GREIFSWALDER GEOGRAPHISCHE ARBEITEN

Band 33

# Leitfaden für die statistische Auswertung geographischer Daten

von

**Tim Hoffmann & Raimund Rödel**

**GEOGRAPHISCHES INSTITUT**
**ERNST-MORITZ-ARNDT-UNIVERSITÄT**
**GREIFSWALD**

**GREIFSWALD 2004**

**GREIFSWALDER GEOGRAPHISCHE ARBEITEN**

Geographisches Institut der Ernst-Moritz-Arndt-Universität Greifswald

Band 33

# Leitfaden für die statistische Auswertung geographischer Daten

herausgegeben von

**Tim Hoffmann & Raimund Rödel**

**GREIFSWALD 2004**

**ERNST-MORITZ-ARNDT-UNIVERSITÄT GREIFSWALD**

Die Übungsdateien und die digitale Version dieses Bandes befinden sich auf der beigefügten CD-ROM. Sie können ebenfalls unter:

http://www.uni-greifswald.de/~geograph/publikationen/gga

http://www.geomodellierung.de

abgerufen werden.

Impressum

ISBN: 3-86006-234-4
Ernst-Moritz-Arndt-Universität Greifswald

Herausgeber: Tim Hoffmann & Raimund Rödel

Redaktion: Tim Hoffmann (Kap. I, II, V & VI),
Raimund Rödel (Einführung, Kap. III, IV & VI)

Layout: Tim Hoffmann & Raimund Rödel

Grafik: Tim Hoffmann & Raimund Rödel

Herstellung: KIEBU-Druck Greifswald

Kontakt: Ernst-Moritz-Arndt-Universität, Geographisches Institut
Friedrich-Ludwig-Jahn-Str. 16,
D-17487 Greifswald
e-mail: geogra@uni-greifswald.de

---

# VORWORT VON PROF. DR. KLAUS D. AURADA

Nach JOSEPH MARIA BOCHEN´SKI (1902-1995) wird unter „Logik“ (griech. logikós) die Lehre von der Folgerichtigkeit (1949) verstanden (formale Logik), die als traditionelle Logik durch eine neue oder moderne Logik, die Logistik (griech. logistikós), insbesondere mit dem Namen GOTTLOB FREGEs (1848-1925) verbunden, weiterentwickelt worden ist. Die Bezeichnung „Logistik“ wurde erstmals 1904 von LOUIS COUTURAT (1868-1914) vorgeschlagen. Die Logistik unterscheidet sich von der Logik vor allem durch Formalisierung und Kalkülbildung; AURADA (1982) hat vor diesem Hintergrund ein „systemtheoretisches Kalkül“ der Geographie zu spezifizieren versucht: „... unter Anwendung des systemtheoretischen Kalküls in der Geographie wird die Übertragung mathematischer Beschreibungsformen von Prozeßabläufen und Systemzuständen auf als Systeme zu behandelnde geographische Objekte (Geosysteme) verstanden ...“ (a.a.O., 241).

Grundlage vieler mathematischer Beschreibungsformen sind statistische Verfahren, die sinnvoll sowohl unter Berücksichtigung vorhandener (gegebenenfalls noch zu erhebender) Datensätze, der angestrebten Zielstellung als auch insbesondere notwendiger Rand- und Gültigkeitsbedingungen auszuwählen und anzuwenden sind. Ihr zweckmäßiger Einsatz und insbesondere auch ihre empfehlenswerte Aufeinanderfolge folgt einer dementsprechenden Logistik als Abfolge von Systemanalyse, Systemidentifikation und Systemsynthese.

PHILIPP BUACHE (1700-1773) hatte bereits 1752 mit dem "charpenture du globe" („Erdgezimmer“) ein Konzept der Erdoberflächengliederung nach von Wasserscheiden begrenzten Gewässereinzugsgebieten entwickelt, das eine durchgängig anwendbare Gliederung der Erdoberfläche ermöglichte. "The sedimentary basin is to geology what the drainage basin is to geomorphology,..." (LEEDER 1997, 229); diese Feststellung kennzeichnet zugleich die aktuelle Entwicklungsrichtung geowissenschaftlicher Forschung.

Die bisher vorliegenden Lehr- und Lernmaterialien gehen von der Priorität statistischer Verfahren aus, die mit Hilfe jeweils als geeignet erscheinender Datensätze erläutert und geübt werden sollen. Dieser Leitfaden versucht nun, wissenschaftliche Fragestellungen in den Vordergrund zu rücken, die unter Verwendung sowohl eines einheitlichen Datensatzes des Ostseeraums als auch problemadäquater Verfahren diskutiert und einer schrittweisen Lösung zugeführt werden können; er resultiert aus entsprechenden Lehrveranstaltungen der Autoren, die die Studenten wohlwollend und motiviert sehr gut aufgenommen haben.

Wenn Sie als Studenten Ihre ersten statistisch gestützten Gehversuche im Rahmen dieses Leitfadens unternehmen, seien Sie sich eingedenk, daß studere „sich bemühen“ heißt. Bemühen Sie sich, in diese Gedankenwelt einzudringen, weil sie mit der Notwendigkeit prägnanten Denkens einen ersten Schritt wissenschaftlichen Arbeitens repräsentiert, der zu einer „Logik und Logistik der naturwissenschaftlichen Geographie“ (AURADA 1993) führt, die sich auch beide Autoren – zu meiner Freude – erfolgreich zu eigen gemacht haben.

Greifwald, Dezember 2004 Prof. Dr. Klaus D. Aurada

# INHALTSVERZEICHNIS

**Zeichenkonventionen**:

> Menü

🖫 Übungsdatei auf CD

■ Übungsbeispiel

`KORREL`- Tabellenfunktion oder Zellenformel in EXCEL

Die Beispiele wurden mit MS EXCEL 2002 und SPSS 12.0 gerechnet

SPSS is a registered Trademark

EXCEL is a registered Trademark of Microsoft

# BESCHREIBEN ODER ENTDECKEN ?– LOGISTIK GEOGRAPHISCHER DATENANALYSE

Um die geographische Wirklichkeit zu erforschen, wird man oft eine größere Anzahl von Daten erheben, messen und auswerten. Dass die Statistik hierbei Hilfsmittel zur Verfügung stellt, mit welchen die verschiedenen Aspekte realer Beobachtungen erfasst werden (SCHÖNWIESE, 1983) können, ist wohl unbestritten. Elementare statistische Arbeitsweisen sind heute ein fester Bestandteil der Ausbildung von Geographen. Doch wie lässt sich dieses „unentbehrliche" Hilfsmittel (BAHRENBERG, GIESE & NIPPER, 1985) so nutzen, dass es möglich wird, ein tieferes Verständnis für die in einem Raum ablaufenden Prozesse zu entwickeln?

An fast allen Geographischen Instituten Deutschlands werden Einführungskurse zu statistischen Arbeitsweisen angeboten. Die Inhalte der meisten dieser Lehrveranstaltungen vermitteln Grundbegriffe, welche zu einer empirischen Beschreibung eines Datensatzes notwendig sind. Man lernt Daten in Tabellen aufzubereiten, Mittelwerte und Streuung zu berechnen. Mit der Schätz- und Teststatistik gewinnen die Studierenden auch einen Einblick in Methoden, „Entscheidungen im Falle von Ungewissheiten zu treffen" (WALD, 1950 zit. in SCHÖNWIESE, 1983), man untersucht ebenso einfache Zusammenhänge, also Korrelationen. Zudem werden an einigen Universitäten im zweiten Teil des Studiums der Geographie ausgewählte Methoden der multivariaten Datenanalyse vermittelt.

Und dennoch bleibt nach solchen Erfahrungen mit der Statistik oft die Frage bestehen, zu welchen geographisch interessanten Antworten man denn nun mit ihrer Hilfe kommen kann. Statistische Methoden, welche ja immer von realen Beobachtungen ausgehen, können oft erst in einem zweiten oder dritten Schritt Aussagen über mögliche Zusammenhänge oder Ursachen geben (SCHÖNWIESE, 1983). Vor diesen muss also meist erst eine interessante Fragestellung formuliert werden, denn keine noch so gute statistische Methode kann auf eine uninteressante, „falsche" Frage eine interessante und spannende Antwort geben (BAHRENBERG, GIESE & NIPPER, 1985).

Also doch nur ein „Beschreiben" mit der Statistik, kein „Entdecken" geographisch interessanter Strukturen? Mit solch einer Reduktion des Potenzials statistischer Methoden für die Geographie mag man sich nicht zufrieden geben. Und tatsächlich ergibt ein gründlicheres Studium diverser Lehrbücher der Statistik, dass die multivariaten

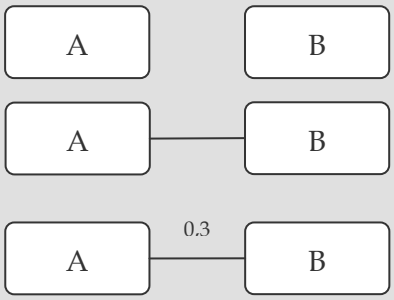
Analysemethoden zum einen struktur-entdeckende Verfahren wie auch struktur-überprüfende Verfahren (BACKHAUS ET. AL, 2000) bereithalten.

Wenn die zu analysierenden Daten reale Sachverhalte beschreiben, sollten sie auch zugleich immer die Wirkung von Vorgängen in der Wirklichkeit widerspiegeln. Folglich lässt sich das Potenzial statistischer Methoden für die geowissenschaftliche Arbeit am sinnvollsten nutzen, wenn mit ihnen nach Ursachen, nach kausalen Zusammenhängen gesucht werden kann (SCHÖNWIESE, 1983 & 2000). Die Betrachtung von Vorgängen oder Sachverhalten als Resultat von Wirkungsmechanismen, welche Eingangsgrößen in Wirkungen transformieren, leitet jedoch direkt in ein systemtheoretisches Denken über. Systemtheoretisches Denken findet sich in der Geographie als Geosystemlehre (KLUG & LANG, 1983) wieder. Um es für die Geographie anwendbar zu gestalten, kann auf der Logik des Systemkonzepts eine Logistik des Systemkonzepts aufgebaut werden (AURADA, 1982 & 1993).

**Logistik des Systemkonzeptes** (vereinfacht und modifiziert nach AURADA, 1982 & 2003)

Beschreibt man ein System bildlich als eine Anzahl von in Boxen verknüpften Zustands- und Prozessgrößen, dann lässt sich die Logistik des Systemkonzepts wie folgt beschreiben:

Zunächst füllt die Systemanalyse die Boxen. Diese Boxen repräsentieren die ausgewählten Variablen. Im zweiten Schritt der Systemidentifikation werden die Zusammenhänge zwischen den Boxen ermittelt. Die Systemsynthese kann schließlich vereinfacht so aufgefasst werden, dass diese Zusammenhänge quantitativ verknüpft werden.




SYSTEMANALYSE

SYSTEMIDENTIFIKATION

SYSTEMSYNTHESE

Kann die Logistik des Geosystemkonzepts also jener Leitfaden sein, welcher dem Anwender statistischer Methoden jene Fragen bei der Auswertung erhobener Daten gestattet, die ein „Entdecken“ ermöglichen? Gibt es also eine logische Abfolge von Schritten der Datenanalyse, die es erlauben, stets hinreichend auf die Kernfrage der Geographie zu fokussieren? Wie sich nämlich aus der Kenntnis räumlicher Strukturen und Muster auf das Verhalten von Systemen oder die in einem Raum ablaufenden Prozesse schließen lässt? Wer also derart fragend einen geographischen Datensatz bearbeitet, wird weiter nicht mehr umhin kommen, „erst denken, dann rechnen“ (BACKHAUS ET. AL, 2000) zu müssen. Indem in einer logischen Reihenfolge Fragen aufgeworfen werden und Möglichkeiten zu deren Beant-

wortung offenliegen, fällt das „Entdecken“ interessanter Wechselwirkungen womöglich leichter.

Solche Fragestellungen zum Verständnis des geographischen Raumes über das Erkennen von gerichteten Zusammenhängen und Wechselwirkungen bis zur quantitativ korrekten Beschreibung dieser Zusammenhänge als Wirkungen innerhalb eines Systems (des Geosystems) können innerhalb der Logistik des Systemkonzepts erfasst werden (AURADA, 1993 & 2003). Die Systemanalyse eröffnet dem forschenden Geographen ein immer tiefergehenderes Verständnis in die Struktur der zu analysierenden Daten und den von ihnen widergespiegelten Eigenschaften des Raumes. Vorhandene kausale Zusammenhänge und möglichen Wirkungsketten lassen sich im Arbeitsschritt der Systemidentifikation aufdecken. Schließlich werden in der anschließenden Systemsynthese Ursachen und Wirkungen mit quantitativen Parametern verknüpft und verifiziert. Jetzt ergibt sich zudem die Möglichkeit, die gefundenen kausalen Zusammenhänge nicht nur zu interpretieren, sondern ihre Empfindlichkeit oder Stabilität zu untersuchen. Szenarien und Simulationen können durchgeführt werden, welche die Bandbreite der möglichen interessanten Fragestellungen noch einmal erhöhen. Theorien können getestet und verworfen werden.

Die folgenden Abschnitte werden versuchen zu beschreiben, welche statistischen Methoden sich besonders sinnvoll innerhalb der Logistik des Systemkonzepts anwenden lassen, mithin den erforderlichen Erkenntnisgewinn zum weiteren Arbeiten ermöglichen. Statistische Methoden anzuwenden, um kausale, systemare Zusammenhänge aufzudecken, heißt aber auch immer, keine vollständig sicheren Aussagen zu treffen. Wenn zwar oft keine streng deterministischen Aussagen möglich sind, so lassen sich aber definitive Wahrscheinlichkeiten ausweisen, mit denen die gefundenen Erkenntnisse zutreffen. Genau dieser Vorteil ist es, den die Statistik als Methode zur Analyse realer Daten einzubringen vermag.

## Datenauswertung im geographischen Raum als logische Abfolge von Arbeitsschritten

Wenn in den nächsten Kapiteln verschiedene statistische Arbeitsweisen vorgestellt werden, soll das gerade nicht geschehen, um den bereits zahllosen Lehrbüchern zur Statistik ein nächstes hinzuzufügen. Am Beispiel eines konsequent zu verwendenden Datensatzes soll vielmehr gezeigt werden, wie Methoden der Statistik und Datenauswertung als „algorithmisierte Anwendung systemtheoretischer“ (AURADA, 1982) Arbeitsweisen angewandt werden können. Die schrittweise Untersuchung eines geographischen Datensatzes soll so erfolgen können, dass sich jeweils aufeinander folgende logische Fragen aufwerfen lassen. Indem man hierzu Antworten erhält, folgt man aber wieder weiteren Fragen. Damit besteht die Möglichkeit, ein immer tiefergehenderes Verständnis für verschiedene

Ursachen und Wirkungen, welche sich in einem Datensatz widerspiegeln können, zu gewinnen.

Unser Interesse soll einem Datensatz gelten, welcher die Flusseinzugsgebiete der Ostsee beschreibt. Diese Daten sind frei erhältlich, zusammen mit weiteren Informationen können sie unter *www.grida.no* eingesehen werden. Für die Arbeit in den nächsten Kapiteln sind diese Daten als EXCEL Datei auf der beigelegten CD-ROM hinzugefügt worden.

GGA_Einführung\BasinData_fromGRIDA.xls

Zu den Einzugsgebieten als Raumeinheiten existieren verschiedene Angaben. Sie umfassen Flächenanteile verschiedener Landnutzungen, Stofffrachten, Bevölkerungsdichte sowie etliche andere. Zusammen stellen sie die Ausprägungen von verschiedenen Variablen in den Flusseinzugsgebieten der Ostsee dar. In Bild 1 werden sie mit den Begriffen struktur- und prozessabbildender Variablen beschrieben. Mit dem Beschreiben der Daten beginnt zugleich der Schritt der Systemanalyse:

# 1 SYSTEMANALYSE

Ein erster Schritt bei der Systemanalyse ist es, Datensätze zu finden, welche die Eigenschaften von Räumen repräsentieren und andererseits Datensätze zu ermitteln, welche eher für die in einem Raum ablaufenden Prozesse charakteristisch sind. Zunächst wird man sich einen ungefähren Überblick über die bekannten Daten verschaffen, welche einen oder mehrere Räume charakterisieren. Methoden der grafischen Datenanalyse können bei dieser Aufgabe hilfreich sein, Informationen über die Größe und den Betrag der zu betrachtenden Daten zu gewinnen, über deren Streubreite und zeitliche Variabilität.

Gibt es Räume, die sich in ihren Eigenschaften einander ähnlich sind? Wie kann man dies ermitteln und diese Ähnlichkeit quantitativ beschreiben?

Damit lässt sich in der Terminologie der Geosystemforschung eine konstitutionelle Beschreibung des geographischen Raumes vornehmen. Solche Eigenschaften des Geosystems haben sich über sehr lange Zeit entwickelt, das somit evolvierende Geosystem stellt ein Abbild der bisherigen Entwicklung des geographischen Raumes dar.

Mit Hilfe von einfachen Kartendarstellungen lässt sich erkennen, wie stark räumliche Unterschiede ausgeprägt sind. Bereits mit einfachen Mitteln der grafischen Datenanalyse wie Sonnenstrahlicons oder Chernoff-Gesichtern lassen sich damit Aussagen über die Ähnlichkeit bestimmter Räume und Raumeigenschaften oder auch ähnlicher Prozesse ermitteln.

Sehr instruktive Ergebnisse können hier auch mit ausgefeilteren Methoden zur Klassifikation von Datensätzen gewonnen werden. Die Hierarchische Clusteranalyse wird sich am Beispiel des zu verwendenden Datensatzes von Einzugsgebieten der Ostsee als vorteilhaft erweisen.

Mit ihr werden sich Einzugsgebiete unterschiedlicher Landnutzung, aber auch unterschiedlicher Stofffrachten herausarbeiten lassen. Laufen in diesen Räumen ähnliche Prozesse ab oder „funktionieren“ diese Räume gleichartig? Was sind überhaupt die Eigenschaften eines Raumes, als was kann man seine Funktionsweise auffassen? Hier wird auch zum ersten Male die Frage nach dem „Funktionieren“ unterschiedlicher geographischer Räume gestellt und beantwortet.

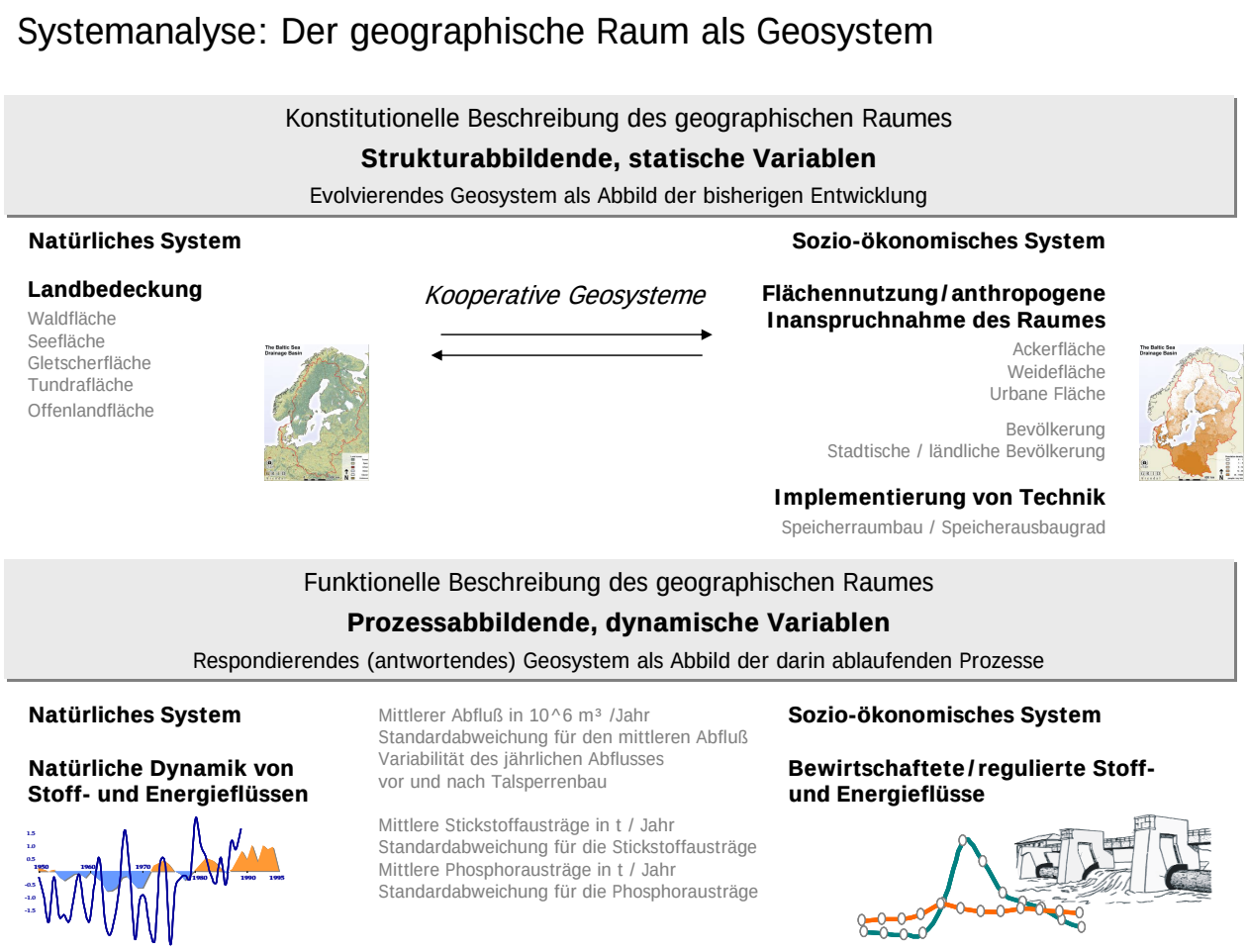



*Bild 1: Einteilung des verwendeten Datensatzes in Variablen für die konstitutionelle und funktionelle Raumbeschreibung*

Unterschiedliche N- und P-Frachten treten in unterschiedlich konstitutionell zu beschreibenden Räumen auf. Der Begriff konstitutionell bekommt daher im Sinne der „Konstitution“ der Räume eine fast selbsterklärende Bedeutung. Einzugsgebiete mit einer, in diesem Sinne „guten“ Konstitution bedingen auch niedrige Stoffausträge. Stoffrachten oder allgemeiner Energie- und Stoffbilanzen können andererseits die respondierenden (Response – Antwort) Eigenschaften eines Geosystems beinhalten.

Respondierende Geosysteme stellen Abbilder der im geographischen Raum ablaufenden Prozesse dar, sie beschreiben eine funktionelle Ebene. Im Wortsinne könnte man ebenso davon sprechen, sie widerspiegelten die Funktionsweise des Raumes.

Wichtige Arbeitsmethoden während der Systemanalyse werden in den Kapiteln I bis III beschrieben.

# 2 SYSTEMIDENTIFIKATION

Durch die Systemanalyse konnten Räume mit unterschiedlicher Struktur und unterschiedlich abgebildeten Prozessen gefunden werden. Im Schritt der Systemidentifikation soll nun auf die direkten Wechselwirkungen und Zusammenhänge von Raumeigenschaften und abgebildeten Prozessen, dem „Funktionieren" des Raumes, fokussiert werden.

Welche Eigenschaften des Raumes steuern einzelne Prozesse unmittelbar? Zwischen welchen Prozessen und welchen dieser Raumeigenschaften bestehen überhaupt enge Zusammenhänge? Sind diese Zusammenhänge zum einen statistisch relevant und stellen zum anderen wirkliche Abhängigkeiten dar? Die Korrelationsanalyse kann solche Zusammenhänge aufdecken helfen.

Um einen funktionalen Zusammenhang zwischen einem Prozess und einem Set von Raumeigenschaften zu finden, müssen kennzeichnende, „führende" Variablen mit einem hohen Erklärungswert ermittelt werden. Ebenso sollen Sets verschiedener Variablen als komplexe Größen möglichst gut, aber durch wenige Variable repräsentiert werden. Eine hierbei also im Mittelpunkt stehende Datenreduktion kann durch die Hauptkomponenten- und Faktorenanalyse bewerkstelligt werden.

Da die Faktorenanalyse jedoch die Korrelationsmatrix, also alle möglichen Zusammenhänge zwischen den Variablen, auswertet, weist sie im Schritt der Systemidentifikation oftmals bereits auf „Potenziale" einzelner Räume hin. Auf deren besondere Bedeutung wird später noch einmal einzugehen sein. Bei sinnvoller Interpretation erschließen die „Potenziale", welche gleichzeitig die Faktoren aus der Faktoranalyse sind, die in einzelnen Räumen unterschiedlich ausgeprägten Effekte des gemeinsamen Wirkens einzelner Variablen.

Mit dem Wissen über enge Zusammenhänge zwischen Variablen lassen sich verbindende Linien zwischen einzelnen Variablen (gedanklich als Boxen repräsentiert) zeichnen. Damit stellen die Korrelationsanalyse und die erweiterte Auswertung der Korrelationsmatrix über die Faktoranalyse wichtige Arbeitschritte bei der Systemidentifikation dar.

Die dabei zunächst aufgedeckten Zusammenhänge sind jedoch statistischer Natur und geben allenfalls Hinweise auf mögliche deterministische Erklärungsansätze. Ein bekanntes Beispiel hierfür ist die enge Korrelation zwischen Geburtenrate und Storchendichte. Ein solcher Zusammenhang führt statistisch zu einer Scheinkorrelation, also einem nicht deterministisch erklärbaren Zusammenhang. Um solche Fehlschlüsse möglichst zu umgehen, muss Systemidentifikation auch heißen, gefundene Zusammenhänge auf deterministische Relevanz untersuchen zu können.

Wie können also Scheinkorrelationen aufgedeckt werden? Bild 2 zeigt, dass der korrekte Zusammenhang zwischen den Systemkomponenten beim „Klapperstorchproblem“ weniger als Zusammenhang zwischen Storchendichte und Geburtenrate, sondern besser über den Anteil der industriellen Produktion am Bruttoinlandsprodukt erklärt werden kann.

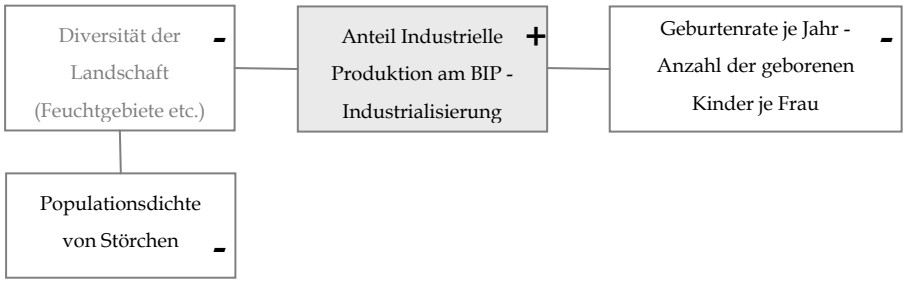



*Bild 2: Wirkungszusammenhänge beim „Klapperstorchproblem“*

Hierzu hält die Statistik mit der Partiellen Korrelationsanalyse und der Pfadanalyse wertvolle Methoden bereit. Wird der Partielle Korrelationskoeffizient für das „Klapperstorchproblem“ berechnet, lässt sich recht schnell die nur extrem geringe Korrelation zwischen Geburtenrate und Storchendichte erkennen. Die Partielle Korrelationsanalyse kann helfen, ein wirklichkeitsgetreueres Wirkungsgefüge zwischen den Systemkomponenten aufzuklären.

Die Pfadanalyse (BAHRENBERG, GIESE & NIPPER, 1992) stellt wiederum das Methodeninventar bereit, die durch die Korrelationsanalyse erhaltenen gedanklichen Verbindungsstriche zwischen den Systemkomponenten („Boxen“) durch gerichtete Pfeile zu ersetzen. Als Modifikation der Regressionsanalyse kann die Pfadanalyse ebenfalls Wirkungsbeziehungen konstruieren und überprüfen helfen. Mit Hilfe der Regressionsanalyse beginnt man dann im folgenden Schritt der Systemsynthese, diese Wirkungsbeziehungen zu quantifizieren.

Die Korrelationsrechnung und die Faktoranalyse als ausgewählte Methoden der Systemidentifikation werden in den Kapiteln IV und V beschreiben.

# 3 SYSTEMSYNTHESE

Schließlich werden die erkannten Zusammenhänge zwischen erklärenden Datensätzen und in ihrer Dynamik zu erklärenden Datensätzen quantitativ zu verknüpfen sein.

Es geht darum, Prozesse oder eben den dynamischen Verlauf eines Datensatzes mit Hilfe steuernder Größen des Raumes zu berechnen, also zueinander in einen möglichst exakt beschreibbaren Zusammenhang zu bringen. Diesem Ansatz folgt in der Statistik üblicherweise die Regressionsanalyse. Als Arbeitsschritt während der Systemsynthese wird sie im Kapitel VI beschrieben.

Während die Regressionsanalyse allerdings versucht, lineare oder auch nichtlineare Beziehungen zwischen metrischen Variablen zu schätzen, kann mit einer anderen Methode auf die wahrscheinlichste Ausprägung von Typen geschlossen werden. Die Diskriminanzanalyse erlaubt es, mit Hilfe metrischer Variablen Kategorien zu schätzen.

Im Arbeitschritt der Systemsynthese können parametrisierte und regional gültige Modelle erstellt werden. Solche Modelle können das Verhalten abhängiger Variablen durch eine oder mehrere unabhängige Variablen über einen mathematischen Zusammenhang beschreiben. Mit Hilfe solcher Modelle können schließlich Szenarien berechnet werden, welche den Einfluss von Änderungen in den unabhängigen, also steuernden Variablen untersuchen. Schließlich sollen die entwickelten Modelle nicht nur zum Selbstzweck erstellt worden sein. Neben „Was wäre Wenn?“ - Analysen sind zudem Prognosen darüber möglich, in welchen verschieden wahrscheinlichen Intervallen sich künftig zu erwartende Veränderungen abspielen können. Wichtig bei solchen Vorhersagen und Analysen bleibt, dass sie stets statistisch gewonnene Aussagen repräsentieren. Diese Aussagen lassen sich „nur“, aber gerade deswegen, mit einer exakt definierten Wahrscheinlichkeit treffen.

Verschiedene statistische Methoden sind somit über Schritte der Systemanalyse, der Systemidentifikation und der Systemsynthese sinnvoll abgearbeitet worden. Die in jeweils einem Arbeitsschritt erhaltenen Ergebnisse sollen Fragen aufwerfen, welche mit Hilfe der folgenden Methoden wieder beantwortet werden können. Damit existiert zum Ende dieser Vorgehensweise ein vertieftes Verständnis für jene geographische Wirklichkeit, welche durch den Datensatz repräsentiert wird.

# KAPITEL I
# GRAPHISCHE UND PARAMETRISIERTE DATENANALYSE

## 1 Statistische Grundlagen

Am Anfang einer statistischen Problembearbeitung in der Geographie steht eine Beschreibung eines Raumes mit seinen Strukturen und Prozessen – die Geosystemanalyse (siehe Kapitel 1).

In diesem ersten Schritt wird es für den Geographen darauf ankommen, sein gesammeltes Datenmaterial in Augenschein zu nehmen, ohne komplizierte Analyse- und Testverfahren anwenden zu müssen. Eine Reihe von Fragen können, allein mit Hilfe der Berechnung einfacher Parameter oder mit einer günstigen Diagrammdarstellung beantwortet werden (siehe Box). Diese Art von Auswertung nennt man deskriptive (beschreibende) Datenanalyse.

**Diese Fragen können mit deskriptiver Statistik beantwortet werden:**

- Welches ist der „typischste", „mittlere", „häufigste" Wert der Verteilung?
- Unterscheiden sich die Werte gering oder sind sie stark gestreut?
- Weichen einzelne Werte stark von der Mehrheit der anderen ab?
- Sind die Werte normal- oder gleichverteilt?
- Ist die Verteilung symmetrisch oder unsymmetrisch?
- Welches ist der „typischste", „mittlere", „häufigste" Wert der Verteilung?

## 1.1 Untersuchungsgegenstand statistischer Methoden

Untersuchungsobjekt der statistischen Datenanalyse wird immer die G r u n d g e s a m t h e i t (z.B. alle Flusseinzugsgebiete) oder deren Teilmenge, die S t i c h p r o b e (z.B. Flusseinzugsgebiete im Ostseeraum) sein. Ein Element der Grundgesamtheit wird als M e r k m a l s t r ä g e r (z.B. Einzugsgebiet des Alan) bezeichnet. Jedem dieser Merkmalsträger sind ein oder mehrere gleiche M e r k m a l e (z.B. Abflussspende oder Fläche) zugewiesen. Jedes dieser Merkmale wiederum kann eine definierte Menge von Werten annehmen. Dies ist die Menge der M e r k m a l s a u s p r ä g u n g e n (z.B. Fläche gleich 12,5 km², 1003 km², …).

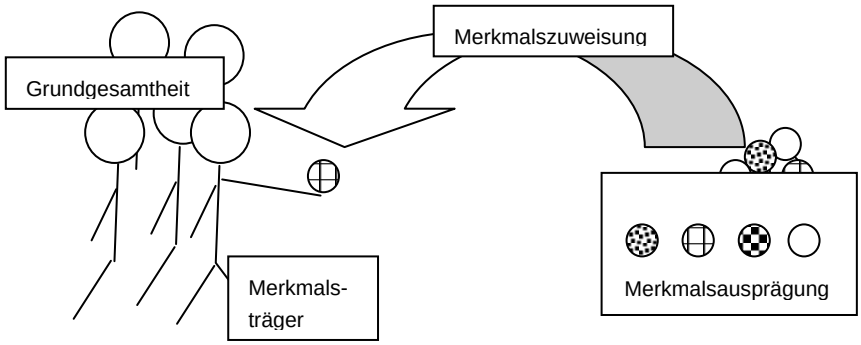



*Bild 1.1 Der Prozess der Merkmalszuweisung*

Programme wie EXCEL und SPSS verwalten statistische Rohdaten in tabellarischer Form (Bild 1.2). In den Zeilen gibt man üblicherweise die Auflistung der einzelnen Merkmalsträger an, die Spalten sind den Merkmalen vorbehalten.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Flußname** | **Form- und Flächenparameter** | | | **Abfluß** | | **Nährstoffzuflüsse** | | | |
| 2 | **Name** | Polygonfläche | Polygon-umfang | Einzugebiets-fläche (km²) | Mittlerer Jahresabfluss ($10^6$ m³ /Jahr) | Standard-abweichung des Jahresabfluss | durchschnittliche N-Fracht (t/Jahr) | Standard-abweichung der N-Fracht | durchschnittliche P-Fracht (t/Jahr) | Standard-abweichung der P-Fracht |
| 3 | **DESCRIPT** | A | u | A | Q | $s_Q$ | $M_N$ | $s_{MN}$ | $M_P$ | $s_{MP}$ |
| 4 | Aa | 8989621000.00 | 474219.70 | 8989.62 | 2100.80 | 563.26 | 5663.56 | 1620.28 | 65.74 | 18.02 |
| 5 | Abyalven, Byskealven | 8932701000.00 | 565102.80 | 8932.70 | 2363.84 | 415.74 | 1067.07 | 215.39 | 53.57 | 13.04 |
| 6 | Alan, Rosan | 935224800.00 | 335745.80 | 935.23 | 843.82 | 207.95 | 413.80 | 104.62 | 21.53 | 5.28 |
| 7 | Angermanalven | 30482570000.00 | 947279.90 | 30482.57 | 16613.03 | 2668.81 | 4608.59 | 1022.05 | 198.79 | 41.66 |
| 8 | Aurajoki | 10013540000.00 | 910232.80 | 10013.54 | 2214.46 | 684.44 | 4121.63 | 1466.64 | 314.16 | 126.29 |
| 9 | Braknean | 2902977000.00 | 326752.70 | 2902.98 | 734.25 | 235.58 | 946.53 | 380.43 | 23.06 | 11.05 |
| 10 | Dalalven | 27979770000.00 | 1024733.00 | 27979.77 | 10656.62 | 2596.88 | 4679.45 | 1416.09 | 184.45 | 54.01 |
| 11 | Danish Straits - Kavlingean, Saxan | 2394684000.00 | 362901.90 | 2394.68 | 746.57 | 213.53 | 1304.73 | 408.43 | 43.98 | 19.94 |
| 12 | Daugava | 91919990000.00 | 2289026.00 | 91919.99 | 20748.25 | 5563.89 | 45471.90 | 12226.22 | 1142.69 | 344.33 |

*Bild 1.2: Tabellarische Anordnung von Rohdaten in EXCEL*

Untersucht man die Einträge der Datentabelle, stellt man meistens fest, dass es äußerst schwierig ist, Eigenheiten und Gemeinsamkeiten dieser Daten durch den bloßen Blick auf deren tabellarische Auflistung zu entdecken. Wir werden uns deshalb in den weiteren Abschnitten mit der Visualisierung und Parameterbeschreibung von Verteilungen auseinandersetzen.

## 1.2 Skalenniveaus statistischer Daten

Beginnt man die Deskription des Datensatzes, wird man schnell auf das Problem stoßen, mit unterschiedlichen Formen der Merkmalsausprägungen der Daten konfrontiert zu sein. Die Bevölkerungszahl der Einzugsgebiete im Ostseeraum beinhaltet z.B. nur ganze, die der Einzuggebietsfläche auch gebrochene Zahlen. Der Name der Einzugsgebiete, welcher zweifellos ebenfalls eine Variable darstellt, eignet sich überhaupt nicht zur Mittelwertsberechnung. Um diese Probleme zu erkennen und angemessen zu behandeln, teilt man Variablen bezüglich der Form ihrer Merkmalsausprägungen in unterschiedliche Skalenniveaus ein (Tabelle 1.1).

| **Nominalskala** | **Ordinalskala** | **Metrische Skala** |
|---|---|---|
| **Qualitative** Angaben ohne Rangordnung (meist verbaler Wertebereich: „männlich“, „weiblich“, „dunkelbraun“ , „humos“) | **Qualitative** Angaben mit **Rangordnung** und **Ordnungsrelation** (Wertebereich Zensuren: „sehr gut“, … oder Einschätzungen: „sehr wichtig“, „wichtig“, …) | **Quantitative** Variable mit diskreter (Wertebereich ganze Zahlen) oder stetiger (Wertebereich reelle Zahlen) Ausprägung |
| Namen der Einzugsgebiete (Boden- oder Klimatyp, Beruf, Nachname) | Rangordnung der Tourismuseignung (Zufriedenheit, Rankings) | Einwohnerzahl (diskret) Flächenangaben (stetig) Abfluss (stetig) |

Tabelle 1.1 Skalenniveaus von Variablen

Häufig werden nominal- oder ordinalskalierte Werte zur besseren Verwaltung in Statistikprogrammen numerisch kodiert. Gleichen Merkmalsausprägungen werden dabei gleiche Zahlenwerte zugewiesen. Während ordinalskalierte Variablen ihrer Rangordnung gemäß kodiert werden, ist die Zuordnung für die Nominalskala völlig

willkürlich. Es macht keinen Unterschied, ob die Ausprägung „männlich“ mit dem Wert 1 oder 2,5 und die Ausprägung „weiblich“ mit 3 oder 10000 verknüpft werden.

Die Kodierung sorgt für eine effizientere Speicherung der Daten, birgt aber auch eine Fehlerquelle. Numerische Werte suggerieren die Möglichkeit statistische Kennwerte zu berechnen, obwohl diese nicht unbedingt sinnvoll sein müssen. Einige Kennwerte und Darstellungsmethoden eignen sich deshalb nicht ohne weiteres für alle Skalenniveaus (Tabelle 1.2)!

| **Nominalskala** | **Ordinalskala** | **Metrische Skala** |
|---|---|---|
| nur **Modus** und **Histogramm** als sinnvolle Verteilungscharakteristik | **Modus** und **Histogramm** **Median**, **Quantile** und **Boxplot** immer als sinnvolle Verteilungscharakteristik (bei metrischer Verschlüsselung auch **Mittelwert** und **Varianz)** | **alle Parameter** oder **Verteilungsdiagramme** (Ausnahme Modus für stetige Skalen) dienen der sinnvollen Verteilungscharakteristik |

Tabellen 1.2 Sinnvolle Parameter auf einzelnen Skalenniveaus

# 2 Parametrische Datenanalyse

Üblicherweise werden Merkmalsvariablen mit Hilfe einfacher Parameter charakterisiert (Tabelle 1.3), um mit wenigen Werten Aussagen über die gesamte Verteilung zu treffen. Man unterscheidet zwei Gruppen von Verteilungsparametern. Soll die Größe des „typischsten“, „mittleren“ oder „häufigsten“ Wert (die Lage der Verteilung) gekennzeichnet werden, spricht man von einem Lageparameter (Zentralmaße). Interessiert man sich für die Abweichung der Werte voneinander (die Variabilität der Verteilung) untersucht man Variabilitätsparameter (Streuungsmaße).

