple Regression zu machen.

6.4.4 Fehlanwendungen der linearen Regression

Auf häufig gemachte Fehlanwendung bezüglich der linearen Regression wird im Folgenden eingegangen. Gegeben ist ein lineares Regressionsmodell:

$$Y = A + B * X + \delta$$

A und B sind Koeffizienten und δ ist ein statistischer Zufallsterm.

Damit dieses lineare Regressionsmodell angewendet werden darf, müssen X und Y definitionsgemäß unabhängige Messgrößen sein. Zusätzlich muss gelten:

- X und Y sind fehlerfrei gemessen
- X und Y sind normal verteilt
- Es existiert eine lineare Abhängigkeit zwischen X und Y
- Der Fehler δ ist normal verteilt, sein Mittelwert ist Null, die Varianz ist konstant und nicht autokorreliert.

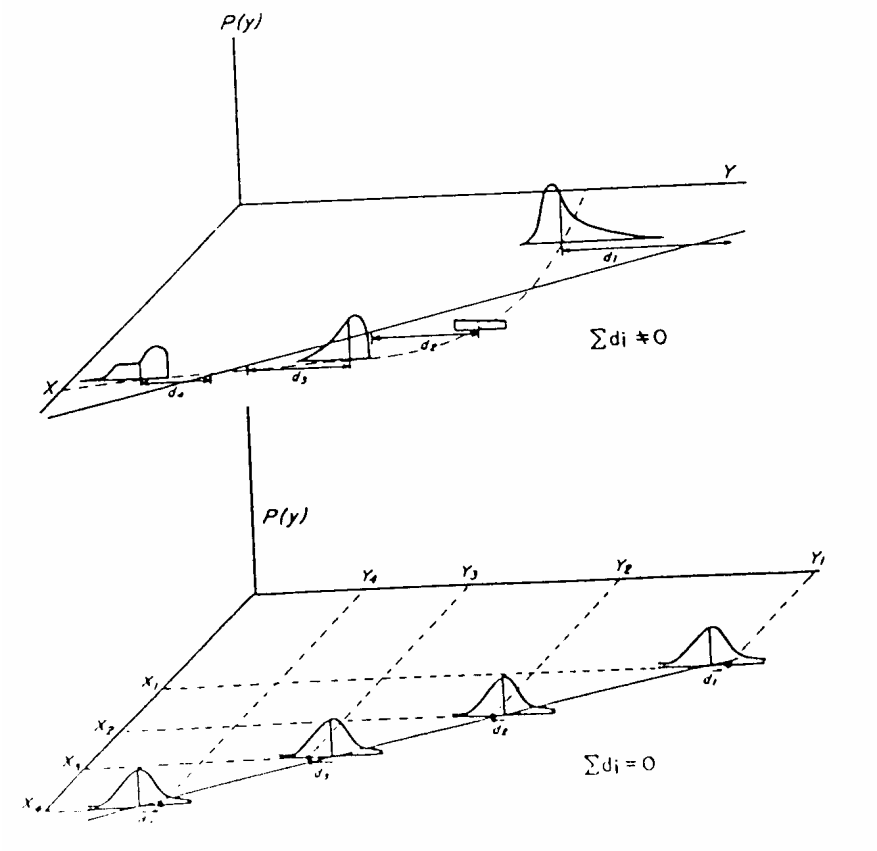


Abb. 13. Im oberen Bild sind die Kriterien alle nicht eingehalten; die lineare Regression ist sonst für diese Daten ungeeignet; im unteren Bild sind demgegenüber die Randbedingungen erfüllt (aus SIZE 1987).

Im Folgenden sind vier Scatterplots von 20 randomisierten X - Y - Paaren dargestellt:

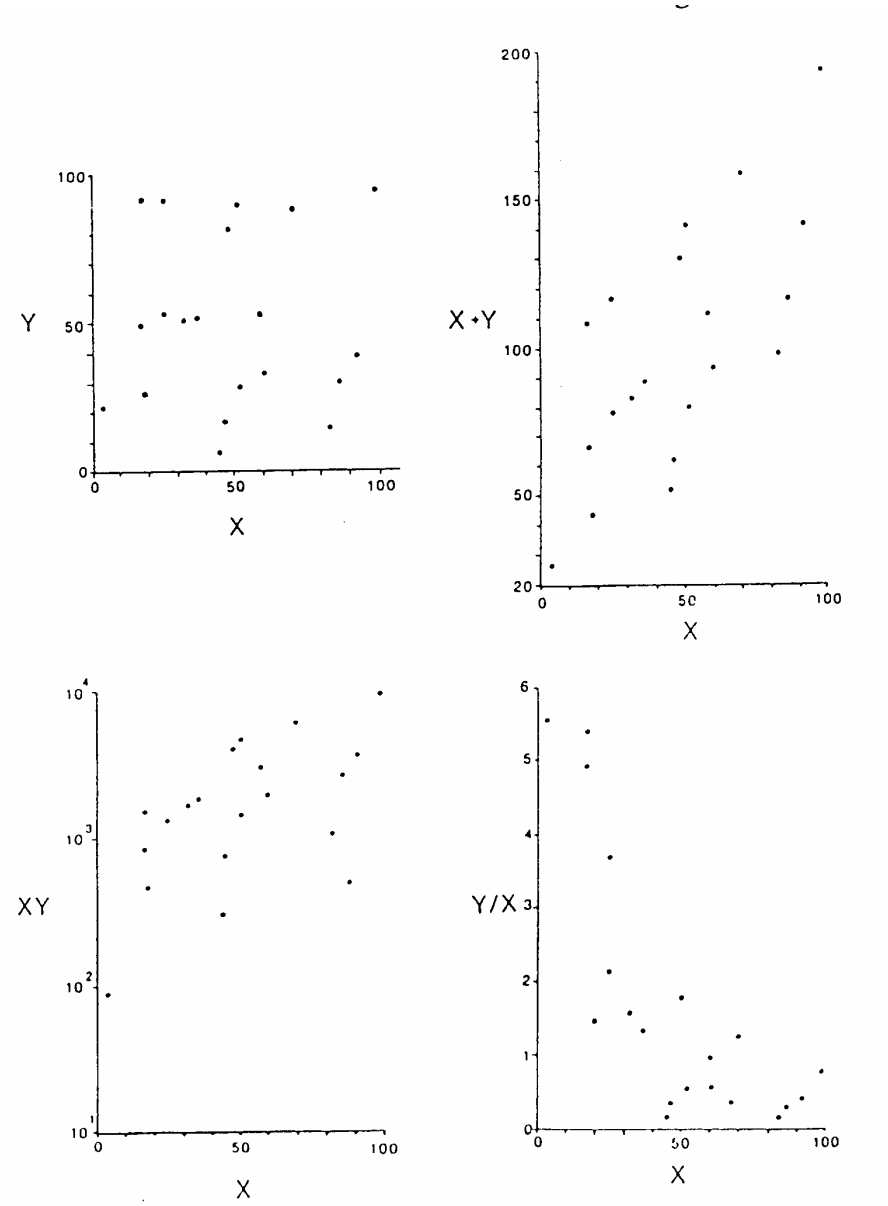


Abb. 14. Scattergramm mit Zufallsdaten und sinnlose Korrelationen (aus SIZE 1987)

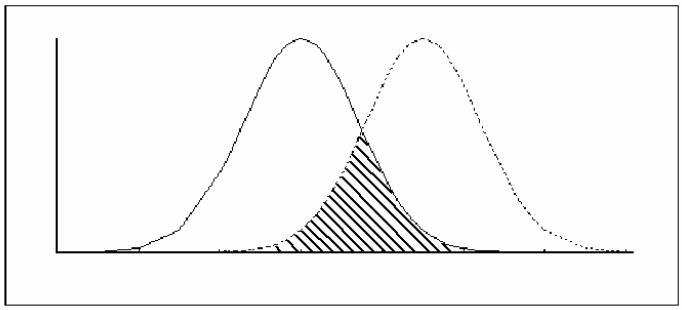
Links oben sind die Zufallszahlen zu sehen Dieser Scatterplot von X gegen Y zeigt erwartungsgemäß keinerlei Abhängigkeit, eine lineare Regression ist somit sinnlos. Die anderen drei Scatterplots wurden hergestellt, indem die Rohdaten durch einfache arithmetische Operation (Y' = XY ; Y' = X + Y ; Y' = Y/X) verändert wurden. Die Plots von X gegen die Summe X + Y oder das Produkt X * Y oder den Quotienten Y/X deuten durchaus lineare oder nichtlineare Zusammenhänge an.

Aber: In allen drei Fällen handelt es sich um völlig sinnlose Korrelationen. Ein Beispiel für die Transformation $X + Y$ ist die Regression der Härte (°dH) mit der Ca-Konzentration.

6.5 Diskriminanzanalyse

Die Diskriminanzanalyse klassifiziert Fälle in eine oder mehrere bekannte Gruppen; als Basis können verschiedene Merkmale dienen. Eine Anwendung der Diskriminanzanalyse ist sinnvoll, wenn 2 Stichproben mit mindestens 2 Messgrößen hinsichtlich der Gruppenzugehörigkeit untersucht werden sollen und t-Teste diesbezüglich versagen. Die Diskriminanzanalyse gibt Auskunft über die Frage, ob

- Stichproben unterschiedlichen Populationen angehören
- zu welchen Populationen Daten im Überlappungsbereich gehören



Als Beispiel dient folgender einfacher Datensatz mit 3 bzw. 4 Fällen:

Probe A		
i	$\mathbf{X_{ai}}$	$\mathbf{Y_{Ai}}$
1	172	130
2	180	124
3	184	130
4	168	124

n = 4

μXA = 176

μYA = 127

Probe B		
i	$\mathbf{X_{aj}}$	$\mathbf{Y_{Aj}}$
1	191	128
2	187	112
3	180	117

n = 3

μXB = 186 $\rightarrow$ = 10

μYB = 119 $\rightarrow$ = 8

Neue zusammen gefasste Eigenschaften ergeben sich durch folgende Linearform:

$Z = ax + by$; wobei *a, b* Konstanten sind

unter der Annahme a = +1 und b = -1, ergibt sich:

Probe A	
i	ZA
1	42
2	56
3	54
4	44

Probe B	
j	ZB
1	63
2	75
3	63

Nach dieser Transformation gibt es keine Überlappungen mehr:

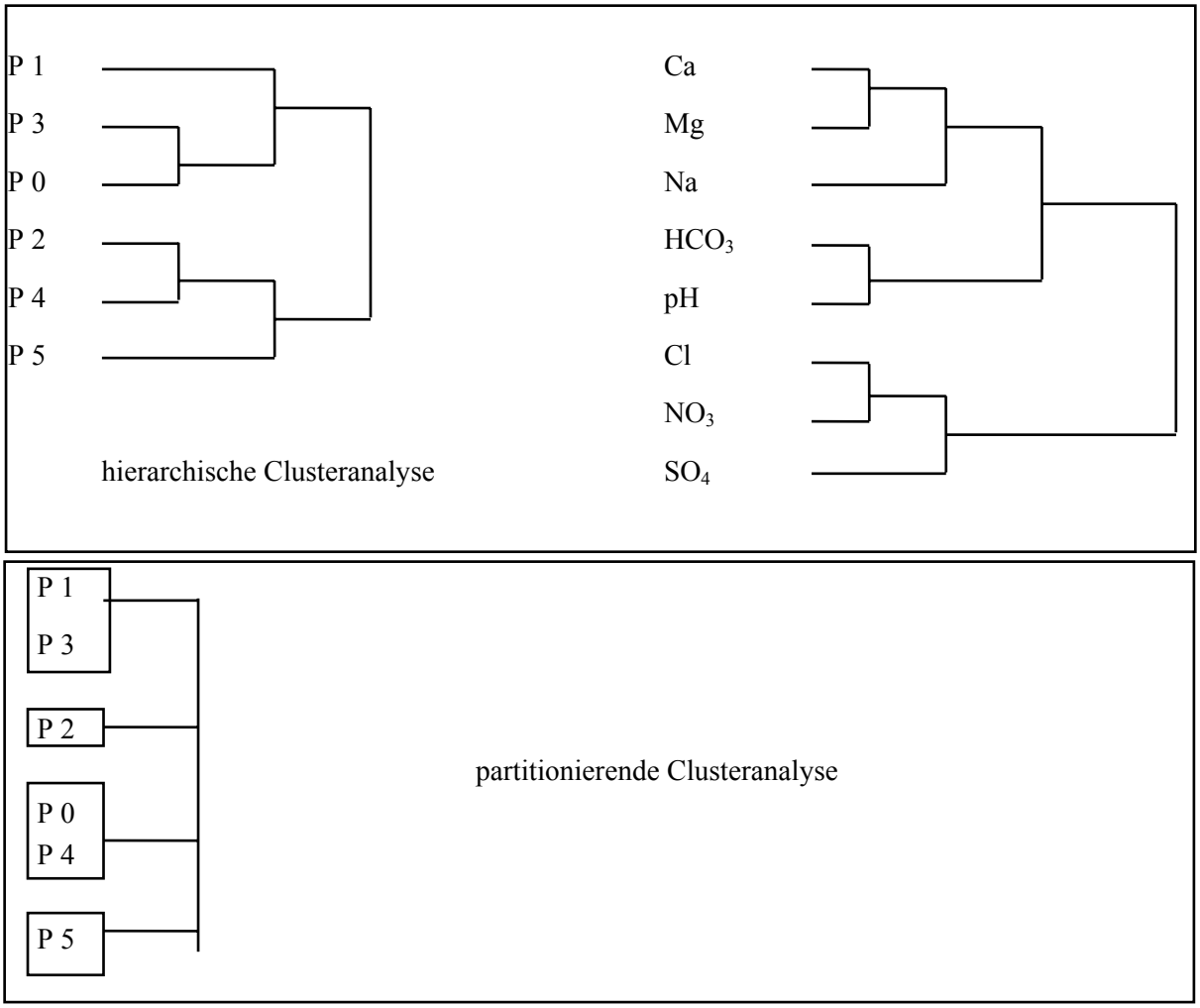
μ Z A = 49 μ Z B = 67

Auch der Abstand der Mittelwerte ist besser:

μ X A - μ X B = 10
μ Y A - μ Y B = 8
μ Z A - μ Z B = 18

Clusteranalyse

Unter dem Begriff Clusteranalyse werden Verfahren verstanden, die in der Lage sind, eine beliebige Datenmenge in Gruppen zu klassifizieren. Das Wort Clusteranalyse steht somit nicht für ein definiertes mathematisches Verfahren wie z.B. bei der Faktorenanalyse, sondern für eine Vielzahl möglicher Strategien. Zwei in sich verschiedene Arten der Gruppenbildung sind zu nennen: hierarchische und partitionierende Clusterbildung.