Häufig wird der Fehler gemacht, zur Kennzeichnung einer Verteilung nur einen Lageparameter (meist das arithmetische Mittel) anzugeben. Zur vollständigen Charakterisierung gehört aber immer auch die Angabe eines Variabilitätsparameters.

| **Lageparameter** | | **Variabilitätsparameter** | |
|---|---|---|---|
| „typischster“, „mittlerer“ oder „häufigster“ Wert (Lage) der Verteilung | | Abweichung der Werte voneinander (Variabilität) | |
| arithmetisches Mittel (Summe durch Anzahl aller Verteilungswerte) | `MITTELWERT()` | Varianz (mittlere quadratische Abweichung der Verteilungswerte vom arithmetischen Mittel) | `VARIANZ()` |
| Median (jeweils die Hälfte aller Verteilungswerte sind kleinergleich bzw. größergleich des Medians) | `MEDIAN()` | Standardabweichung oder Streuung (Pos. Wurzel der Varianz) | `STABW()` |
| Modalwert oder Modus (Häufigster Wert der Verteilung) | `MODALWERT()` | x %-Quantil (x % aller Verteilungswerte sind kleinergleich, 100-x % größergleich als das Quantil) (Quartil für x=25,50,75) | `QUANTIL()` |
| | | Spannweite (Differenz aus größtem und kleinstem Wert der Verteilung) | `MAX()-MIN()` |

Tabelle 1.3 statistische Kenngrößen und dazugehörige EXCEL-Funktionen

Angesichts der Vielzahl unterschiedlicher Kennwerte, stellt sich die Frage, welcher Parameter für welche Sachverhalte am besten zu verwenden ist. Zum einen existiert eine Einschränkung durch das Skalenniveau (Tabelle 1.2), zum anderen treffen die einzelnen Lage- und Variabilitätsparameter unterschiedliche Aussagen. Sie sind darum je nach Intension und Fragestellung einzusetzen.

Wir untersuchen dieses Thema etwas später am Beispiel.

# 3 Graphische Datenanalyse

Die einfachsten statistischen Verteilungsdiagramme sind Histogramm (Häufigkeitsdiagramm) und Boxplot. Sie dienen der kumulierten Darstellung eines Merkmals (z.B. jährlicher Abfluss aller Einzugsgebiete), ohne die Merkmalsausprägung eines einzelnen Merkmalsträgers (z.B. jährlicher Abfluss des Alån-Einzugsgebiets gleich $844*10^6$ $m^3$/Jahr) explizit wiederzugeben.

Legt man auf Letzteres Wert, sollte man auf die Verwendung von Sonnenstrahlicons oder Chernoff-Gesichtern zurückgreifen. Diese haben zudem den Vorteil, gleichzeitig mehrere Merkmale graphisch abzubilden.

## 3.1 Histogramm

Die Grundlage eines Histogrammes bildet eine Klasseneinteilung der Werte. Im Diagramm wird die Anzahl der Werte pro Klasse (Häufigkeit) eingetragen. Die Klassifikationsmethode hängt von der Aufgabenstellung ab und kann sehr unterschiedliche Ergebnisse in der Histogrammdarstellung zur Folge haben. Aus diesem Grund muss beim Vergleich von Häufigkeitsdiagrammen immer auf die Klassifikationsmethode geachtet werden. Dies kann als Nachteil der Histogrammdarstellung gewertet werden. Histogramme finden für alle Skalentypen Verwendung.

**Wie erstellt man ein Histogramm mit EXCEL?**

| | |
|---|---|
| 0 | Extras > Add-Ins > „Analyse-Funktionen“ anschalten |
| 1 | Extras > Analysefunktionen… (wenn nicht aufgelistet dann Schritt 0) |
| 2 | Histogramm |
| 3 | Eingabebereich und eventuell Klassenbereich (sonst Standardklassifikation) festlegen |
| 4 | „Diagrammdarstellung“ anschalten |
| 5 | OK |

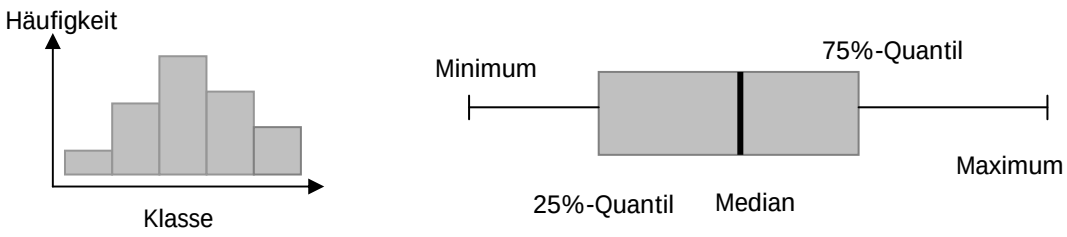



*Bild 1.3 Histogramm und Boxplot*

## 3.2 Boxplot

Der Boxplot ist dagegen unabhängig von Klassifikationsmethoden, enthält aber im Allgemeinen weniger Informationen als ein Histogramm, da in ihm nur Quantile, Maximum und Minimum dargestellt werden. Einige Statistikprogramme (z.B. SPSS) weisen Ausreißer im Diagramm gesondert aus. Ein Boxplot kann nur für nicht nominal-skalierte Variablen erstellt werden.

**Wie erstellt man ein Boxplot mit SPSS?**

1 Grafiken> Boxplot…
2 „Einfach“ auswählen > „Auswertung über verschiedene Variablen“ auswählen
3 > Definieren
4 Variablen in Feld „Box entspricht“ eintragen
5 OK

## 3.3 Chernoff-Gesichter

Die Chernoff-Gesichter verwendet man zur anschaulichen Wiedergabe mehrerer Merkmale eines Merkmalsträgers. Verschiedene Kopf-, Brauen-, … oder Nasenformen geben an, in welcher Klasse die Merkmalsausprägung eingeordnet ist. Jedes Gesicht repräsentiert dabei einen Merkmalsträger.

Als nachteilig erweist sich, das nur eine begrenzte Anzahl verschiedener Merkmale und Klassen darstellbar ist, dies wird aber durch Anschaulichkeit und Originalität der graphischen Methode ausgeglichen.

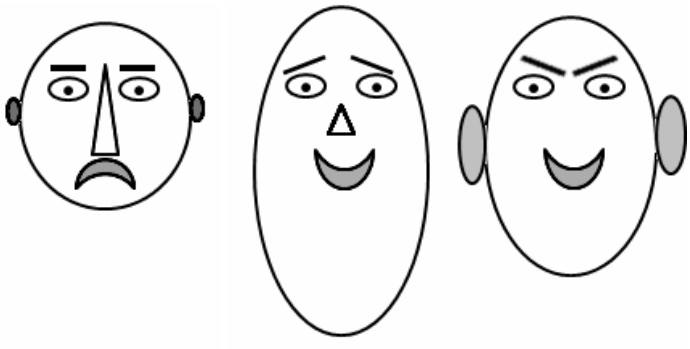


*Bild 1.4 Mit dem auf CD beiliegenden Programm generierte Chernoff-Gesichter*

**Wie erstellt man ein Chernoff-Gesicht?**

1 bis zu 5 verschiedene Merkmale klassifizieren (max. 4 verschiedene Klassen je Merkmal)
2 Chernoff.exe auf der CD starten
3 Klassen den Gesichtsformen zuweisen
4 > Bild kopieren und zum Beispiel in den Auswertungstext in MS Word einfügen

GGA_Chernoff\Chernoff.exe

## 3.4 Sonnenstrahl-Icons

Die Anwendung von Sonnenstrahl-Icons erstreckt sich auf den gleichen Bereich, wie die der Chernoff-Gesichter. Es können ebenfalls mehrere Merkmale eines Merkmalsträgers in einem Diagramm visualisiert werden. Die jeweilige Merkmalsausprägung wird auf sternförmig angeordneten Achsen eingetragen und zumeist verbunden. Da der Wertebereich unterschiedlicher Merkmale sehr stark differieren kann, sollte zuvor eine Standardisierung (zum Beispiel am Maximum) erfolgen.

Als Vorteil der Sonnenstrahl-Icons gegenüber den Chernoff-Gesichtern gilt, dass die Menge der wiederzugebenden Merkmale nicht beschränkt und deren Wertebereich stetig ist.

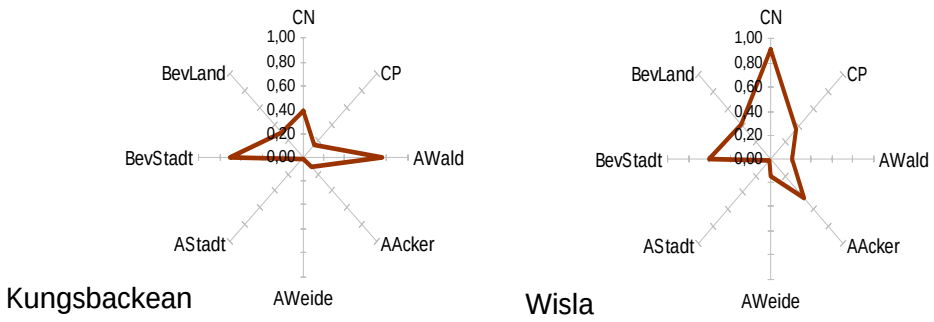



*Bild 1.5 Sonnenstrahl-Icons zum Nährstoffeintrag zweier Ostsee-Einzugsgebiete*

**Wie erstellt man ein Sonnenstrahl-Icons mit EXCEL?**

1 Merkmale standardisieren
2 Einfügen > Diagramm
3 Diagrammtyp „Netz“ auswählen
4 Zuweisung der Diagrammparameter > Fertig stellen
5 eventuell Gitternetzlinien entfernen

# 4 Parametrisierte und Graphische Datenauswertung am Beispiel

In diesem Beispiel möchten wir uns mit unserem Ostseedatensatz vertraut machen, deswegen berechnen wir einfache statistische Parameter und erstellen Verteilungsdiagramme. Uns interessiert der Waldanteil pro Einzugsgebiet, die Bevölkerungsdichte und die Abflussspende (Abfluss je km²/Jahr) jedes Einzugsgebietes. Diese Variablen sind im Ausgangsmaterial noch nicht enthalten, lassen sich aber leicht aus den Rohdaten ableiten. (Bild 1.6)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | **Name** | Einzugebiets-fläche ($km^2$) | Mittlerer Jahresabfluss ($10^6$ $m^3$ /Jahr) | Waldfläche ($km^2$) | Gesamt-bevölkerung | Abflussspende ($10^6$ $m^3$/ (Jahr*$km^2$) | Waldanteil | Bevölkerungs-dichte |
| 3 | **DESCRIPT** | A | Q | $A_{Wald}$ | Bev | Q/A | $A_{Wald}$/A | Bev/A |
| 4 | Aa | 8989.62 | 2100.80 | 5186.00 | 232288.00 | 0.234 | 0.577 | 25.840 |
| 5 | Abyalven, Byskealven | 8932.70 | 2363.84 | 7843.00 | 38366.00 | 0.265 | 0.878 | 4.295 |
| 6 | Alan, Rosan | 935.23 | 843.82 | 583.00 | 30765.00 | 0.902 | 0.623 | 32.896 |

*Bild 1.6 Abgeleitete Variablen Waldanteil, Bevölkerungsdichte und Abflussspende aus dem GRIDA-Ostseedatensatz*

GGA_Parameter_Graph\Übung_1.xls

■ Zur Charakterisierung der Merkmalsverteilungen berechnen wir einige Lage- und Variabilitätsparameter (Bild 1.7) und untersuchen ihre Eigenschaften:

| Lageparameter | Abflussspende ($10^6$ $m^3$/ (Jahr*$km^2$) | Waldanteil | Bevölkerungs-dichte |
|---|---|---|---|
| | Q/A | $A_{Wald}$/A | Bev/A |
| Anzahl | 61 | 61 | 61 |
| Summe | 20.435 | 39.072 | 2219.184 |
| Minimum | 0.105 | 0.052 | 1.355 |
| Maximum | 0.902 | 0.904 | 271.590 |
| Mittelwert | 0.335 | 0.641 | 36.380 |
| Median | 0.311 | 0.674 | 24.903 |
| Modalwert | 0.604 | 0.798 | 44.390 |

*Bild 1.7 Statistische Lageparameter*

Vergleichen wir die Lageparameter der Verteilungen, stellen wir fest, dass (arithmetischer) Mittelwert, Median und Modalwert stark voneinander abweichen. Wie kommt es zu dieser Abweichung und welcher Wert ist zur Charakterisierung der günstigste?

Die Berechnung eines Modalwertes, somit die Suche nach dem häufigsten Werte, ist für eine stetige Verteilung ohne Klassifikation nicht sinnvoll (Tabelle 1.2). Die Gleichheit zweier Werte entsteht häufig durch Rundung und somit zufällig. Trotzdem findet EXCEL einen Modalwert. Ein Vergleich mit den Ausgangsdaten zeigt, dass Angermanalven und Gadean genau dieselben Merkmalsausprägungen für Einzugsgebietsfläche, Jahresabfluss … aufweisen. Wir haben vermutlich einen Fehler im Ausgangsdatensatz zutage gefördert.

Der Mittelwert ist für die Abflussspende leicht größer und für den Waldanteil leicht kleiner als der Median. Das deutet darauf hin, dass die erste Verteilung leicht linkssteil und letztere leicht

rechtssteil ist (Bild 1.8). Sehen wir uns das in der graphischen Auswertung genauer an und wenden uns vorerst dem letzten Merkmal, der Bevölkerungsdichte zu.

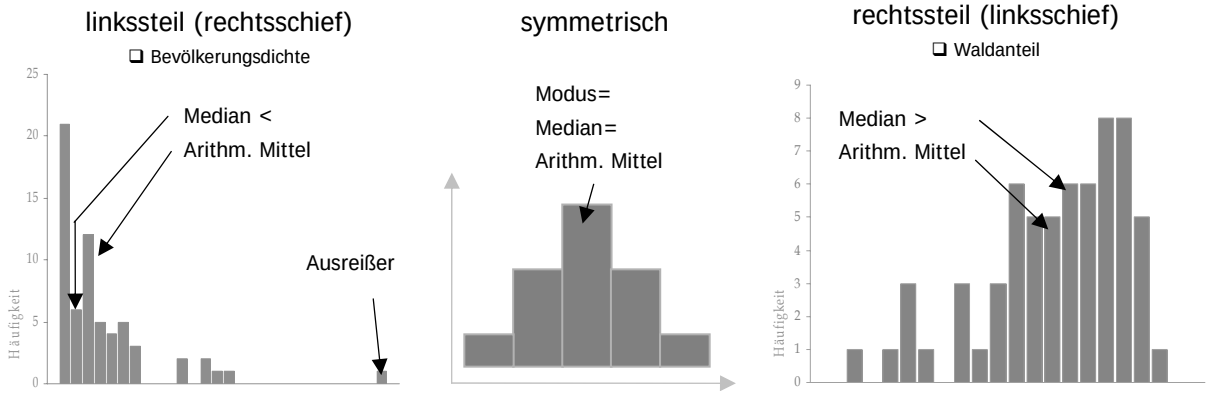



*Bild 1.8 Lageparameter in Abhängigkeit der Steilheit von Verteilungen*

Die Abweichung zwischen Median und Mittelwert ist hier wesentlich größer. Dies ist zum einen mit der Linkssteilheit der Verteilung begründbar. Skandinavische Einzugsgebiete mit geringer Bevölkerungsdichte sind weitaus häufiger im Datensatz enthalten, als Einzugsgebiete der Ostseesüdküste mit mittlerer und hoher Bevölkerungsdichte. Zum anderen weist das Merkmal für das Gebiet Danish Straits mit 271,6 Einwohner/km² eine Merkmalsausprägung auf, die stark vom Median 24,9 Einwohner/km² abweicht. Merkmalsträger, deren Abweichung vom mittleren Wert so gravierend ist, sind Ausreißer (Bild 1.8).

Diese stören zumeist die Beurteilung eines Merkmals und werden bei vielen statistischen Analysen im Voraus entfernt. Ein mögliches Kriterium dafür ist die Differenz vom Mittelwert. Beträgt sie mehr als die zweifache Standardabweichung, charakterisiert sie einen Ausreißerwert. Andere Erkennungsverfahren werden in der Clusteranalyse (Kapitel III) angesprochen. Wichtig für die Entscheidung, welcher Parameter den typischsten Wert am besten wiedergibt, ist die Ausreißerempfindlichkeit (siehe Box). Die Verwendung des ausreißerunempfindlichen Medians, ist der des arithmetischen Mittelwertes immer vorzuziehen.

**Ausreißerempfindlichkeit der Lageparameter?**

Der arithmetische Mittelwert wird durch einen Extremwert viel stärker beeinflusst als Median und Modalwert. Ursache: Das **Mittel** wird mit der an der Anzahl gewichteten Summe der Merkmalsausprägungen gebildet, also haben besonders große oder kleine Werte einen überproportionalen Einfluss auf das Ergebnis.

Beim **Median** zählt nur die Rangfolge der Werte. Der größte Wert hat zwar auch die größte Rangnummer, sein Abstand zum Nachbarn mit der zweitgrößten Rangnummer ist aber unerheblich. Der Median verschiebt sich nicht, selbst wenn ein Wert besonders groß ist.

Vorsichtig muss mit dem **Modalwert** operiert werden. Einzelne Ausreißer ändern nichts am häufigsten Wert. Kommt es aber zu einer Gruppierung von Ausreißern, so kann die maximale Häufigkeit überschritten werden. Der Modus wird zu dieser Ausreißergruppe verlagert und verliert seine Repräsentationsfunktion völlig.

Wenden wir uns an dieser Stelle den Variabilitätsparametern zu. Der einfachste Parameter in dieser Kategorie ist die Spannweite, die aus Differenz von Maximum und Minimum berechnet wird. Die Spannweite beträgt für den Waldanteil 0,85. Bei einem Anteilsmerkmal, kann das Minimum nicht kleiner als 0, das Maximum nicht größer als 1 werden. Die Spannweite liegt also immer zwischen 0 und 1. Deutet damit der Wert von 0,85 auf eine hohe Streuung hin?

| Varibilitätsparameter | Abflussspende ($10^6$ m³/ (Jahr*km²)) | Waldanteil | Bevölkerungs-dichte |
|---|---|---|---|
| | Q/A | $A_{Wald}$/A | Bev/A |
| Varianz | 0.024 | 0.039 | 2156.098 |
| Standardabweichung | 0.156 | 0.198 | 46.434 |
| Spannweite | 0.797 | 0.851 | 270.236 |
| unteres Quartil | 0.221 | 0.544 | 6.725 |
| oberes Quartil | 0.428 | 0.798 | 44.390 |
| Quartilsabstand | 0.207 | 0.254 | 37.665 |

*Bild 1.9 Statistische Variabilitätsparameter*

Nicht unbedingt, denn die Spannweite ist extrem ausreißerempfindlich! Schon ein besonders großer oder kleiner Wert verfälscht diesen Parameter so stark, das keine Aussage über die innere Streuung der Werte getroffen werden kann (siehe auch Bevölkerungsdichte). Die Spannweite sollte also nie unbedacht verwendet werden.

Besser verträglich mit Extremwerten sind Varianz oder als Wurzel der Varianz, die Standardabweichung. Sie geben die mittlere quadratische Abweichung vom Mittelwert an. Aus Dimensionsgründen ist die Angabe der Standardabweichung vorzuziehen. Eventuell vorhandene Einheiten stimmen dann mit denen der Ausgangswerte überein.

Möchte man die Standardabweichung zweier Verteilungen vergleichen, so berechnet man am besten den Variationskoeffizient $v = \frac{\sigma}{\mu}$ ($\mu$.- Mittelwert $\sigma$ - Standardabweichung). Dieser gibt an, wie stark die Werte um den Mittelwert konzentriert sind.

Für die Bevölkerungsdichte liegt die Standardabweichung bei 46,3 EW/km² und ist bei einem Mittelwert von 36,4 EW/km² ziemlich groß (Variationskoeffizient $v = \frac{46,3}{36,4} = 1,27$). Für das Merkmal Waldanteil beträgt sie 0,19 (bei einem Mittel von 0,64) und ist, entgegen der mit Hilfe der Spannweiten getroffenen Aussage, als klein einzustufen (Variationskoeffizient $v = \frac{0,19}{0,64} = 0,30$).

Ein Nachteil von Varianz und Standardabweichung ist ihre Abhängigkeit vom arithmetischen Mittel. Wird es durch Ausreißer verfälscht, so werden es auch die davon abgeleiteten Streuungsmaße.

Als Alternative dient die Verwendung des oberen und unteren Quartil. Diese Werte geben an, für welche Schranke 25% bzw. 75% der Verteilungswerte kleiner sind und sie besitzen wie der Median eine große Toleranz gegenüber Ausreißern. Ein weiterer Streuungsparameter ist der Quartilsabstand, die Differenz aus den beiden Quartilen.

GGA_Parameter_Graph\Übung_2.sav

■ Aufgabe: Erstellen Sie mit Hilfe des Statistikprogramms SPSS Boxplots der abgeleiteten Merkmale (Bild 1.10). Welche statistischen Parameter sind direkt abzulesen, welche lassen sich ableiten? Vergleichen Sie diese Diagramme mit Histogrammdarstellungen (Bild 1.8).

Welches sind die Vor- und Nachteile der verschiedenen Diagrammtypen?

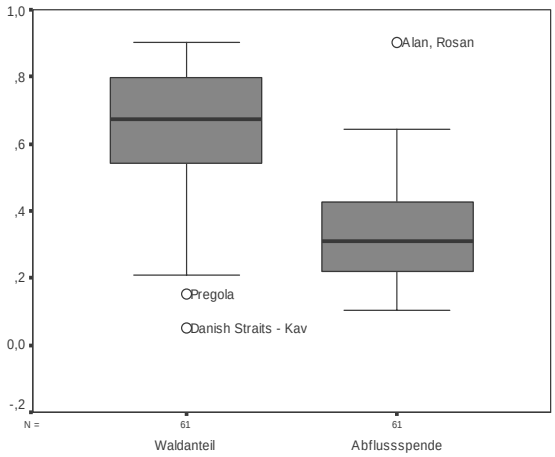



*Bild 1.10 SPSS-Boxplot*

GGA_Parameter_Graph\Übung_3.xls

GGA_Parameter_Graph\Einzugsgebietskarte.jpg

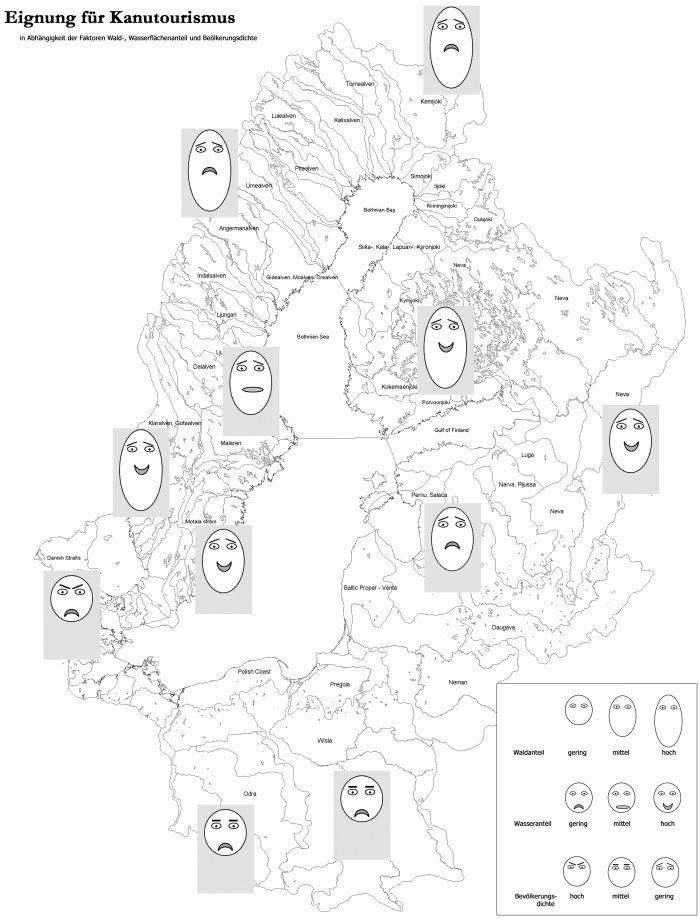



*Bild 1.11 Chernoff-Gesichter zur Einzugsgebiets-Charakterisierung*

■ Wir möchten die Eignung von Einzugsgebieten für einsamen Natur-Kanutourismus darstellen. Dafür erscheinen uns ein hoher Wald- und Wasserflächenanteil und geringe Bevölkerungsdichte als wichtig!

1. Ordnen Sie die drei Merkmale jeweils 3 Klassen zu!

2. Erstellen Sie für 10 Einzugsgebiete entsprechende Chernoff-Gesichter!

3. Kopieren Sie diese mit Hilfe eines Grafikbearbeitungsprogramms in die vorbereitete Karte und weisen Sie günstige und ungünstige Gebiete aus!

# KAPITEL II
# WAHRSCHEINLICHKEITSRECHNUNG

## 1 Einführung

In statistischen Jahrbüchern und anderen Datensammlungen werden oftmals Angaben zu Mittelwert und Standardabweichung verschiedener Variablen gemacht. Aber gerade die Angabe der Standardabweichung wird bei der Datenanalyse kaum verwendet.

Wie könnten die hiermit charakterisierten unterschiedlichen Verteilungen aussehen? Welche zusätzlichen Aussagen lassen sich daraus ableiten? Wie korrespondiert überhaupt das Konzept der Normalverteilung und der Berechnung von Wahrscheinlichkeiten mit derartigen Angaben?

## 2 Grundlagen

### 2.1 Dichte- und Verteilungsfunktionen

Im Kapitel I wurde der Diagrammtyp Histogramm vorgestellt. Er dient der Abbildung von Häufigkeiten. Bei der Betrachtung einer Variablen mit diskreter metrischer Skala, verwendet man die Häufigkeit einer Merkmalsausprägung *k*. Normiert man jedoch die Variable, so dass die Summe aller Häufigkeiten 1 ergibt, erzeugt man eine diskrete Wahrscheinlichkeitsfunktion *f(k)* (Bild 2.1). Die Merkmalshäufigkeit ist hier zu einer Eintrittswahrscheinlichkeit des Merkmals transformiert worden. Kumuliert man die Klassenwahrscheinlichkeiten, also addiert man die Werte der kleineren oder gleichen Merkmalsausprägungen auf, erzeugt man eine diskrete Verteilungsfunktion *F(k)* (Bild 2.1).

Mit ihrer Hilfe kann leicht ermittelt werden, wie groß die Eintrittswahrscheinlichkeit von Merkmalsausprägungen eines ganzen Intervalls ist.

Es gilt: $P(Intervall) = F(obere\ Intervallsgrenze) - F(untereIntervallsgrenze)$

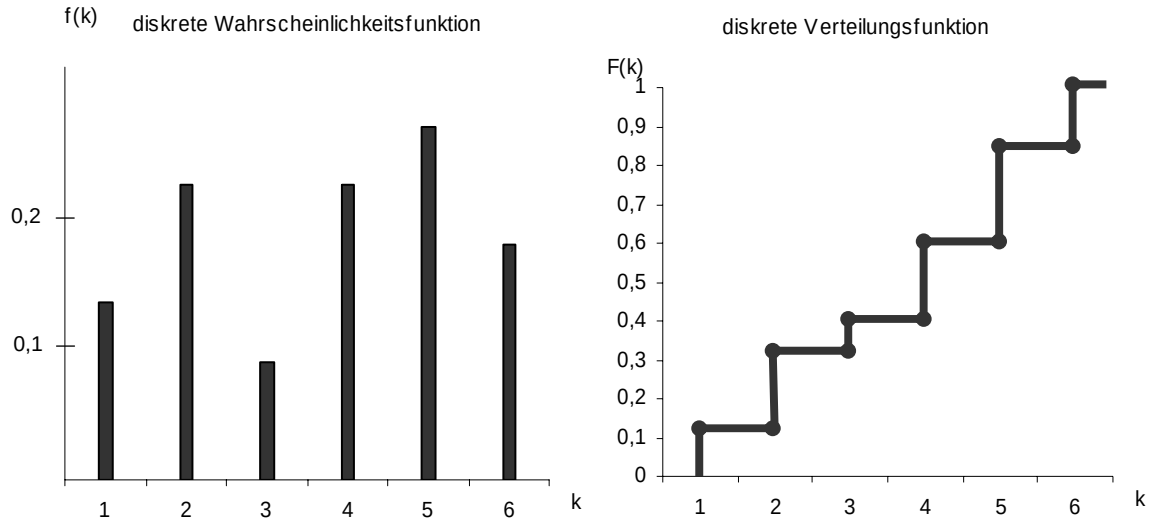



*Bild 2.1 Diskrete Wahrscheinlichkeits- und Verteilungsfunktionen*

Im stetigen Fall betrachtet man als Äquivalent zur diskreten Wahrscheinlichkeitsfunktion eine stetige Dichtefunktion (Bild 2.2). Die Wahrscheinlichkeit des Auftretens einer einzelnen Merkmalsausprägung ist verständlicherweise Null. Es lässt sich allerdings vorhersagen, wie groß die Wahrscheinlichkeit ist, dass die Ausprägung eines Merkmals innerhalb eines Intervalls $(x_0; x_1)$ liegt. Diese Wahrscheinlichkeit entspricht gerade der Fläche unter der Dichtefunktion im Intervall.

Um diese Fläche im allgemeinen Fall zu ermitteln, berechnet man die stetige Verteilungsfunktion (Bild 2.2) als bestimmtes Integral der Dichtefunktion $F(x) = \int_{-\infty}^{x} f(z)dz$ (Integration als Pendant der Kumulation im diskreten Fall). Mit ihrer Hilfe kann wieder leicht die Eintrittswahrscheinlichkeit von Merkmalsausprägungen eines Intervalls abgeleitet werden. Es gilt wie im diskreten Fall: $P(x_0, x_1) = F(x_1) - F(x_0)$.

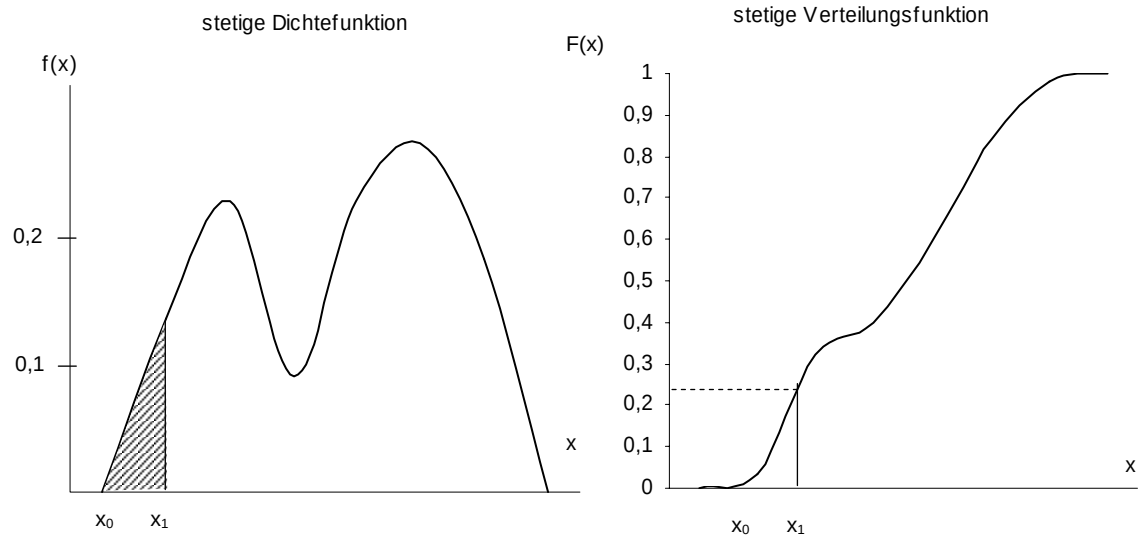



*Bild 2.2 Stetige Dichte- und Verteilungsfunktionen*

## 2.2 Die Normalverteilung

Die Gaußsche Normalverteilung ist eine wichtige Verteilungsfunktion der Statistik und wird unter anderem für die Beschreibung des Mittelswerts von Stichprobenverteilungen, als Modell empirischer Zufallsvariablen oder als Grenzwert der Binomialverteilung verwendet.

Wegen der glockenartigen Form ihrer Dichtefunktion wird sie auch Gaußsche Glockenkurve genannt. Sie wird als so bedeutend angesehen, dass die ihre Dichtefunktion als Vermächtnis des großen deutschen Mathematikers Carl Friedrich Gauß (1777-1855) auf dem alten Zehn-DM-Schein abgebildet worden ist (Bild 2.3). Dort lässt sich auch die Formel der Dichtefunktion ablesen:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu$ – arithmetri sches Mittel
$\sigma$ – Standardab weichung
$e$ – Eulersche Zahl


*Bild 2.3 Alter Zehn-DM-Schein mit Bild der Gaußschen Normalverteilung*

Gauß hatte diese Verteilung ursprünglich entwickelt, um aufgetretene Messfehler zu beschreiben (Gaußsche Fehlerkurve). Heute verwendet man die Verteilung aber auch als allgemeine Näherung für beliebige unbekannt verteilte Größen. Dabei ist zu beachten das die Eigenschaften der Normalverteilung (Eingipfligkeit, Symmetrie) durch die zu nähernde Zufallsvariable nicht offenkundlich verletzt wird. Die Anpassung einer Normalverteilung mag z.B. für Häufigkeiten von Niederschlagsereignissen im Monat Mai an einer festen Messstation zulässig sein, für die Verteilung von Extremhochwässern ist sie aber ungeeignet.

**Eigenschaften der Normalverteilung**

- die Verteilung ist eingipflich und symmetrisch
- der arithmetische Mittelwert, Median der Verteilung und Maximum der Dichtefunktion sind gleich
- die Ränder der Verteilung nähern sich asymptotisch der Null
- die Wendepunkte der Dichtefunktion liegen bei σ und -σ
- ist Mittelwert der Verteilung gleich 0 und Standardabweichung gleich 1, dann heißt sie Standardnormalverteilung (SNV)

# 3 Über- und Unterschreitungswahrscheinlichkeit

## 3.1 Vertrauensintervall

Häufig ist bei der Untersuchung von Merkmalen nicht nur die Frage interessant, welche Ausprägung im Mittel angenommen wird, sondern auch in welchem Intervall besonders viele der Werte liegen. Sollte man eine weitere Messung des Merkmals vornehmen, so würde man darauf vertrauen können, dass der neue Wert mit einer hohen Wahrscheinlichkeit in diesem festen Intervall liegt. Man bezeichnet dieses Intervall deswegen auch Vertrauensintervall. Oft wird gefordert, dass die Wahrscheinlichkeit für die Lage im Vertrauensintervall 95% betragen soll. Für bestimmte Analysen z.B. in der Medizin oder der Physik kann diese Grenze aber auch höher oder geringer gesetzt werden.

Die Lage der Grenzen des Vertrauensintervalls hängen von der dem Merkmal zugrunde liegenden Verteilung ab. In der Praxis ist es allerdings gar nicht so einfach, eine unbekannte Verteilung zu bestimmen. Man muss sich deswegen häufig mit einer näherungsweisen Verteilung begnügen.

## 3.2 Das Vertrauensintervall der Normalverteilung

In Abschnitt 2.2 wird die Normalverteilung als eine häufig verwendete Näherung für Verteilungen beschrieben. Verletzt die Merkmalsverteilung keine der grundlegenden Normalverteilungseigenschaften, ist deren Wahl bei nicht bekannter Verteilung allen anderen vorzuziehen. Bedeutsam ist dabei die Tatsache, dass sich eine Normalverteilung allein durch die Angabe von Mittelwert und Standardabweichung parametrisieren lässt und dadurch voll bestimmt ist.

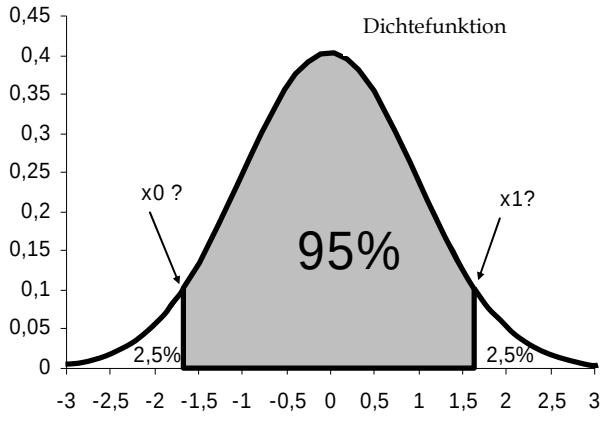



*Bild 2.4 Vertrauensintervall einer Standardnormalverteilung*

Wie wird für eine Standardnormalverteilung das 95%-Vertrauensintervall $(x_0; x_1)$ berechnet?

Gesucht sei zuerst mit $x_0$ diejenige Stelle der Dichtefunktion, die ein Flächenstück mit Inhalt 0,025 nach rechts begrenzt. Um Flächeninhalte unter Funktionen exakt zu berechnen, dient dem Mathematiker die Integralrechnung.

Da das bestimmte Integral der ohnehin komplizierten Normalverteilungsfunktion noch wesentlich komplizierter ist, soll in diesem Buch auf einen Abdruck desselben verzichtet und auf die Literatur verwiesen werden. Dem Anwender statistischer Methoden stehen mit dem Programm EXCEL oder anderer Mathematiksoftware ausreichend Möglichkeiten zur Verfügung, die gesuchten Intervallsgrenzen aufzuklären.

Dazu wenden wir uns der Verteilungsfunktion $F(x)$ zu. Sie ist bereits die Integralfunktion der Normalverteilung. Gesucht wird $x_0$ für das $F(x_0) = 0{,}025$ also $x_0 = F^{-1}(0{,}025)$ (Bild 2.5). Zu berechnen ist somit ein Funktionswert der Inversen der Verteilungsfunktion. Mit Hilfe der EXCEL-Funktion `NORMINV` erhält man etwa den Wert -1,96 für die untere und $x_1 = F^{-1}(0{,}975) \approx 1{,}96$ für die obere Vertrauensintervallsgrenze (Siehe Box). Für ein standardnormalverteiltes Merkmal lässt sich feststellen, dass 95% aller Werte im Intervall (-1,96; 1,96), also ungefähr dem zweifachen Standardabweichungsintervall, zu finden sind.