Die hierarchische Clusterbildung klassifiziert Messgrößen und liefert als grafisches Ergebnis ein Dendogramm und zeigt die Beziehung der einzelnen Messgrößen zueinander auf.

Demgegenüber werden durch partitionierende Algorithmen die Fälle eines Datensatzes klassifiziert. So können z.B. Wasseranalysen aus einem Untersuchungsgebiet in Subgruppen unterteilt werden. Das prinzipielle Vorgehen partitionierender Algorithmen wird im Folgenden kurz besprochen. Ausgangspunkt ist in jedem Fall die Korrelationsmatrix der betreffenden Messgrößen:

Var 1	1					
Var 2	0,5	1				
Var 3	0,7	0,3	1			
Var 4	0,4	0,2	0,9	1		
Var 5	0,1	0,4	0,3	0,6	1	
Var 6	0,99	0,5	0,4	0,5	0,8	1
	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6

Zunächst wird jedes Element als ein Cluster angenommen. Anschließend werden in der Ähnlichkeitsmatrix die Cluster mit größter Ähnlichkeit gesucht und daraus neue Cluster gebildet. Für diese neuen Cluster werden neue Ähnlichkeitsmaße gemäß $C_{pq},i = ½ \ (S_{pi} + S_{qi})$ mit p, q = neue Cluster, i = restliche Cluster berechnet (Unweighted Average Linkage Method).

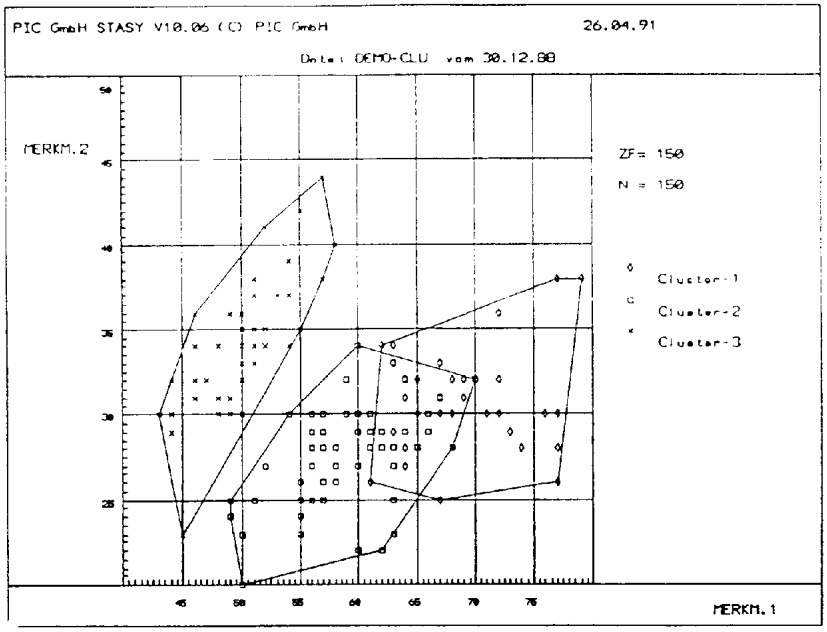
Vorgehen bei iterativ partitionierender Verfahren (MacQueen's K-Means-Verfahren):

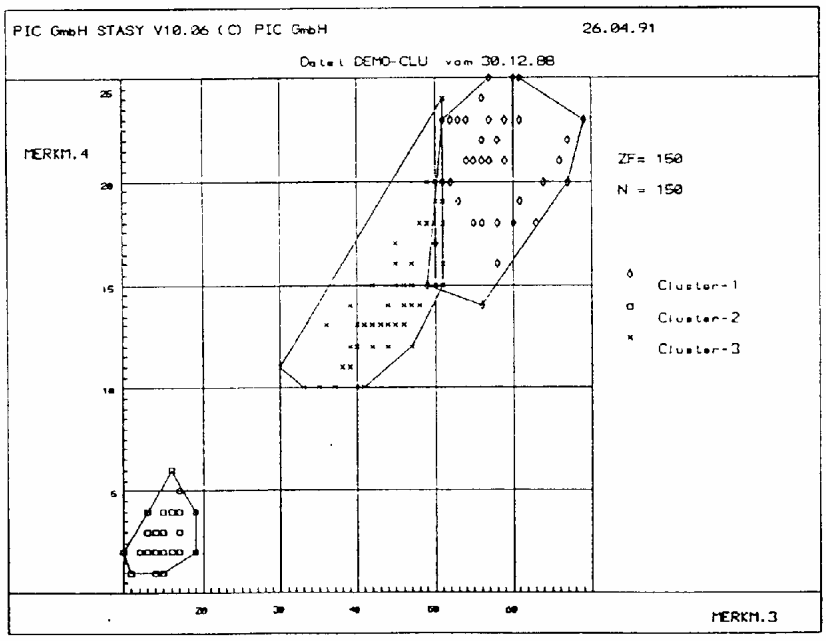
- wähle Anzahl der Cluster
- die Cluster mit den ersten Fällen vor besetzen Cluster 1 = Fall 1, Cluster 2 = Fall 2, ...
- der Reihe nach die restlichen Fälle dem Cluster zuordnen, zu dessen Schwerpunkt der Abstand am kleinsten ist
- Schwerpunkt neu berechnen
- Schwerpunkt fixieren
- Zuordnung aller Fälle auf diese Schwerpunkte

Konvergentes K-Means Verfahren:
Nach Zuordnung auf Schwerpunkte, Neuberechnung der Schwerpunkte und Wiederholung, bis kein Klassenwechsel mehr erfolgt

Forgy's Methode:
wie Konvergentes K-Means, aber in der ersten Phase keine Neuberechnung der Schwerpunkte

Beispiel: grafische Darstellung des Ergebnisses einer Cluster-Analyse


PIC GmbH STASY V10.06 (C) PIC GmbH
26.04.91
Datei DEMO-CLU vom 30.12.88
MERKM.2
ZF= 150
N = 150
Cluster-1
Cluster-2
Cluster-3
MERKM.1


PIC GmbH STASY V10.06 (C) PIC GmbH
26.04.91
Datei DEMO-CLU vom 30.12.88
MERKM.4
ZF= 150
N = 150
Cluster-1
Cluster-2
Cluster-3
MERKM.3

6.6 Faktorenanalyse

Die Faktorenanalyse dient der Bestimmung grundlegender, unabhängiger Dimensionen innerhalb des betrachteten Messgrößenraumes. Durch Linearkombinationen wird eine große Zahl von ursprünglichen Messgrößen zu so genannten Faktoren zusammengefasst.

Diese Faktoren können als neue Messgrößen aufgefasst werden. Ziel einer Faktorenanalyse ist es somit, eine Vielzahl von Messgrößen auf einen Minimalsatz von neuen Messgrößen zu reduzieren. Aufgabe des Benutzers ist es, diese neuen Messgrößen sinnvoll zu interpretieren, d.h. ihnen "Namen" zu geben.

Es existiert eine Vielzahl von Lösungsverfahren und Varianten zur Bestimmung und Weiterverarbeitung der Faktoren. Es handelt sich aber um eine mathematisch definierte Aufgabe, so dass alle Verfahren letztlich zum gleichen Ergebnis führen müssen. Die mathematische Methode der Hauptkomponentenextraktion ist für PC besonders geeignet, da sie mit vergleichsweise geringer Rechenzeit auskommt. Bei der Hauptkomponentenanalyse (Principal Component Analysis, PCA) wird versucht, alle Varianz soweit als möglich auf den ersten Faktor zu laden. Durch orthogonale Rotation der Faktoren lassen sich verbleibende Freiheitsgrade nutzen.

Übliche Methoden sind:

- Varimax 1 Faktor wird möglichst nur durch eine Variable erklärt
- Quartimax Alle Messgrößen kommen möglichst nur in einem Faktor mit einem hohen Korrelationskoeffizienten vor
- Equimax Kombination von Varimax und Quartimax

Ob Rotationen sinnvoll sind, muss der Anwender entscheiden. Die Factor Scores sind die Werte der neuen Messgrößen, die sich aus den Faktor Score Koeffizienten berechnen lassen. Diese können als neue Variablen in einen Datensatz zurückgeschrieben werden.

Die Faktorladungen können auf verschiedene Weise grafisch dargestellt werden (Scatterplot oder 3d-Visualisierung). Ein Manko der Faktoranalyse ist die fehlende Prüfung auf statistische Signifikanz.

7 Zeitreihenverfahren

Unter Zeitreihen werden Daten verstanden, die entlang einer Achse (Zeit) aufgenommen wurden. Bei der Achse muss es sich nicht zwingend um die Zeit handeln. So können z.B. Daten, die bei der Niederbringung einer Bohrung erfasst wurden, ebenfalls mit Methoden der Zeitreihenanalytik analysiert werden. (Die Ablagerung von Sedimenten ist letztlich natürlich auch wieder ein zeitabhängiges Ereignis.)

7.1 Äquidistante Reihen

Zeitreihenverfahren basieren in der Regel darauf, dass die Daten in äquidistanten Abständen vorliegen. Da dies messtechnisch oder aus anderen Gründen oft nicht der Fall ist, ist es vielfach unumgänglich, dass nicht äquidistante Messdaten in eine äquidistante Zeitreihe umgewandelt werden. Dies kann z.B. durch lineare Interpolation oder Interpolation mittels Splines erfolgen. Bei größeren Lücken in der Reihe der Rohdaten (gaps) kann es allerdings zu erheblichen Problemen kommen. Per Definition verläuft die interpolierte Linie immer exakt durch die Datenpunkte (Stützstellen). Man spricht von einem Modell, auch wenn der Linienzug dies nicht tut.

7.2 Filter

Zeitreihen enthalten wie andere Daten möglicherweise Fehler (Drift) und systembedingte Schwankungen (Rauschen). Andererseits können periodische Variationen auftreten, die separat untersucht werden sollen. Zur Eliminierung bzw. Erkennung solcher Effekte kann eine Filterung der Daten im Amplituden- oder im Frequenzbereich vorgenommen werden. Ein sehr einfaches Verfahren zur Eliminierung ist die Faltung mit einem Rechteckfenster (gleitende Mittelwerte). Als Beispiel dient die 50 Tage - Linie von Aktienkursen. Voraussetzung ist, dass die Daten auf äquidistanten Stützstellen vorliegen. Durch die Berechnung des Fourierspektrums einer Zeitreihe und Wahl einer geeigneten Filterfunktion kann aus der Rücktransformation eine im Frequenzbereich gefilterte Zeitreihe erstellt werden. Als Filter kommen Tiefpass, Hochpass und Bandpass in Frage, wobei als Filtertypen zwischen Rechteck, Trapez, Butterworth und Exponentialfilter zu unterscheiden ist. Als schneller Algorithmus für die Fourieranalyse kann z.B. die Fast Fourier Tranformation (FFT) eingesetzt werden, um kurz- oder langwellige Variationen zu separieren.

7.3 Zeitreihenzerlegung

Ein mögliches Vorgehen ist die Zerlegung einer Zeitreihe in die einzelnen Bestandteile. Dazu wird z.B. erste der lineare Trend aus einer Zeit mittels linearer Regression bestimmt und die Zu- oder Abnahme aus den Originaldaten rausgerechnet. Anschließend wird mittels der Autoregressionsfunktion der autokorrelierte Anteil (wenn einer vorhanden ist) aus der Zeitreihe rausgerechnet. Mittels Hoch- und Tiefpass-Filterung z.B. mittels FFT werden dann periodische Anteile ermittelt und entfernt. Letztlich verbleibt eine Zeitreihe, die nur noch den Zufallsanteil und das weiße Rauschen der Messwerterfassung enthält.

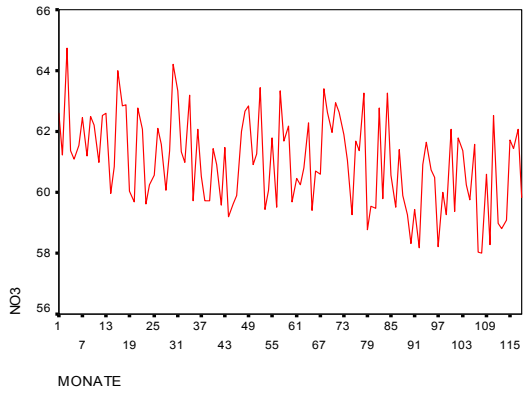
7.4 Autokorrelation

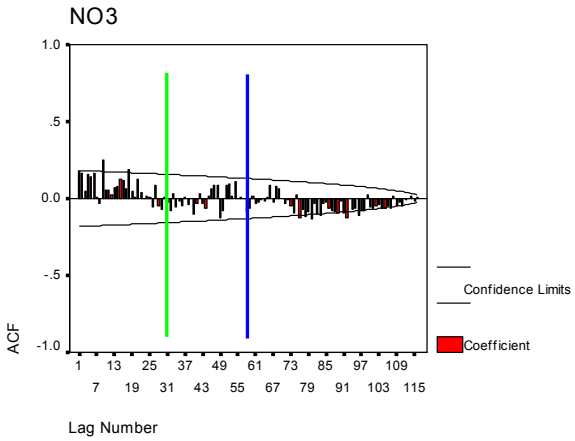
Bei Zeitreihen ist oft das Phänomen zu beobachten, dass die Werte der Zeitreihe zeitverzögert mit sich selbst korreliert sind, also ein Zusammenhang zwischen den zu unterschiedlichen Zeitpunkten beobachteten Werten besteht und aus einer heutigen Beobachtung Schlüsse auf spätere Beobachtungen gezogen werden können. Die an einem bestimmten Tag gemessene Wasser-Temperatur eines Flusses lässt somit Rückschlüsse auf die Temperatur des Folgetages zu. Besteht ein Zusammenhang zwischen den unmittelbar aufeinanderfolgenden Beobachtungen einer Zeitreihe, spricht man von Autokorrelation **erster** Ordnung. Ebenso können sich aber auch Autokorrelationen mit größerer Zeitverzögerung ergeben. Bei Quartalsdaten spricht man von Autokorrelation **vierter** Ordnung, wenn eine Auto-Korrelation zwischen den jeweils vier Perioden voneinander entfernten Beobachtungen vorliegt.