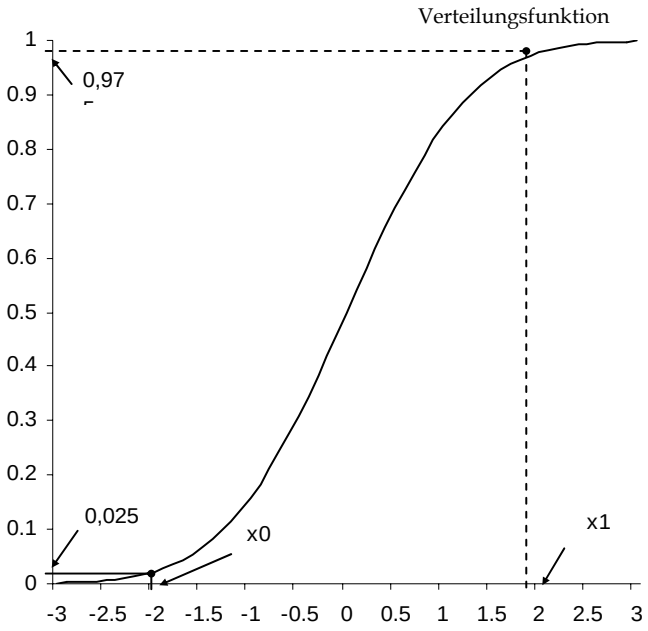



*Bild 2.5 Vertrauensintervall der Verteilungsfunktion der Standardnormalverteilung*

**Wie berechnet man das 95%-Vertrauensintervall der SNV mit EXCEL?**

1. Einfügen > Funktion…
2. Kategorie auswählen: „Statistik“
3. Funktion auswählen: `NORMINV`
4. Wahrsch „0,025“ (unteres Vertrauensintervall) oder „0,975“ (oberes Vertrauensintervall)
5. Mittelwert „0“ (oder entsprechend)
6. Standabwn „1“ (oder entsprechend)

# 4 Berechnung von Vertrauensintervallen am Beispiel

GGA_Wahrscheinlichkeit\Uebung_2.xls und

GGA_Wahrscheinlichkeit\Uebung_3.xls

■ Zur Übung werden wir uns mit der Verteilung der jährlichen Abflüsse in unserem Untersuchungsgebiet befassen. Uns interessiert in diesem Beispiel nicht mehr die Verteilung der Werte über alle Einzugsgebiete. Wir wollen in dieser Übung wissen, wie sich die Schwankungen des Abflusses in einem bestimmten Einzugsgebiet verhalten, welche Werte er mit einer bestimmten Wahrscheinlichkeit über- oder unterschreitet, kurz, welches Vertrauensintervall für die Variable anzunehmen ist.

Zuvor müssen wir eine Annahme bezüglich der Art der Verteilung der Abflüsse treffen. Dazu erstellen wir uns für eine Beispielstation (Datei: Uebung_3.xls) mit bekannten Werten ein Histogramm (Bild 2.6).

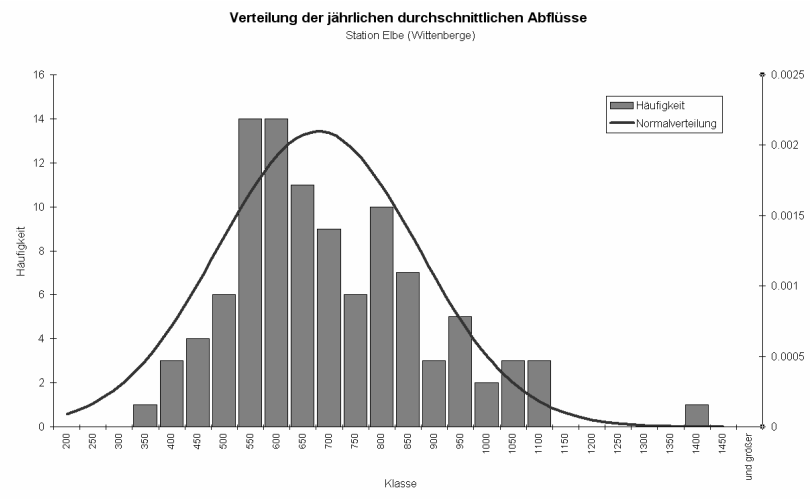



*Bild 2.6 Histogramm einer Zeitreihe von Abflüssen*

Jährliche Abflussverteilungen sind oftmals leicht linksschief. Dies wird auch für die Station Wittenberge an der Elbe bestätigt. Alles in allem ist die Abweichung zwischen Normalverteilung und gemessener Verteilung aber so gering, dass wir eine Normalverteiltheit für unbekannte Abflussverteilungen idealisieren können.

Zurück zum baltischen Datensatz (Datei: Uebung_2.xls). In ihm wurde für jedes Einzugsgebiet Mittelwert und Standardabweichung des jährlichen Durchschnittsabfluss angegeben. Diese beiden Parameter reichen aus, um die Normalverteilung der Variable für diese Einzugsgebiete berechnen zu können.

Um herauszufinden wie groß das 95%-Vertrauensintervall der Variable ist benötigen wir die EXCEL-Tabellenfunktion `NORMINV` mit den folgenden Parametern:

- Wahrsch (oder Alpha) = 0.025 (untere Grenze) Wahrsch = 0.975 (obere Grenze)
- Mittelwert = Wert aus Spalte B in der gleichen Zeile
- Standabwn =Wert aus Spalte C in der gleichen Zeile

Wir können nun in den Spalten D und E die untere und obere Intervallsgrenze für unsere Einzugsgebiete ablesen. Mit einer Wahrscheinlichkeit von 95%, also statistisch relativ sicher, liegt ein zufällig heraus gegriffener Jahreswert innerhalb dieses Intervalls.

Um die Einzugsgebiete bezüglich der Breite ihres Vertrauensintervalls vergleichen zu können stellen wir dieses für eine Auswahl in einem gemeinsamen Diagramm dar.

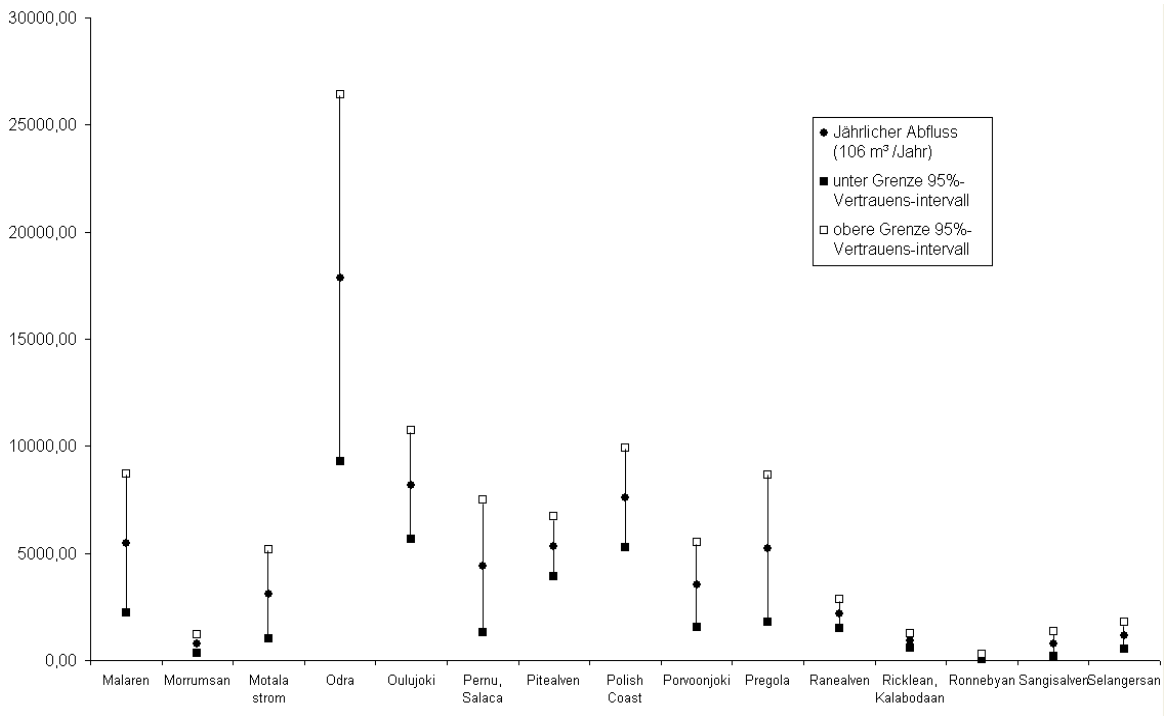



*Bild 2.7 Mittlerer Abfluss in $10^6$ m³/Jahr und Abflüsse, welche in 5% der Fälle unterschritten bzw. überschritten werden.*

Wir erkennen, dass das Einzugsgebiet mit dem größten jährlichen Durchschnittsabfluss (Oder) auch das größte Vertrauensintervall aufweist. Mit geringerem Abflussmittelwert nimmt in der Regel auch der Bereich der Schwankungen ab. Es gibt aber auch Ausnahmen: So besitzen Pernu, Salaca und Pregola einen geringeren Jahresdurchschnitt aber eine wesentlich höhere Schwankungsbreite als Abflüsse des Piteälv und Oulojoki. Eine Interpretation dieses Phänomens, ist die Hauptaufgabe des Geographen, und sollte nicht vergessen werden. Je weiter östlich die Einzugsgebiete liegen, umso größer wirken sich Veränderung im kontinentalen Strömungssystem aus. Westliche Einzugsgebiete werden nahezu in jedem Jahr von hohe Wassermengen transportierenden Zyklonen erreicht. Östliche Einzugsgebiete werden aber in Einzeljahren nicht oder weniger oft überstrichen und sind somit viel trockener. Die Streuung der Jahreswerte wird größer.

GGA_Wahrscheinlichkeit\Uebung_1.xls

Mit Hilfe dieser Übungsdatei kann selbstständig der Übergang von der Häufigkeit in einer Verteilung zur Wahrscheinlichkeit des Eintretens eines Ereignisses nachvollzogen werden.

# KAPITEL III
# CLUSTERANALYSE

Bislang konnten wir im Rahmen der Systemanalyse einige Kenntnisse über die unterschiedliche Ausprägung von Eigenschaften (=Variablen) in den Einzugsgebieten der Ostsee gewinnen. Erinnert sei noch einmal an die unterschiedlichen Histogramme für die Waldbedeckung. Hier gab die deutliche Rechtsschiefe einen Hinweis darauf, daß in vielen Einzugsgebieten ein hoher Waldanteil zu finden ist. Demgegenüber wies die linksschiefe Verteilung der Bevölkerungsdichte auf zahlreiche dünn besiedelte Einzugsgebiete im betrachteten Datensatz hin. Mit Hilfe der Berechnung von Quantilen und deren Darstellung in Karten erschloss sich zudem ein Überblick, ob in einem einzelnen Einzugsgebiet eine Eigenschaft eher gering, mittel oder stark ausgeprägt ist.

Diese Analysen wiesen darauf hin, dass sich etliche der betrachten Einzugsgebiete ähnlich sind, sich aber auch voneinander unterscheiden. Einen ersten Eindruck hiervon konnten wir durch die grafische Datenanalyse gewinnen. Chernoff-Gesichter und Sonnenstrahl-Icons waren hilfreich, unterschiedliche und gleiche Eigenschaften von Einzugsgebieten zu visualisieren. Hier wurde die Systemanalyse bereits um einen Schritt erweitert. Die Merkmale aller Einzugsgebiete wurden mit diesen einfachen Verfahren der grafischen Datenanalyse nicht nur jeweils für sich, sondern bereits als komplexe Merkmalsausprägung behandelt. Wir haben somit bereits eine erste multiple Datenanalyse durchgeführt.

## 1 Bildung von Gruppen durch Schwellenwerte

Wie können aber Ähnlichkeiten oder Unterschiede in einem Datensatz mit einer größeren Anzahl von Variablen gefunden werden? Zunächst wollen wir nochmals lediglich zwei Variablen gemeinsam untersuchen. Wird der Wald– und Seenanteil in den Einzugsgebieten der Ostsee in einem XY-Diagramm dargestellt, können bereits ähnliche Gruppen in Bezug zu diesen beiden Variablen erkannt werden. Es werden jetzt nicht nur eindimensional Unterschiede in der Waldbedeckung betrachtet, vielmehr entscheidet das Zusammenspiel von Wald- und Seenanteil über die Gruppenzugehörigkeit.

Anhand des XY-Diagramms lassen sich optisch recht gut 3 Gruppen unterschiedlicher Wald-Seen-Verteilung erkennen. Wald- und seenarme Einzugsgebiete in Polen und dem Baltikum haben einen Seenanteil unter 4%, andererseits übersteigt der Waldanteil nicht 60%. Die waldreiche, aber seenärmere zweite Gruppe weist Waldanteile über 40% und einen

Seenanteil zwischen 4% und 10% auf. Zu dieser Gruppe gehören zusätzlich noch Einzugsgebiete mit einem Waldanteil über 60%, aber Seeflächenanteilen unter 10%. Die Seenplatte Finnlands mit ihren ausgedehnten Wäldern wird dagegen durch einen Seenflächenanteil über 10% repräsentiert und findet sich wie die anderen beiden Gruppen ebenfalls gut im Kartenbild wieder.

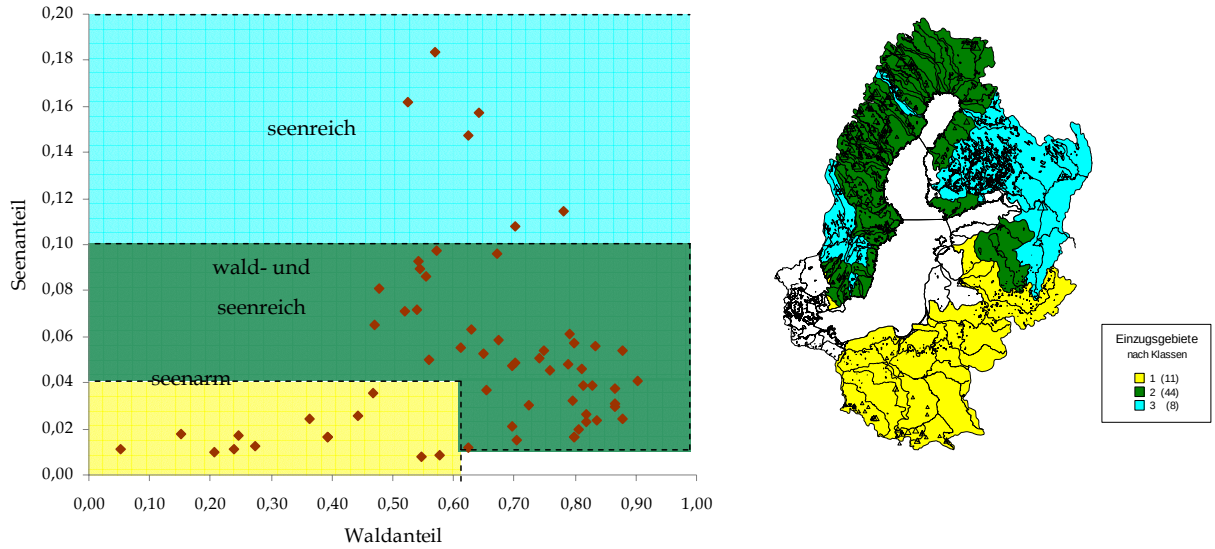



*Bild 3.1: Gruppen unterschiedlichen Wald- und Seenreichtums in den Einzugsgebieten der Ostsee*

Die hier gerade vorgenommene Unterteilung unterscheidet die Gruppen anhand von Schwellenwerten. Schwellenwerte müssen meist subjektiv festgelegt werden und sollten zudem plausibel zu begründen sein. Hier wurden sie allerdings aufgrund unterschiedlicher „Klumpen" im XY-Diagramm festgelegt, eine Methode, welche im folgenden durch die Clusteranalyse verbessert und verfeinert werden soll.

**Anwenden der Filterfunktion in EXCEL**

1. Markieren der Spaltenköpfe.
2. Menü Daten > Filter > Autofilter aktiviert die Filterfunktion.
3. Wie in Bild 3.2 wird jetzt ein benutzerdefinierter Autofilter ausgewählt.
4. Zunächst werden für den Anteil der Waldfläche Fälle kleiner 0,6 abgefiltert, danach wird der Autofilter für die Seenfläche angewandt und die Fälle kleiner 0,04 abgefiltert. In die Spalte rechts neben Anteil Seenfläche kann jetzt jeweils eine 1 eingetragen werden.
5. Das Vorgehen wird für die Gruppen 2 und 3 wiederholt, jetzt werden aber für den Waldanteil Werte > 0,4 und Seenanteile zwischen 0,04 und 0,1 abgefiltert, zusätzlich noch Waldanteile > 0,6 und Seenanteile < 0,1. Die jetzt abgefilterten Einzugsgebiete repräsentieren die Gruppe 2. Für die 3. Gruppe sind die Seenanteile größer 0,1. Nach dem Filtern sind in die Spalte für die Klassen (links neben der Spalte Seenflächenanteil) entsprechend die Zahlen 2 und 3 einzutragen.

■ Übungsbeispiel: Stellen Sie die unterschiedlichen Wald- und Seenflächenanteile für die Einzugsgebiete der Ostsee ebenfalls anhand eines XY-Diagramms dar! Extrahieren Sie die mit Hilfe von Schwellenwerten festgelegten Gruppen anschließend mit Hilfe der Filterfunktion von EXCEL!

GGA_Clusteranalyse\Uebung_1.xls

Arbeitsschritte: Markieren Sie die Spalten Anteil Waldfläche und Anteil Seenfläche in der Datei Uebung_1.xls! Erstellen Sie mit dem Diagrammassistenten ein Punkt(XY)-Diagramm (Einfügen > Diagramm, Diagrammtyp „Punkt(XY)“). Sie erhalten eine Darstellung ähnlich der in Bild 3.1. Nun sollen die 3 hier erkennbaren unterschiedlichen Gruppen des Wald- und Seenflächenanteils den jeweiligen Einzugsgebieten zugeordnet werden. Hierzu wollen wir die Filterfunktion von EXCEL nutzen. Bild 3.2 zeigt, wie vorgegangen wird. Die erhaltene Klassenzuordnung kann jetzt auch auf einer Karte dargestellt werden (Bild 3.1). Neben der gerade angeführten, inhaltlich sinnvollen Gruppenunterscheidung, finden sich hier auch räumlich unterschiedliche Gruppen wieder.

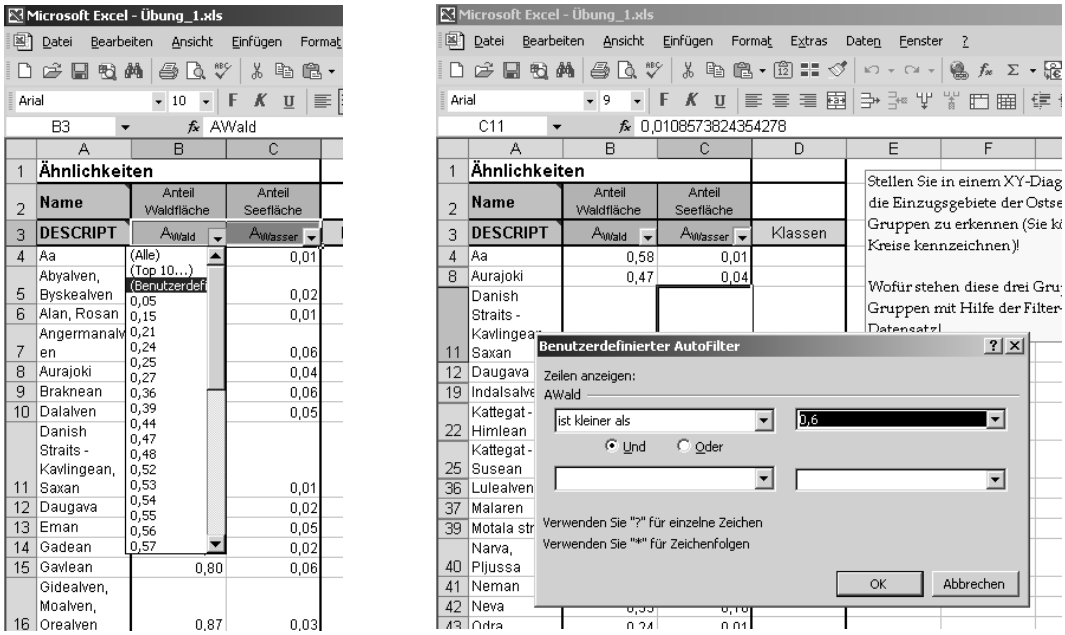



*Bild 3.2: Anwenden eines benutzerdefinierten Autofilters*

# 2 Vergleich von mehreren Merkmalen – Ähnlichkeits– oder Distanzmaße?

Die gerade ermittelten 3 Gruppen verglichen die unterschiedlichen Einzugsgebiete lediglich nach ihrem Wald- und Seenanteil, also hinsichtlich zweier Merkmale. Diese Aufgabe ließ sich noch grafisch lösen. Es könnten aber sicherlich noch mehr Kennzeichen für unterschiedliche Raumeinheiten in die Analyse eingehen. In unserem Datensatz lassen sich für die Ostseeeinzugsgebiete zudem Aussagen über Abfluss, Nährstofffrachten, Acker-

flächenanteil…, also zahlreiche andere Variablen treffen. Wie können in einem solchen Fall die Einzugsgebiete untereinander verglichen werden?

Hier muss die Ähnlichkeit oder die Unterschiedlichkeit der Variablen zwischen mindestens zwei verschiedenen Gebieten gemessen werden. Ähnlichkeit oder Distanz sind damit auch genau jene Schlüsselworte, welche wir zunächst klären wollen. Bild 3.3 soll zeigen, was mit diesen beiden Begriffen gemeint ist.

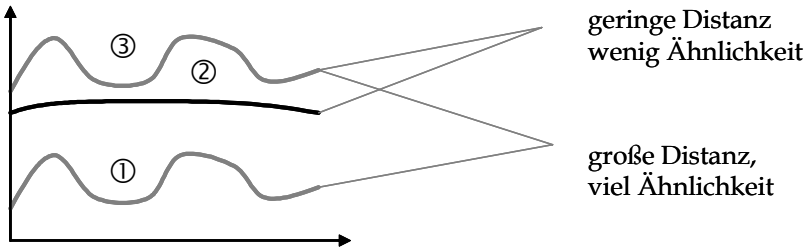



*Bild 3.3: Distanz-oder Ähnlichkeitsmaß?*

**Ähnlichkeits- oder Distanzmaße? (Bild 3.3)**

Unterschiede werden oft als Distanzen gemessen. Eine Distanz misst dabei den direkten Abstand zwischen den Werten in den zu vergleichenden Fällen, hier als Linien mit den Nummern 1 bis 3 dargestellt. So ist der Abstand zwischen ① und ③ sicherlich größer als jener zwischen ① und ②. In ① und ③ findet sich aber ein Kurvenverlauf, der in ② so nicht zu finden ist. Die Linien 1 bis 3 werden damit hinsichtlich Distanz und Ähnlichkeit unterschiedlich zu bewerten sein. Während in der Clusteranalyse zumeist Distanzmaße verwendet werden, kann bei verschiedenen Fällen durchaus auch ein Vergleich über die Ähnlichkeit den Daten angemessen sein. Bei Temperatur- oder Aktienverläufen könnte somit der Kurvenverlauf von stärkerem Interesse sein als die jeweiligen Differenzen.

Generell gilt:

D i s t a n z m a ß e werden oft dann verwendet, wenn der absolute Abstand der Objekte relevant ist und die Unähnlichkeit der Objekte umso größer ist, je weiter die Objekte voneinander entfernt liegen (Absolutbeträge von z.B. Nährstofffrachten differieren).

Besonders das Blockmaß, weniger die Euklidische Distanz betonen die Unterschiede zwischen Fällen. Damit entstehen zwar homogene Klassen, die „Ausreißer"-klassen sind jedoch nur gering besetzt. Es treten starke „Normalklassen" und geringbesetzte „Randcluster" auf!

Ä h n l i c h k e i t s m a ß e werden oft dann verwendet, wenn mit der Ähnlichkeit die generelle Tendenz der Werte gemessen werden soll und das Lageniveau weniger von Interesse ist (Relationen der Werte zueinander werden betont, z. B. mit einem hohen Waldanteil korrespondiert ein hoher Seenanteil).

Ähnlichkeitsmaße nivellieren Singularitäten und ordnen diese „echten" Klassen zu. Die Klassen werden damit recht stark, aber nicht vollständig homogen!

## 2.1 Anwendung verschiedener Distanz- und Ähnlichkeitsmaße

Mit Abstands- oder Distanzmaßen und Ähnlichkeitsmaßen kann gemessen werden, wie stark sich Fälle untereinander unterscheiden. Als Fälle werden im besprochenen Beispiel stets die Einzugsgebiete angesehen. Wie stark variiert aber die gemessene Unterschiedlichkeit, wenn verschiedene Distanzmaße und Ähnlichkeitsmaße verwendet werden?

Zunächst wieder ein Beispiel. Das einfachste Distanzmaß ist die Blockdistanz oder Manhattan-Distanz. Sie berechnet sich als

$$D(x,y) = \sum \left| X_i - Y_i \right| \quad \text{(Gl. 1: Block- oder Manhattan-Distanz)}$$

und kann als „rechtwinkliger" Weg zwischen Werten zweier unterschiedlichen Fälle aufgefasst werden, ähnlich der Wegstrecke in einem typischen nordamerikanischen Straßennetz. Bild 3.4 illustriert dies.

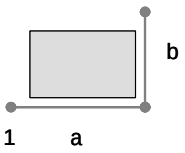



*Bild 3.4: Block- oder Manhattan-Distanz*

■ Übungsbeispiel: Berechnen Sie die Blockdistanz aller Einzugsgebiete der Ostsee gegenüber der Daugava! Verwenden Sie dabei zunächst nur die Variablen Abflussspende (auf die Fläche bezogener Abfluss) sowie die Anteile verschiedener Landnutzungen in den jeweiligen Einzugsgebieten!

GGA_Clusteranalyse\Uebung_2.xls

Arbeitsschritte: In die Zelle L4 wird der Abstand zwischen der Abflussspende in den Einzugsgebieten der Aa und der Daugava eingetragen. Dazu wird folgende Formel verwendet: `=ABS(B4-B$12)`. Das $-Zeichen bewirkt, dass immer die Zeile 12, also hier die Zeile der Daugava, subtrahiert wird. Dies wird beim Erweitern der Formeln noch nützlich sein. Anschließend werden diese Abstände für alle Variablen berechnet. In der Spalte Gesamtabstand (Spalte U) werden schließlich alle diese Einzelabstände summiert, so wie es auch Gl. 1 erfordert.

Somit wurde mit der Blockdistanz ein Maß für die Unterschiedlichkeit aller Einzugsgebiete gegenüber der Daugava berechnet, der Abstand der Daugava zu sich selbst ist konsequenterweise 0. So gehen auch die gängigen Statistikprogrammpakete vor. Hier wird später lediglich eine Ähnlichkeitstabelle erstellt, wo jeder Fall (jedes Gebiet) mit jedem hinsichtlich seiner Ähnlichkeit bewertet wird.

Wir wollen nun die Abstände der Abflussspende mit den summierten Gesamtabständen vergleichen (Bild 3.4)! Sofort fällt auf, dass sich die Blockdistanzen in beiden Reihen (Spalte `L` und Spalte `U` in Bild 3.4) fast gleich verhalten, die Abflussspende also die Unterschiede in den Variablen der Landnutzung überdeckt. Dies ist auch nicht weiter verwunderlich, wegen der hohen Werte der Abflussspende gegenüber den jeweiligen Anteilen der Landnutzung dominieren ja auch die Differenzen zwischen den Abflussspenden den Gesamtabstand. Damit fällt eine unabdingbare Vorraussetzung zur Berechnung von Distanz- und Ähnlichkeitsmaßen auf: Variablen müssen vor der Berechnung von

Distanz- und Ähnlichkeitsmaßen standardisiert, d.h. sie müssen auf ein einheitliches Niveau angeglichen werden!

| | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Distanzmaß: Block-Distanz** | | | | | | | | | |
| eil utzte dfläche | | Distanz von Daugava | | mit | $D(x,y) = \sum \lvert X_i - Y_i \rvert$ | | | | Absolutwerte der Distanzen | | Gesamt-abstand |
| ren | | $A_F$ | $A_{Wald}$ | $A_{Acker}$ | $A_{Weide}$ | $A_{Stadt}$ | $A_{Wasser}$ | $A_{Gletscher}$ | $A_{Tundra}$ | $A_{Offen}$ | **Σ** |
| 0,13 | | 7,9710 | 0,2134 | 0,0718 | 0,0454 | 0,0000 | 0,0161 | 0,0000 | 0,0000 | 0,0807 | **8,3984** |
| 0,09 | | 38,9069 | 0,5145 | 0,2617 | 0,1317 | 0,0049 | 0,0001 | 0,0000 | 0,0000 | 0,1174 | **39,9372** |
| 0,27 | | 676,5434 | 0,2599 | 0,2566 | 0,1312 | 0,0022 | 0,0127 | 0,0000 | 0,0000 | 0,0663 | **677,2724** |
| 0,20 | | 319,2802 | 0,2657 | 0,2518 | 0,1302 | 0,0027 | 0,0388 | 0,0000 | 0,0648 | 0,0066 | **320,0409** |
| 0,14 | | 4,5742 | 0,1043 | 0,0462 | 0,1212 | 0,0026 | 0,0109 | 0,0000 | 0,0000 | 0,0661 | **4,9255** |
| 0,00 | | 27,2092 | 0,4705 | 0,2057 | 0,1186 | 0,0051 | 0,0316 | 0,0000 | 0,0000 | 0,2042 | **28,2449** |
| 0,15 | | 155,1480 | 0,3957 | 0,2416 | 0,1303 | 0,0004 | 0,0209 | 0,0000 | 0,0011 | 0,0587 | **155,9968** |
| 0,19 | | 86,0406 | 0,3113 | 0,3292 | 0,0889 | 0,0722 | 0,0137 | 0,0000 | 0,0000 | 0,0132 | **86,8690** |
| 0,21 | | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | **0,00** |
| 0,00 | | 107,1218 | 0,5156 | 0,2197 | 0,1175 | 0,0012 | 0,0294 | 0,0000 | 0,0000 | 0,2063 | **108,2115** |
| 0,07 | | 377,9629 | 0,4346 | 0,2323 | 0,1286 | 0,0551 | 0,0080 | 0,0000 | 0,0000 | 0,1362 | **378,9577** |
| 0,07 | | 87,6246 | 0,4350 | 0,2331 | 0,1302 | 0,0238 | 0,0325 | 0,0000 | 0,0000 | 0,1329 | **88,6120** |
| 0,07 | | 61,5536 | 0,5026 | 0,2504 | 0,1306 | 0,0009 | 0,0052 | 0,0000 | 0,0000 | 0,1367 | **62,5799** |
| 0,09 | | 47,0653 | 0,2868 | 0,0934 | 0,1027 | 0,0008 | 0,0284 | 0,0000 | 0,0000 | 0,1222 | **47,6996** |
| 0,11 | | 128,1883 | 0,4260 | 0,2317 | 0,1298 | 0,0039 | 0,0235 | 0,0000 | 0,0000 | 0,1013 | **129,1045** |
| 0,30 | | 284,3222 | 0,1572 | 0,2562 | 0,1296 | 0,0026 | 0,0463 | 0,0000 | 0,0775 | 0,0936 | **285,0852** |

*Bild 3. 4:. Notwendigkeit der Standardisierung von Variablen – das Blockmaß repräsentiert hier fast nur den Einfluss der Abflussspende (Uebung_2.xls)*

In der Lösungstabelle von Uebung_2.xls sind die Blockdistanzen dann auch mit einer standardisierten Abflussspende berechnet worden. Hier lässt sich der Einfluss der Standardisierung gut erkennen, alle Variablen gehen jetzt gleich in die Messung der Distanzen ein.

**Standardisierung von Variablen**

Variablen müssen vor der Berechnung von Distanz- und Ähnlichkeitsmaßen standardisiert werden! Standardisierte Variablen sind dimensionslos!

Die Standardisierung kann auf unterschiedlichem Wege geschehen. In den verwendeten Beispieldaten ist die unterschiedliche Landnutzung bereits als Anteil angegeben, die Daten damit auf eine Summe von 1 oder 100% standardisiert. Prozentuale Anteile müssen also nicht nochmals standardisiert werden.

Werte können weiterhin an der Summe oder gegenüber dem Maximum standardisiert werden. Bei der Standardisierung gegenüber dem Maximum wird der größte Wert mit 1 gewichtet, alle anderen liegen zwischen 0 und < 1. Die Standardisierung gegenüber der Summe bewirkt dagegen kleinere Werte, 1 wird nicht erreicht. Entsprechend geringer fallen auch die Abstände (Differenzen) aus. Prinzipiell lässt sich mit der Wahl der Standardisierungsmethode immer auch eine Gewichtung des Einflusses der betreffenden Variablen auf den Gesamtabstand vornehmen.

Die in der Statistik ebenfalls angewandte Normalstandardisierung wird im Kapitel zur Faktorenanalyse angesprochen, einen Überblick hierzu geben auch BAHRENBERG, GIESE & NIPPER, 1985 .

Neben der Blockdistanz ist das am häufigsten verwendete Distanzmaß die Quadrierte Euklidische Distanz. Während die Block- oder Manhattandistanz den Abstand zwischen zwei Datenpunkten quasi über Kathete und Ankathete misst, definiert bei der Quadrierten Euklidischen Distanz die Hypothenuse den Abstand.

$$D\ (x, y)\ =\ \sum_{i=1}^{n}\ (Y_i - X_{i)})^2 \quad \text{(Gl. 2: Quadrierte Euklidische Distanz)}$$

Als weiteres Distanzmaß kann noch die Tschebyscheff-Distanz genannt werden, hier werden wie bei der Blockdistanz wieder die Differenzen zwischen den Datenpunkten berechnet. Anders als bei der Blockdistanz werden deren Beträge dann aber nicht summiert, sondern lediglich der größte Abstand als Tschebyscheff-Distanz verwendet

$$D(x, y) = \underset{i}{Max}\left(\left|X_i - Y_i\right|\right) \qquad \text{(Gl. 3: Tschebyscheff-Distanz)}$$

Über weitere Distanzmaße kann man sich u.a. bei BAHRENBERG, GIESE & NIPPER, 1992, Kapitel 7, informieren.

Nicht als Distanz- sondern als Ähnlichkeitsmaß kann der Pearson´sche Korrelationskoeffizient angewendet werden. Wenngleich es ungewöhnlich erscheint, müssen auch bei Verwendung des Korrelationskoeffizienten die Variablen standardisiert werden. Andernfalls werden die Abweichungen in großen Variablen zu gering bewertet. Dies veranschaulicht am besten wieder ein kurzes Beispiel:

■ „100“ und „110“ gelten beide als deutliche Abweichungen von überwiegenden Werten um 0,9. Die Wertepaare (110; 1; 0,9) und (100; 0,9; 0,9) [r= 0,99] korrelieren damit aber enger als (1,1; 1; 0,9) und (1; 0,9; 0,9) [r=0,86], obwohl bei der Berechnung des zweiten Korrelationskoeffizienten die Werte 100 und 110 in beiden Fällen gleich durch 100 dividiert wurden!

Schließlich wollen wir uns anschauen, wie die gerade besprochenen Distanz- und Ähnlichkeitsmaße die Einzugsgebiete der Ostsee bewerten. Damit versuchen wir, deren Unterschiede zum Einzugsgebiet der Daugava zu messen. Jedes Einzugsgebiet wurde durch die Variablen Abflussspende, Anteil verschiedener Landnutzungen, Bevölkerungsdichte und Anteil städtischer/ländlicher Bevölkerung sowie Speicherausbaugrad charakterisiert. Abfluss und Bevölkerungsdichte mussten zuvor noch am Maximum standardisiert werden, die restlichen Variablen blieben als Anteile unstandardisiert.

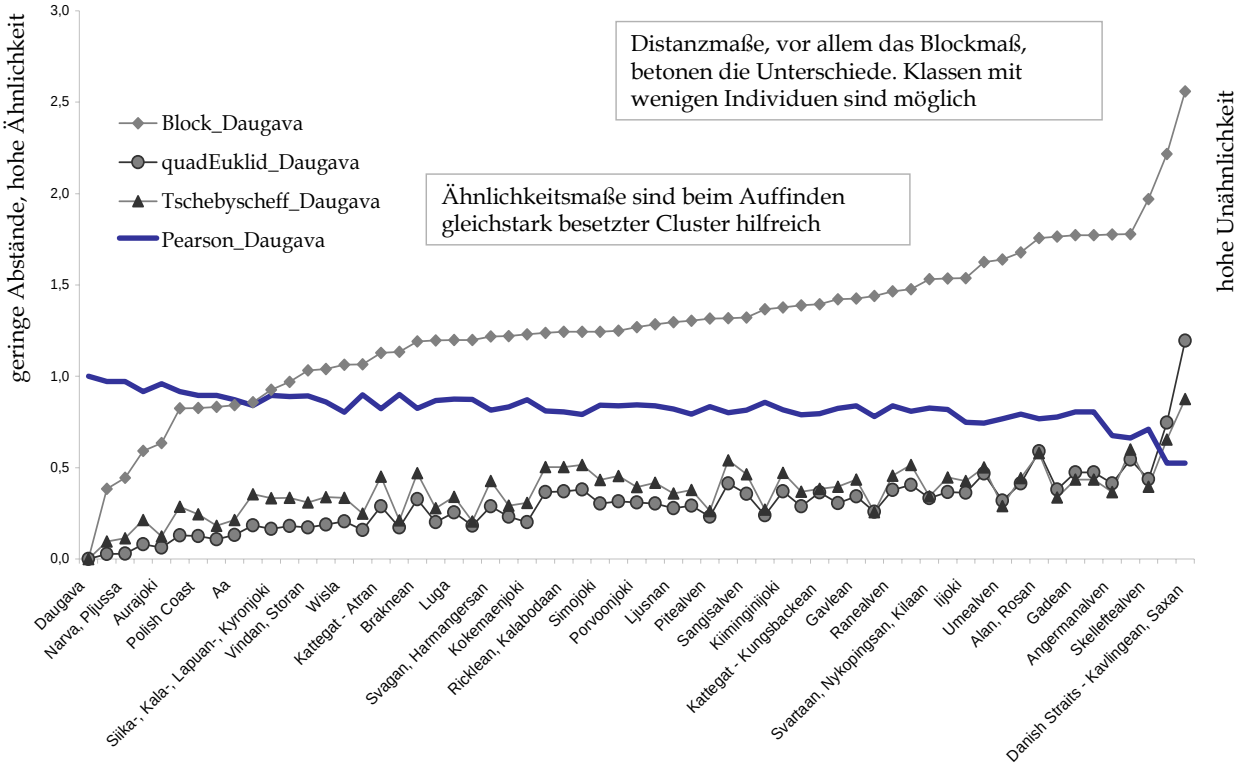



*Bild 3.5: Ähnlichkeiten oder Abstände der Einzugsgebiete der Ostsee gegenüber dem Einzugsgebiet der Daugava – Wirksamkeit verschiedener Distanz- und Ähnlichkeitsmaße*

In Bild 3.5 sind hierfür die Distanzen oder Ähnlichkeiten eines jeden Einzugsgebietes gegenüber der Daugava dargestellt. Die PEARSON-Ähnlichkeit läuft entgegengesetzt den Kurven der Distanzmaße. Ein hoher Korrelationskoeffizient bedeutet ja schließlich auch eine hohe Ähnlichkeit, während die Distanz dann eher gering ist. Deutlich fällt auf, daß vor allem die Block- oder Manhattandistanz die Unterschiede gegenüber der Daugava stark betont, es werden nur sehr wenige Einzugsgebiete mit einer geringen Distanz ausgeschieden. Beim PEARSON-Korrelationskoeffizient kann dagegen eine umgekehrte Tendenz beobachtet werden. Die Quadrierte Euklidische Distanz und das Tschebyscheff-Maß dagegen lassen am zuverlässigsten deutlich und weniger unterschiedliche Einzugsgebiete erkennen.

Wichtiger als die unterschiedlichen Distanz- und Ähnlichkeitsmaße sind bei einer erfolgreichen Klassifizierung aber die Linkage-Verfahren, also jene Vorgehensweise, nach denen Fälle von Daten aufgrund ihrer Abstände oder Ähnlichkeiten zueinander schließlich zusammengefasst werden. Diese Linkage-Verfahren werden im Folgenden dargestellt.

## 2.2 Die Distanzmatrix – Unterschiedlichkeit zwischen allen Fällen

Bevor sich die einzelnen Fälle zu Gruppen zusammenfassen lassen, muss zunächst in einer Distanzmatrix die Ähnlichkeit –oder Distanz eines jeden Falles mit den anderen Fällen ermittelt werden. Auch für unseren Datensatz soll die Ähnlichkeit oder die Unterschiedlichkeit von jedem Einzugsgebiet mit allen anderen berechnet werden.

■ Übungsbeispiel: Anhand der Distanzmatrix soll bewertet werden, welche Einzugsgebiete der Ostsee sich besonders ähnlich sind oder einen möglichst geringen Abstand zueinander aufweisen. Diese Einzugsgebiete sollen später zu einer Gruppe zusammengefasst werden. Wir verwenden nun das Programm SPSS.

GGA_Clusteranalyse\Uebung_3.sav

Öffnen Sie die Datei Uebung_3.sav in SPSS. Um sich mit der Methode der Clusteranalyse vertraut zu machen, soll zunächst die Distanzmatrix für alle Einzugsgebiete berechnet und interpretiert werden! Dazu wird die Clusteranalyse über das Menü Analysieren > Klassifizieren > Hierarchische Cluster… aufgerufen (Bild 3.6). Danach wird zunächst das Feld `Descript` (Bild 3.6, ①) mittels des Pfeils (Bild 3.6, ②) als Fallbeschriftung ausgewählt, damit lassen sich später in der Distanzmatrix die Werte auch wieder den Einzugsgebieten zuordnen. Ebenso werden alle links befindlichen Felder nach „Variable(n)“ übertragen, um sie für die Clusteranalyse zu verwenden. Danach wird mittels des Button Statistik (Bild 3.6, ③) noch die Distanzmatrix angefordert, ein „OK“ startet die Clusteranalyse und gibt im SPSS-Viewer unter anderem die Distanzmatrix (bei SPSS als als Näherungsmatrix bezeichnet) aus.

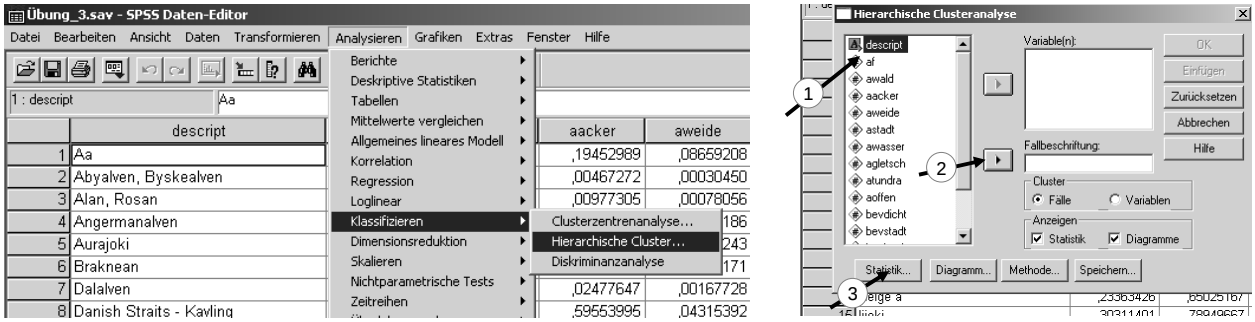
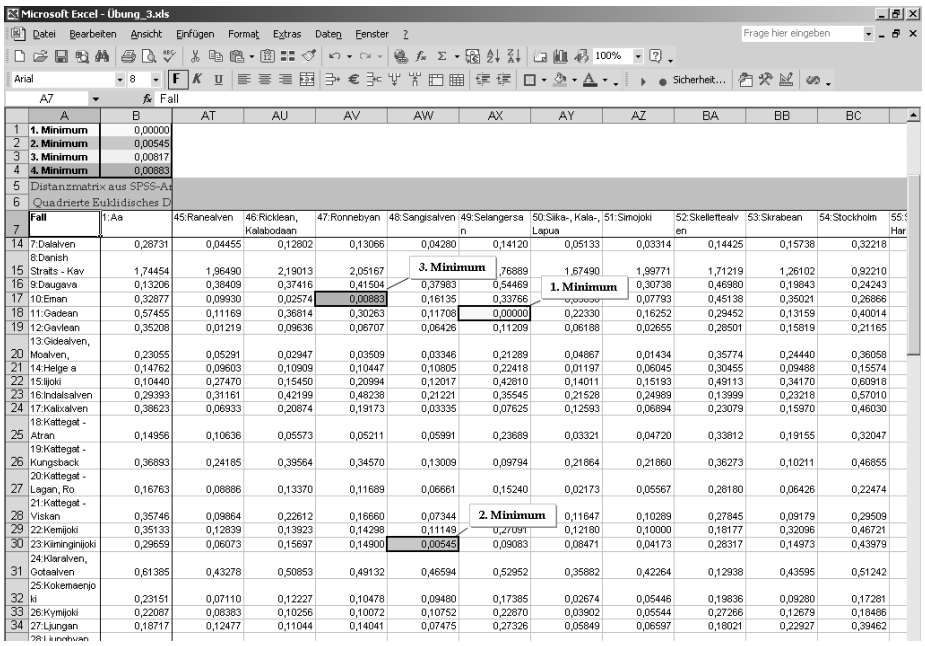



*Bild 3.6: Aufruf der Clusteranalyse in SPSS und Ausgeben einer Distanzmatrix*

Die Distanzen wurden zunächst als Quadrierte Euklidische Distanzen berechnet. Wie wir später noch sehen werden, lassen sich auch die anderen in Kap. 3.2 besprochenen Distanz- oder Ähnlichkeitsmaße verwenden. Welche Einzugsgebiete werden nun aber als kaum unterschiedlich ausgewiesen? Hierzu suchen wir in der Distanzmatrix nach den kleinsten Werten. Wir können dies entweder in SPSS selbst bewerkstelligen oder die Übungsdatei GGA_Clusteranalyse\Uebung_3.xls nutzen.

Hier ist die berechnete Distanzmatrix nochmals enthalten, die Suche nach den Minima gestaltet sich in EXCEL etwas einfacher (Bild 3.7).




| | A | B | AT | AU | AV | AW | AX | AY | AZ | BA | BB | BC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1. Minimum | 0,00000 | | | | | | | | | | |
| 2 | 2. Minimum | 0,00545 | | | | | | | | | | |
| 3 | 3. Minimum | 0,00817 | | | | | | | | | | |
| 4 | 4. Minimum | 0,00883 | | | | | | | | | | |
| 5 | Distanzmatrix aus SPSS-Ar | | | | | | | | | | | |
| 6 | Quadrierte Euklidisches D | | | | | | | | | | | |
| 7 | Fall | 1:Aa | 45:Ranealven | 46:Ricklean, Kalabodaan | 47:Ronnebyan | 48:Sangisalven | 49:Selangersan | 50:Sika-, Kala-, Lapua | 51:Simojoki | 52:Skellefteälven | 53:Skrabean | 54:Stockholm |
| 14 | 7:Dalalven | 0,28731 | 0,04455 | 0,12802 | 0,13066 | 0,04280 | 0,14120 | 0,05133 | 0,03314 | 0,14425 | 0,15738 | 0,32218 |
| 15 | 8:Danish Straits - Kav | 1,74454 | 1,96490 | 2,19013 | 2,05167 | 3. Minimum | ,76889 | 1,67490 | 1,99771 | 1,71219 | 1,26102 | 0,92210 |
| 16 | 9:Daugava | 0,13206 | 0,38409 | 0,37416 | 0,41504 | 0,37983 | 0,54469 | 1. Minimum | 0,30738 | 0,46980 | 0,19843 | 0,24243 |
| 17 | 10:Eman | 0,32877 | 0,09930 | 0,02574 | 0,00883 | 0,16135 | 0,33766 | | 0,07793 | 0,45138 | 0,35021 | 0,26866 |
| 18 | 11:Gadean | 0,57455 | 0,11169 | 0,36814 | 0,30263 | 0,11708 | 0,00000 | 0,22330 | 0,16252 | 0,29452 | 0,13159 | 0,40014 |
| 19 | 12:Gavlean | 0,35208 | 0,01219 | 0,09636 | 0,06707 | 0,06426 | 0,11209 | 0,06188 | 0,02655 | 0,28501 | 0,15819 | 0,21165 |
| 20 | 13:Gideälven, Moälven, | 0,23055 | 0,05291 | 0,02947 | 0,03509 | 0,03346 | 0,21289 | 0,04867 | 0,01434 | 0,35774 | 0,24440 | 0,36058 |
| 21 | 14:Helge a | 0,14762 | 0,09603 | 0,10909 | 0,10447 | 0,10805 | 0,22418 | 0,01197 | 0,06045 | 0,30455 | 0,09488 | 0,15574 |
| 22 | 15:Iijoki | 0,10440 | 0,27470 | 0,15450 | 0,20994 | 0,12017 | 0,42810 | 0,14011 | 0,15193 | 0,49113 | 0,34170 | 0,60918 |
| 23 | 16:Indalsalven | 0,29393 | 0,31161 | 0,42199 | 0,48238 | 0,21221 | 0,35545 | 0,21528 | 0,24989 | 0,13999 | 0,23218 | 0,57010 |
| 24 | 17:Kalixalven | 0,38623 | 0,06933 | 0,20874 | 0,19173 | 0,03335 | 0,07625 | 0,12593 | 0,06894 | 0,23079 | 0,15970 | 0,46030 |
| 25 | 18:Kattegat - Atran | 0,14956 | 0,10636 | 0,05573 | 0,05211 | 0,05991 | 0,23689 | 0,03321 | 0,04720 | 0,33812 | 0,19155 | 0,32047 |
| 26 | 19:Kattegat - Kungsback | 0,36893 | 0,24185 | 0,39564 | 0,34570 | 0,13009 | 0,09794 | 0,21864 | 0,21860 | 0,36273 | 0,10211 | 0,46855 |
| 27 | 20:Kattegat - Lagan, Ro | 0,16763 | 0,08886 | 0,13370 | 0,11689 | 0,06661 | 0,15240 | 0,02173 | 0,05567 | 0,28180 | 0,06426 | 0,22474 |
| 28 | 21:Kattegat - Viskan | 0,35746 | 0,09864 | 0,22612 | 0,16660 | 0,07344 | 2. Minimum | 0,11647 | 0,10289 | 0,27845 | 0,09179 | 0,29509 |
| 29 | 22:Kemijoki | 0,35133 | 0,12839 | 0,13923 | 0,14298 | 0,11149 | 0,27091 | 0,12180 | 0,10000 | 0,18177 | 0,32096 | 0,46721 |
| 30 | 23:Kiminginijoki | 0,29659 | 0,06073 | 0,15697 | 0,14900 | 0,00545 | 0,09083 | 0,08471 | 0,04173 | 0,28317 | 0,14973 | 0,43979 |
| 31 | 24:Klaralven, Gotaalven | 0,61385 | 0,43278 | 0,50853 | 0,49132 | 0,46594 | 0,52952 | 0,35882 | 0,42264 | 0,12938 | 0,43595 | 0,51242 |
| 32 | 25:Kokemaenjoki | 0,23151 | 0,07110 | 0,12227 | 0,10478 | 0,09480 | 0,17385 | 0,02674 | 0,05446 | 0,19836 | 0,09280 | 0,17281 |
| 33 | 26:Kymijoki | 0,22087 | 0,08383 | 0,10256 | 0,10072 | 0,10752 | 0,22870 | 0,03902 | 0,05544 | 0,27266 | 0,12679 | 0,18486 |
| 34 | 27:Ljungan | 0,18717 | 0,12477 | 0,11044 | 0,14041 | 0,07475 | 0,27326 | 0,05849 | 0,06597 | 0,18021 | 0,22927 | 0,39462 |

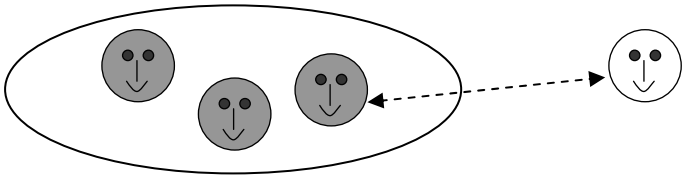
*Bild 3.7: Distanzmatrix – welches sind die Einzugsgebiete mit dem geringsten Abstand zueinander?*

Die beiden ähnlichsten Einzugsgebiete scheinen Selangersån und Gadeån zu sein. Dabei weist der Distanzwert von 0 allerdings darauf hin, dass die Merkmale in diesen Einzugsgebieten identisch sind, eine Unstimmigkeit im verwendeten Datensatz. Wirklich aussagefähig ist in unserem Beispiel erst das 2. Minimum, hier wird ersichtlich, dass die Werte von Ljungbyån/Alsterån und Emån kaum nennenswerte Unterschiede aufweisen, also zu einer Gruppe zusammengefasst werden können. Als nächste Einzugsgebiete ließen sich Ronnebyån und Emån wählen, auch diese weisen lediglich geringe Unterschiede bezüglich der betrachteten Variablen auf.

# 3 Zusammenfassen zu Gruppen – Die Linkage-Verfahren der Clusteranalyse

Mit Hilfe der Distanzmatrix können also Fälle oder hier Einzugsgebiete erkannt werden, welche nur geringe Unterschiede in den ihnen zugeordneten Werten aufweisen. Solche Einzugsgebiete können zu einer Gruppe zusammengefasst werden. Bislang funktioniert das jedoch nur, wenn zwei Einzugsgebiete zusammengefasst werden. Diese neu entstandene Gruppe müsste nun eigentlich auch danach bewertet werden, wie deutlich sie sich gegenüber weiteren Einzugsgebieten (oder Fällen) unterscheidet. Um zu beurteilen, wie stark sich also eine solche Gruppe von einem weiteren Einzugsgebiet unterscheidet, könnte der kleinste Abstand zwischen den Distanzwerten in der Gruppe und dem interessierenden Einzugsgebiet genutzt werden (Bild 3.8).

Genau nach einem solchen oder ähnlichen Verfahren arbeiten die Linkage-Verfahren der Clusteranalyse. Hier wird entschieden, ob zwei Fälle (oder in unserem Beispiel Einzugsgebiete) zusammengefasst werden sollen, weil sie zueinander nur einen geringen Abstand aufweisen, oder ob eine bestehende Gruppe um ein neues Einzugsgebiet (oder einen neuen Fall) vergrößert werden soll. Noch mehr als bei der Verwendung unterschiedlicher Distanz – oder Ähnlichkeitsmaße ändern sich die Ergebnisse einer Clusteranalyse, wenn unterschiedliche Linkage-Verfahren genutzt werden. Gleichzeitig dienen diese unterschiedlichen Linkage-Verfahren aber dazu, eine den jeweiligen Daten entsprechende sinnvolle Gruppierung zu finden. Wie dies funktioniert und wie die einzelnen Ergebnisse einer Clusteranalyse dann zu interpretieren sind, wollen wir wieder am Beispiel unseres Datensatzes der Einzugsgebiete der Ostsee demonstrieren. Hierbei lernen wir auch zu entscheiden, wie viele Klassen wir sinnvollerweise bilden sollten.




.

*Bild 3.8: Nearest Neighboor Ansatz für den Abstand zu einer Gruppe*

GGA_Clusteranalyse\Uebung_3.sav – Fortsetzung ■ Wieder wird eine Clusteranalyse in SPSS analog Bild 3.6 durchgeführt und wir wählen unter Methode (dritte Schaltfläche unten, Bild 3.6 rechts) die Clustermethode „Nächstgelegener Nachbar“ aus. Zusätzlich fordern wir über die Schaltfläche Diagramm (zweite Schaltfläche von links) noch ein Dendrogramm an (Häkchen bei „Dendrogramm“ setzen). Wir bekommen abermals zahlreiche Angaben im Ausgabefenster angezeigt. Neben der Distanzmatrix interessieren uns nun besonders die Zuordnungsmatrix und das Dendrogramm.

Das Dendrogramm wie auch die Zuordnungsübersicht illustrieren den Prozess der Klassenbildung, bis schließlich alle Fälle soweit gruppiert worden sind, dass lediglich noch zwei Klassen vorhanden sind. Wenngleich zwei Klassen meist eine zu starke Zusammenfassung darstellen, lässt sich über diesen Prozess der Klassenbildung recht gut erkennen, wie stark die Daten aggregiert werden sollten.

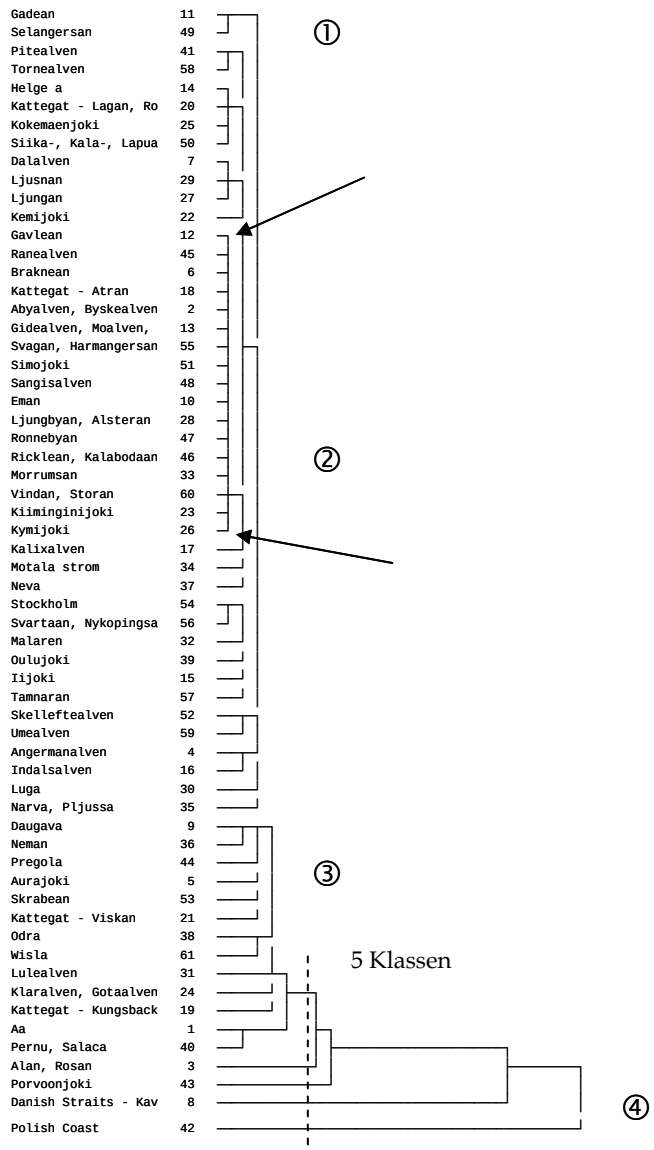



*Bild 3.9: Interpretation eines Dendrogramms – siehe Text (Distanzmaß: Quad. Euklid. Distanz, Nearest Neighboor -Linkage)*

Werfen wir zunächst einen Blick auf das Dendrogramm (Bild 3.9). Dieses ist von links nach rechts zu lesen und zeigt, wie Einzugsgebiete mit geringen Abständen zwischen den Variablenwerten zusammengefasst werden. Selangersån und Gadeån, Ljungbyån-Alsterån und Emån sowie Ronnebyån und Emån sind jene Einzugsgebiete, die bereits bei der Auswertung der Distanzmatrix als einander recht gleichartig aufgefallen waren. Sie werden bei der Klassenbildung in der Clusteranalyse frühzeitig zusammengefasst. Genau diese Aussage lässt sich auch dem Dendrogramm entnehmen. Die Einzugsgebiete im oberen Bereich der Abbildung weisen nur geringe Abstände in den charakterisierenden Variablen auf.

① Selangersån und Gadeån sind im ersten Schritt zusammengefasst worden, zu einer großen Gruppe sind zudem auch die Einzugsgebiete im oberen Drittel (② - Pfeile) des Dendrogramms vereinigt worden, unter anderem Ljungbyån/Alsterån und Emån sowie Ronnebyån und Emån. Sie wurden aber nicht nur paarweise, sondern bereits folgend über das entsprechende Linkage-Verfahren (hier Nächster Nachbar Linkage) in einer Ebene zusammengefasst.

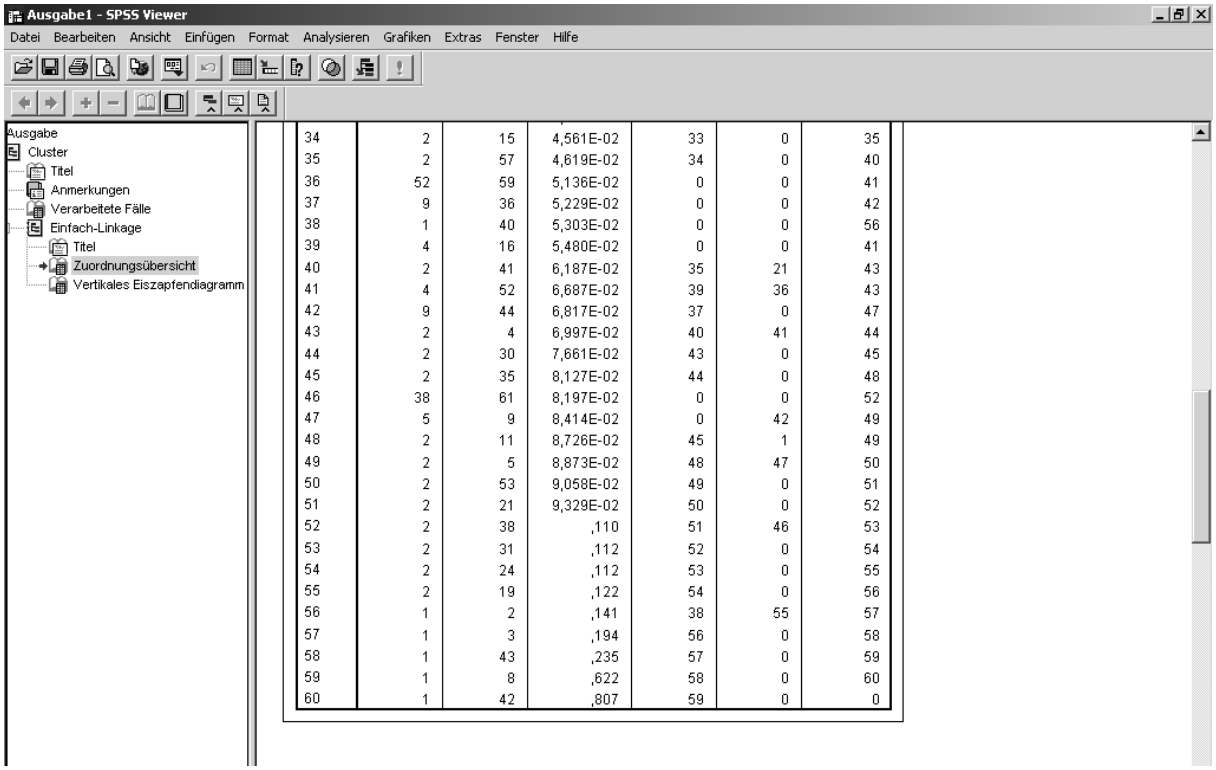
③ Einzugsgebiete wie Eman, Pregola und Odra unterscheiden sich von den erstgenannten Einzugsgebieten dagegen deutlich, können untereinander aber zusammengefasst werden.

④ Schließlich sind am Ende des Dendrogramms jene Einzugsgebiete zu finden, welche sich deutlich von allen andern unterscheiden und wie im Fall Polish Coast und Danish Straits - Kavlingeån, Saxån keiner Gruppe mehr zuzuordnen waren.

Bei spätestens ④ wird dem Betrachter des Dendrogramms relativ leicht einsichtig, an welchem Punkt abgebrochen werden sollte, um weitere Klassen zusammenzufassen. Die

waagerechten Linien im Dendrogramm repräsentieren die Abstände eines Falles zu den jeweils bereits zusammengefassten Gruppen. Wenn diese Abstände immer größer werden, werden also immer deutlicher inhomogene Fälle hinzugefügt.

Genau dies lässt sich übrigens auch aus der Zuordnungsübersicht (Bild 3.10), welche von SPSS ausgegeben wird, ablesen.




| | | | | | | |
|---|---|---|---|---|---|---|
| 34 | 2 | 15 | 4,561E-02 | 33 | 0 | 35 |
| 35 | 2 | 57 | 4,619E-02 | 34 | 0 | 40 |
| 36 | 52 | 59 | 5,136E-02 | 0 | 0 | 41 |
| 37 | 9 | 36 | 5,229E-02 | 0 | 0 | 42 |
| 38 | 1 | 40 | 5,303E-02 | 0 | 0 | 56 |
| 39 | 4 | 16 | 5,480E-02 | 0 | 0 | 41 |
| 40 | 2 | 41 | 6,187E-02 | 35 | 21 | 43 |
| 41 | 4 | 52 | 6,687E-02 | 39 | 36 | 43 |
| 42 | 9 | 44 | 6,817E-02 | 37 | 0 | 47 |
| 43 | 2 | 4 | 6,997E-02 | 40 | 41 | 44 |
| 44 | 2 | 30 | 7,661E-02 | 43 | 0 | 45 |
| 45 | 2 | 35 | 8,127E-02 | 44 | 0 | 48 |
| 46 | 38 | 61 | 8,197E-02 | 0 | 0 | 52 |
| 47 | 5 | 9 | 8,414E-02 | 0 | 42 | 49 |
| 48 | 2 | 11 | 8,726E-02 | 45 | 1 | 49 |
| 49 | 2 | 5 | 8,873E-02 | 48 | 47 | 50 |
| 50 | 2 | 53 | 9,058E-02 | 49 | 0 | 51 |
| 51 | 2 | 21 | 9,329E-02 | 50 | 0 | 52 |
| 52 | 2 | 38 | ,110 | 51 | 46 | 53 |
| 53 | 2 | 31 | ,112 | 52 | 0 | 54 |
| 54 | 2 | 24 | ,112 | 53 | 0 | 55 |
| 55 | 2 | 19 | ,122 | 54 | 0 | 56 |
| 56 | 1 | 2 | ,141 | 38 | 55 | 57 |
| 57 | 1 | 3 | ,194 | 56 | 0 | 58 |
| 58 | 1 | 43 | ,235 | 57 | 0 | 59 |
| 59 | 1 | 8 | ,622 | 58 | 0 | 60 |
| 60 | 1 | 42 | ,807 | 59 | 0 | 0 |

*Bild 3.10: Ausgabe der Zuordnungsübersicht bei SPSS*

In der mittleren Spalte wird dabei jener Abstand tabelliert, welcher den gerade mit in die Zusammenfassung einbezogenen Fall von den bereits zusammengefassten Fällen trennte. Spätestens wenn dieser Abstand sprunghaft ansteigt, sollte also ein Fall nicht mehr mit weiteren Fällen zusammengefasst werden.

In unserem Datenbeispiel wäre zwischen dem drittletzten und dem vorletzten Schritt ein solcher Sprung zu beobachten. Um dies zu verdeutlichen, sind die letzten 12 Abstände aus der SPSS - Zuordnungsübersicht in Bild 3.11 als Säulendiagramm dargestellt worden. Wir haben diesen Sprung bereits im Dendrogramm beobachten können (Bild 3.9 - ④, Bild 3.11 – großer Pfeil). Mit den Einzugsgebieten von Polish Coast und Danish Straits - Kavlingeån, Saxån wird die bis dahin zusammengefasste Gruppe der übrigen Einzugsgebiete der Ostsee extrem inhomogen.

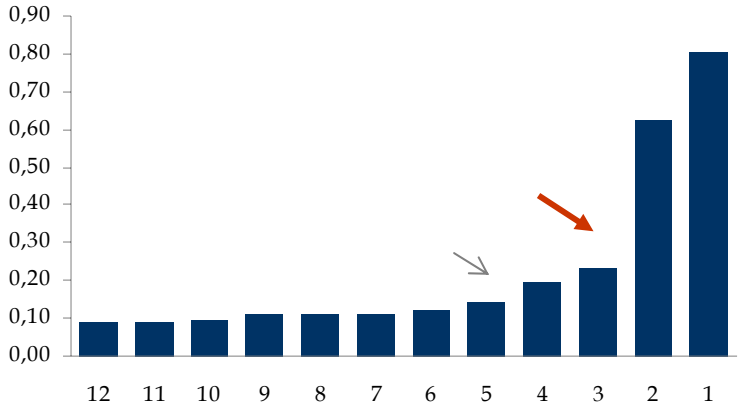



*Bild 3.11: Die letzten 12 Abstände der Zuordnungsübersicht (Distanzmaß: Quad. Euklid. Distanz, Nearest Neighboor -Linkage). Die Pfeile zeigen"Sprünge" in den Abständen bei der Zusammenfassung zu Klassen. (siehe Text)*

Wir sollten also mindestens 3 Klassen ausweisen, in jeweils einer sind dann Polish Coast und Danish Straits -Kavlingeån, Saxån als „Ausreißer" enthalten, die dritte fasst alle weiteren Einzugsgebiete als Fälle zusammen. Eine sicherlich noch nicht befriedigende Lösung, sollte das Ziel der Klassenbildung mit Hilfe der Clusteranalyse doch folgendes sein:

**Clusteranalyse – Welche Anzahl von Klassen ist sinnvoll?**

- Datenreduktion auf wenige Klassen!
- Klassen sollten etwa gleichstark besetzt sein!
- Eine sinnvolle Unterteilung soll nicht erst durch eine hohe Anzahl von Klassen erreicht werden!

Zunächst wollen wir einen weiteren Blick auf die Abstände in der Zuordnungsübersicht werfen. Zwischen der fünften und vierten Stufe der Zusammenfassung zu Klassen lässt sich ein weiterer „Sprung" beobachten (Bild 3.11 – kleiner Pfeil), welcher ebenfalls wieder im Dendrogramm auffällt. Auf der Höhe Klarälv/Götaälv (Bild 3.9) verzweigt sich das Dendrogramm, so dass Porvoonjoki und Alån, Roxån als ebenfalls einzeln besetzte „Klassen" abgetrennt werden können. Alle weiteren Einzugsgebiete sind aber immer noch in einer Klasse vereinigt. Wir haben hier eine typische „Kettenbildung" vor uns.

**Kettenbildung bei der Clusteranalyse**

Die „Kettenbildung“ ist besonders stark für die Clustermethode „Nächstgelegener Nachbar“ kennzeichnend. Trotz des Zusammenfassens entstehen keine klar zu trennenden Gruppen, gegen Ende werden allerdings stark abweichende Fälle erkennbar. Typische Dendrogramme ähneln dem aus Bild 3.9.

Eine Kettenbildung kann allerdings helfen, „Ausreißer“ zu erkennen, die bei anderen Verfahren der Gruppenzusammenfassung möglicherweise mit auf verschiedene Klassen verteilt worden wären.

# 4 Anwendung verschiedener Linkage-Verfahren

Um die verschieden Linkage-Methoden und die durch sie vorgenommene Unterteilung der Flusseinzugsgebiete der Ostsee näher untersuchen zu können, findet sich auf der beigelegten CD-ROM ein EXCEL-Arbeitsblatt zur vereinfachten Auswertung von Clusteranalysen aus SPSS.

GGA_Clusteranalyse\Arbeitsblatt_1.xls

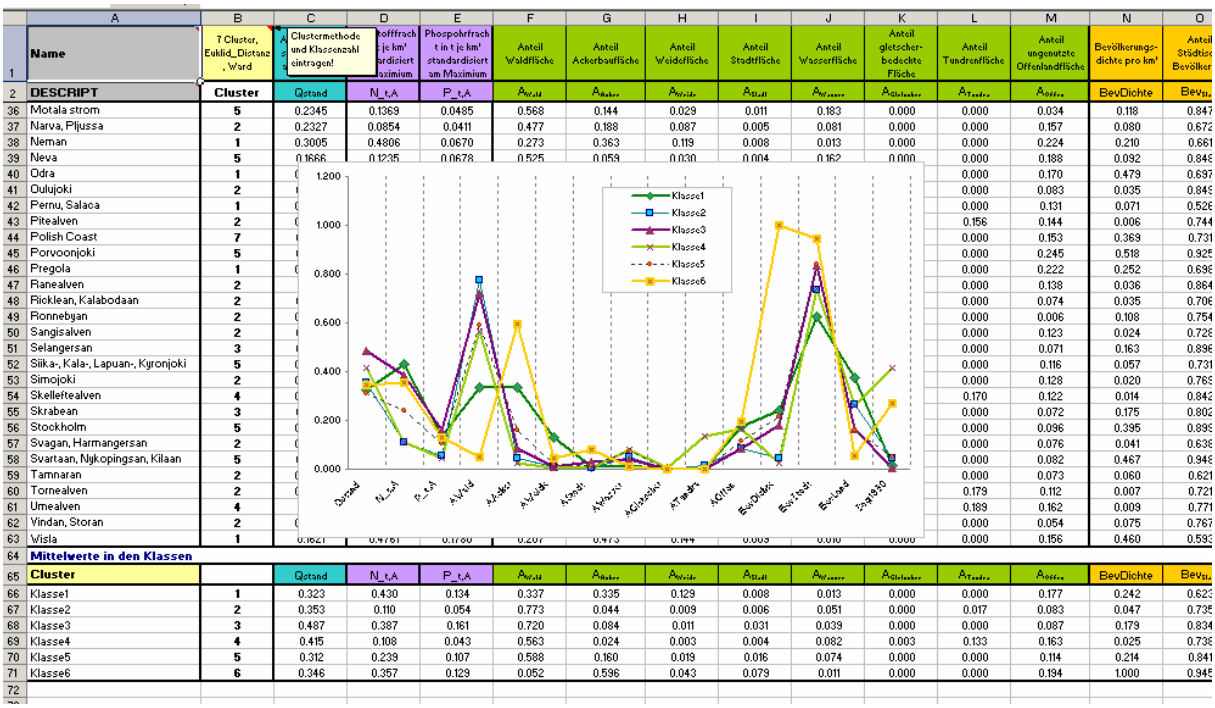
■ Öffnen Sie die Datei Uebung_3.sav in SPSS und führen wiederum, wie im dritten Kapitel beschrieben, eine Clusteranalyse durch. Zunächst nutzen wir bei den Methoden (Button Methode… in Bild 3.6) wieder die Voreinstellungen. SPSS führt hierbei (wie im vorigen Kapitel) eine Clusteranalyse mit der Quadrierten Euklidischen Distanz und dem Nächste-Nachbar Linkage aus. Anhand unserer Kenntnisse der Interpretation des Dendrogramms und der letzten Abstände in der Zuordnungsmatrix aus dem vorigen Kapitel, legen wir eine Zahl zu wählender Klassen (Cluster)! Hier hatten wir festgestellt, dass entweder 3 oder auch 5 Klassen möglich sein könnten. Die letzten Werte der Zuordnungsmatrix legen die Wahl von 3 Cluster nahe (roter Pfeil in Bild 3.11) Wir wollen also 3 Cluster auswählen. (HINWEIS: In Bild 3.9 ist der „Sprung“ im Dendrogramm allerdings so gelegen, dass wir uns mit Hilfe der gestrichelten Linie auch für 5 Klassen entscheiden könnten.)

Hierzu betätigen wir in SPSS im Dialog zur Clusteranalyse ( wieder wie in Bild 3.6 ) den rechtesten Button für „Speichern“. Hier wählen wir eine einzelne Lösung von 3 Klassen. OK bestätigt unsere Wahl SPSS klassifiziert neu. Im Datenblatteditor von SPSS findet sich nun ganz rechts von unseren Variablen eine Spalte mit der Bezeichnung „clu_3“. Hier ist nun eingetragen worden, zu welchen Klassen unsere einzelnen Fälle (einzelne Einzugsgebiete) zugeordnet worden sind. Wir kopieren diese Spalte aus SPSS in die Datei

GGA_Clusteranalyse\Arbeitsblatt_1.xls

in die Spalte B (mit der Bezeichnung „Cluster“). Das Tabellenblatt ist so programmiert, das im unteren Tabellenteil die mittleren Werte für jedes Cluster und jede Variable berechnet werden. Anschließend werden die Werte für jedes Cluster als Linienverlauf im ebenfalls enthaltenen Diagramm dargestellt.

Mit dessen Hilfe können die einzelnen Cluster inhaltlich interpretiert werden. Hierzu lässt sich überblicksartig feststellen, wie sich beispielsweise die mittleren Waldanteile oder die mittleren Ackerflächenanteile in einem einzelnen Cluster untereinander bzw. im Vergleich zu anderen Clustern verhalten.



| | Name | 7 Cluster, Euklid_Distanz, Ward | Clustermethode und Klassenzahl eintragen! | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | DESCRIPT | Cluster | Qotsnd | N_t,A | P_t,A | Anteil Waldfläche | Anteil Ackerbaufläche | Anteil Weidefläche | Anteil Stadtfläche | Anteil Wasserfläche | Anteil gletscherbedeckte Fläche | Anteil Tundrenfläche | Anteil ungenutzte Offenlandfläche | Bevölkerungsdichte pro km² | Anteil Städtische Bevölker |
| 36 | Motala strom | 5 | 0.2345 | 0.1369 | 0.0485 | 0.568 | 0.144 | 0.029 | 0.011 | 0.183 | 0.000 | 0.000 | 0.034 | 0.118 | 0.847 |
| 37 | Narva, Pljussa | 2 | 0.2327 | 0.0854 | 0.0411 | 0.477 | 0.188 | 0.087 | 0.005 | 0.081 | 0.000 | 0.000 | 0.157 | 0.080 | 0.672 |
| 38 | Neman | 1 | 0.3005 | 0.4806 | 0.0670 | 0.273 | 0.363 | 0.119 | 0.008 | 0.013 | 0.000 | 0.000 | 0.224 | 0.210 | 0.661 |
| 39 | Neva | 5 | 0.1666 | 0.1235 | 0.0678 | 0.525 | 0.059 | 0.030 | 0.004 | 0.162 | 0.000 | 0.000 | 0.188 | 0.092 | 0.848 |
| 40 | Odra | 1 | | | | | | | | | | 0.000 | 0.170 | 0.479 | 0.697 |
| 41 | Oulujoki | 2 | | | | | | | | | | 0.000 | 0.083 | 0.035 | 0.849 |
| 42 | Pernu, Salaca | 1 | | | | | | | | | | 0.000 | 0.131 | 0.071 | 0.526 |
| 43 | Pitealven | 2 | | | | | | | | | | 0.156 | 0.144 | 0.006 | 0.744 |
| 44 | Polish Coast | 7 | | | | | | | | | | 0.000 | 0.153 | 0.369 | 0.731 |
| 45 | Porvoonjoki | 5 | | | | | | | | | | 0.000 | 0.245 | 0.518 | 0.925 |
| 46 | Pregola | 1 | | | | | | | | | | 0.000 | 0.222 | 0.252 | 0.696 |
| 47 | Ranealven | 2 | | | | | | | | | | 0.000 | 0.138 | 0.036 | 0.864 |
| 48 | Ricklean, Kalabodaan | 2 | | | | | | | | | | 0.000 | 0.074 | 0.035 | 0.706 |
| 49 | Ronnebyan | 2 | | | | | | | | | | 0.000 | 0.006 | 0.108 | 0.754 |
| 50 | Sangisalven | 2 | | | | | | | | | | 0.000 | 0.123 | 0.024 | 0.728 |
| 51 | Selangersan | 3 | | | | | | | | | | 0.000 | 0.071 | 0.163 | 0.896 |
| 52 | Siika-, Kala-, Lapuan-, Kyronjoki | 5 | | | | | | | | | | 0.000 | 0.116 | 0.057 | 0.731 |
| 53 | Simojoki | 2 | | | | | | | | | | 0.000 | 0.128 | 0.020 | 0.769 |
| 54 | Skellefteälven | 4 | | | | | | | | | | 0.170 | 0.122 | 0.014 | 0.842 |
| 55 | Skrabean | 3 | | | | | | | | | | 0.000 | 0.072 | 0.175 | 0.802 |
| 56 | Stockholm | 5 | | | | | | | | | | 0.000 | 0.096 | 0.395 | 0.899 |
| 57 | Svagan, Harmangersan | 2 | | | | | | | | | | 0.000 | 0.076 | 0.041 | 0.638 |
| 58 | Svartaan, Nykopingsan, Kilaan | 5 | | | | | | | | | | 0.000 | 0.082 | 0.467 | 0.948 |
| 59 | Tamnaran | 2 | | | | | | | | | | 0.000 | 0.073 | 0.060 | 0.621 |
| 60 | Tornealven | 2 | | | | | | | | | | 0.179 | 0.112 | 0.007 | 0.721 |
| 61 | Umealven | 4 | | | | | | | | | | 0.189 | 0.162 | 0.009 | 0.771 |
| 62 | Vindan, Storan | 2 | | | | | | | | | | 0.000 | 0.054 | 0.075 | 0.767 |
| 63 | Wisla | 1 | 0.1621 | 0.4761 | 0.1780 | 0.207 | 0.473 | 0.144 | 0.003 | 0.010 | 0.000 | 0.000 | 0.156 | 0.460 | 0.593 |
| 64 | **Mittelwerte in den Klassen** | | | | | | | | | | | | | | |
| 65 | Cluster | | Qotsnd | N_t,A | P_t,A | Anteil Waldfläche | Anteil Ackerbaufläche | Anteil Weidefläche | Anteil Stadtfläche | Anteil Wasserfläche | Anteil gletscherbedeckte Fläche | Anteil Tundrenfläche | Anteil ungenutzte Offenlandfläche | BevDichte | Bevst |
| 66 | Klasse1 | 1 | 0.323 | 0.430 | 0.134 | 0.337 | 0.335 | 0.129 | 0.008 | 0.013 | 0.000 | 0.000 | 0.177 | 0.242 | 0.623 |
| 67 | Klasse2 | 2 | 0.353 | 0.110 | 0.054 | 0.773 | 0.044 | 0.009 | 0.006 | 0.051 | 0.000 | 0.017 | 0.083 | 0.047 | 0.735 |
| 68 | Klasse3 | 3 | 0.487 | 0.387 | 0.161 | 0.720 | 0.084 | 0.011 | 0.031 | 0.039 | 0.000 | 0.000 | 0.087 | 0.179 | 0.834 |
| 69 | Klasse4 | 4 | 0.415 | 0.108 | 0.043 | 0.563 | 0.024 | 0.003 | 0.004 | 0.082 | 0.003 | 0.133 | 0.163 | 0.025 | 0.736 |
| 70 | Klasse5 | 5 | 0.312 | 0.239 | 0.107 | 0.588 | 0.160 | 0.019 | 0.016 | 0.074 | 0.000 | 0.000 | 0.114 | 0.214 | 0.841 |
| 71 | Klasse6 | 6 | 0.346 | 0.357 | 0.129 | 0.052 | 0.596 | 0.043 | 0.079 | 0.011 | 0.000 | 0.000 | 0.194 | 1.000 | 0.945 |
| 72 | | | | | | | | | | | | | | | |



*Bild 3.12: Arbeitsblatt_1.xls zur Auswertung der Clusterlösung aus SPSS. Es werden die Mittelwerte der einzelnen Variablen für die jeweiligen Cluster berechnet und als Liniendiagramm dargestellt.*

In Bild 3.12 sind damit mittlere Variablenwerte für das Nächster-Nachbar-Linkage angegeben. Wegen der Kettenbildung sind alle Einzugsgebiete zu einem Cluster zusammengefasst worden (siehe auch noch einmal Bild 3.9). Lediglich Polish Coast und Danish Straits - Kavlingeån, Saxån sind als Ausreißereinzugsgebiete abgetrennt. Die Klassen 2 und 3 widerspiegeln damit auch nur die Variablenwerte der letzten beiden. Eine weitere Interpretation ist hier logischerweise nicht sinnvoll.