Mit Hilfe der Autokorrelation wird geprüft, inwieweit Werte $Y(t_i)$ einer Zeitreihe von den vorangegangenen Werten $Y(t_{i-j})$ dieser Zeitreihe abhängig sind. Die Berechnung der Autokorrelationsfunktion erfordert äquidistante Zeitreihen. Mit der Autokorrelationsanalyse wird eine Messwertreihe auf wiederkehrende Muster untersucht. Dabei wird ein Datensatz mit sich selbst verglichen. Bildlich betrachtet legt man einen Datensatz und eine Kopie desselben nebeneinander und vergleicht für jeden t-Wert die Funktionswerte Y(t). Anschließend wird die Kopie um einen Zeitschritt (lag, time-lag) verschoben und mit dem Original verglichen. Bei jeder Verschiebung nimmt die Zahl der Fälle n um 1 ab. Trifft man nach einem bestimmten Verschiebungsversatz auf übereinstimmende Werte Y(t) der Kopie mit dem Original und wiederholt sich dieser Fakt, weist die Messwertreihe ein wiederkehrendes Grundmuster auf. Die Periodizität einer Sinus-Funktion ist hier ein ideales Beispiel.

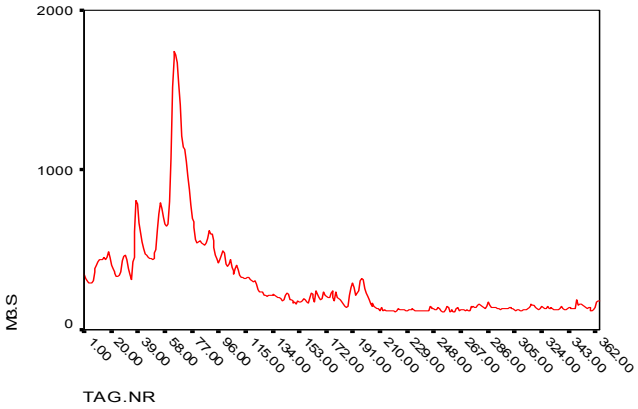
Für Lags, bei denen die ausgewiesene Korrelation über diesen Signifikanzgrenzen liegt, ist mit einer Wahrscheinlichkeit von über 95% somit tatsächlich Autokorrelation vorhanden.

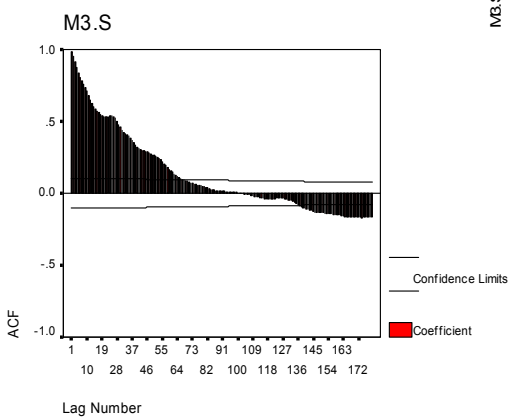
Im Folgenden ist der Verlauf der Nitrat-Konzentrationen links dargestellt (Monatswerte Grundwasser über einen Zeitraum von 10 (120 Monate) Jahren. Die rechte Abbildung zeigt das Autokorrelogramm dazu über den gesamten Zeitraum. Time-Lags über n/2, also in diesem Beispiel ab Lag 60 sollten nicht mehr interpretiert werden (blaue Linie), da die Wahrscheinlichkeit zunimmt, dass es sich nur noch um Zufallskorrelationen handelt; Statistik-Puristen empfehlen sogar Lags > n/4 nicht mehr zu berücksichtigen (grüne Linie). Die Aussage ist somit, dass die vorliegende Zeitreihe der Nitratwerte im Grundwasser nicht autokorreliert ist (lediglich bei zwei Lags ist der Korrelationskoeffizient signifikant).

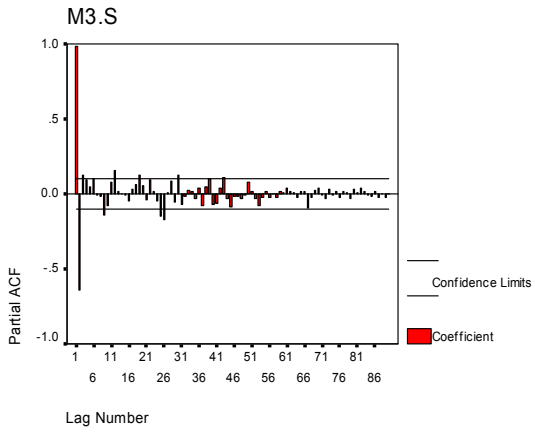




Die folgenden drei Abbildungen zeigen den Abfluss der Elbe im Jahr 1999 sowie das Autokorrelogramm und das partielle Autokorrelogramm.







Die Korrelationen nehmen mit zunehmenden Time-Lags ab und werden nach 100 Lags negativ. Für Lags bis 71 werden signifikante positive Korrelationen angegeben. Dabei stellt sich aber die Frage, ob diese Korrelationen tatsächlich daraus resultieren, dass der an einem Tag gemessene Abfluss sich auf Abflüsse in 71 Tagen noch auswirkt, oder ob sich die berechneten Korrelationen lediglich aus den

Korrelationen erster Ordnung ergeben, die in jeweils abgeschwächter Form auf größere Time-Lags zurückwirken. Die Korrelationen größerer Lags, die jeweils um die Korrelationen geringerer Lags bereinigt wurden, werden als partielle Autokorrelationen bezeichnet.

Das PACF (partial autocorrelation factor)-Diagramm zeigt ein vollkommen anderes Bild als das ACF-Diagramm. Die Autokorrelation erster Ordnung wird mit ca. 0.95 ausgewiesen und liegt damit weit über der 95%-Signifikanzgrenze. Die Autokorrelation zweiter Ordnung ist dagegen mit ca. -0.65 negativ, also invers und signifikant korreliert. Danach wechseln sich positive und negative partielle Korrelationen zufällig und wenig plausibel ab; auch das Signifikanz-Niveau wird selten erreicht. Dies legt den Schluss nahe, dass für den Abfluss eine klare Autokorrelation erster und zweiter Ordnung (1. und 2. Tag) und keine Autokorrelationen höherer Ordnung vorliegen. Das Bild sähe sicher anders aus, wenn eine längere Zeitreihe (z.B. über 10 Jahre) vorliegen würde.
Die Autokorrelation liefert letztlich auch Informationen über die Trägheit des untersuchten Systems. Beispielsweise sind Niederschlagsereignisse wesentlich weniger autokorreliert als die Abflüsse eines Flusses, da letzterer auch Tage nach einem Hochwasserdurchgang noch vergleichsweise hohe Wasserstände hat. Mit Hilfe der partiellen Autokorrelation kann bei trägen Systemen schärfer analysiert werden.