Für jede der weiteren Linkage-Methode sind die erhaltenen mittleren Verläufe der Variablen in den einzelnen Clustern folgend jeweils mit dargestellt. Sie werden in den nächsten Unterkapiteln entsprechend kurz interpretiert

Vor der Auswahl der möglichst sinnvollen Clusteranzahl sollten stets die Dendrogramme analysiert werden und die Zuordnungsmatrix ausgewertet werden. Hierzu besteht in Arbeitsblatt_1.xls ebenfalls die Möglichkeit, sich die letzten 12 Tabellenwerte der Zuordnungsmatrix sich als Diagramm anzeigen zu lassen. Diese letzten 12 Werte müssen allerdings vorher ebenfalls aus SPSS kopiert und in die EXCEL-Tabelle eingefügt werden.

## 4.1 Nächster Nachbar-Linkage und Abstände als Block–Distanzen

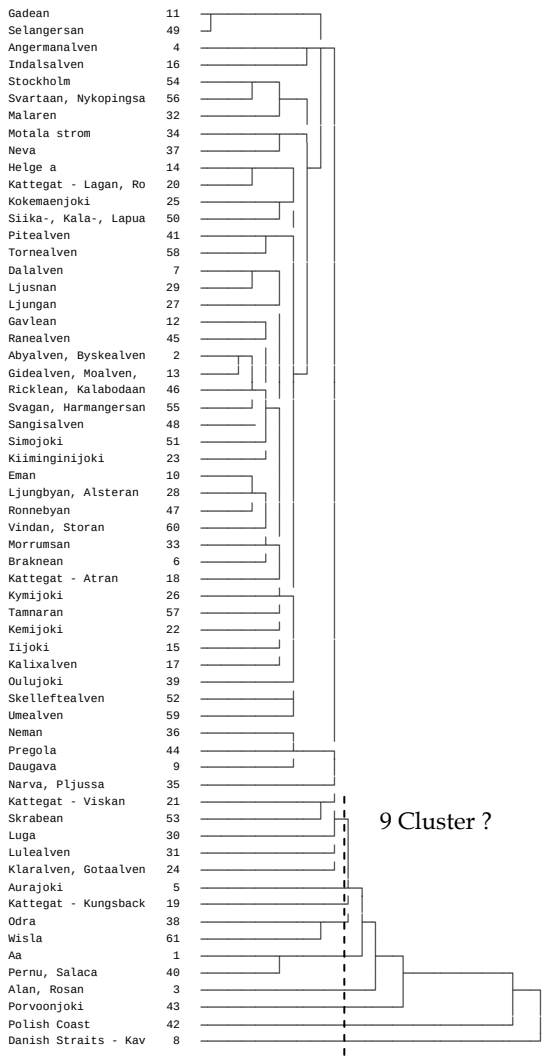



*Bild 3.13: Dendrogramm (Distanzmaß: Block. Distanz, Nächster Nachbar -Linkage)*

Wird das Nächste-Nachbar-Linkage zusammen mit der Block-Distanz angewendet, tritt die Kettenbildung weiterhin sehr deutlich auf. Die Zuordnungsmatrix legt nahe, 3 Cluster zu verwenden. Allerdings weisen diese nur einzelne Einzugsgebiete auf. Selbst bei 9 Klassen sind alle Klassen außer der Klasse 1 sehr gering besetzt. Klasse 1 wird hingegen nicht differenziert. Trotz recht hoher Distanzen zwischen den Fällen bleiben diese in diesem Cluster zusammengefasst.

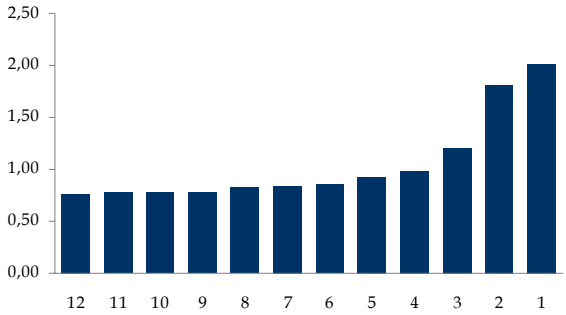



*Bild 3.14: Die letzten 12 Abstände der Zuordnungsübersicht ((Distanzmaß: Block. Distanz, Nächster Nachbar -Linkage).*

Auf die Darstellung der Variablenmittelwerte in den Clustern kann hier verzichtet werden. Diese repräsentieren jeweils nur einzelne Einzugsgebiete und einen großen mittleren Wert.

## 4.2 Nächster Nachbar-Linkage und Ähnlichkeiten als PEARSON-Korrelationen

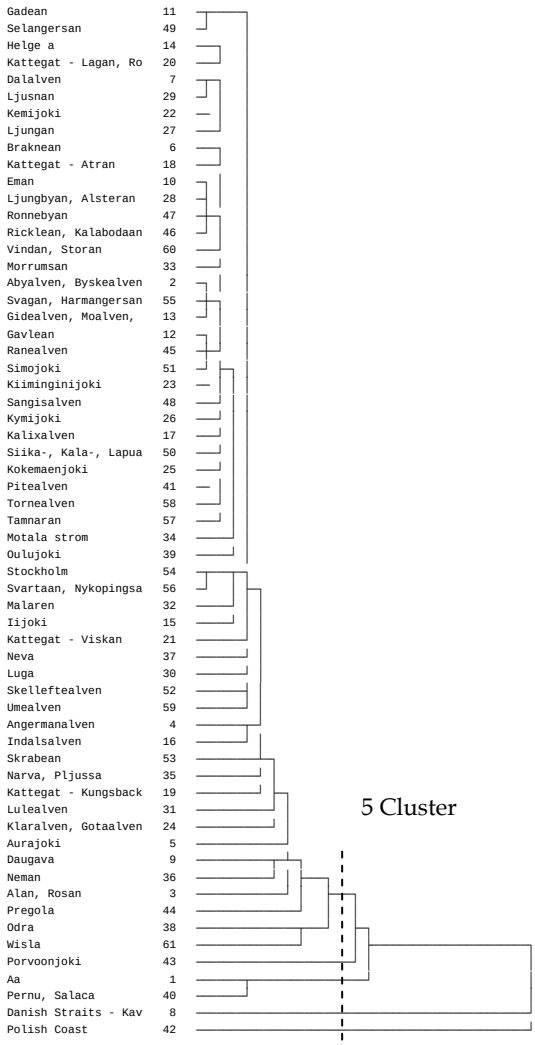



*Bild 3.15: Dendrogramm (Pearson-Korrelation, Nächster Nachbar -Linkage)*

Beim Nächster Nachbar-Linkage und dem PEARSON-Korrelationskoeffizienten als Ähnichkeitsmaß wird anfangs eine große Gruppe von Fällen als sehr ähnlich betrachtet (Dendrogramm in Bild 3.15). Wiederum überwiegt die Kettenbildung deutlich. Die Wahl unterschiedlicher Distanz- oder Ähnlichkeitsmaße scheint also diesen Effekt nicht deutlich zu beeinflussen.

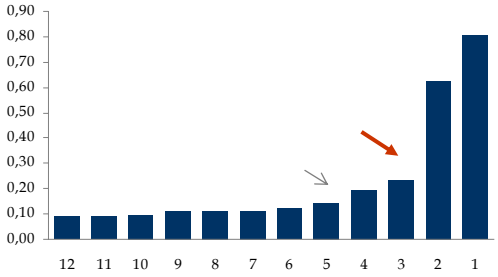



*Bild 3.16: Die letzten 12 Abstände der Zuordnungsübersicht (Pearson-Ähnlichkeit, Nearest Neighboor -Linkage).*

Die Zuordnungsmatrix weist einen „Sprung“ ab den letzten 3 Fällen auf. Es könnten also 3 Cluster als sinnvoll erachtet werden. Diese würden wiederum nur die Einzugsgebiete Polish Coast und Danish Straits - Kavlingeån, Saxån repräsentieren. Auch wenn man nach Maßgabe des Dendrogramms eine höhere Klassenanzahl wählt, sind die letzten Cluster oft nur mit einem oder zwei Einzugsgebieten besetzt.

Wir sprechen hierbei davon, dass keine „echten“ Klassen gebildet werden. Dieses Verhalten haben wir mit den gerade gezeigten Beispielen besonders für das Nächste Nachbar-Linkage als kennzeichnend feststellen können. Bei diesem Linkage-Verfahren wird die Distanz zu bereits bestehenden Cluster analog Bild 3.8 festgelegt.

## 4.3 Median(Average)-Linkage und Abstände als Quadrierte Euklidische Distanzen

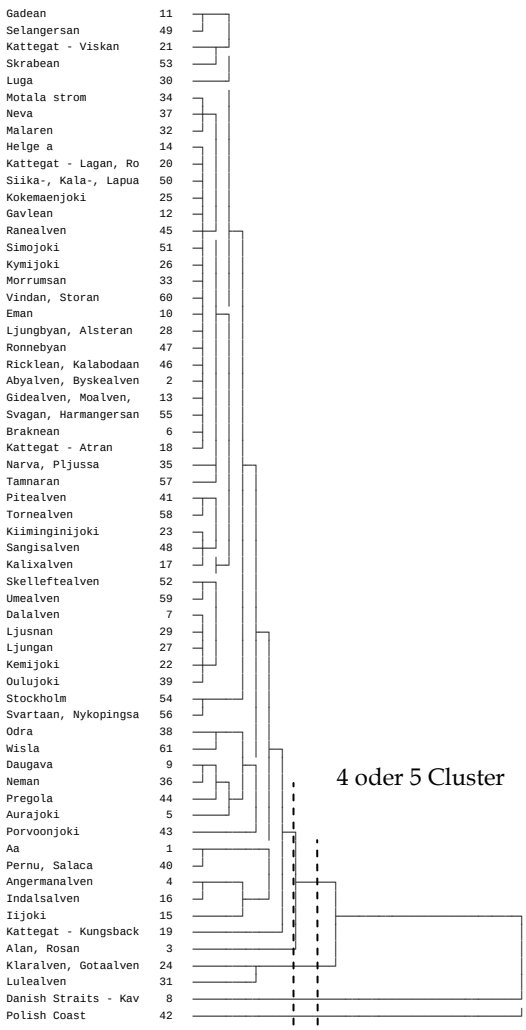



*Bild 3.17: Dendrogramm (Distanzmaß: Quad. Euklid. Distanz, Median -Linkage)*

Beim Median(Average)-Linkage wird die die Distanz zu bereits bestehenden Cluster nicht analog zu Bild 3.8 aufgrund des nächsten Nachbarn festgelegt.

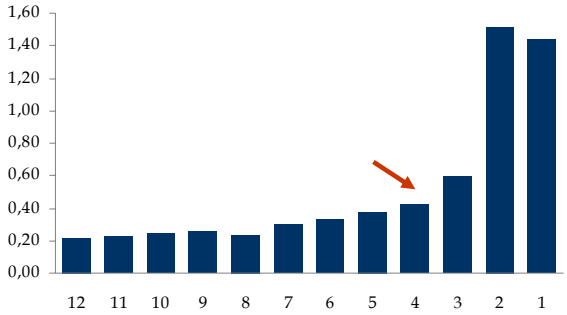



*Bild 3.18: Die letzten 12 Abstände der Zuordnungsübersicht ((Distanzmaß: Quad. Euklid. Distanz, Median -Linkage).*

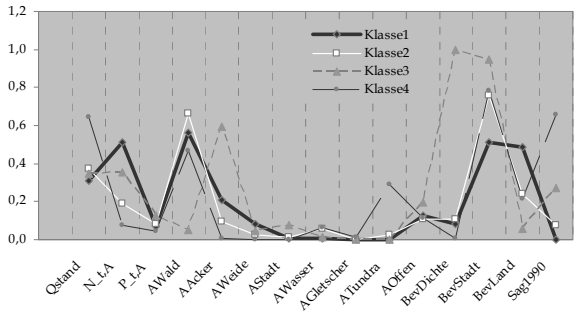



*Bild 3.19: Klassenbildung (Distanzmaß: Quad. Euklid. Distanz, Median -Linkage)*

Hier entscheidet der Mittelwert der Distanz aller Elemente zwischen dem bestehenden Cluster und einem oder mehreren neu aufzunehmenden Elementen über die geringste Distanz.

Wir stellen jetzt 4 sinnvolle Klassen fest (Bild 3.18), allerdings ist immer noch eine deutliche Kettenbildung zu verzeichnen. Sinnvoll wären möglicherweise noch 5 Klassen (siehe zweite gestrichelte Line in Bild 3.17). Damit erhalten wir (in Arbeitsblatt_1.xls) für die Mittelwerte der Variablen eine Darstellung wie in Bild 3.19. Neben 3 einzelnen Einzugsgebieten werden zusätzlich eine Klasse mit waldreichen und eine Klasse mit abflussreichen Einzugsgebieten extrahiert.

## 4.4 Entferntester Nachbar-Linkage und Abstände als Quadrierte Euklidische Distanzen

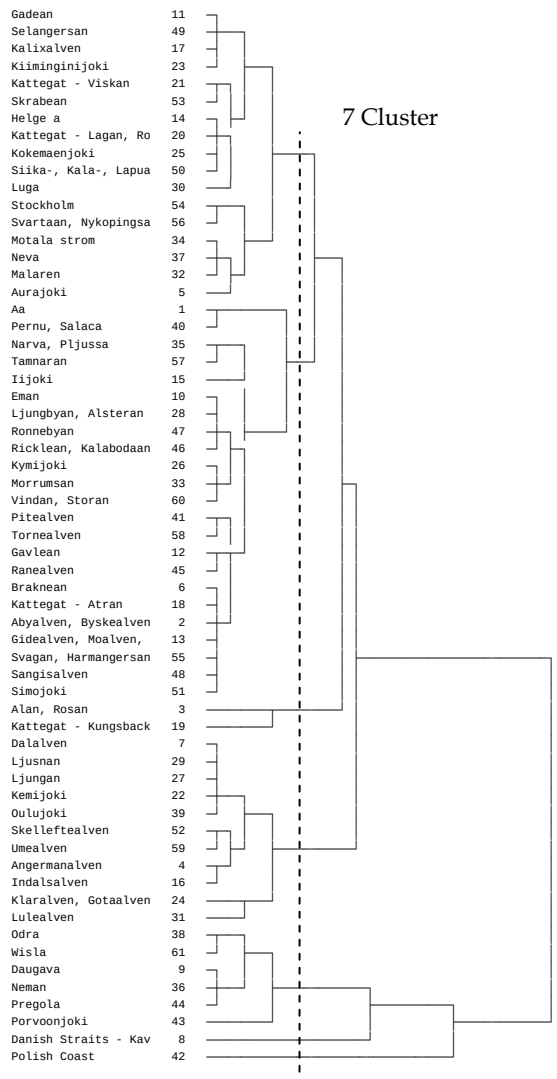



*Bild 3.20: Dendrogramm (Distanzmaß: Quad. Euklid. Distanz, Entferntester Nachbar -Linkage)*

Beim Entfernteste Nachbar-Linkage (auch Complete-Linkage) erhalten wir nun zum ersten Mal ein Dendrogramm ohne Kettenbildung. Wir können im Dendrogramm erkennen, dass „echte“ Klassen gebildet worden sind.

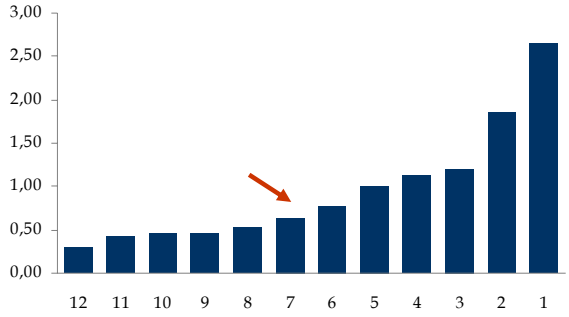



*Bild 3.21: Die letzten 12 Abstände der Zuordnungsübersicht ((Distanzmaß: Quad. Euklid. Distanz, Entferntester Nachbar -Linkage)*

Das Entfernteste Nachbar-Linkage beurteilt die Distanz eines bestehenden Clusters zu einem neu aufzunehmenden Fall hinsichtlich der größten Distanz zwischen einem Clusterelement zu dem neu aufzunehmenden Fall. Damit wird implizit erreicht, dass die Abstände der Fälle zueinander in den Cluster möglichst klein sind. Ein ähnliches Ziel verfolgt das im nächsten Kapitel vorgestellte WARD-Linkage.

Die Zuordnungsmatrix und auch das Dendrogramm (Bild 3.20 und Bild 3.21) zeigen, dass entweder 6 oder 7 Klassen unterteilt werden sollten. Beide Varianten sind möglich. Damit zeigt sich übrigens sehr eindrücklich, dass die Clusteranalyse keine „eindeutigen“ Ergebnisse liefert. Ihre Resultate geben dem Bearbeiter allerdings die Möglichkeit, eine möglichst „objektive“ Klasseneinteilung zu finden!

Wird die Clusterlösung in SPSS mit 7 Clustern berechnet, ergibt sich mit Arbeitsblatt_1.xls folgende inhaltliche Interpretation der Klassen (Bild 3.22): Neben 2 Ausreißerklassen (wieder Polish Coast und Danish Straits - Kavlingeån, Saxån) existieren 5 sinnvolle Klassen. Klasse 1

enthält Waldeinzugsgebiete mit einer geringen Bevölkerungsdichte. Klasse 2 repräsentiert abflussreiche Waldeinzugsgebiete mit aber nur 2 Fällen.

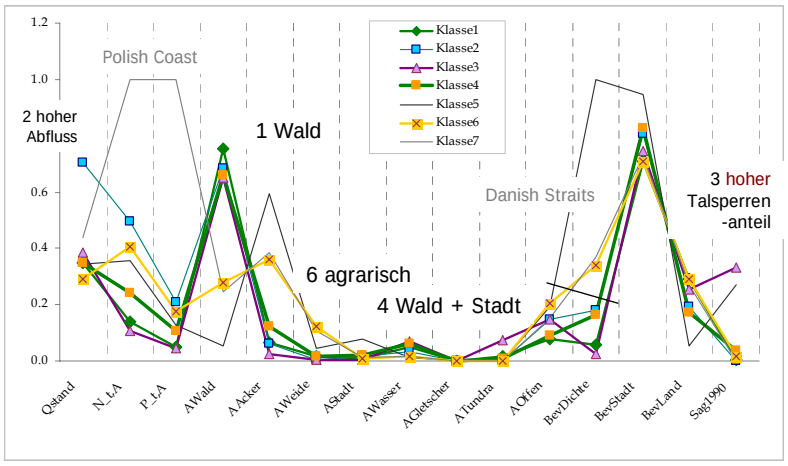



*Bild 3.22: Klassenbildung (Distanzmaß: Quad. Euklid. Distanz, Entferntester Nachbar -Linkage)*

In der Klasse 3 finden sich talsperrenreiche Waldeinzugsgebiete mit sehr geringen Stofffrachten, Klasse 4 fasst Waldeinzugsgebiete mit hohen Stofffrachten und hoher Bevölkerungsdichte (hoher Anteil städtischer Bevölkerung) zusammen. Die Klasse 6 widerspiegelt die agrarisch genutzten südbaltischen Einzugsgebiete mit ebenfalls hohen Stofffrachten.

## 4.5 WARD-Linkage und Abstände als Quadrierte Euklidische Distanzen

Die WARD-Methode wird oft als „bestes" Linkage-Verfahren bezeichnet. Hier werden die neu in ein Cluster aufzunehmenden Fälle schrittweise so ausgewählt, dass die Abweichungen aller Werte zueinander - also deren Varianzen – in einer Klasse möglichst klein sind. Eine mathematisch detailliertere Beschreibung des Verfahrens kann man in Lehrbüchern zur Multivariaten Statistik (z.B. BAHRENBERG, GIESE & NIPPER, 1991, Kap. 7) nachlesen.

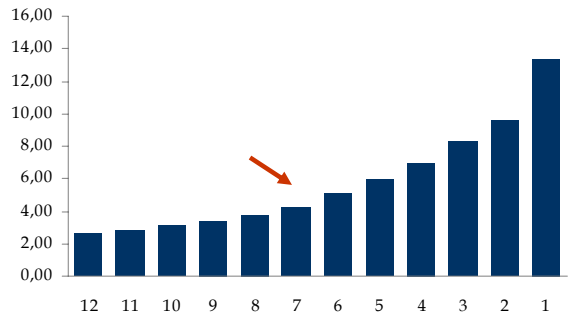



*Bild 3.23: Die letzten 12 Abstände der Zuordnungsübersicht (Quad. Euklid. Distanz, WARD -Linkage).*

Wir beschränken uns hier wieder auf die Interpretation der Ergebnisse. Bei der WARD-Methode fällt meistens auf, dass die letzten Werte der Zuordnungsmatrix keinen deutlichen „Sprung" ausführen.

Vielmehr flacht der Abfall der Werte kontinuierlich ab. Die Anzahl der sinnvoll zu wählenden Klassen sollte daher nicht nur anhand der Zuordnungsmatrix erfolgen. Vielmehr sollte das Dendrogramm zur Entscheidung herangezogen werden.

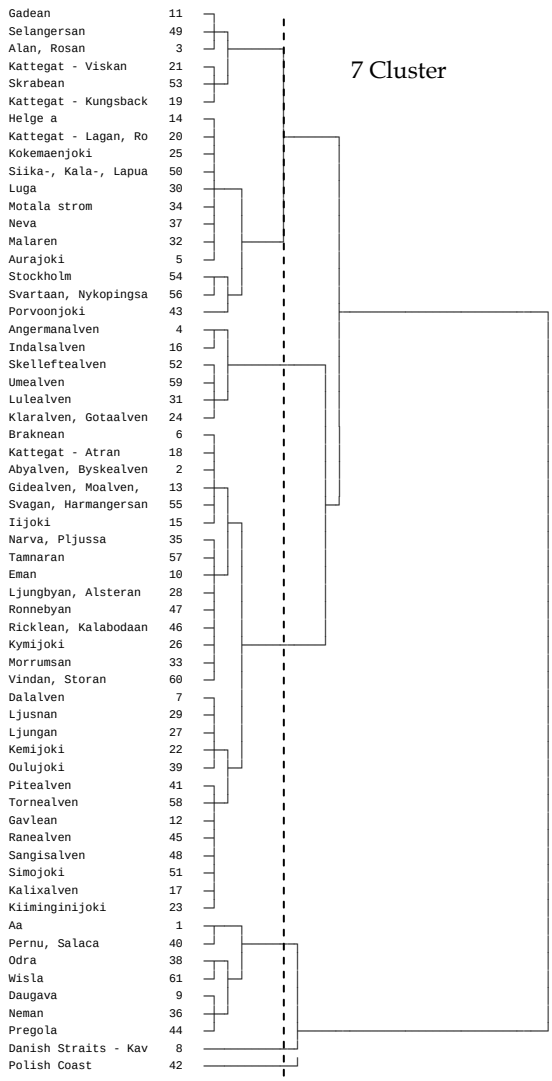



*Bild 3.24: Dendrogramm (Distanzmaß: Quad. Euklid. Distanz, WARD -Linkage)*

Bild 3.24 zeigt wie beim vorangegangenen Verfahren des Entferntester Nachbar-Linkage wiederum ein Dendrogramm mit deutlich ausgeprägten, „echten“ Klassen. Wir erkennen, dass es sinnvoll ist, 7 Klassen auszuwählen. Der ab dem 7. Wert (rückwärts gezählt) flacher werdende Abfall in der Zuordnungsmatrix (Bild 3.23) bestätigt dann diese Wahl.

Wir haben mit diesem Dendrogramm die bislang beste Unterteilung in verschiedene Klassen gefunden. Noch mehr als beim Entferntester Nachbar-Linkage sind die Distanzen in den Klassen sehr gering (linker Teil des Dendrogramms), die Abstände zwischen den Klassen aber sehr groß (rechter Teil des Dendrogramms ab der gestrichelten Linie). Das verwundert auch nicht, soll das WARD-Linkage doch genau eine Varianzminimierung in den Klassen und eine Varianzmaximierung zwischen den Klassen bewirken.

# 5 Systemanalyse: Räumliche und inhaltliche Interpretation der Cluster

Welche Erkenntnisse lassen sich gewinnen, wenn wir die Clusterlösung des WARD-Linkage im Folgenden noch auswerten wollen? Hierzu sind in Bild 3.25 zum einen die Ausprägungen der mittleren Werte der Variablen in den einzelnen Klassen dargestellt. Ebenfalls sind diese Klassen in einer Karte den jeweiligen Einzugsgebieten der Ostsee zugeordnet worden.

Wir erkennen mit der Klasse 1 die südbaltischen, agrarisch genutzten Einzugsgebiete mit hohen Stofffrachten. Die Klasse 2 stellt Waldeinzugsgebiete mit geringen Bevölkerungs-

dichten dar, Klasse 3 kleine Waldeinzugsgebiete mit hohen Abflussspenden. Diese kennzeichnen vor allem küstennahe Regionen im westlichen Ostseeeinzugsgebiet. In einer weiteren Klasse können wir ähnliche Einzugsgebiete finden. Klasse 4 zeigt gleichfalls abflussreiche, aber zusätzlich durch wasserwirtschaftliche Energiegewinnung genutzte Einzugsgebiete.

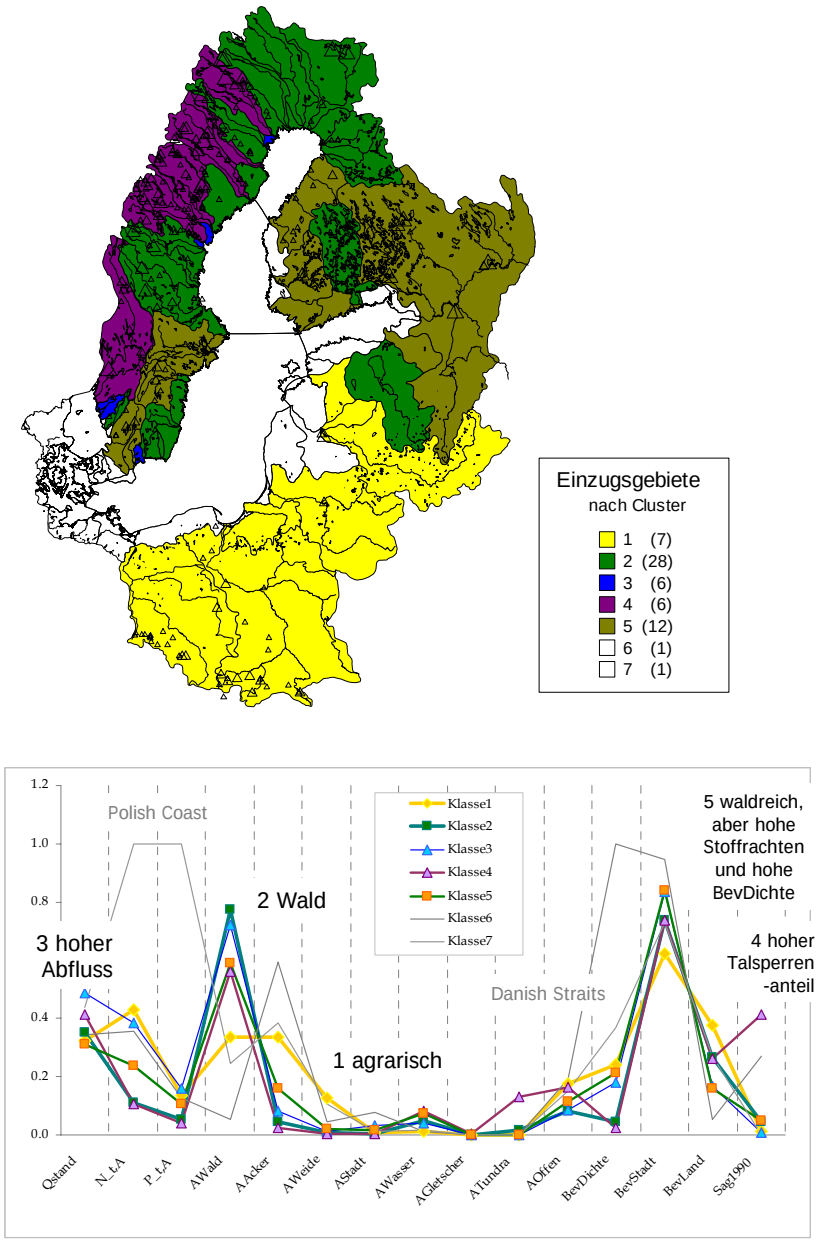



*Bild 3.25: Ergebnisse der Clusterlösung, WARD –Linkage (Distanzmaß: Quad. Euklid. Distanz)*

Diese Einzugsgebiete umfassen Gebirgs- und Hochgebirgsanteile der Skanden und sind zusätzlich durch geringe Stofffrachten gekennzeichnet.

Klasse 5 repräsentiert Waldeinzugsgebiete mit hohen Bevölkerungsdichten und hohen Anteilen städtischer Bevölkerung, ebenfalls aber hohen Stofffrachten. Wir finden hier Einzugsgebiete in der Nähe der großen Städte und Hauptstädte Skandinaviens sowie westfinnische Einzugsgebiete. Diese sind infolge der hier vorhandenen und Zellstoff- und Papierindustrie durch höhere Nährstofffrachten zur Ostsee gekennzeichnet.

Beeinflusst die räumliche Lage der Klassen 2 und 4 (und die damit verbunden Merkmale) die Ausprägung geringer Nährstofffrachten?

Auf eine ähnliche Fragestellung könnte immerhin auch die erste und die fünfte Klasse deuten. Hier sind all jene Einzugsgebiete zusammengefasst worden, welche entweder

intensiv agrarisch genutzt werden (hoher Acker- und Weideflächenanteil) oder hohe Bevölkerungsdichten aufweisen. In beiden Klassen sind hohe Stofffrachten zu beobachten.

Wir haben also mit der räumlichen Charakterisierung der Clusterlösung bereits einige Ansatzpunkte für daraus ableitbare interessante Fragestellungen gefunden. Mit dieser Interpretation der günstigsten Clusterlösung wollen wir dieses Kapitel abschließen. Der Schritt der Systemanalyse in der Datenauswertung sollte uns in die Lage versetzen, ein möglichst detailliertes Bild über die Ausprägung der verschieden Variablen im betrachteten Datensatz zu erhalten. Wenngleich die Clusteranalyse bereits voraussetzt, sich durch zunächst eine explorative Datenanalyse mit den Dateninhalten vertraut gemacht zu haben (Kap. I und II), ermöglicht sie doch auch zahlreiche neue Einblicke in die Strukturen eines verwendeten Datensatzes.

Aus der Sicht des Geographen soll dabei auch möglichst herausgefunden werden, wie sich unterschiedliche Variablen in unterschiedlichen Fällen verhalten. Da die unterschiedlichen Fälle meistens Raumeinheiten darstellen (hier die von uns verwendeten Einzugsgebiete), lässt sich so auch etliches über das unterschiedliche Verhalten im Raum erfahren. Wo sind bestimmte Merkmale wie ausgeprägt? Mit diesen regionalen Kenntnissen kann man sich dann gut informiert den nächsten Schritten der Systemidentifikation und Systemsynthese zuwenden.

# KAPITEL IV
# KORRELATIONSANALYSE

Mit Kapitel IV verlassen wir den Schritt der Systemanalyse. Bildlich hatten wir diese damit beschrieben, dass die unterschiedlichen Merkmale der Flusseinzugsgebiete der Ostsee zu untersuchen waren und nun sozusagen als „Kästen“ oder „Boxen“ bekannt sind. Während der Clusteranalyse war jedoch schon aufgefallen, dass in einigen Einzugsgebieten Merkmale, wie hohe Nährstofffrachten und der Anteil der Ackerfläche, in einer Gruppe auftraten. Zu Beginn der Clusteranalyse hatte uns sogar interessiert, ob sich Gruppen von Einzugsgebieten mit hohen Seen- und Waldflächenanteilen bilden lassen (Kap. III. 1.).

## 1 Analyse von Zusammenhängen - Streudiagramme

Betrachten wir also zunächst noch einmal die Variablen Acker- und Waldflächenanteil sowie die Stickstofffrachten und den Ackerflächenanteil.

Mit Hilfe des Datensatzes auf der CD (🖫 BasinData_fromGRIDA.xls, (Einfügen > Diagramm, Diagrammtyp „Punkt(XY)“) können Sie hierzu einen Scatterplot wie in Bild 4.1. erstellen.

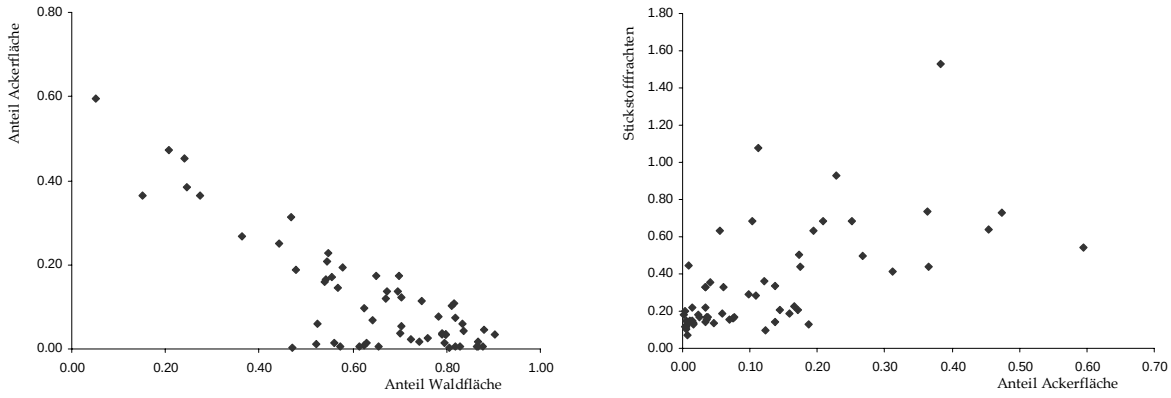



*Bild 4.1: Gut sichtbare Zusammenhänge: Wald- und Ackerflächenanteil (links) sowie Ackerflächenanteil und Stickstofffrachten in den Einzugsgebieten der Ostsee*

Links in Bild 4.1 ist deutlich erkennbar, dass die beiden Merkmale Acker- und Waldflächenanteil einander zu bedingen scheinen. Das ist auch verständlich, immerhin werden hier Anteilswerte betrachtet. In Einzugsgebieten mit hohem Waldflächenanteil werden also nur

geringe Ackerflächenanteile zu beobachten sein. Viel wichtiger als diese banale Feststellung ist aber, dass der aufgestellte Zusammenhang sich offensichtlich gut in einem diagonalen Verlauf der Punktwolke in Bild 4.1 widerspiegelt. Ein ebensolcher, wenn auch vielleicht nicht ganz so deutlich sichtbarer Zusammenhang ist im rechten Teil von Bild 4.1 zu erkennen. Es scheint also ein Zusammenhang zwischen dem Ackerflächenanteil und den Stickstofffrachten zu existieren. Auch hier kann man sich wieder einen diagonalen Verlauf der Punktwolke vorstellen, hohe Ackerflächenanteile würden damit auch hohe Stickstofffrachten bedingen.

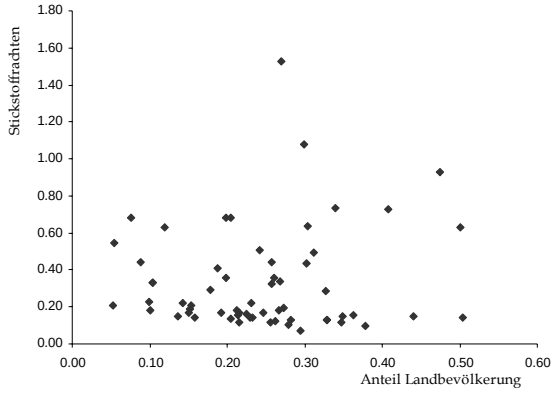



*Bild 4.2: Schlechter Zusammenhang: Anteil der Landbevölkerung und der Stickstofffrachten in den Einzugsgebieten der Ostsee*

Eine gegensätzliche Darstellung findet sich in Bild 4.2 Wir wollen vermuten, dass der Anteil der Landbevölkerung einen Einfluss auf die Stickstofffrachten in den Einzugsgebieten der Ostsee haben könnte. Wird hierzu jedoch ein Scatterplot erstellt, so zeigen die recht waagerecht verlaufenden Streifen der Punktwolke, dass weder bei einem hohen noch bei einem geringen Anteil der Landbevölkerung eine deutliche Änderung der Stickstofffrachten aufzutreten scheint.

Je deutlicher also der Punkteverlauf in einem Scatterplot einer diagonalen Linie folgt, desto eher wird man von einem Zusammenhang zwischen den beiden betrachteten Variablen sprechen können. Einen solchen Zusammenhang wollen wir im Folgenden als lineare Korrelation kennen lernen.

## 2 Korrelation – Messung eines Zusammenhanges

Wie lassen sich in einem Scatterplott „Linien“ innerhalb der Punktwolken beschreiben? Hierzu schauen wir uns zunächst Bild 4.3 an. Auf beiden Diagrammen ist noch einmal der Zusammenhang zwischen dem Anteil der Ackerfläche und den Stickstofffrachten dargestellt. Auf dem linken Diagramm sind die Stickstofffrachten nur auf der X-Achse abgetragen, auf dem rechten Diagramm auf der Y-Achse. In beiden Fällen kann durch die Punktwolke eine Gerade gelegt werden.

■ Übungsbeispiel: Erstellen Sie einen Scatterplot der Stickstofffrachten und der Ackerflächenanteile! Wie man einen Scatterplot erstellt, können Sie im Kapitel III - Clusteranalyse noch einmal nachlesen. Über das Menü Diagramm > Trendlinie > Linear (siehe auch Bild ① und ② in Uebung_1.xls) können

Sie eine Linie an die Punktwolke anpassen. Über die Registerkarte Optionen soll hier noch zusätzlich die Ausgabe einer Gleichung angefordert werden.