Durch die Berechnung der Autoregressionsfunktion besteht die Möglichkeit, eine Zeitreihe bezüglich des autokorrelierten Anteils zu bereinigen.

7.5 Kreuzkorrelation

Mit Hilfe der Kreuzkorrelation wird geprüft, inwieweit zwei Zeitreihen $Y_1(t_i)$ und $Y_2(t_i)$ voneinander im Hinblick auf eine zeitliche Verschiebung abhängig sind. Die Daten müssen auf äquidistanten Stützstellen vorliegen und mindestens Intervallskalenniveau haben. Es wird für verschiedene Versatzbeträge ein Korrelationskoeffizient berechnet, der anzeigt, von welcher Position ab die größte Übereinstimmung zwischen zwei Messwertreihen zu beobachten ist bzw. mit welcher Verzögerung die eine auf die andere Zeitreihe (z.B. Abfluss auf Niederschlag) reagiert (time-lag).

8 Regionalisierung von Punktdaten

Wie bei den Zeitreihenverfahren besteht ein wesentliches Problem bei der zwei-dimensionalen Interpolation in der ungleichförmigen Verteilung der Stützstellen mit bekanntem Z-Wert in einem von X und Y aufgespannten Feld. Die Stützstellen können durch ein umspannendes Polygon abgegrenzt werden. Außerhalb dieses Polygonzuges kann (falls gewollt) extrapoliert werden. Mit Hilfe der Dreiecksvermaschung kann für jeden beliebigen Punkt innerhalb des Polygonzuges ein Wert Z aus den Flächenfunktionen der gebildeten Dreiecke geschätzt werden. Dies entspricht einer linearen Interpolation. Die anschließende Glättung (smoothen) solcher linear interpolierten Isolinien ist problematisch, da es sehr schnell zu Isolinienverschneidungen kommen kann. Möglich ist aber die Berechnung nichtlinearer Flächenfunktionen für die einzelnen Dreiecke, wie die aus den Verfahren der finiten Elemente bekannt ist. Eine optimale Dreiecksvermaschung (VORONOI-, THIESSEN-Polygone) insbesondere großer Datenbereiche ist eine mathematisch anspruchsvolle Aufgabe. Die resultierende Datenstruktur heißt TIN (Triangulated Irregular Network) und wird vor allem in Geo-Informationssystemen (GIS) verwendet.

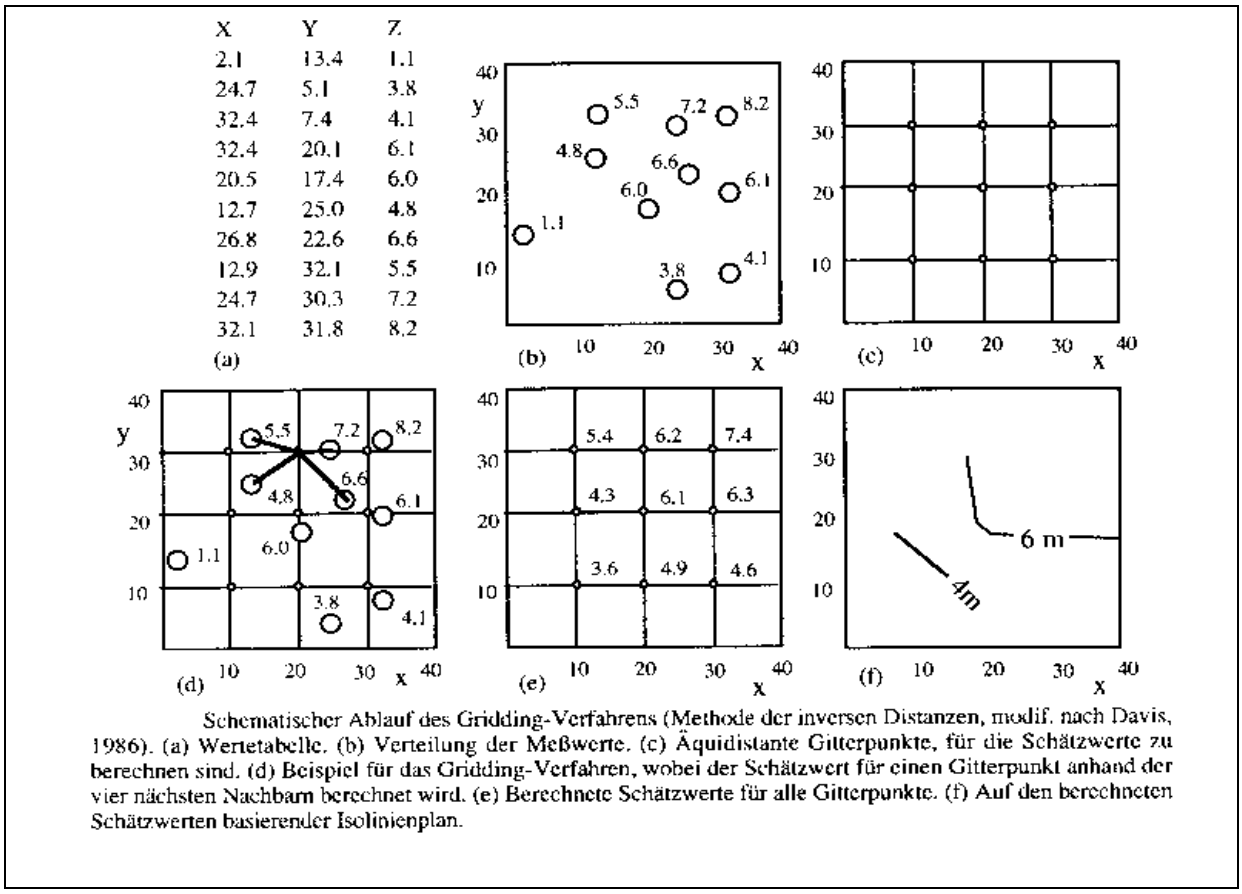
Die meisten Isolinien-Programme wählen aber den Weg über die Abbildung der Rohdaten auf ein rechtwinkliges Gitter. Liegt innerhalb der durch einen Gitterpunkt bestimmten Rasterfläche mehr als ein Datenpunkt, kann das Kriterium der exakten Interpolation bereits nicht mehr eingehalten werden. Insgesamt führt die Abbildung der Daten auf Rasterflächen zu einer gewissen Unschärfe. In einem zweiten Schritt werden dann auf Basis dieses äquidistanten Gitters die Isolinien entwickelt. Für den ersten Schritt (Abbildung auf das Gitter) stehen verschiedene Verfahren zur Verfügung, von denen hier zwei gegenüber gestellt werden:

- Inverse Distanzen
- Inverse Distanzen (gewichtet)
- nächste Nachbarn
- stückweiser polynominaler Fit
- Minimum Curvature
- Dreiecksvermaschung
- Kriging
- Polynom 2. bis n.ter Ordung [Trendfläche]
- Fourier-Transformation (z.B. FFT) [Trendfläche]

8.1 Methode der inversen Distanzen

Die Methode der inversen Distanzen ist ein relativ einfaches Verfahren für die Umrechnung von unregelmäßig verteilten Messstellen auf ein äquidistantes Gitternetz. Grundlage für die Methode ist die Annahme, dass näher am zu berechnenden Gitterpunkt liegende Messpunkte eine höhere relative Übereinstimmung als weiter entfernte aufweisen.

Mit mehreren um einen Gitterpunkt nächstgelegenen Messpunkten wird ein gewichteter Mittelwert berechnet. Die Wichtungskoeffizienten sind umgekehrt proportional zu den Abständen zwischen den jeweiligen Messstellen und dem Gitterpunkt ($\rightarrow$ Inverse Distanzen). Die Anzahl der Messpunkte für das zu berechnende gewichtete Mittel kann manuell bestimmt werden.



Schematischer Ablauf des Gridding-Verfahrens (Methode der inversen Distanzen, modif. nach Davis, 1986). (a) Wertetabelle. (b) Verteilung der Meßwerte. (c) Äquidistante Gitterpunkte, für die Schätzwerte zu berechnen sind. (d) Beispiel für das Gridding-Verfahren, wobei der Schätzwert für einen Gitterpunkt anhand der vier nächsten Nachbarn berechnet wird. (e) Berechnete Schätzwerte für alle Gitterpunkte. (f) Auf den berechneten Schätzwerten basierender Isolinienplan.

Abb. 15. aus Schlüter M (1996)

Die Methode der inversen Distanzen liefert zwar "nette" Isolinien, allerdings darf deren Aussagekraft nicht überschätzt werden.

8.2 Variogrammanalyse

Das Verfahren der Variogrammanalyse berücksichtigt die Tatsache, dass Messwerte neben der deterministischen Komponente immer eine stochastische, zufällig verteilte Komponente beinhalten. D.h. zwischen den Messwerten besteht eine regionale Verknüpfung. Im Vergleich zum Gridding-Verfahren ist mit Hilfe der Variogrammanalyse die Definition eines statistisch abgesicherten Einzugsbereiches möglich.

Der statistische Zusammenhang kann mit Hilfe der Variogrammanalyse quantifiziert werden. Für die Erstellung des Variogramms werden zunächst die Quadrate der Differenz zwischen im Abstand x befindlichen Messwerten berechnet. Mit der folgenden Gleichung werden die Werte γ für das Variogramm berechnet:

$$\gamma = \frac{1}{2n} * \sum_{x=1}^{n} \left(z_i - z_{i+h} \right)^2$$

γ: Variogramm-Wert
z_i: Messwert an der Stelle i
z_{i+h}: Messwert an der Stelle i+h
n: Anzahl der Messstellen
x: Abstandsintervall

Zur Erstellung der Variogramm-Kurve werden die mittleren Varianzen als γ-Werte gegen die Entfernung x (wenn gewünscht auch richtungsabhängig) aufgetragen.

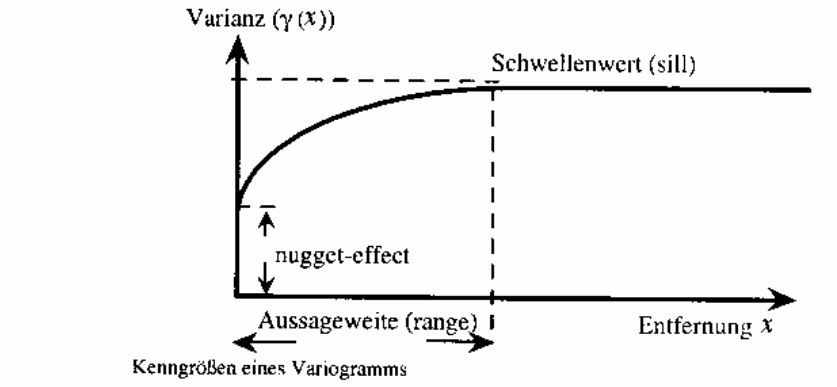


Abb. 16. Variogrammkurve

Der Anstieg und die Entfernung, ab der die γ-Werte den Schwellenwert (sill) erreichen, spiegeln die räumliche Korrelation der Messstellen, d.h. die Entfernung, ab der sich die Varianz nicht mehr ändert, wider (Aussageweite (range) oder Einflußweite (= Length)). Der Schwellenwert beschreibt die Varianz der außerhalb der Aussageweite liegenden Messwerte.

Der so genannte nugget-effect ist ein Maß für die Beschreibung für die Varianz an der Stelle x = 0. Er kann zum einen als unvermeidlicher Messfehler, zum anderen als kleinräumige Variabilität interpretiert werden. Ursache können z.B. Geländemerkmale wie Flusstäler sein. (Der Begriff nugget-effect stammt aus der Lagerstättenprospektion im Goldbergbau.)

Lässt sich keine Funktion an die Daten des Variogramms sinnvoll anpassen, so ist dies ein klarer Hinweis, dass diese Daten nicht regionalisierbar sind! Von der Erstellung einer Isolinienkarte, gleich mit welchem Verfahren, sollte dann Abstand genommen werden. Unterschiedliche Kenngrößen eines Variogramms deuten auf richtungsabhängige, anisotrope Verteilungen von regionalisierten Messwerten hin.

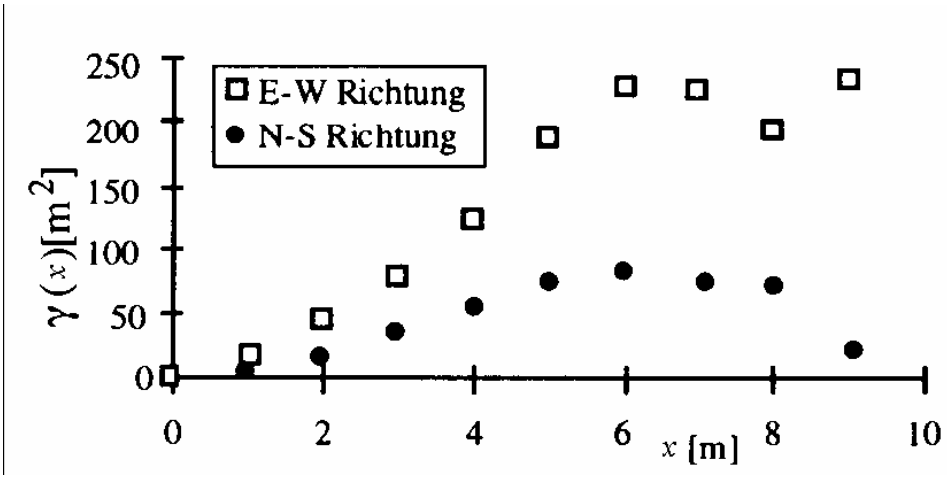


Abb. 17. Richtungsabhängige Verteilung von regionalisierten Messwerten

8.3 Kriging

Die Besonderheit des Krigings besteht darin, dass die Ergebnisse des Variogramms, d.h. die ermittelte Variogrammkurve, die die Varianz der Abweichungen in Abhängigkeit von der Entfernung des Messpunkte untereinander darstellt, in Form einer Funktion (z.B. linear, exponential, sphärisch, Gauss) errechnet werden.