GGA_Korrelation\Uebung_1.xls

Mit dem eigentlichen Schritt der Anpassung einer Geraden werden uns in Kapitel VI noch ausführlicher befassen. An dieser Stelle wollen wir uns zunächst nur für den Anstieg der beiden Geraden interessieren. Je enger der Winkel zwischen den beiden Geraden, umso stärker ist der Zusammenhang zwischen den beiden Variablen Stickstofffrachten und Ackerflächenanteil.

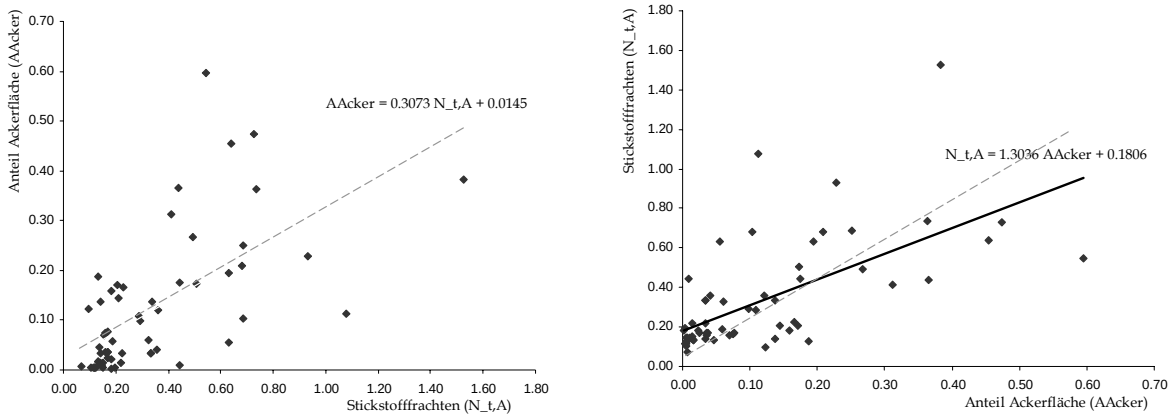



*Bild 4. 3: Scatterplot zwischen Ackerflächenanteil und Stickstofffrachten sowie umgekehrt zwischen Stickstofffrachten und Ackerflächenanteil: Der Winkel zwischen den beiden Geraden (im Bild rechts) zeigt den Zusammenhang zwischen Ackerflächenanteil und Stickstofffrachten an.*

Wir haben nun zwei Variablen in einem jeweils achsenvertauschten Scatterplott (Bild 4.3) dargestellt. Der Zusammenhang zwischen den beiden Variablen ist gleich dem Winkel zwischen den beiden durch die Punktwolken verlaufenden Geraden. Ein Ausdruck für diesen Winkel lässt sich ermitteln, indem die Anstiege der beiden durch die Punktwolken verlaufenden Geraden multipliziert werden.

Der Zusammenhang r lässt sich damit als

$$r = \sqrt{Anstieg_{X,Y} * Anstieg_{Y,X}} \qquad \text{(Gl. 1: Zusammenhang als Produkt der Anstiege)}$$

ausdrücken und widerspiegelt damit genau die Winkelöffnung zwischen den beiden Geraden wie in Bild 4.3.

■ Für Bild 4.3 bedeutet diese Aussage, dass die Anstiege aus den Formeln im rechten und linken Bildteil zu multiplizieren sind. Wenn wir in Gl. 1 die Anstiege aus Bild 4.3 einsetzen, so erhalten wir

$r = \sqrt{0{,}3703 * 1{,}3036} = 0{,}63$. Jetzt berechnen wir zusätzlich in

GGA_Korrelation\Uebung_1.xls

- links oben auf dem Tabellenblatt - als Zusammenhangsmaß den Korrelationskoeffizienten. Dieser wird durch die Funktion `KORREL` erhalten, welche sich über den Funktionsassistenten über die Statistikfunktionen aufrufen lässt. Der von der Funktion `KORREL` ausgegebene Wert ist identisch mit dem von uns - anhand des Winkels zwischen den beiden Geraden aus Bild 4.3 - errechneten Zusammenhang.

Mit Hilfe des soeben durchgeführten Beispiels haben wir sehen können, dass sich ein Zusammenhang zwischen zwei Variablen als Winkelabstand zwischen diesen beiden auffassen lässt. Genau eine solche Aussage trifft ebenfalls der lineare Korrelationskoeffizient. Als Zahl mit Werten zwischen „-1“ und „+1“ kann mit seiner Hilfe die Stärke eines Zusammenhanges zwischen zwei Variablen angegeben werden. Dabei bedeutet ein Wert nah an „1“ einen möglichst guten Zusammenhang, nah an „0“ einen nicht vorhandenen Zusammenhang und nah an „-1“ einen umgekehrt proportionalen Zusammenhang. Umgekehrt proportional heißt hier, dass zu einem großen X-Wert ein kleiner Y-Wert gehört.

■ Betrachten wir noch einmal Bild 4.1 und Bild 4.2. Für den Zusammenhang zwischen den Stickstofffrachten und dem Ackerflächenanteil (Bild 4.1 rechts) ließe sich damit ein Korrelationskoeffizient von 0.63 angeben. Jener zwischen Ackerflächenanteil und Waldflächenanteil ist zum einen negativ, befindet sich mit -0,84 allerdings auch recht nah an „-1“. Anhand des negativen Wertes müssen wir annehmen, dass bei hohen Waldanteilen oft geringe Ackerflächenanteile vorzufinden sind. Diese Aussage klingt sehr plausibel, daher verwundert die hohe (negative) Korrelation nicht. Für Bild 4.2 ergibt der Korrelationskoeffizient einen Wert von 0,10 und untermauert damit, dass sich auch grafisch kein sinnvoller Zusammenhang zwischen den Stickstofffrachten und dem Anteil der Landbevölkerung in den Einzugsgebieten der Ostsee erkennen lässt.

Indem wir den Korrelationskoeffizienten gerade als Winkelabstand zwischen zwei durch die Punktwolke der Variablen gelegte Geraden kennen gelernt haben, können wir auch einige seiner wichtigsten Eigenschaften näher betrachten.

In Gl. 1 wurden die Anstiege der Geraden aus Bild 4.3 multipliziert. Dabei ist es völlig egal, ob erst $\text{Anstieg}_{X,Y}$ mit $\text{Anstieg}_{Y,X}$ oder umgekehrt multipliziert wird. Diese Eigenschaft wird die Symmetrieeigenschaft des Korrelationskoeffizienten genannt. Damit können nur Zusammenhänge zwischen zwei Variablen angegeben werden, der Korrelationskoeffizient selbst kann nicht festlegen, ob die Stickstofffrachten die Ackerflächenanteile oder umgekehrt die Ackerflächenanteile die Stickstofffrachten beeinflussen.

SCHÖNWIESE, 2000 nennt noch weitere Vorraussetzungen, welche möglichst vor der Berechnung eines Korrelationskoeffizienten erfüllt sein sollten. Zum einen sollten Stichproben mit mindestens 30 Elementen verwendet werden. Zum anderen sollten die miteinander zu

korrelierenden Datenreihen annähernd normalverteilt sein. Hinter dieser Forderung verbirgt sich die Tatsache, dass der Korrelationskoeffizient streng genommen nur für lineare Zusammenhänge aussagefähig ist. Für nichtlineare Zusammenhänge unterschätzt man sonst recht leicht die Korrelation! Diese wichtige Einschränkung des Korrelationskoeffizienten wollen wir ebenfalls anhand eines Beispiels näher erläutern:

■ Bild 4.4 zeigt die Punktwolke und damit einen möglichen Zusammenhang zwischen dem Anteil der Seenfläche und dem Waldflächenanteil im Einzugsgebiet der Ostsee. Es ist deutlich erkennbar, dass sich durch die Punktwolke keine Gerade legen lässt. Dennoch legt die Punktwolke einen möglichen Zusammenhang zwischen den beiden Merkmalen dar. Offensichtlich ist der Seenflächenanteil in den sehr waldreichen gegenüber den waldarmen Einzugsgebieten etwas erhöht.

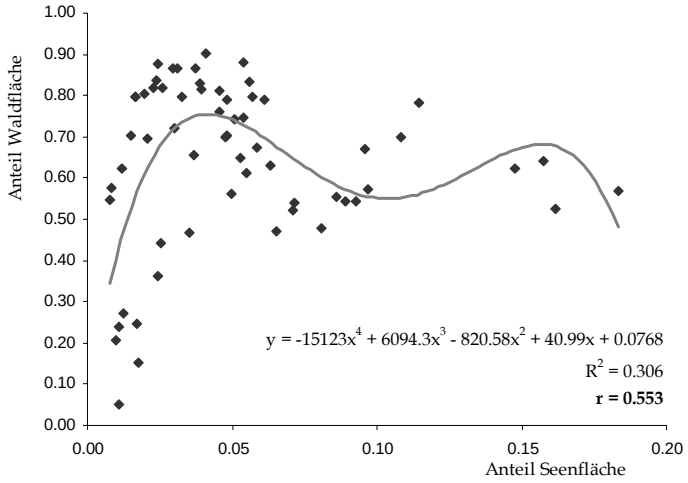



*Bild 4.4: Nichtlinearer Zusammenhang zwischen Wald- und Seenflächenanteil in den Einzugsgebieten der Ostsee*

Je höher der Seenflächenanteil, je mehr nimmt der Waldflächenanteil aber wieder ab. Eine wiederum plausible Aussage, können doch Einzugsgebiete mit einem Seenflächenanteil über 15% ohnehin kein Waldanteil über 85% besitzen.

Der Zusammenhang lässt sich als nicht linear beschreiben. Folglich weist der errechnete lineare Korrelationskoeffizient mit 0,05 auch auf keinen Zusammenhang hin. Stellen wir jedoch einen Zusammenhang zwischen dem Waldflächen- und dem Seenflächenanteil auf der Basis einer als Polynom 3. Ordnung dargestellten Kurve her (Gleichung in Bild 4.4), so lässt sich anhand dieser Kurve ein Korrelationskoeffizient von 0,55, somit also ein deutlich besserer Zusammenhang ausweisen. Wir haben an dieser Stelle allerdings nicht geklärt, wie realistisch dieser in Bild 4.4 dargestellte Kurvenverlauf ist. Mit solchen Fragenstellungen soll sich Kapitel VI - Regressionsanalyse beschäftigen.

In der folgenden Box sind die gerade getroffenen Aussagen zusammengefasst worden. Wichtig ist zudem, dass ein durch den Korrelationskoeffizienten aufgefundener Zusammenhang zunächst nur für die untersuchte Stichprobe gilt. Zudem ist es wichtig, zu betonen, dass durch die Korrelation Ähnlichkeiten und keine gleichen Distanzen erfasst werden.

Hinsichtlich der Unterschiede zwischen Distanzen und Ähnlichkeiten ist auch noch einmal auf Kapitel III – Clusteranalyse – dort Kapitel 3.2 zu verweisen. Der Korrelationskoeffizient kann damit sehr instruktiv auch als Phasenähnlichkeitsindex (SCHÖNWIESE, 2000) aufgefasst werden.

**Eigenschaften und Voraussetzungen zur Berechnung des Korrelationskoeffizienten**

1. Korrelationskoeffizienten nah am Wert „1“ und „-1“ weisen auf eine hohe Ähnlichkeit zwischen zwei Merkmalen hin.
2. Es kann nur die Stärke eines Zusammenhanges vermutet werden, die Richtung der Abhängigkeit zwischen zwei Variablen gibt der Korrelationskoeffizient selbst nicht an.
3. Für sinnvolle Aussagen sollten zur Berechnung des Korrelationskoeffizienten Stichproben mit weniger 30 als Elementen nicht verwendet werden.
4. Vermutete Zusammenhänge sollten sich linear beschreiben lassen. Wird erfüllt, wenn die beiden zu korrelierenden Variablen annähernd normalverteilt sind.

   HINWEIS: Bei nichtlinearen Zusammenhängen können die Variablen durch die FISHER-Transformation „normalisiert“. Das genaue Vorgehen ist bei SCHÖNWIESE, 2000, Kapitel 11.4 beschrieben. Die dabei notwendige Berechnung des Tangens hyperbolicus lässt die EXCEL-Funktion `ARCTANHYP` zu. (Aufzurufen unter Einfügen > Funktion > Math. & Trigonom.)
5. Es sollte möglichst eine Datenunabhängigkeit zwischen den Stichproben gegeben sein. Eine vorhandene Datenabhängigkeit lässt sich durch Partielle Korrelationsanalyse auffinden, weiter hierzu in Kap. 4. 4 und Kap. 4. 5.

# 3 Test des Korrelationskoeffizienten

Je größer die Stichproben sind, je mehr Werte also verglichen werden, desto eher kann auch bei einem nicht so deutlich an eins heranreichenden Korrelationskoeffizienten angenommen werden, dass dieser einen vorhandenen Zusammenhang anzeigt. Diese Aussage, dass einem statistischen Maß mit zunehmender Größe der Stichprobe auch mit höherer Wahrscheinlichkeit vertraut werden kann, ist allgemein als Signifikanz bekannt.

Die Signifikanz des Korrelationskoeffizienten lässt sich mittels des t-Test abschätzen. Hierfür wird nach Gl. 2 für den Korrelationskoeffizienten ein t-Wert berechnet.

$$t = \frac{r * \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

Gl. 2 Berechnung des t-Wertes

$$mit \quad n - 2 \quad FG$$

n sind dabei die Anzahl der Stichprobenelemente, also die Anzahl Werte in jeweils einer Variablen. Damit wird deutlich, dass große Stichproben offensichtlich anders in die Berechnung des t-Wertes eingehen als kleine.

GGA_Korrelation\Uebung_2.xls, Registerblatt KorrelN

■ Berechnen Sie in den Korrelationskoeffizienten zwischen allen Variablen und den Stickstofffrachten in den Einzugsgebieten der Ostsee. Verwenden Sie hierfür die EXCEL-Tabellenfunktion `KORREL` und setzen Sie diese in Zeile 66 ein. Berechnen Sie anschließend den t-Wert für die ermittelten Korrelationskoeffizienten nach Gl. 2. Tragen Sie dazu ab Spalte B beginnend die Formel

`=B66*WURZEL(61-2)/WURZEL(1-B66^2)` (Großbuchstaben geben Spaltenbezug an)

in die jeweiligen Zellen in Zeile 68 ein! Der Ausdruck `61 - 2` in der ersten Klammer sind die Freiheitsgrade, 61 Stichprobenelemente minus 2. Mit dieser Schreibweise wird Gl. 2 umgesetzt.

Die t-Funktion ist wie die Normalverteilung (Kap II) eine Wahrscheinlichkeitsfunktion. Zu einem ermittelten t-Wert können dabei die Quantile der t-Funktion ermittelt werden. Diese geben die Irrtumswahrscheinlichkeit an, mit welcher der errechnete Korrelationskoeffizient auf einen vorhandenen Zusammenhang hinweist. Die Quantile der t-Funktion können in EXCEL (ähnlich wie bei der Normalverteilung in Kap. II) wiederum mit einer Tabellenfunktion abgefragt werden. Dazu wird die Funktion `TVERT` verwendet (Bild 4.5) Die Anzahl der Freiheitsgrade ist dabei wieder n (Anzahl der Elemente in der Stichprobe) minus 2. Wir wollen 2-seitig testen, geben bei Seiten also 2 an.

GGA_Korrelation\Uebung_2.xls, Registerblatt KorrelN

■ Berechnen Sie die Quantile der t-Verteilung mit der Funktion TVERT (Bild 4.5). In Zeile 71 wird dabei automatisch angegeben, ob der Korrelationskoeffizient bei der errechneten Irrtumswahrscheinlichkeiten auf einen signifikanten Zusammenhang hinweist. Dabei werden Korrelationskoeffizienten als signifikant aufgefasst, wenn der t-Wert eine Irrtumswahrscheinlichkeit für die Korrelation von weniger oder gleich 0,05 (5%) ergibt.

So korrelieren offensichtlich die Wald-, Acker – und Weideflächenanteile signifikant mit den Stickstofffrachten. Hier weist also der Korrelationskoeffizient auf einen Zusammenhang hin.

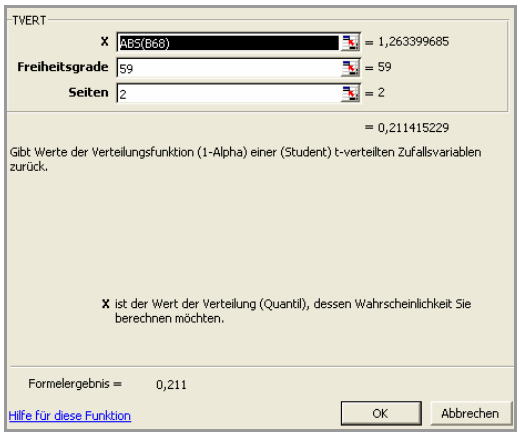



*Bild 4.5: Funktion* TVERT

Umgekehrt lässt sich auch die Frage stellen, ab welchem Wert denn ein Korrelationskoeffizient als signifikant angesehen werden kann? Diese Aussage muss selbstverständlich ebenfalls vom Stichprobenumfang abhängen.

Dazu wird die Formel aus Gl. 2 so umgestellt, dass sich der Korrelationskoeffizient zu einem vorgegebenem Quantil der t-Funktion bei entsprechend zu prüfendem Signifikanzniveau (meist zweiseitig 95%-Sicherheitswahrscheinlichkeit; 5% Irrtumswahrscheinlichkeit) berechnen lässt. Dieser

Korrelationskoeffizient ist jener ab dem Stichproben vom Umfang n miteinander signifikant korrelieren sollten.

GGA_Korrelation\Uebung_2.xls, Registerblatt SigKorr

■ Auf diesem Tabellenblatt ist genau dieser Schritt vollzogen worden. Indem Sie über das Drehfeld die Größe der Stichprobe variieren, können Sie einen Überblick darüber bekommen, ab welchem Korrelationskoeffizienten Sie laut Aussage des t-Tests mit signifikanten Korrelationen zu rechnen haben. Erwartungsgemäß werden damit bei steigender Stichprobenanzahl schon bei sehr kleinen Korrelationen signifikante Zusammenhänge als möglich angezeigt. Geben Sie versuchsweise eine Stichprobengröße von 1000 ein. Ab welchem Korrelationskoeffizient wäre ein Zusammenhang bei einer solchen Stichprobe signifikant?

Das gerade vorgestellte Beispiel der Signifikanzprüfung von Korrelationskoeffizienten zeigt nun aber eindrücklich, dass gleichzeitig davor gewarnt werden muss, selbst bei einem positivem Signifikanztest aber eben einem kleinen Korrelationskoeffizienten unkritisch einen Zusammenhang zu postulieren. Bei einer Stichprobe von 1000 Elementen, aber einem Korrelationskoeffizienten von nur wenig über 0,07 muss unbedingt weiter die Frage bestehen bleiben, welcher Zusammenhang hier anzunehmen wäre! Damit muss zwingend noch ein Kapitel über die Interpretation des Korrelationskoeffizienten angefügt werden.

# 4 Interpretation des Korrelationskoeffizienten

Die Korrelationsanalyse soll ein wichtiges Hilfsmittel bei der Systemidentifikation sein. Wir wollen also nicht nur bloße Zusammenhänge zwischen einzelnen Merkmalen oder Eigenschaften innerhalb des Datensatzes feststellen. Vor dem Hintergrund einer systemaren Betrachtungsweise müssen vielmehr Wirkungszusammenhänge aufgedeckt werden. Werden die Ergebnisse der Korrelationsanalyse aber unkritisch betrachtet, können durch sie auch leicht Fehleinschätzungen zustande kommen.

Zunächst ist es wichtig, Folgendes festzuhalten: Durch den Korrelationskoeffizienten können keine Wirkungszusammenhänge begründet, wohl aber vermutete Zusammenhänge quantitativ genauer eingegrenzt werden (vgl. hierzu auch SCHÖNWIESE, 2000, Kap. 11)! Die Korrelationskoeffizienten zwischen verschiedenen Variablen lassen sich allerdings als Indizien auffassen, welche auf mögliche Abhängigkeiten hinweisen. Gerade bei komplexen Wirkungsgefügen kann man damit wertvolle Hinweise auf deterministisch bislang noch wenig verstandene Zusammenhänge erhalten. Insofern sollen Möglichkeiten der Korrelationsanalyse nicht unterschätzt, ihre Ergebnisse aber stets inhaltlich hinterfragt und interpretiert werden.

Damit kann aber auch durch die Signifikanzprüfung (vorheriges Kapitel) ein Zusammenhang nicht „bewiesen“ werden. Ein geringer Korrelationskoeffizient weist stets auch nur auf

einen möglichen Zusammenhang hin. Hierbei bietet es sich zudem an, statt des Korrelationskoeffizienten r dessen Quadrat $R^2$ zu anzugeben. $R^2$ ist das Bestimmtheitsmaß für den untersuchten Zusammenhang und gibt an, welcher Anteil der Streuung der Werte der einen Variablen sich in der zweiten Variablen auch beobachten lässt.

**Bestimmheitsmaß: erklärbarer Anteil eines Zusammenhangs**

Zusätzlich zum Korrelationskoeffizienten sollte auch dessen Quadrat zur Bewertung eines vermuteten Zusammenhanges betrachtet werden. Auch als Bestimmtheitsmaß bezeichnet, lässt sich dieser Wert direkt als Anteil jener Streuung der Werte interpretieren, welcher sich in beiden Variablen wieder finden lässt. Damit ermöglicht das Bestimmtheitsmaß eine kritischere Interpretation des Korrelationskoeffizienten.

■ Für die Korrelation zwischen den Stickstofffrachten und dem Ackerflächenanteil hatten wir einen Korrelationskoeffizienten von 0,63 berechnet. Das Quadrat von $0{,}63^2 = 0{,}40$ – das Bestimmtheitsmaß - bedeutet somit, dass 40% der Werteabweichungen der Ackerflächenanteile sich ebenso in den Variationen der Stickstofffrachten beobachten lassen. Ein geringer Wert, aber immer noch ein Indiz für einen möglichen Zusammenhang. Dieser lässt sich inhaltlich begründen. Die Düngemittelgaben auf Ackerflächen und damit auch der Anteil der von Düngemittelgaben beeinflussten Fläche der Einzugsgebiete (die Ackerfläche) sollten in irgendeiner Weise die Nährstofffrachten beeinflussen. Die gering übereinstimmenden Varianzen, lediglich 40%, geben aber einen Hinweis auf den möglichen Einfluss weiterer steuernder, bislang aber nicht betrachteter Größen.

So werden Bodenart und Bodentyp, Art der Feldfrüchte usw. sicher ebenfalls einen Einfluss auf die Höhe der wieder ausgetragenen Stickstofffrachten ausüben. Solche Größen sind aber bisher nicht betrachtet worden. Insofern ist der Zusammenhang zwischen dem Ackerflächenanteil und Stickstofffrachten mit 40% gleicher Varianz wiederum schon recht bemerkenswert.

Dass Wirkungen oft nicht durch den Einfluss einer Variablen erklärt werden können, zeigt einen weiteren kritischen Aspekt bei der Interpretation des Korrelationskoeffizienten auf. Zusammenhänge sind oft nur über das Wirken mehrer Größen erklärbar. Dieses Phänomen wurde in Kapitel 4.2 bereits kurz als Datenunabhängigkeit (bzw. gefordertes Vorhandensein von Datenunabhängigkeit) genannt. Steuert eine dritte, aber unbekannte Variable einen Wirkungszusammenhang, erscheint dieser auch zwischen den beiden ersten Variablen als deutlich ausgeprägt.

Eine Möglichkeit, solche Fehler zu umgehen, wäre, nur Variablen zu korrelieren, welche bekanntermaßen eine Abhängigkeit aufweisen. Über diesen Weg würde es aber nur eingeschränkt möglich sein, unbekannten Zusammenhängen nachzugehen.

Um Fehler durch den Einfluss dritter oder vierter Variablen zu minimieren, kann zusätzlich zur Korrelationsanalyse aber auch die Partielle Korrelationsanalyse angewandt

werden. Mit Hilfe der Partiellen Korrelationsanalyse wird es leichter, den der Systemidentifikation innewohnenden Ansatz weiterzuverfolgen, Wirkungszusammenhänge aufzuspüren.

# 5 Partielle Korrelation

Um Wirkungszusammenhänge im Sinne eines echten Wirkungsdiagramms aufzuspüren, hält die Statistik mit der Partiellen Korrelationsanalyse (und weiter der Pfadanalyse, siehe hierzu BAHRENBERG, GIESE & NIPPER, 1991) eine wertvolle Methode bereit.

Die Kernfrage der Partiellen Korrelationsanalyse ist die folgende: Welcher Zusammenhang zwischen zwei Variablen bleibt bestehen, wenn zur Erklärung eine dritte oder vierte Variable hinzugenommen wird? Anders formuliert, welcher Zusammenhang zwischen zwei Größen in einem komplexen Wirkungsgefüge lässt sich finden, wenn die „störende" Wirkung anderer Einflüsse möglichst klein gehalten wird. Durch ein solches Vorgehen werden quasi experimentelle Bedingungen geschaffen.

**Partielle Korrelationsanalyse – ein experimenteller Ansatz für die Datenauswertung**

Mit Hilfe der Partiellen Korrelationsanalyse erhält man bei einer empirischen Untersuchung die bemerkenswerte Möglichkeit, Zusammenhänge experimentell zu untersuchen. Damit lässt sich ein für die Naturwissenschaften typischer Versuchsaufbau simulieren. Ein Zusammenhang wird untersucht, indem alle möglichen weiteren Störgrößen möglichst ausgeschaltet werden. Erst dann kann der wirkliche Einfluss einer Größe auf die zweite möglichst realitätsgetreu gemessen werden!

Mit der Partiellen Korrelationsanalyse lassen sich damit auch sehr gut Scheinkorrelationen untersuchen. Ebenso lässt sich so klären, ob der Zusammenhang zwischen zwei Größen möglichst unabhängig ist, also nicht ebenfalls durch eine dritte Größe erklärt werden kann.

Diese Aspekte lassen sich sehr gut anhand des „Klapperstorchproblems" illustrieren.

GGA_Korrelation\Uebung_3.xls, Registerblatt PartielleKorrelation

■ Dass zwischen der Geburtenrate und der Anzahl der Störche pro Fläche kein Zusammenhang besteht, ist hinlänglich bekannt. Dennoch lässt mit Hilfe des linearen Korrelationskoeffizienten hier recht leicht ein Zusammenhang postulieren. Für das fiktive Beispiel aus Uebung_3.xls (das Zahlenbeispiel ist MONKA & VOß, 2002, Kap. 21 entnommen, hier findet sich auch eine sehr instruktive Einführung in die Partielle Korrelationsrechnung) ergibt sich mit 0,85 ein sehr guter Zusammenhang. Es ist aber auch bekannt, dass die industrielle Entwicklung eines Landes die Geburtenrate wie auch die Storchenanzahl (pro Fläche) gegenläufig beeinflusst. Es sollte also vielmehr ein Wirkungsgefüge

wie in Bild 4.6 dargestellt angenommen werden. Damit ist die „Industrialisierung“ jene Größe, welche den Zusammenhang zwischen Geburtenrate und Storchenanzahl „stört“!

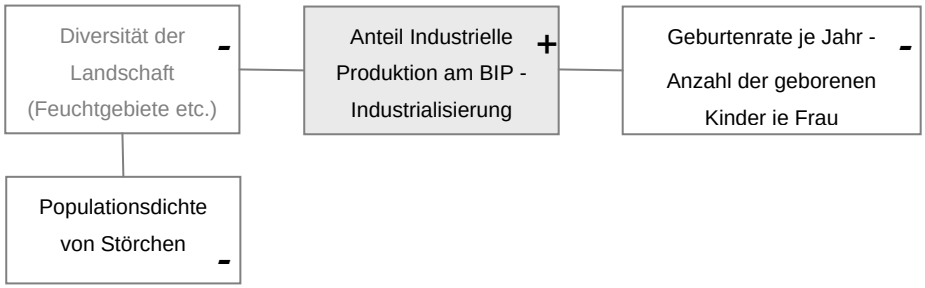



*Bild 4. 6: Wirkungszusammenhänge beim „Klapperstorchproblem“*

GGA_Korrelation\Uebung_3.xls, Registerblatt PartielleKorrelation

■ Berechnen wir für den Zusammenhang zwischen Anteil industrieller Produktion am BIP und der Storchendichte sowie Anteil industrieller Produktion am BIP und der Geburtenrate die Korrelationskoeffizienten, so weisen beide auf einen deutliche (aber negativen) Korrelation hin.

Jetzt wollen wir nochmals die Korrelation zwischen der Geburtenrate und der Storchendichte berechnen, diesmal aber als Partiellen Korrelationskoeffizienten. Hierbei unterdrücken wir den Einfluss der „Industrialisierung“.

Der Partielle Korrelationskoeffizient lässt sich ohne große Mühe aus den linearen Korrelationskoeffizienten der beteiligten Variablen berechnen. Hierbei gilt Gl. 3:

$$r_{xy,z} = \frac{r_{xy} - r_{xz} * r_{yz}}{\sqrt{1 - r_{xz}^{\ 2}} * \sqrt{1 - r_{yz}^{\ 2}}}$$

Gl. 3 Der Partielle Korrelationskoeffizient

GGA_Korrelation\Uebung_3.xls, Registerblatt PartielleKorrelation

Wird der Partielle Korrelationskoeffizient für das „Klapperstorchproblem“ berechnet, lässt sich recht schnell die nur extrem geringe Korrelation zwischen Geburtenrate und Storchendichte erkennen. Berechnen Sie den Partiellen Korrelationskoeffizienten indem Sie in Tab.1 (Uebung_3.xls) zunächst die Korrelationskoeffizienten zwischen allen drei Variablen berechnen und dann gemäß Gl. 3 einsetzen! Alternativ können Sie auch die in diesem Tabellenblatt mittels VBA-Script beigefügte Tabellenfunktion `PartKorr` nutzen (Bild 4.7)! AusschaltZ nimmt dabei jene Größe in die Berechnung auf, deren Einfluss „experimentell“ ausgeschlossen werden soll!

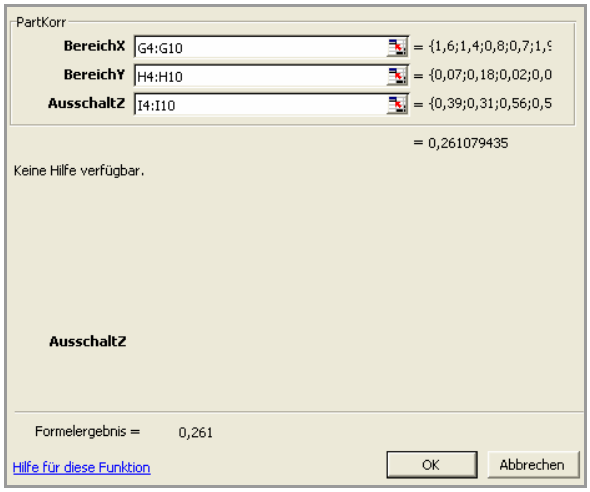



*Bild 4.7: Funktion* `PartKorr` *für die Berechnung des „Klapperstorchproblems“ in Uebung_3.xls, aufzurufen aus Einfügen > Funktion > Benutzerdefiniert*

Wir erhalten nur noch eine Korrelation von 0,26, damit wären gerade 7 % der Streuung beider Größen identisch (Bestimmtheitsmaß!). Wir haben also eine Scheinkorrelation aufgedeckt. Gleichzeitig müssen wir feststellen, dass wir den linearen Korrelationskoeffizienten formal hätten gar nicht berechnen dürfen. Durch den Einfluss des Anteils der industriellen Produktion am BIP bestand ja immerhin keine Datenunabhängigkeit (Kap. 4.2).

# 6 Partielle Korrelation bei der Systemidentifikation

Der Partielle Korrelationskoeffizient eignet sich in anschaulicher Weise, die Aussagen des linearen Korrelationskoeffizienten zu überprüfen und zu verifizieren. Diese Eigenschaft hatte uns in Kap. 4.5 ermöglicht, den offensichtlich unsinnigen Zusammenhang zwischen der Geburtenrate und der Storchendichte auch statistisch zu widerlegen. Gemeinsam mit dem linearen Korrelationskoeffizienten wollen wir nun beide Verfahren dazu verwenden, detaillierter mögliche Wirkungsmechanismen zwischen einzelnen Merkmalen in den Einzugsgebieten der Ostsee zu untersuchen.

GGA_Korrelation\Uebung_3.xls, Registerblatt PartKorrelN

■ Auf dem Tabellenblatt sind wieder die Korrelationskoeffizienten für die Zusammenhänge zwischen den Stickstofffrachten und allen anderen Variablen aus Uebung_3.xls mit den dazugehörigen Signifikanzniveaus angegeben. In der darunter befindlichen Tabelle wollen wir nun die gefundenen linearen Korrelationen „experimentell“ überprüfen. Hierzu beschränken wir uns allerdings auf die 4 stärksten signifikanten Korrelationen und schließen sukzessive jeweils eine der 3 anderen Variablen aus. Mit diesem Vorgehen wollen wir der Frage nachgehen, welche Variable am ehesten ein wirkliches Erklärungspotenzial für die Höhe der Stickstofffrachten besitzen.

Zur Berechnung der Partiellen Korrelationskoeffizienten nutzen wir wieder die Tabellenfunktion `PartKorr` (Bild 4.7), alternativ wäre die Partielle Korrelationsanalyse auch mit SPSS möglich. Die Ergebnisse finden sich zum Vergleich rechts neben der Tabelle nochmals eingetragen. Die Ergebnisse können wir zusammen mit den linearen Korrelationskoeffizienten gemäß der folgenden Box zu interpretieren versuchen.

**Zusammenhang zwischen dem linearen und dem partiellen Korrelationskoeffizienten (SCHÖNWIESE, 2000, S. 182-183)**

1. **|partieller Korrelationskoeffizient| < |Korrelationskoeffizient|:** Korrelation zwischen zwei Variablen wird wegen des Einflusses einer Drittvariablen überschätzt. Ist der partielle Korrelationskoeffizient (bei Ausschluss der Drittvariablen) im Gegensatz zum Korrelationskoeffizienten nicht mehr signifikant, liegt sehr wahrscheinlich eine Scheinkorrelation vor ➔ „Klapperstorchproblem".
2. **|partieller Korrelationskoeffizient| > |Korrelationskoeffizient|:** Korrelation zwischen zwei Variablen wird wegen des zusätzlichen Einflusses einer „störenden" dritten Variablen unterschätzt, der partielle Korrelationskoeffizient beschreibt den Zusammenhang besser.
3. Mehrere **partiellen Korrelationskoeffizienten** unter Ausschluss von Zweit- oder Drittvariablen sind **signifikant**. Hinweis auf multiple Korrelationsrechnung.

Um die Ergebnisse besser zu veranschaulichen, wollen wir die vermuteten Zusammenhänge und die durch die Partielle Korrelationsanalyse näher überprüften Zusammenhänge einmal als Boxen mit dazugehörigen Verbindungen darstellen (Bild 4.8, Partielle Korrelation in Bild 4.9).

Als erstes wollen wir hierbei den Zusammenhang zwischen den Stickstofffrachten, der Bevölkerungsdichte und dem Ackerflächenanteil genauer überprüfen. Die Bevölkerungsdichte korreliert ihrerseits sehr eng mit dem Ackerflächenanteil (r=0,80 *). Liegt damit eine „Scheinkorrelation" des Ackerflächenanteils oder der Bevölkerungsdichte mit den Stickstofffrachten vor?

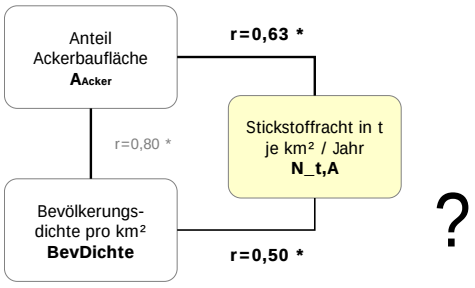



*Bild 4.8: Korrelationen zwischen den Stickstofffrachten und der Bevölkerungsdichte sowie dem Ackerflächenanteil (signifikante Werte sind durch * markiert).*

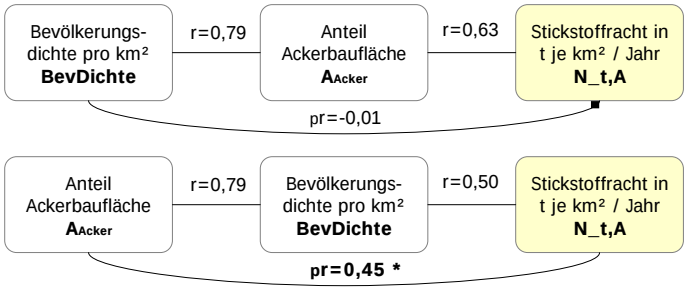



*Bild 4.9: „Experimenteller" Ansatz zur Überprüfung der Korrelationen zwischen den Stickstofffrachten und der Bevölkerungsdichte sowie dem Ackerflächenanteil.* **pr** *ist der Partielle Korrelationskoeffizient unter Ausschluss der Variable in der Mitte der Grafik.*

In Bild 4.9 sind die Ergebnisse der Partiellen Korrelationsanalyse veranschaulicht worden. Die partielle Korrelation zwischen dem Ackerflächenanteil und den Stickstofffrachten bleibt auch unter Ausschluss der Bevölkerungsdichte noch signifikant ($_pr$=0,45 *). Der Zusammenhang zwischen Bevölkerungsdichte und den Stickstofffrachten stellt sich als „Schein-

korrelation“ heraus ($_{p}$r=0,01), hervorgerufen durch die Wirkung des Ackerflächenanteils. Da Ackerflächenanteil und Bevölkerungsdichte ihrerseits recht eng korrelieren, wurde bezüglich der Stickstofffrachten lediglich ein Zusammenhang vorgetäuscht.

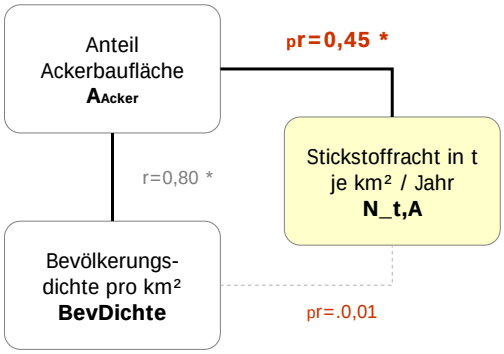



*Bild 4.10: Verbessertes Wirkungsdiagramm zu Bild 4.8*

Somit müssen wir die zunächst vermuteten Zusammenhänge aus Bild 4.8 verwerfen. Die jetzt am ehesten möglichen Zusammenhänge sind in Bild 4.10 dargestellt.

Der Zusammenhang der Stickstofffrachten mit der Bevölkerungsdichte lässt sich empirisch nicht weiter aufrechterhalten. Bislang scheinen sich die Stickstofffrachten daher noch am ehesten durch die Ackerflächenanteile erklären zu lassen.

Als nächstes wollen wir noch die Weideflächenanteile in unsere Betrachtung einbeziehen. Auch zwischen diesen und den Stickstofffrachten existiert zunächst ein signifikanter Korrelationskoeffizient.

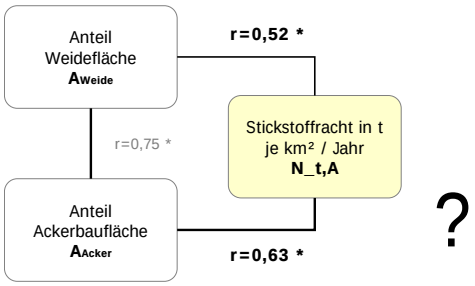



*Bild 4.11: Korrelationen zwischen den Stickstofffrachten und dem Weideflächenanteil sowie dem Ackerflächenanteil (signifikante Werte sind durch * markiert).*

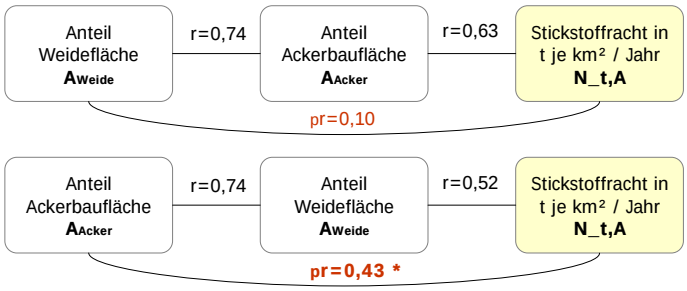



*Bild 4.12: „Experimenteller“ Ansatz zur Überprüfung der Korrelationen zwischen Stickstofffrachten und Weideflächenanteil sowie dem Ackerflächenanteil.* **pr** *ist der Partielle Korrelationskoeffizient unter Ausschluss der Variable in der Mitte der Grafik.*

Auch in diesem Beispiel wird deutlich, dass die signifikante Korrelation (r=0,52 *) zwischen Weideflächenanteil und Stickstofffrachten wieder durch eine enge Korrelation des Weideflächenanteils mit dem Anteil der Ackerfläche (r=0,75 *) überbewertet wird (Abb. 3). In einem Wirkungsdiagramm wäre eine Verbindung zwischen Weideflächenteil und Stickstofffrachten bei einem partiellen Korrelationskoeffizienten von $_{p}$r=0,10 zu streichen (Bild 4.13).

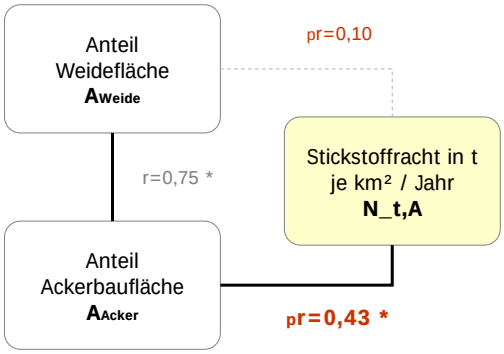



*Bild 4.13: Verbessertes Wirkungsdiagramm zu Bild 4.11.*

Bislang haben wir also feststellen können, dass die Variation der Stickstofffrachten noch am ehesten einen Zusammenhang mit der Variation der Ackerflächenanteile aufweist. In Uebung_3.xls, Registerblatt PartKorrelN, können wir zudem erkennen, dass auch der Waldflächenanteil, der ja im Wesentlichen zum Ackerflächenanteil invers ist, offensichtlich nur eine „Scheinkorrelation" mit den Stickstofffrachten aufweist. Der partielle Korrelationskoeffizient unter „experimentellem" Ausschluss des Ackerflächenanteils fällt auf einen Wert von $_{p}r = 0{,}11$ ab.

Nachdem wir nun vier mögliche Zusammenhänge zwischen einer prozessabbildenden Variablen und drei strukturabbildenden Variablen (siehe hierzu auch das Kapitel Systemanalyse in der Einführung) mit Hilfe der Linearen wie auch der Partiellen Korrelationsanalyse untersucht haben, wollen wir diese Betrachtungen mit einem etwas noch komplexeren Modell eines Wirkungsdiagramms abschließen.

In Bild 4.14 ist versucht worden, die Variation der Stickstofffrachten durch das Wirken aller drei strukturabbildenden Variablen Acker- und Weideflächenanteil sowie Bevölkerungsdichte zu erklären. Die Lineare Einfachkorrelation (Kap. 4. 2) wurde dabei um das Verfahren der Multiplen Korrelationsanalyse erweitert. Eine Einführung hierzu geben BAHRENBERG, GIESE & NIPPER, 1991, Kap. 2, ein guter Überblick findet sich ebenfalls bei MONKA & VOß, 2002 in Kapitel. 21. Alle Berechnungen wurden jetzt mit SPSS ausgeführt.

Im oberen Teil der Grafik sind die Partiellen Korrelationskoeffizienten zwischen Acker- und Weideflächenanteil sowie der Bevölkerungsdichte angegeben. Hier fällt auf, dass diese drei Variablen alle sehr gut miteinander korrelieren. Der multiple Korrelationskoeffizient ist mit $_{m}r=0{,}92$ * sehr hoch, ebenfalls die Partiellen Korrelationskoeffizienten. Faktisch ist damit bereits keine Datenunabhängigkeit gegeben, eine Forderung, welche eigentlich an den linearen Korrelationskoeffizienten zu stellen ist.

Wir haben diesen daher im Beispiel auch nicht nochmals berechnet. Trotzdem wollen wir uns die Partielle Korrelation zwischen den dem Ackerflächenanteil und den Stickstofffrachten unter „experimentellem" Ausschluss von Weideflächenanteil und Bevölkerungsdichte (siehe auch unterer Teil von Bild 4.14) nochmals ansehen. Jetzt fällt $_{p}r$ mit 0,22 unter die Signifikanzgrenze. Mit Hilfe der Korrelationsanalyse können wir also eigentlich auch keinen Zusammenhang zwischen Ackerflächenanteil und den Stickstofffrachten „beweisen".

„Beweisen" ist dabei das Schlüsselwort, um hier noch einmal auf den Wert der Korrelationsanalyse hinzuweisen. Mit der Korrelationsanalyse lassen sich nämlich nur sehr schlecht wirkliche deterministische Zusammenhänge „beweisen". Indem wir aber bereits die

möglichen Zusammenhänge zwischen den von uns betrachteten fünf Variablen (Beispiel aus Uebung_3.xls Registerkarte PartKorrelN) genauer untersucht haben, können wir nun unsere Aussagen über Variablen mit einem möglichen Erklärungspotenzial wesentlich präziser fassen. Alle betrachteten vier Variablen korrelierten anfangs signifikant mit den Stickstofffrachten. Letztlich lohnt es sich aber noch am ehesten, dem Zusammenhang zwischen dem Ackerflächenanteil und den Stickstofffrachten weiter nachzugehen. Wenngleich wir auch hier noch nach besseren erklärenden Merkmalen (z. B. Düngemittelgaben und/oder Bodentypen) suchen müssten. Weitere Möglichkeiten zur Systemidentifikation lässt der von uns verwendete Datensatz aber momentan nur schlecht zu.

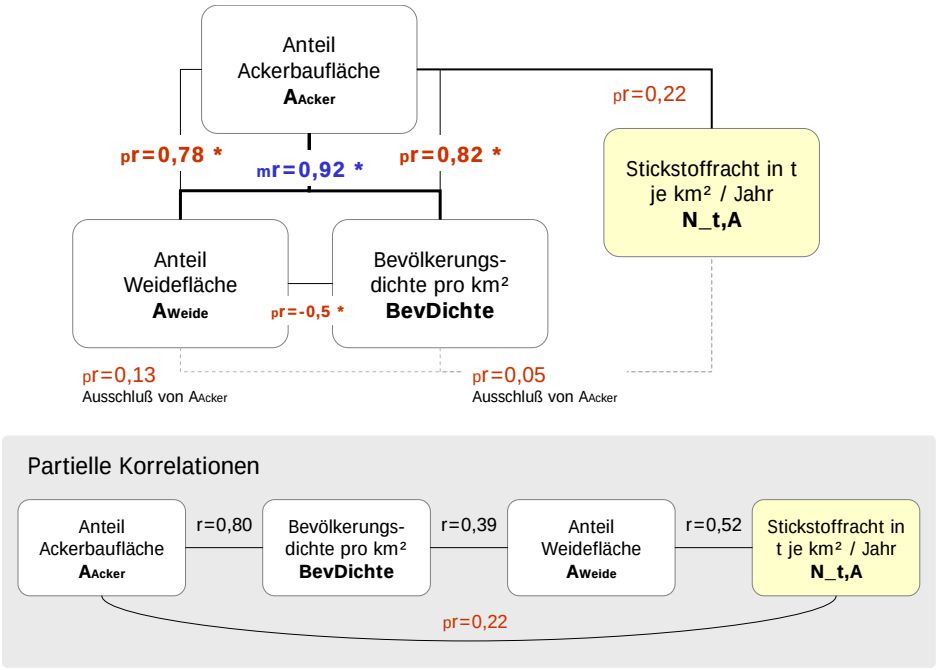



*Bild 4.14:Höherkomplexes Wirkungsdiagramm zur möglichen Erklärung der Stickstofffrachten in den Einzugsgebieten der Ostsee.* **pr** *sind die Partiellen,* **mr** *sind die Multiplen Korrelationskoeffizienten.*

Insofern hat sich unser Wissen über Wirkungszusammenhänge bezüglich unserer betrachteten Variablen aber doch deutlich erweitert. Mit der Extraktion von Variablen mit einem möglichst guten Erklärungspotenzial innerhalb eines Datensatzes werden wir uns deshalb im folgenden Kapitel der Faktorenanalyse noch intensiver befassen.

Mit dem letzten Beispiel sind wir aber auch einen Schritt gegangen, welcher sich in der Pfadanalyse fortführen ließe. Eine Einführung hierzu findet sich bei BAHRENBERG, GIESE & NIPPER, 1991, Kap. 3. Bei der Pfadanalyse werden die Zusammenhänge innerhalb eines Wirkungsnetzes von Variablen näher untersucht. Dabei werden vor allem die wechselseitigen Einflüsse dar Variablen untersucht. Zusätzlich zu unseren Betrachtungen mit Hilfe der Partiellen Korrelationskoeffizienten versucht man bei der Pfadanalyse aber auch die

Richtung der Zusammenhänge („Pfeile“ zwischen den „Kästen“ des Systems) mit anzugeben. Damit leitet dieser Ansatz bereits in die Systemsynthese (Kap. 6) über, wenngleich wir die Systemsynthese in dieser Veröffentlichung nur als Anwendung der Regressionsanalyse einführen werden.

# KAPITEL V
# HAUPTKOMPONENTEN- UND FAKTORENANALYSE

## 1 Fragestellung und Grundlagen

Wie kann die Anzahl von Variablen in einer Untersuchung sinnvoll reduziert werden?

Die Hauptkomponenten- und Faktorenanalyse versucht, Informationen über die Korrelation von Variablen zu ihrer Bündelung zu nutzen. So ist zum Beispiel zu erwarten, dass bei Messungen von Wolkenbedeckung, Niederschlagsmenge, Luftfeuchtigkeit, Windgeschwindigkeit und Sonnenscheindauer an Klimastationen auffällig signifikante Zusammenhänge zwischen den Merkmalen bestehen. Ein geschicktes Zusammenziehen und Entfernen der Variablen verkleinert den Informationsgehalt nur unwesentlich, z.B. mit 50 Prozent der Variablen könnte immerhin noch 80 Prozent der Gesamtvarianz erklärt werden.

(Die Gesamtvarianz ist die Summe der Varianzen der einzelnen Merkmale. Sie repräsentiert hier den Gesamt-Informationsgehalt einer Untersuchung).

**Bedingungen für die Anwendung der Hauptkomponenten- und Faktorenanalyse:**

- Analyse einer Untersuchung (Probenahme, Befragung) mit mehreren Merkmalen
- metrisches Skalenniveau aller Merkmale
- nicht alle Merkmalsvariablen sind paarweise unkorreliert
- alle Variablen annähernd normalverteilt (diese Bedingung ist bei geographischen Betrachtungen selten zu erfüllen, sollte aber diskutiert werden)

## 1.1 Z-Standardisierung

Mit Hilfe von Standardisierungsverfahren werden Merkmale in ein gleiches Größenverhältnis transformiert. Neben der Standardisierung am Maximum oder der Summe einer Merkmalsverteilung bzw. an einem anderen festen Faktor ist die Z-Standardisierung, auch Z-Normalisierung genannt, ein häufig verwendetes Verfahren.

Die Berechnungsvorschrift für die Normalisierung von Variablen lautet:

$$X_{stand} = \frac{X - m_{Vert}}{s_{Vert}}$$

$m_{Vert}$ – Mittelwert der Verteilung

$s_{Vert}$ – Standardabweichung der Verteilung

Diese Methode bietet gegenüber anderer Standardisierungsverfahren eine Reihe von Vorteilen.

**Vor- und Nachteile der Z-Standardisierung**

+ Mittelwert der Verteilung ist nach der Standardisierung gleich 0, die Standardabweichung beträgt 1
+ Am Vorzeichen des standardisierten Wertes ist zu erkennen, ob die Merkmalsausprägung größer oder kleiner als der Mittelwert ist
+ fast alle standardisierten Werte liegen in einem Intervall von -2 bis 2
+ ist der Betrag nach der Standardisierung größer als 2, so handelt es sich um einen Ausreißer
- ursprünglich nichtnegative Variablen können negative Wert aufweisen
- eine Deutung des standardisiertes Wertes im Merkmalskontext ist schwierig

GGA_Faktorenanalyse\Uebung_1.xls

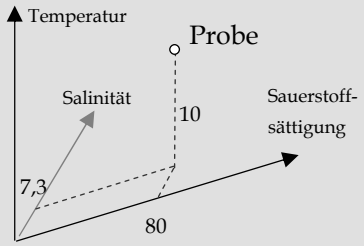
## 1.2 Unabhängige Einflussfaktoren

Die Reduzierung von Merkmalsvariablen basiert auf der Annahme, dass hinter einem ganzen Set gemessener Größen nur eine kleine Anzahl von grundsätzlichen Einflussfaktoren steht, die nicht miteinander korrelieren. Kennt man diese, kennt man die gesamte Information einer Untersuchung. Ziel der Hauptkomponenten und Faktorenanalyse ist es, diese unabhängigen Einflussfaktoren zu bestimmen.

Diese Einflussfaktoren wirken sich meist auf mehrere Größen aus. Beeinflusst ein unabhängiger Faktor mehrere Messgrößen hinlänglich stark, so drückt sich dies durch die

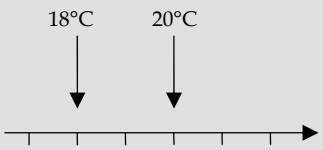
Korrelation derselben aus. Man wird dementsprechend die Korrelationsmatrix der Untersuchung auswerten.

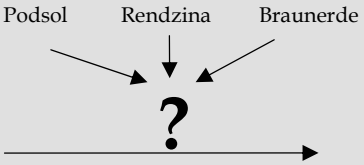
**Darstellung eines Merkmalsträgers im Koordinatensystem**




Wurden an einer Wasserprobe eine Temperatur von 10°C, eine Salinität von 7,3 ppt und eine Sauerstoffsättigung von 80% gemessen, wird die Probe in einem Koordinatensystem mit drei Achsen durch den Punkt (10; 7,3; 80) repräsentiert.
Man sagt auch: der Zustandsraum für die Proben ist dreidimensional.

Hier wird klar, warum die Bedingung des metrischen Skalenniveaus gefordert werden muss. Nominalskalierte Variable wie Bodentyp oder Staat lassen sich nicht auf einer Achse lokalisieren und somit nicht in ein Koordinatensystem eintragen.




Grundlage der Suche ist die Vorstellung, dass das Ergebnis einer Untersuchung mit *n* gemessenen Merkmalen durch Punkte in einem Koordinatensystem mit *n* Achsen dargestellt werden kann (siehe Box). Um Effekte durch unterschiedliche Skalenausprägungen zu vermeiden, müssen auch hier die Variablen zuvor standardisiert werden.

Warum werden die Merkmalsausprägungen in ein schiefwinkliges Koordinatensystem eingetragen? Besitzen zwei Messgrößen einen gemeinsamen Korrelationskoeffizient ungleich 0, so gibt es ein Abhängigkeitsverhältnis zwischen ihnen (Man sagt: Sie liegen nicht orthogonal zueinander). Im Koordinatensystem wird dies durch den nicht rechten Winkel zwischen den Achsen angedeutet.

## 2 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse hat als Ziel, alle unabhängigen Einflussfaktoren zu bestimmen. Da diese Faktoren nicht korreliert sein sollen, müssen ihre Achsen im entsprechenden

Koordinatensystem senkrecht aufeinander stehen. Ist das der Fall, repräsentiert jede der Achsen eine Hauptkomponente.

Die Orthogonalisierung eines Bezugssystems ist ein komplexer mathematischer Vorgang und basiert auf dem Schmidtschen Orthogonalisierungsverfahren. Glücklicherweise übernehmen die gängigen Statistikprogramme diese Transformationen. Bild 5.1 bis 5.3 stellen dieses Verfahren schematisch dar.

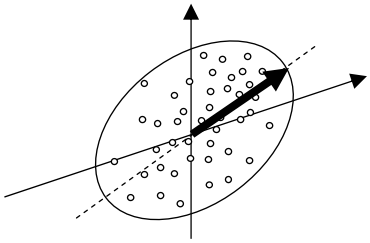


*Bild 5.1 Bestimmung der Komponente mit größter Varianzaufklärung*

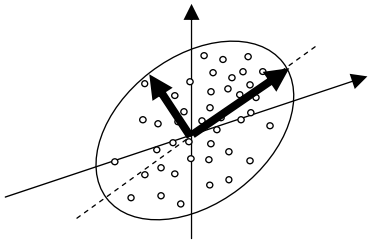


*Bild 5.2 Festlegungen der 2. Achse orthogonal zur ersten (mit größtmöglicher Varianz). Ebenso 3.-n. Achse.*

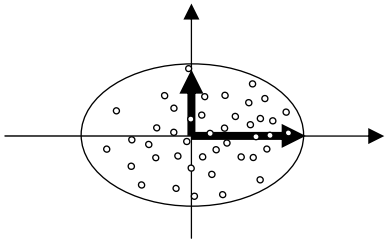


*Bild 5.3 Endergebnis der Analyse: ein orthogonales Koordinatensystem*

# 3 Faktorenanalyse

## 3.1 Ausgangssituation

Die Varianz jedes einzelnen Ausgangsmerkmals ist nach der Standardisierung gleich 1, da die Standardabweichung des Merkmals gleich 1 ist. (Die Gesamtvarianz aller Ausgangsmerkmale beträgt demnach n·1=n)

Nach Abschluss der Hauptkomponentenanalyse mit *n* Merkmalen erhalten wir im Allgemeinen ein Set von *n* unabhängigen Einflussfaktoren. Der erste besitzt die größte Varianz, der zweite die zweitgrößte, usw.

## 3.2 Berechnung von Faktoren

Die Faktorenanalyse versucht anschließend, zwischen „wichtigen“ und „unwichtigen“ Hauptkomponenten zu unterscheiden, um die Anzahl der Merkmale zu reduzieren. Ein Faktor ist dann „wichtig“, wenn er eine große Differenzierung und somit eine große Varianz der Untersuchung erklärt. Wie groß die, durch die Hauptkomponente erklärte Varianz ist, wird durch den Eigenwert der Komponente angegeben. Je geringer der Eigenwert, desto geringer die Aussagekraft der Komponente. Liegt der Wert über 1 ist er überproportional,

liegt er unter 1 ist er unterproportional wichtig zur Erklärung der Gesamtvarianz. Die Hauptkomponenten mit einer hohen Varianzaufklärung nennt man Faktoren (Scores). Das Entfernen von „unwichtigen" Hauptkomponenten entspricht der Transformation des Zustandsraumes in einen Raum geringerer Dimension. (Bild 5.4-5.6)

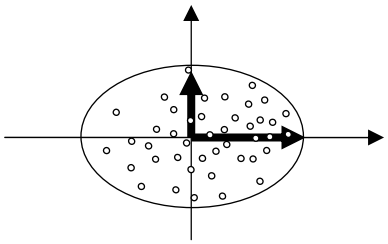


*Bild 5.4 Orthogonale Komponenten mit unterschiedlicher Varianzaufklärung*

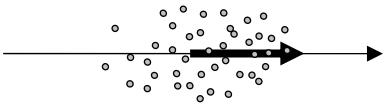


*Bild 5.5 Entfernen der wenig bestimmenden Faktoren*

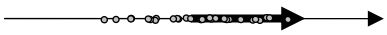


*Bild 5.6 Endergebnis: Projektion*

## 3.3 Kriterien zur Faktorenabtrennung

Mit Hilfe recht unterschiedlicher Kriterien (siehe Box Seite 76) kann festgelegt werden, wie viele Faktoren am Ende einer Untersuchung übrig bleiben. Alle Komponenten, die diese Kriterien nicht erfüllen, werden aus dem Modell entfernt.

Die Ergebnisse der Auswahlverfahren können sehr stark differieren. Es liegt in der Verantwortung des Geographen, zu entscheiden, welches Verfahren für diesen Kontext am besten geeignet ist. Eine Empfehlung kann daher nicht ausgesprochen werden.

**Ausschlusskriterien „unwichtiger" Faktoren**

- Auswahl der aller Komponenten mit Eigenwert größer als 1

  Dieses Kriterium wählt alle Komponenten aus, deren Varianz (Eigenwert) gegenüber der Varianz der Ausgangsmerkmale gestiegen ist. Das Kriterium wird von SPSS standardmäßig verwendet.
  Ein Nachteil ist die strenge Grenzziehung der Abtrennung, denn eine Komponente mit Eigenwert 0.95 besitzt eine hohe Aussage im Faktorenmodell und wird trotzdem aussortiert.

- Kaiserkriterium: $\frac{\textit{Summe der Eigenwerte}}{\textit{Gesamtvarianz}} > 90\ \%\ (80\%,..)$

  Dieses Kriterium stellt sicher, dass die Gesamtvarianz des Faktorenmodells eine feste Schwelle nicht unterschreitet. Das Kriterium ist zumeist das härteste und hat deswegen den Nachteil die Merkmalszahl nur gering zu reduzieren.

- Screekriterium

  Hier wird der Screeplot (Scree = Halde) der Komponenten untersucht. In diesem Diagramm werden die Eigenwerte aller Faktoren in fester Reihenfolge angezeigt.

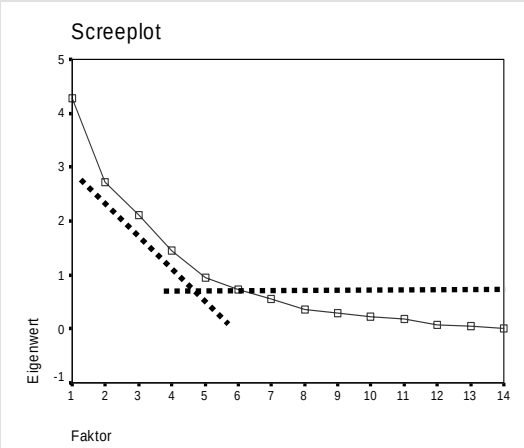



*Bild 5.7 SPSS-Screeplot*

Die Komponente am Fuß der Halde wird der letzte Faktor im Modell (die fünfte Komponente in Bild 5.7). Dieses Kriterium hat den Nachteil, häufig zu subjektiv zu sein, da der Haldenfuß unterschiedlich festgelegt werden kann.

## 3.4 Wie gut spiegelt das Faktorenmodell die Untersuchung wieder?

Zur Bewertung des Modells stehen zwei Gütekriterien zur Verfügung. Zu allererst sollte man prüfen, ob die Gesamtvarianz aller Faktoren (Summe aller Eigenwerte) nicht zu gering ist. Beträgt diese weniger als 50-70% der Ausgangsvarianz, ist das Modell durch die Aufnahme neuer Faktoren zu verbessern.

**Kommunalitäten**

| | Anfänglich | Extraktion |
|---|---|---|
| QSTAND | 1,000 | ,629 |
| N_T,A | 1,000 | ,735 |
| P_T,A | 1,000 | ,570 |
| AWALD | 1,000 | ,888 |
| AACKER | 1,000 | ,935 |
| AWEIDE | 1,000 | ,766 |
| ASTADT | 1,000 | ,580 |
| AWASSER | 1,000 | ,595 |
| AGLETSCH | 1,000 | ,710 |
| ATUNDRA | 1,000 | ,817 |
| BEVDICHT | 1,000 | ,862 |
| BEVSTADT | 1,000 | ,918 |
| BEVLAND | 1,000 | ,918 |
| SAG1990 | 1,000 | ,631 |

Extraktionsmethode: Hauptkomponentenanalyse.

*Bild 5.8 SPSS-Kommunalitätentabelle*

Ein weiterer Aspekt der Kontrolle sollte die Tabelle der Kommunalitäten sein (Bild 5.8).

In ihr wird angezeigt, wie gut jedes einzelne Ausgangsmerkmal im Modell repräsentiert wird. Sinkt der Wert eines bedeutsamen Merkmals nach der Faktorenextraktion unter eine feste Schwelle (z.B. 0.5) muss das Faktorenmodell auch hier durch die Aufnahme weiterer Faktoren angepasst werden.

## 3.5 Ergebnis der Analyse und Rotation

Das Ergebnis der Hauptkomponenten- und Faktorenanalyse sind k Faktoren $H_j$, die mit dem Ausgangsmerkmal $Z_i$ über die Faktorladungen $a_{ij}$ im Zusammenhang stehen.

Die Summanden $U_i$ repräsentieren die nicht erklärten Anteile des Modells. Die Faktorladungen zeigen an, welcher Zusammenhang zwischen Faktor und Ausgangsmerkmal besteht und sind wie ein Korrelationsfaktor (hoher positiver [negativer] Betrag → großer direkter [indirekter] Zusammenhang, geringer Betrag → kein Zusammenhang) zu interpretieren. Die hiermit gewonnenen Informationen dienen später der Namensgebung der neuen Faktoren.

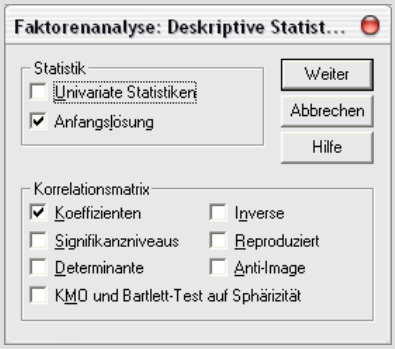
$$
\begin{aligned}
Z_1 &= a_{11}H_1 + a_{12}H_2 + \ldots + a_{1k}H_k + U_1 \\
Z_1 &= a_{11}H_1 + a_{12}H_2 + \ldots + a_{1k}H_k + U_1 \\
Z_2 &= a_{21}H_1 + a_{22}H_2 + \ldots + a_{2k}H_k + U_2 \\
&\ldots \\
Z_n &= a_{n1}H_1 + a_{n2}H_2 + \ldots + a_{nk}H_k + U_n
\end{aligned}
$$

Es ist wichtig zu wissen, dass die Merkmalsausprägung der Faktoren für jeden Merkmalsträgers in z-standardisierter Form in der Datentabelle abgespeichert wird.

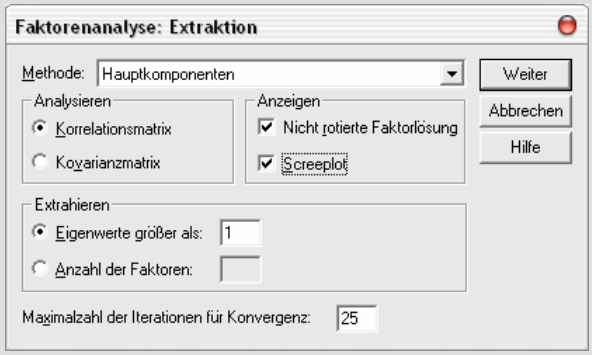
SPSS zeigt die Faktorladungen wieder in einer Tabelle an. Es wird dennoch schwer sein, Aussagen über einzelne Faktoren zu treffen. Die Beträge aller Faktorladungen liegen im Bereich 0.34 – 0.77. Für eine verbesserte Endaussage muss das Koordinatensystem noch einmal angepasst werden. Man führt demzufolge eine Rotation der Faktoren durch. Je nach Rotationsmethode (siehe Box) verändert sich die Parametrisierung des Modells ein wenig, die Faktorladungen erhalten jedoch eindeutigere Beträge.

**Hauptkomponenten und Faktorenanalyse mit SPSS**

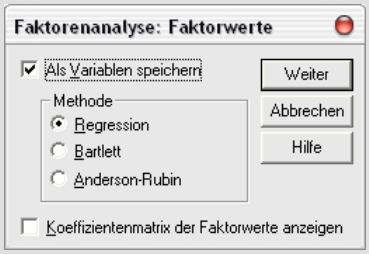
1. Analysieren> Dimensionsreduktion> Faktorenanalyse…
2. metrisch skalierte Variablen für die Analyse auswählen> mit „►“-Taste übertragen
3. Deskriptive Statistik>



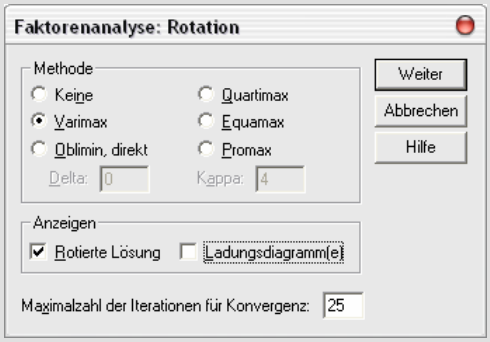

4. Extraktion>



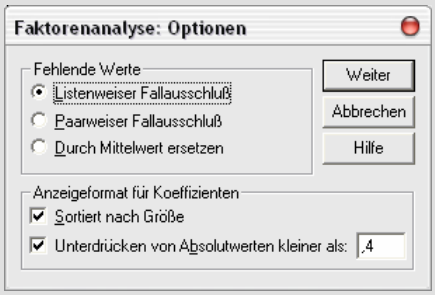

5. Werte>




6. Rotation>




7. Optionen>




8. OK>

**Rotationsverfahren in der Faktorenanalyse**

- Varimaxrotation (empfohlen)
  Die Faktoren werden so rotiert, dass die Varianz der quadratischen Ladungen maximal ist. Der Betrag der Faktorladungen wird somit maximal klein oder maximal groß. Vorteil dieses Rotationsverfahrens: Die gesamte aufgeklärte Varianz wird durch die Rotation nicht verändert, lediglich die die Verteilung auf die Faktoren. Die Orthogonalität der Faktoren bleibt erhalten.
- Quartimax
  Das Koordinatensystem wird so rotiert, dass die Anzahl der Faktoren eines Ausgangsmerkmals minimal ist.
- Equamax
  Mischung aus Varimax und Quartimax
- Oblimin und Promax
  sind oblique Rotationen. Das heißt die mühsam erworbene Unabhängigkeit (Orthogonalität) der Faktoren geht wieder verloren. Dafür werden die Faktorladungen besser getrennt.

# 4 Hauptkomponenten- und Faktorenanalyse am Beispiel

Unser Augenmerk gilt wiederum dem aus früheren Kapiteln bekannten Datensatz der Ostseeeinzugsgebiete vom GRIDA. Folgende Merkmale sollen in die Analyse eingehen (in Klammern stehen die SPSS-Variablennamen):

Einzugsgebietsname als Bezeichner (DESCRIPT), der standardisierte mittlere Jahresabfluss ($Q_{STAND}$), die mittlere Stickstofffracht pro Jahr ($N_{\_TA}$), die mittlere Phosphorfracht pro Jahr ($P_{\_TA}$), der Flächenanteil Wald ($A_{WALD}$), der Flächenanteil Ackerland ($A_{ACKER}$), der Flächenanteil Weideland ($A_{WEIDE}$), der Flächenanteil Stadt ($A_{STADT}$), der Wasserflächenanteil ($A_{WASSER}$), der Gletscherflächenanteil ($A_{GLETSCH}$), der Flächenanteil Tundra ($A_{TUNDRA}$), die Bevölkerungsdichte (BEVDICHT), der Bevölkerungsanteil Stadt ($BEV_{STADT}$), der Bevölkerungsanteil Land ($BEV_{LAND}$), der Talsperrenspeicherausbaugrad des Jahres 1990 ($SAG_{1990}$).

GGA_Faktorenanalyse\Uebung_2.sav

■ Wir führen die Faktorenanalyse mit SPSS wie in der Box auf Seite 79 beschrieben wurde durch. Das Statistikprogramm stellt nach der automatischen Durchführung der Analysen einen Protokollplot zur Verfügung. Sehen wir uns die einzelnen Tabellen genauer an.

Korrelationsmatrix

| | QSTAND | N_T,A | P_T,A | AWALD | AACKER | AWEIDE | ASTADT | AWASSER | AGLETSCH | ATUNDRA | BEVDICHT | BEVSTADT | BEVLAND |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| QSTAND | | | | | | | | | | | | | |
| N_T,A | 0.07 | | | | | | | | | | | | |
| P_T,A | 0.10 | **0.75** | | | | | | | | | | | |
| AWALD | 0.12 | **-0.48** | **-0.38** | | | | | | | | | | |
| AACKER | -0.23 | **0.63** | **0.46** | **-0.84** | | | | | | | | | |
| AWEIDE | -0.14 | **0.53** | **0.31** | **-0.73** | **0.75** | | | | | | | | |
| ASTADT | -0.02 | 0.13 | 0.04 | -0.20 | **0.30** | 0.01 | | | | | | | |
| AWASSER | -0.27 | **-0.33** | -0.25 | 0.05 | -0.23 | -0.25 | -0.09 | | | | | | |
| AGLETSCH | -0.02 | -0.14 | -0.07 | -0.11 | -0.15 | -0.10 | -0.11 | 0.04 | | | | | |
| ATUNDRA | 0.14 | -0.26 | -0.13 | -0.10 | **-0.30** | -0.21 | -0.21 | 0.05 | **0.70** | | | | |
| BEVDICHT | -0.13 | **0.50** | **0.39** | **-0.66** | **0.80** | **0.39** | **0.65** | -0.18 | -0.13 | -0.27 | | | |
| BEVSTADT | 0.03 | -0.10 | 0.02 | -0.01 | -0.04 | **-0.31** | **0.47** | 0.26 | 0.02 | 0.00 | **0.35** | | |
| BEVLAND | -0.03 | 0.10 | -0.02 | 0.01 | 0.04 | **0.31** | **-0.47** | -0.26 | -0.02 | 0.00 | **-0.35** | **-1.00** | |
| SAG1990 | 0.03 | -0.22 | -0.15 | -0.14 | -0.14 | -0.19 | -0.08 | **0.31** | **0.45** | **0.51** | -0.09 | 0.06 | -0.06 |

*Bild 5.9 Korrelationsmatrix mit signifikanten Koeffizienten*

Zunächst erfolgt die Wiedergabe der Korrelationsmatrix. In jeder Zelle wird der Pearsonsche Korrelationskoeffizient der beiden Variablen an Zeilen- und Spaltenanfang angezeigt. Die Hauptkomponentenanalyse baut auf der Korrelationsmatrix auf. Das ist wichtig, denn so lässt sich die Merkmalsreduktion vorhersagen. Je besser Merkmale korrelieren, desto wahrscheinlicher ist es, dass sie in einer Komponente zusammengefasst werden. Betrachten wir unsere Daten, können wir unter anderem konsternieren, dass es eine Merkmalsgruppe mit Bevölkerungsdichte, Stadt- und Landanteil oder eine weitere mit Stickstoff-, Phosphorfracht, Acker-, Weide- und Waldanteil sowie Bevölkerungsdichte gibt.

**Kommunalitäten**

| | Anfänglich | Extraktion |
|---|---|---|
| QSTAND | 1,000 | ,646 |
| N_T,A | 1,000 | ,742 |
| P_T,A | 1,000 | ,579 |
| AWALD | 1,000 | ,885 |
| AACKER | 1,000 | ,936 |
| AWEIDE | 1,000 | ,767 |
| ASTADT | 1,000 | ,577 |
| AWASSER | 1,000 | ,571 |
| AGLETSCH | 1,000 | ,710 |
| ATUNDRA | 1,000 | ,817 |
| BEVDICHT | 1,000 | ,862 |
| BEVSTADT | 1,000 | ,917 |
| BEVLAND | 1,000 | ,917 |
| SAG1990 | 1,000 | ,621 |

Extraktionsmethode: Hauptkomponentenanalyse.

*Bild 5.10 Kommunalitätentabelle des Beispiels*

Schauen wir zunächst auf die Kommunalitätentabelle (Bild 5.10)! Sie gibt an, wie gut jedes Ausgangsmerkmal durch das Faktorenmodell repräsentiert wird.

Die Varianz jedes Merkmals beträgt aufgrund der z-Standardisierung anfänglich 1 und wird durch die Extraktion verringert. Die Kommunalität der Variablen $P_{\_TA}$, $A_{STADT}$, $A_{WASSER}$ sind mit 0,5 - 0,6 zwar gering, aber mit Werten über 0,5 noch tolerierbar. Besonders gut werden die Variablen $A_{ACKER}$, BEVLAND und $BEV_{STADT}$ erklärt. Ihre Varianz geht im Modell fast überhaupt nicht verloren.

Addieren wir die Summe aller Kommunalitäten, erhalten wir die erklärte Gesamtvarianz des Faktorenmodells. Diese ist in der nächsten Tabelle (Bild 5.11) besser abzulesen.

**Erklärte Gesamtvarianz**

| | Anfängliche Eigenwerte | | | Summen von quadrierten Faktorladungen für Extraktion | | | Rotierte Summe der quadrierten Ladungen | | |
|---|---|---|---|---|---|---|---|---|---|
| Komponente | Gesamt | % der Varianz | Kumulierte % | Gesamt | % der Varianz | Kumulierte % | Gesamt | % der Varianz | Kumulierte % |
| 1 | 4,280 | 30,572 | 30,572 | 4,280 | 30,572 | 30,572 | 3,940 | 28,143 | 28,143 |
| 2 | 2,720 | 19,428 | 50,000 | 2,720 | 19,428 | 50,000 | 2,716 | 19,398 | 47,541 |
| 3 | 2,114 | 15,097 | 65,097 | 2,114 | 15,097 | 65,097 | 2,308 | 16,483 | 64,024 |
| 4 | 1,434 | 10,241 | 75,338 | 1,434 | 10,241 | 75,338 | 1,584 | 11,314 | 75,338 |
| 5 | ,941 | 6,723 | 82,061 | | | | | | |
| 6 | ,731 | 5,220 | 87,280 | | | | | | |
| 7 | ,576 | 4,116 | 91,396 | | | | | | |
| 8 | ,373 | 2,664 | 94,060 | | | | | | |
| 9 | ,302 | 2,155 | 96,214 | | | | | | |
| 10 | ,228 | 1,626 | 97,840 | | | | | | |
| 11 | ,188 | 1,340 | 99,180 | | | | | | |
| 12 | 7,049E-02 | ,504 | 99,684 | | | | | | |
| 13 | 4,428E-02 | ,316 | 100,000 | | | | | | |
| 14 | -1,91E-16 | -1,362E-15 | 100,000 | | | | | | |

Extraktionsmethode: Hauptkomponentenanalyse.

*Bild 5.11 Komponententabelle des Beispiels*

Die Tabelle der erklärten Gesamtvarianz listet die berechneten Komponenten und die extrahierten Faktoren auf. Unter „Anfängliche Eigenwerte“ erscheinen die Varianzkalkulationen aller Komponenten. Die erste Komponente erklärt mit einem Eigenwert von 4,282 mehr als das Vierfache eines Ausgangsmerkmals und damit 30,6% der Gesamtvarianz. Auch die Komponenten 2 bis 4 weisen eine Varianzzunahme mit Eigenwerten größer als 1 auf. Alle weiteren Komponenten besitzen dagegen eine

geringere Varianz. Da für SPSS das Eigenwertkriterium zur Auswahl der Faktoren voreingestellt ist, werden die ersten vier Komponenten für das Faktorenmodell ausgewählt. Gibt es ein besseres Auswahlkriterium?

In der vierten Spalte ist abzulesen, dass die Komponenten 1 bis 4 immerhin 75,3 % der anfänglichen Informationen erklären. Würde wir das Kaiserkriterium (siehe Box Seite 76) mit einer Schwelle 90% anwenden wollen, müssten wir allerdings 7 Komponenten in unser Modell aufnehmen, was die Merkmalsreduktion erheblich verringerte.

Für die dritte Variante, das Screekriterium, sehen wir uns den Screeplot an. Der „Haldenfuß" wird durch die fünfte Komponente gebildet.

Wir entscheiden uns letztendlich trotzdem für das Eigenwertkriterium, weil diese fünfte Komponente in unserem Beispiel schwer erklärt werden kann.

Die Aufgabe des Geographen muss die Bewertung und Veranschaulichung der Analyseergebnisse sein. Hier sollte er klare Vorteile zum reinen Statistiker besitzen, denn er weist das nötige Hintergrundwissen auf.