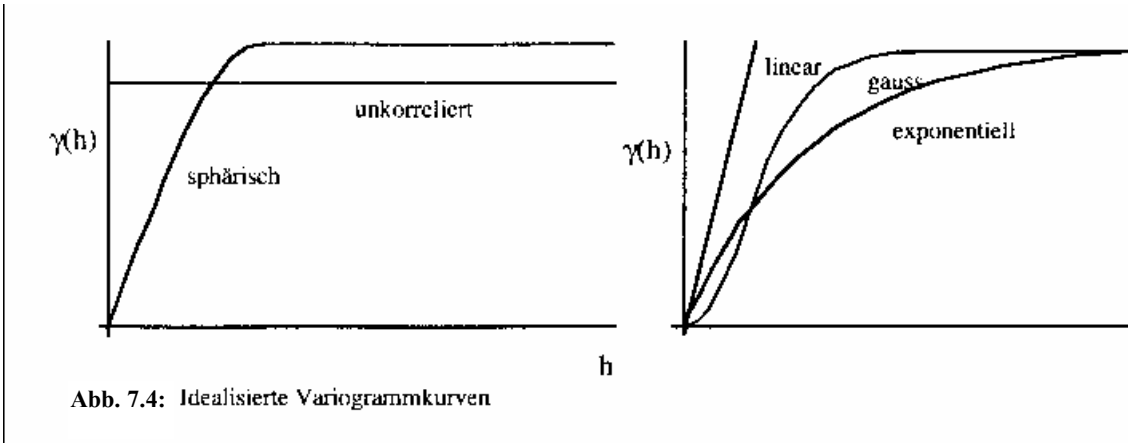


Abb. 7.4: Idealisierte Variogrammkurven

Durch die Verknüpfung der Variogrammanalyse mit dem Kriging-Verfahren lassen sich Wichtungsfaktoren und die jeweiligen Einzugsbereiche über die Messstellen berechnen bzw. bestimmen. Die so ermittelten Schätzwerte sind entsprechend statistisch abgesichert und mit Vertrauensgrenzen belegt. Die aus Variogrammkurven ermittelten Modellfunktionen ermöglichen die Erstellung von Isolinienplänen und darüber hinaus die Festlegung von statistischen Vertrauensgrenzen für berechnete Isolinienpläne.

Im Vergleich mit dem Gridding Verfahren sind beispielsweise für die Darstellung von Geländemorphologien bessere Übereinstimmungen zu erzielen, da u.a. nicht so viele isoliert liegende Senken und Mulden auftreten.

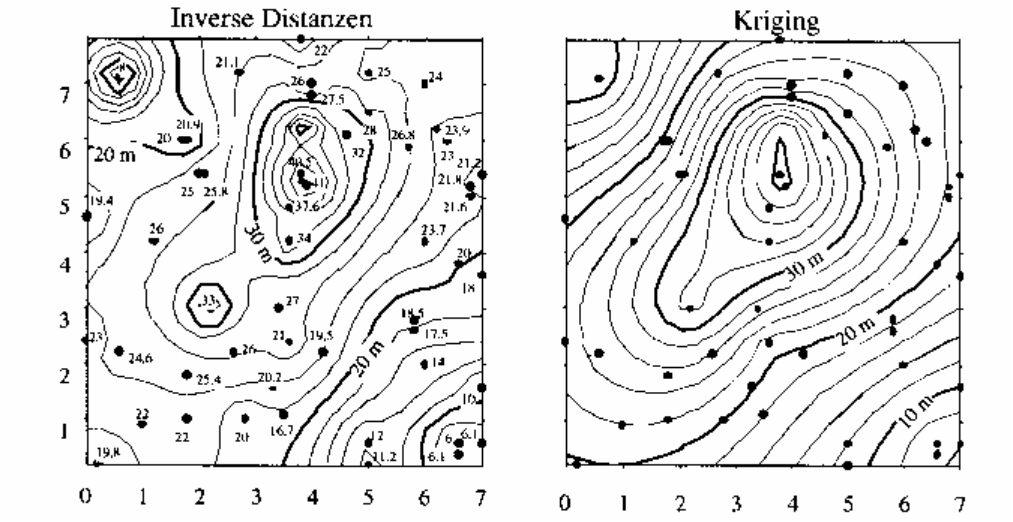


Abb. 18. Gegenüberstellung beider Methoden

Bezüglich diverser Varianten des Kriging-Verfahrens wird auf die Literatur verwiesen.

9 Literatur

AKIN A & SIEMES H (1988): Praktische Geostatistik - Eine Einführung für den Bergbau und die Geowissenschaften. - Springer Verlag, Berlin, Heidelberg, London, New York.

BOSCH K (1987): Formelsammlung Statistik. - R. Oldenbourg Verlag GmbH, München.

DAVID M (1977): Geostatistical Ore Reserve Estimation. - Elsevier Scientific Public. Company, Amsterdam.

DUTTER R (1985): Geostatistik. - B.G. Teubner, Stuttgart.

DVWK Regeln 128 (1992): Entnahme und Untersuchungsumfang von Grundwasserproben. Verlag Paul Parey

FUNK et. Al (1985): Statistische Methoden in der Wasseranalytik. - VCH-Verlag.

HOSCHEK J & LASSER D (1989): Grundlagen der geometrischen Datenverarbeitung. - B.G. Teubner, Stuttgart.

KECKLER D. (1995): Surfer for Windows. - Golden Software Inc., Golden (Colorado, USA).

LAXEN D.P.H. (1977): A specific conductance method for quality control in water analysis. Water research 11, 91-94

LUTZ T (1976): Datenbanken. - Scientific Research Association, Stuttgart, Chikago, Sydney, Paris.

MAIER D. & GROHMANN A. (1977): Bestimmung der IOnenstärke natürlicher Wässer aus deren leitfähigkeit. Z. Wasser- und Abwasserforsch., 10, 9-12

MERKEL B & SPERLING B (1990): Statistik für Microcomputer. - G. Fischer Verlag, Stuttgart.

PANNATIER Y (1996): VARIOWIN: Software for Spatial Data Analysis in 2D. - Springer Verlag, Berlin, New York, London.

ROSSUM J.R. (1975): Checking the accuracy of water analysis through the use of conductivity. J.Am. water Works Assoc., 67, 104-205, Washington

SACHS L (1974): Angewandte Statistik. - Springer Verlag, Berlin, New York, London.

SACHS L (1988): Statistische Methoden: Planung und Auswertung. - Springer Verlag, Berlin, Heidelberg, New York, London.

SCHILCHER M & FRITSCH D (1989): Geo-Informationssysteme. - Wichmann Verlag, Stuttgart.

SCHLÜTER M (1996): Einführung in geomathematische Verfahren und deren Programmierung. - F. Enke Verlag, Stuttgart.

SCHUBÖ W & VEHLINSER HH (1986): SPSS - Statistik, Programmsystem für die Sozialwissenschaften. - G. Fischer Verlag, Stuttgart.

SIZE WB (1987): Use and Abuse of Statistical Methods in the Earth Sciences. - Oxford University Press, New York, Oxford.

STORM R (1988): Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle. - VEB Fachbuchverlag, Leipzig.

WIEDERHOLD G (1980): Datenbanken, Analyse - Design - Erfahrungen. Oldenbourg Verlag, Müchen, Wien.

WIENER BJ (1970): Statistical principlesin experimental design. - Graw-Hill, London, New York.