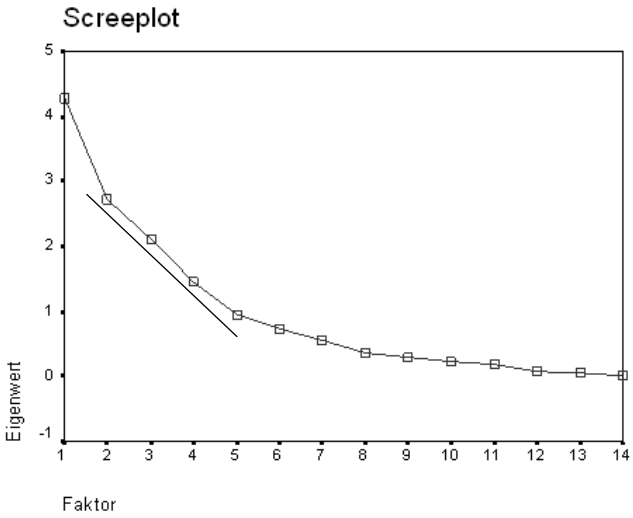



*Bild 5.12 Screeplot des Beispiels*

**Rotierte Komponentenmatrix[a]**

| | Komponente | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| AACKER | ,950 | | | |
| AWALD | -,904 | | | |
| AWEIDE | ,812 | | | |
| BEVDICHT | ,751 | ,525 | | |
| N_T,A | ,701 | | | ,464 |
| P_T,A | ,542 | | | ,524 |
| BEVSTADT | | ,949 | | |
| BEVLAND | | -,949 | | |
| ASTADT | | ,692 | | |
| ATUNDRA | | | ,884 | |
| AGLETSCH | | | ,842 | |
| SAG1990 | | | ,756 | |
| QSTAND | | | | ,754 |
| AWASSER | | | | -,678 |

Extraktionsmethode: Hauptkomponentenanalyse.
Rotationsmethode: Varimax mit Kaiser-Normalisierung.
a. Die Rotation ist in 5 Iterationen konvergiert.

*Bild 5.13 Rotierte Komponentenmatrix des Beispiels*

Unsere letzte Handlung wird nun die Bewertung der vier extrahierten Faktoren sein. Wir betrachten dazu die rotierte Komponentenmatrix (Bild 5.13).

**Erster Faktor:**

Dieser Faktor setzt sich aus den Ausgangsmerkmalen $A_{ACKER}$, $A_{WALD}$, $A_{WEIDE}$, $BEV_{DICHT}$, $N_{\_TA}$ und $P_{\_TA}$ zusammen. Dabei wird nur die Variable AWALD entgegengesetzt verknüpft, d.h. je größer der Wert für $A_{WALD}$ umso kleiner der Faktorwert.

Eine hohe Besetzung der restlichen Variablen ist Ursache und Wirkung einer hohen Emission von Nitrat und Phosphor im Einzugsgebiet. Wir nennen diesen Faktor also „Potenzial zu Nährstoffausträgen".

Bild 5.14 zeigt, dass die Gebiete mit hohem potentiellem Nährstoffeintrag vor allem im Süden des Ostseeeinzugsgebietes liegen, welche durch intensive Landwirtschaft und hohe Bevölkerungszahlen geprägt ist. Im Norden dominiert Forst- und extensive Landwirtschaft, die Bevölkerungsdichte und somit der Nährstoffeintrag ist hier gering.

**Zweiter Faktor:**

Dieser Faktor setzt sich aus den Ausgangsmerkmalen $BEV_{STADT}$, $BEV_{LAND}$ (mit negativem Vorzeichen), $A_{STADT}$, BEVDICHT zusammen. Er hat einen hohen Wert, wenn ein Einzugsgebiet viele Einwohner hat und ein großer Teil der Bevölkerung in Städten lebt. Der Faktor wird „Bevölkerung und Verstädterungsgrad“ benannt. Im Bild 5.15 sehen wir, dass die Verteilung der Faktorwert im Ostseeraum differenzierter ist als beim ersten Faktor. Hohe Werte findet man in den skandinavischen Hauptstadtregionen, die Bevölkerungsdichte und Verstädterung ist hier groß. Obwohl im hohen Norden eine geringe Bevölkerungsdichte angenommen werden kann, treten aber auch hier hohe Faktorwerte auf. Dies ist darauf zurückzuführen, dass anders als an der Ostseesüdküste, fast alle Bewohner in Städten leben. Das Leben auf dem Land ist wesentlich beschwerlicher.

**Dritter Faktor:**

In diesem Faktor werden die Ausgangsmerkmale $A_{TUNDRA}$, $A_{GLETSCH}$ und $SAG_{1990}$ zusammengefasst. Der Faktor besitzt einen hohen Wert, wenn das Einzugsgebiet teilweise vergletschert und mit Tundra bedeckt ist. Diese Eigenschaft besitzen vor allem die skandinavischen Gebirgseinzugsgebiete. Dementsprechend hoch ist auch der Speicherausbaugrad, denn diese Gebiete sind aufgrund höherer Reliefenergie für den Talsperrenbau geeignet. Doch im Bild 5.16 werden noch weitere, südliche Einzugsgebiete mit hohen Faktorwerten dargestellt. Offensichtlich ist hier der Speicherausbau der Gewässer vor allem in den südlichen Mittelgebirgen erfolgt. Der Faktor erhält den Namen „Potenzial zur Wasserkraftnutzung“.

**Vierter Faktor:**

Diesem Faktor ordnet die Analyse die Ausgangsvariablen N_TA und P_TA, sowie $Q_{STAND}$ und $A_{WASSER}$ (mit negativem Vorzeichen) zu. Eine Interpretation dieser Zusammenfassung erscheint gar nicht so einfach. Hohe Faktorwerte deuten auf hohe Abflussspenden, eine geringe Wasserfläche, eine hohe Stickstofffracht und eine etwas höhere Phosphorfracht hin.

Steigt die Abflussspende eines Einzugsgebiets, kann sich auch die Fracht der gelösten Stickstoff- und der partikulär verlagerten Phosphorionen erhöhen. Es befindet sich wenig Sedimentfallen auf dem Transportweg der Ionen, da außerdem die Seenfläche gering und die mittlere Fließgeschwindigkeit relativ hoch ist. Abbauprozesse für Stickstoff und die Absetzung des Phosphorgehaltes durch Sedimentation kommen also kaum in Gang (BRYDSTEN, 1990). Im Einzugsgebiet könnte also potentiell viel Nährstofffracht bis zum Gebietsauslass transportiert und dort gemessen werden. Der Faktor wird deswegen „Potenzial zu Nährstoffmobilisierung“ genannt.

Wenn die Einträge gering sind, muss dieses Potenzial übrigens nicht mit der tatsächlichen Fracht korrespondieren. Dies ist im Bild 5.17 bei den nördlichen Einzugsgebieten festzustellen.

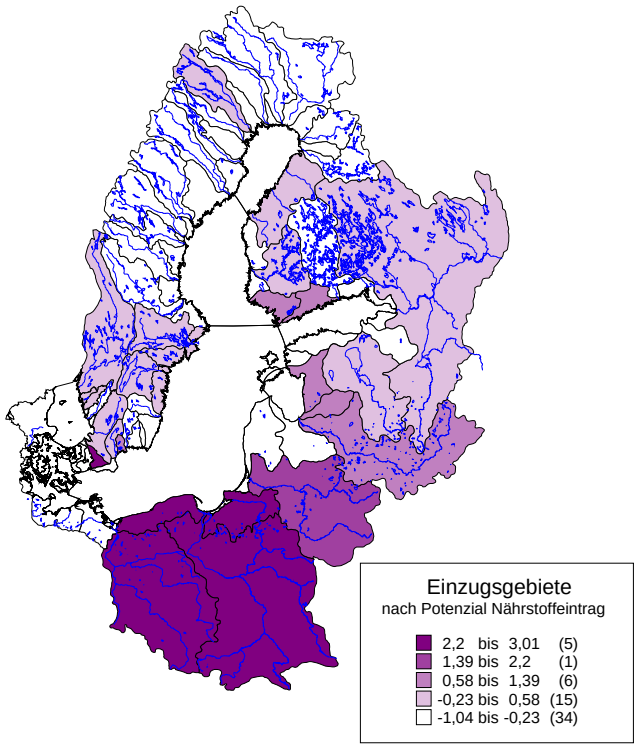



*Bild 5.14 Faktor I: Potenzial Nährstoffeintrag*

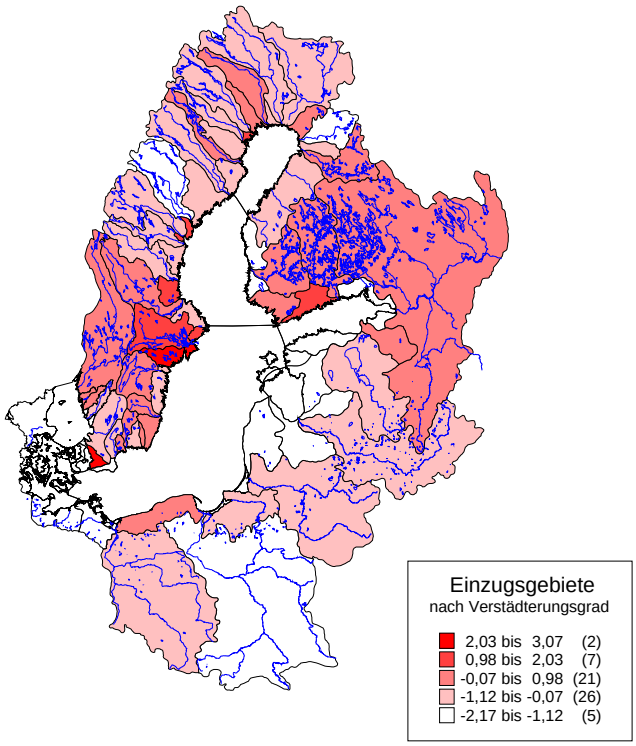



*Bild 5.15 Faktor II: Verstädterungsgrad*

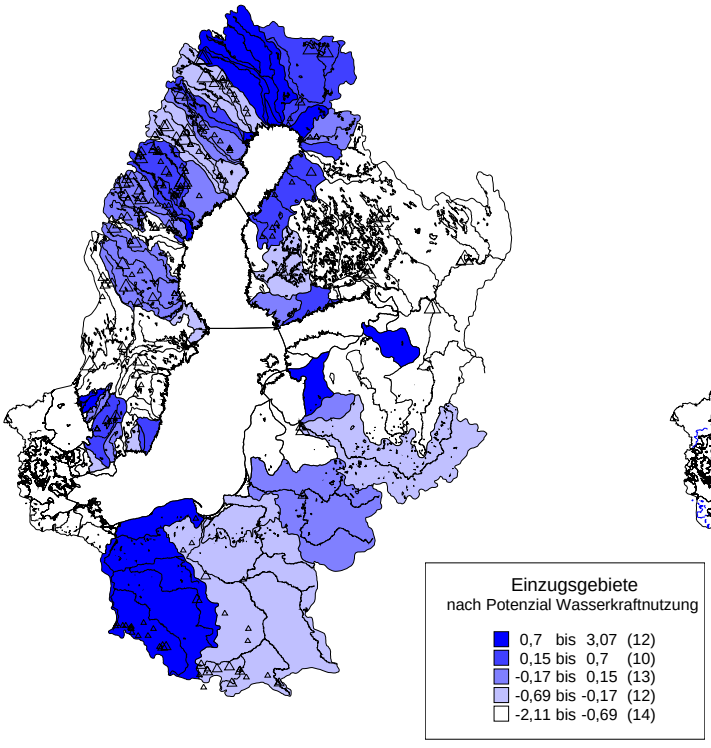



*Bild 5.16 Faktor III: Potenzial Wasserkraftnutzung*

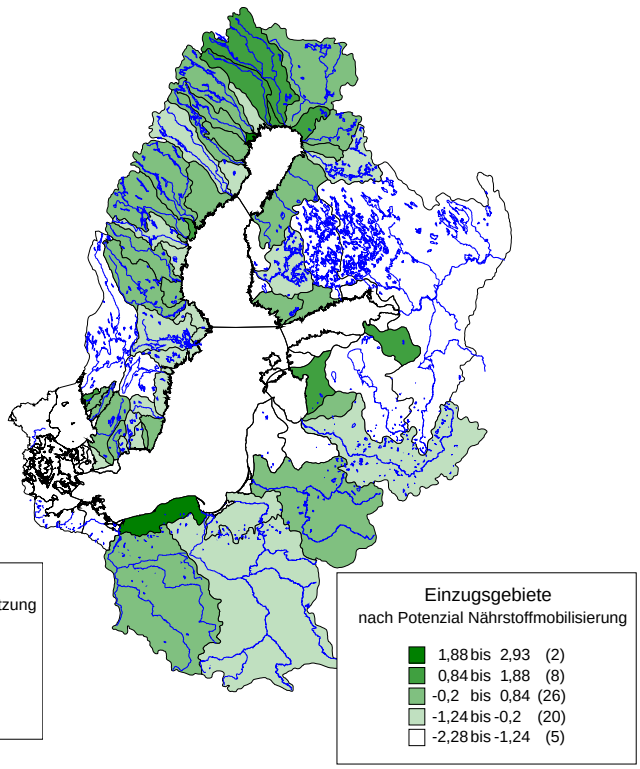



*Bild 5.17 Faktor IV: Potenzial Nährstoffmobilisierung*

# KAPITEL VI
# REGRESSIONSANALYSE

Mit der Systemsynthese schließen wir die Auswertung unseres Datensatzes verschiedener Merkmale der Einzugsgebiete der Ostsee ab. In den vorangegangenen Kapiteln ging es darum, sich einen Überblick über die Ausprägung der Variablen in verschieden Räumen zu verschaffen sowie Hinweise auf mögliche Zusammenhänge zwischen den Variablen zu erhalten. Jetzt soll der Versuch unternommen werden, Prozesse oder den dynamischen Verlauf eines Datensatzes mit Hilfe steuernder Größen des Raumes zu berechnen, also zueinander in einen möglichst exakt beschreibbaren Zusammenhang zu bringen.

Bildlich gesprochen hatten wir in der Systemanalyse die im Datensatz enthaltenen Variablen selbst, also innerhalb von „Boxen" analysiert. Mit der Systemidentifikation hatten wir mögliche Verbindungen zwischen diesen „Boxen" als Linien markiert. Nun soll die Systemsynthese diese Linie durch parametrisierte Pfeile ersetzen. Damit können wir Aussagen über die Richtung dieser Zusammenhänge treffen und diese Abhängigkeiten in quantitative Modelle überführen. Um welchen Betrag muss sich eine Variable ändern, damit sie eine festgelegte Änderung einer anderen Variablen bewirkt?

Ziel ist es, parametrisierte und regional gültige Modelle als Abbilder der natürlichen Systeme zu erstellen. Mit diesen Modellen soll es möglich sein, Szenarien zu berechnen, welche den Einfluss von Änderungen in den unabhängigen, also steuernden Variablen untersuchen. (WAS WÄRE WENN? – ANALYSEN). Außerdem sollen Prognosen für zukünftig zu erwartende Veränderungen abgegeben werden.

■ Eine solche Was Wäre Wenn – Analyse wollen wir im zweiten Kapitel besprechen. Wir werden den Zusammenhang zwischen den Phosphorfrachten und dem Speicherausbaugrad der Einzugsgebiete durch Talsperren in ein quantitatives Modell fassen. Danach wollen wir die Talsperren kurzzeitig aus dem Modell „virtuell" entfernen und die, unter der Bedingung nicht vorhandener Talsperren, geänderten Phosphorfrachten in den betreffenden Einzugsgebieten schätzen.

# 1 Lineare Einfachregression

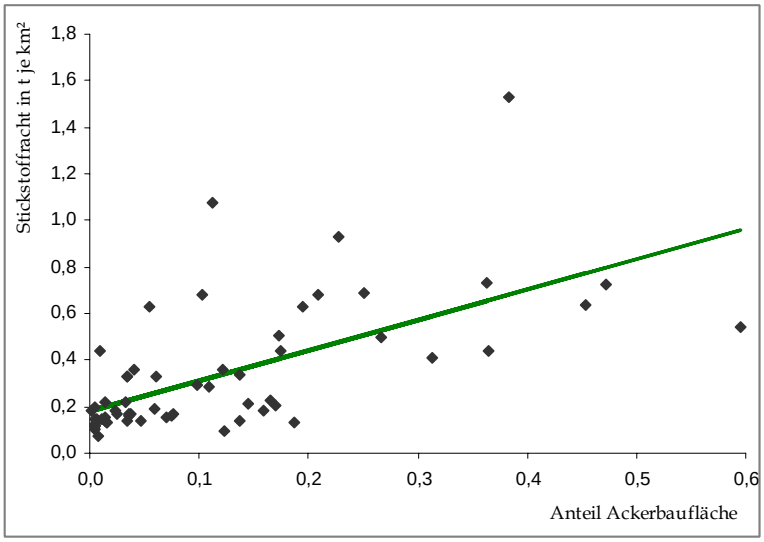



*Bild 6.1: Scatterplot und vermuteter linearer Zusammenhang zwischen dem Ackerflächenanteil und den Stickstofffrachten*

Die Partielle Korrelationsanalyse wie auch die Analyse der Faktorladungen aus der Faktorenanalyse hatten uns gezeigt, dass es einen möglichen Zusammenhang der Nährstofffrachten und dem Anteil der Ackerfläche im Ostseeeinzugsgebiet gibt. Bild 6.1 stellt diesen vermuteten Zusammenhang nochmals mittels eines Scatterplots dar.

Lässt sich die Abhängigkeit der Stickstofffrachten vom Anteil der Ackerfläche je Flusseinzugsgebiet linear beschreiben? Ein solcher Zusammenhang müsste durch eine Gerade in der Punktwolke dargestellt werden können!

GGA_Regression\Uebung_1.xls Registerkarte Linear

■ Versuchen Sie, durch Probieren eine Gerade analog y = mx + n zu finden, welche die Daten möglichst gut repräsentiert. Hierzu erstellen Sie in Spalte D (mit der Überschrift 1. Schätzung für N_t,A) eine Formel, welche auf ein Feld mit veränderbaren m und n verweist. Wir schreiben dabei folgende Formel in die 1. Zelle: `=F$1*B3+F$2` (Bild 6.2).

F1 verweist dabei auf den Wert für m, F2 auf den Wert für n. Das $-Zeichen bewirkt in EXCEL, dass diese Bezüge auch beim Kopieren der Formel in tiefere Zeilen bestehen bleiben. Nachdem wir diese Formel nun auch in die darunter liegenden Zeilen kopiert haben, ändern wir m und n schrittweise durch Ausprobieren!

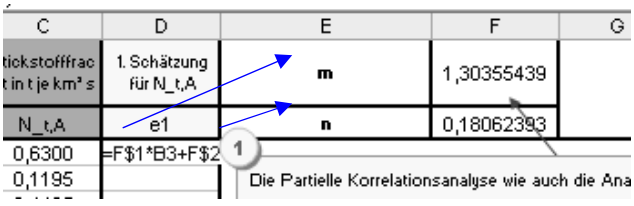



*Bild 6.2: Uebung_1.xls – Eintrag der Schätzformel für den linearen Zusammenhang*

Dabei orientieren wir uns an der Gerade in dem beigefügten Scatterplot in Uebung_1.xls (siehe auch Bild 6.1) und versuchen optisch eine möglichst gut angepasste Gerade zu finden. Die Gleichung dieser Geraden soll den von uns gesuchten Zusammenhang zwischen dem Ackerflächenanteil (x) und den Stickstofffrachten (y) repräsentieren. Nach Bild 6.2 könnten wir sie mit:

*Stickstofffrachten = 1,3 * Ackerflächenanteil + 0,18*

notieren. Wir haben somit eine Formel gefunden, mit dessen Hilfe wir den Modellzusammenhang zwischen den Stickstofffrachten und dem Ackerflächenanteil quantifizieren können: Die Stickstofffrachten könnten abgeleitet werden, wenn der Ackerflächenanteil eines Einzugsgebietes mit 1,3 multipliziert wird und hierzu ein Wert von 0,18 addiert wird.

Der gerade beschriebene Weg stellt einen Versuch dar, eine Gerade als lineare Anpassung an die Punktwolke für den Zusammenhang zwischen Ackerflächenanteil und den Stickstofffrachten zu finden. Anders als bei der Korrelationsanalyse lassen sich übrigens x und y nicht vertauschen. Wir haben es stets mit einer abhängigen Y-Variable zu tun, deren Werte über einen mathematischen Zusammenhang mit der unabhängigen X-Variable erklärt werden sollen.

Bis jetzt können wir allerdings überhaupt nicht einschätzen, wie gut unsere gefundene Gerade den möglichen Zusammenhang überhaupt wiedergibt. Daher wollen wir die Anpassung der Gerade und damit die Schätzung der linearen Abhängigkeit weiter optimieren.

**Korrelationsanalyse versus Regressionsanalyse?**

Die Regressionsanalyse versucht wie die Korrelationsanalyse einen Zusammenhang zwischen mehreren Variablen aufzudecken. Nicht ohne Grund ist die Korrelationsanalyse jedoch bereits im Rahmen der Systemidentifikation, und damit vor der Regressionsanalyse, eingeführt worden.
Formal kann zwischen allen möglichen Variablen ein mathematischer Zusammenhang abgeleitet werden. Vor der Ableitung eines mathematischen Zusammenhanges (eines Regressionsmodells) sollte jedoch unbedingt die statistische und inhaltliche Überprüfung des Zusammenhanges stehen.
Den Zusammenhang zwischen den Ackerflächenanteilen und den Stickstofffrachten hatten wir im Kapitel IV – Korrelationsanalyse - bereits untersucht. Wir hatten zwar keinen exzellenten Zusammenhang finden können, konnten jedoch festhalten, dass aufgrund der inhaltlicher Relevanz und der ersten Ergebnisse der Partiellen Korrelationsanalyse eine hinreichende Abhängigkeit zwischen beiden Variablen besteht.

## 1.1 Methode der kleinsten quadratischen Abstände

Die Anpassung für m und n können wir sinnvollerweise so optimieren, dass jene Ausgleichsgerade gefunden wird, bei der die Summe aller Abweichungen der geschätzten Werte von den beobachteten Werten möglichst klein ist. Es müssen also zunächst alle Abweichungen (Distanzen) zwischen der Geraden berechnet werden. Anschließend werden m und n so gewählt, dass diese Abweichungen möglichst klein werden.

Der Statistiker zieht hierbei sogar die Summe der quadratischen Abweichungen vor, dabei werden besonders hohe Abweichungen entsprechend stärker gewichtet. Diese Ab-

weichungen sollen ja dann durch eine optimale Kurve auch entsprechend minimiert werden. Die Summe der quadratischen Abweichungen lässt sich gemäß folgender Formel berechnen:

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

Gl. 1: Summe der quadratischen Abweichungen für den Zusammenhang zwischen y und x

GGA_Regression\Uebung_1.xls Registerkarte Linear

■ Berechnen Sie in Spalte J die quadratischen Abweichungen zwischen Ihrer geschätzten Kurve und den Originalwerten! Tragen Sie hierzu in Zelle J3 die Formel `=(D3-C3)^2` ein. Damit subtrahieren wir zunächst den Originalwert der Stickstofffracht vom geschätzten Wert in Spalte D und quadrieren diese Differenz anschließend. Kopieren Sie die Formel dann nach unten bis zur Zeile 63. In Zelle I2 finden Sie eine Summenfunktion eingetragen, welche die von Ihnen soeben berechneten quadratischen Abstände summiert. Dieser Wert ist die Summe der quadratischen Abweichungen zwischen der von uns geschätzten Geraden in Spalte D und den Originalwerten der Stickstofffrachten.

Wir können jetzt durch „Ausprobieren“ von neuen Werten in m und n erreichen, dass die Abweichungen noch geringer werden.

Mit der Summe der quadratischen Abweichungen haben wir ein objektives Maß erhalten, mit welchem wir eine Gerade (oder auch Kurven) möglichst optimal an eine Punktwolke anpassen können. Lediglich das „Ausprobieren“ von günstigen Werten sollte noch optimiert werden. Für dieses "Ausprobieren" stellt EXCEL ein nützliches Instrument zur Verfügung, den Solver. Mit Hilfe dieses Tools können Werte so verändert werden, dass ein entsprechender Zielwert möglichst gut erreicht wird.

GGA_Regression\Uebung_1.xls Registerkarte Linear

■ Starten Sie den Solver über das Menü Extras/Solver (wenn der Eintrag dort nicht vorhanden ist, müssen Sie den Solver zunächst über den Menüeintrag Add-Ins aktivieren)!

Wir setzen die Einträge für den Solver so, wie es Bild 6.3 zeigt. Als Zielzelle geben wir die von Ihnen berechnete Summe der quadratischen Abweichungen an. Diese sollen möglichst klein werden, entsprechend wählen wir für Zielwert Minimum. Veränderbar (Veränderbare Zellen) sind die beiden Werte für m und n. Diese sollen mit Hilfe des Solvers so gewählt werden, dass die Summe der quadratischen Abweichungen ein Minimum erreicht.

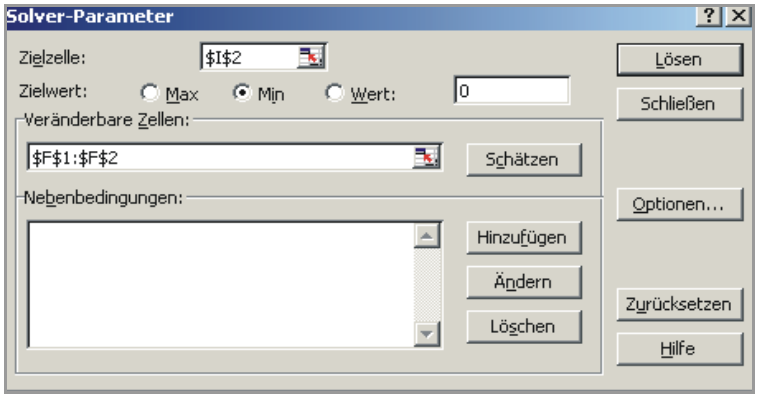



*Bild 6.3: Uebung_1.xls – Minimierung der quadratischen Abweichungen mit Hilfe des Solvers*

Das Tool Solver wird nach Betätigen der Lösen-Taste versuchen, iterativ eine Lösung für das angegebene Problem zu finden. Dazu werden die Werte der veränderbaren Zellen so lange variiert, bis die Zielzelle ein Minimum erreicht (wie entsprechend bei Zielwert angegeben, Bild 6.3). Hier wird also genauso wird durch "Ausprobieren" (diesmal eben nur durch den Rechner wesentlich schneller) eine numerische Lösung gefunden. Letztlich ist die später verwendete analytische Lösung des Statistikers keine andere, auch sie orientiert sich wieder an einer Anpassung mit möglichst kleiner Summe der quadratischen Abweichungen.

Mit einem Wert von 2,75 haben wir die geringste Summe quadratischer Abweichungen zwischen der Schätzgeraden und den Originalwerten der Stickstofffrachten gefunden. Gleichzeitig wurde in dem Übungsbeispiel in Zelle H2 noch die Korrelation zwischen der Schätzgeraden und den Originalwerten der Stickstofffrachten mit ausgegeben. Diese Korrelation ist mit 0,63 als mittelgut zu bezeichnen. Allerdings wird die Schätzgerade in der Statistik nicht mit dem Korrelationskoeffizienten bewertet. Wie schon im Kapitel Korrelation angesprochen, verwendet man hier besser das Quadrat des Korrelationskoeffizienten, also das Bestimmtheitsmaß. Der von uns gefundene lineare Zusammenhang:

*Stickstofffrachten = 1,3 * Ackerflächenanteil + 0,18*

kann also 40% der Varianz der Stickstofffrachten erklären. Hier finden wir wieder genau jenen Wert, welchen wir bereits bei der Korrelationsanalyse erhalten hatten. Damit wird deutlich, dass der einfache lineare Korrelationskoeffizient ebenfalls nur die Güte eines linearen Zusammenhangs bewertet hatte.

Wir müssen diese nicht besonders gute Anpassung im Gedächtnis behalten, wenn wir die Ergebnisse der weiteren regressionsanalytischen Untersuchung interpretieren wollen. Wenn wir fortführende Aussagen mit Hilfe der hier aufgestellten Abhängigkeit der Stickstofffrachten von den Ackerflächenanteilen treffen wollen, sind diese stets mit einer größeren Unsicherheit behaftet. Die Ackerflächenanteile erklären schlicht nur einen Anteil von höchstens 40% der Ausprägung der Stickstofffrachten in den Einzugsgebieten der Ostsee.

## 1.2 Lineare Einfachregression mit EXCEL

Indem wir eine optimale lineare Ausgleichsgerade für eine Punktwolke gefunden haben, haben wir uns intuitiv das Konzept der linearen Einfachregression erschlossen. Bei der Anwendung der Regressionsanalyse - im Wortsinn bedeutet Regression etwa soviel wie „rückwärtiges Erschließen" – wird ebenso die wahrscheinlich beste Anpassung einer mathematisch zu definierenden Kurve an eine Punktwolke gesucht. Die lineare Einfachregression „erschließt" entsprechend eine Gerade.

GGA_Regression\Uebung_1.xls Registerkarte LinRegression

■ Berechnen Sie mit der EXCEL Tabellenfunktion RGP die Kennwerte einer linearen Regression! Positionieren Sie den Ausgabebereich für RGP dazu in die Zellen H6:I9 (Bild 6.4, alle 2x4 Zellen markieren)! Achtung: RGP ist eine Matrixfunktion und wird mit den Tasten STRG + Umschalt + Enter abgeschlossen! Mit KONSTANTE 1 wird der Schnittpunkt mit der Y-Achse berechnet, STATS ermöglicht die Ausgabe der weiter unten aufgeführten statistischen Kennwerte zum Regressionsmodell.

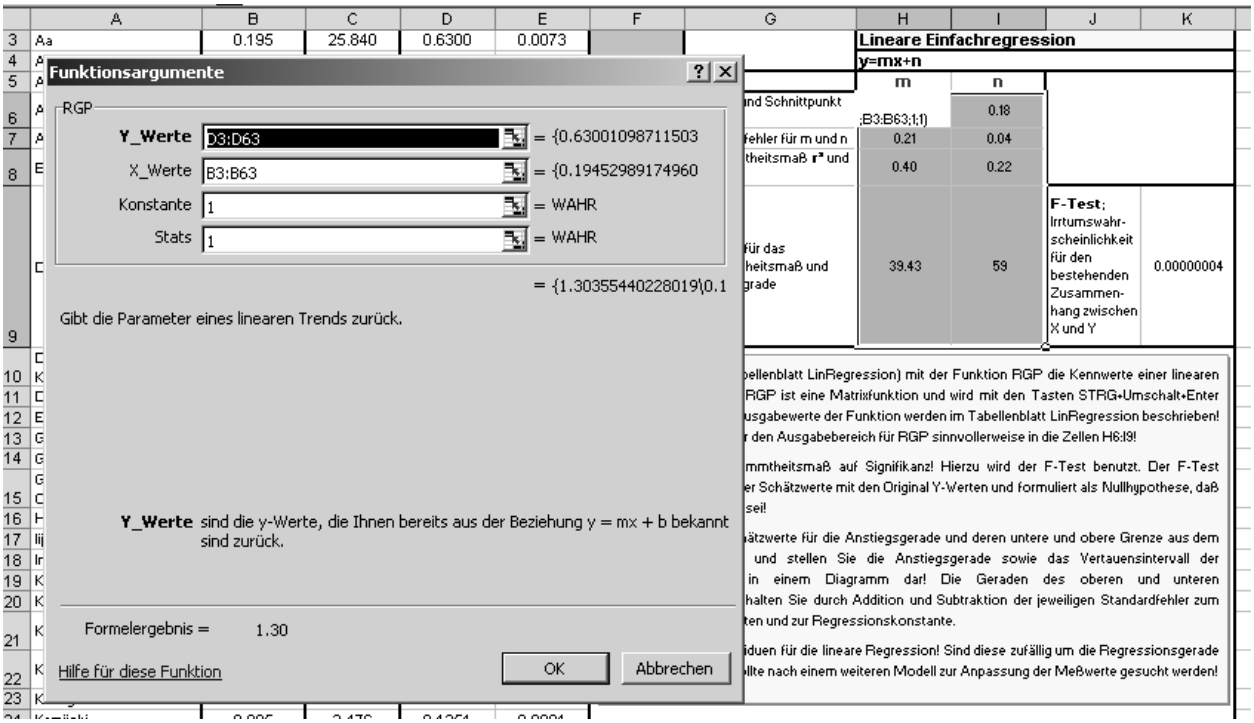



*Bild 6.4: Uebung_1.xls – Anwendung der Funktion RGP für die lineare Einfachregression*

Die Funktion RGP führt innerhalb von EXCEL eine Lineare Regressionsanalyse gemäß der Formel $y = mx + n$ durch. Dabei werden in der ersten Zeile der Anstieg der Funktion ($m$) und der Schnittpunkt mit der Y-Achse ($n$) ausgegeben. Die zweite Zeile enthält die Angaben für die Standardfehler *se(m)* und *se(n)* des Anstiegs und des Schnittpunktes mit der Y-Achse. In der dritten Zeile wird schließlich das Bestimmtheitsmaß für die Schätzung des Regressionsmodells und der Standardfehler für die Schätzung Y-Werte *se(y)* angegeben. Die

vierte und letzte Zeile gibt die Werte aus, mit deren Hilfe die Güte des Regressionsmodells über einen F-Test bewertet werden kann.

**Angabe der Standardfehler im Regressionsmodell**

Wenngleich die Regressionsanalyse eine „exakte" Formel für einen Zusammenhang liefert, stellt sie doch nur eine Schätzung dieses Zusammenhanges dar. Der Verlauf der Regressionsgerade kann aber mit einer bestimmten Wahrscheinlichkeit innerhalb eines anzugebenden Intervalls erwartet werden.

Dieses Intervall wird meistens in Form der Standardfehler angegeben. Mit Hilfe des Standardfehlers *se(m)* für den Anstieg *m* wird die Standardabweichung des Anstiegs berechnet. Durch den Standardfehler *se(n)* wird zusätzlich die Standardabweichung des Schnittpunktes mit der y-Achse angegeben. Die Standardabweichung (zum Begriff siehe auch Kapitel I) ist dadurch charakterisiert, dass in dem durch sie überdeckten Bereich 68% aller Fälle liegen. Damit kann jenes Intervall konstruiert werden, in welchem die Regressionsgerade mit einer Sicherheit von 68% verlaufen wird.

Durch die beiden Formeln

$$Us = [m+se(m)] + [n-se(n)] \qquad \text{Gl. 2}$$

$$Os = [m-se(m)] + [n+se(n)] \qquad \text{Gl. 3}$$

können die Grenzgeraden des Intervalls konstruiert werden, innerhalb dessen die Regressionsgeraden mit einer Wahrscheinlichkeit von 68% liegen wird (Bild 6.5). Diese beiden Grenzgeraden kreuzen sich im Punkt der Mittelwerte von *x* und *y*. Es gibt also einen Punkt der Regressionsgeraden, an welchem die Schätzung als sehr sicher anzusehen ist. Dieser liegt entsprechend an der Stelle $(\overline{x}, \overline{y})$ und ist als eine Art Drehpunkt aufzufassen. Links und rechts des Drehpunktes wird der wahre Verlauf der Regressionsgerade deutlich unsicherer.

## GGA_Regression\Uebung_1.xls Registerkarte LinRegression

■ Berechnen Sie die Schätzwerte für die Regressionsgerade und für die untere und obere Grenzgerade des Intervalls für den Standardfehler des Anstiegs! Mit Hilfe des geschätzten *m* und *n* (Bild 6.4) tragen Sie dazu in Spalte M die Formel `=H$6*B3+I$6` ein und kopieren diese nach unten! Für die Grenzgeraden nutzen Sie den Standardfehler für *m* und *n* aus den Ergebnissen der Funkton `RGP`. Stellen Sie die Regressionsgerade sowie das Vertrauensintervall von deren Anstieg in einem Diagramm dar (Bild 6.5)! Die untere Grenzgerade (*Us*) erhalten Sie gemäß der Formel *Us = [m+se(m)] + [n-se(n)]*, die obere Grenzgerade (*Os*) nach *Os = [m-se(m)] + [n+se(n)]*. Damit lautet die Schätzgleichung für die untere Grenzgerade des Intervalls für den Standardfehler *y = (1.3-0,21)x+(0.18+0,04)*, die für das obere Grenzgerade *y = (1.3+0,21-)x+(0.18-0,04)*.

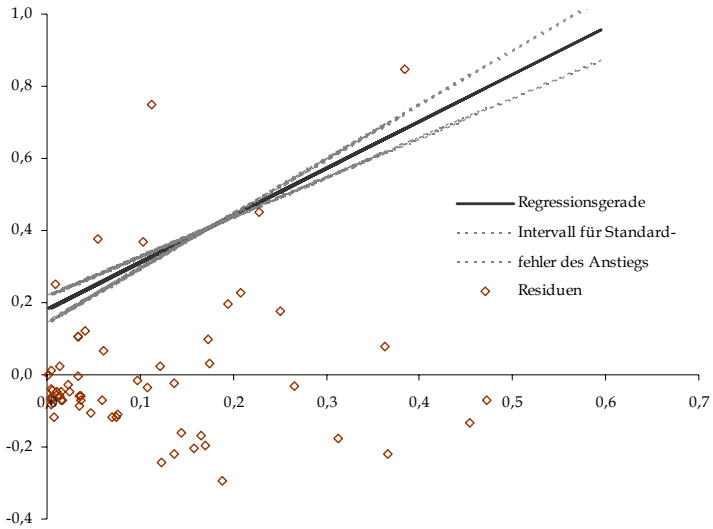



*Bild 6.5: Uebung_1.xls –Regressionsgerade für die Regression der Stickstofffrachten. Zusätzlich sind das Intervall für den Standardfehler des Anstiegs sowie die Resiuduen dargestellt.*

Nachdem wir uns also die Intervallgrenzen für die lineare Regression der Stickstofffrachten aus den Ackerflächenanteilen angeschaut haben, wollen wir nochmals das Bestimmtheitsmaß betrachten. Dieses widerspiegelt ja die Güte zwischen der Schätzgerade und den Originalwerten der Y-Achse. Hierzu wird ein F-Test verwendet. Der F-Test vergleicht die Varianz der Schätzwerte mit den Original Y-Werten und formuliert als Nullhypothese, dass die Varianz nicht gleich sei.

Zu dem ermittelten F-Wert aus der Ausgabe von `RGP` kann dabei das Quantil der F-Funktion ermittelt werden. Dieses Quantil gibt die Irrtumswahrscheinlichkeit an, mit welcher die anhand der Regressionsgleichung geschätzten Werte (in unserem Beispiel die in Spalte M geschätzten Stickstofffrachten) nicht dieselbe Varianz wie die Originalwerte aufweisen. Die Quantile der F-Funktion können in EXCEL wiederum mit einer Tabellenfunktion abgefragt werden. Dazu wird die Funktion `FVERT` verwendet. Die Anzahl der Freiheitsgrade wird dabei von der Funktion `RGP` mit ausgegeben.

GGA_Regression\Uebung_1.xls Registerkarte LinRegression

■ Berechnen Sie das Quantil der F-Verteilung mit der Funktion FVERT in Zelle K9 und testen Sie das Bestimmtheitsmaß auf Signifikanz! Bei einer Einfachregression ist für die Funktion die Zahl der Freiheitsgrade1 = 1. Der Parameter X und die Zahl der Freiheitsgrade2 wurde durch die Funktion RGP in Zelle H9 und I9 ausgegeben.

Die Irrtumswahrscheinlichkeit für eine korrekte Aussage des Bestimmtheitsmaßes ist mit 0.00000004 (siehe auch Bild 6.4) verschwindend gering, damit kann das Bestimmtheitsmaß als korrekt angesehen werden.

Der F-Test für das Bestimmtheitsmaß darf allerdings nicht darüber hinwegtäuschen, dass das Bestimmtheitsmaß selbst dem Regressionsmodell in unserem Beispiel (Uebung_1.xls) kein allzu großes Erklärungspotenzial attestiert. Auch wenn das Bestimmtheitsmaß sehr sicher auf gleichen Varianzen in den Original- und Schätzwerten aufbaut, sagt sein Wert von 0,4 doch nur, dass im Regressionsmodell lediglich 40% der Gesamtvarianz der Stickstofffrachten erklärt werden können!

## 1.3 Residuenanalyse

Wir haben nun die Güte des von uns erstellten Regressionsmodells hinsichtlich der Standardfehler und des Bestimmtheitsmaßes einschätzen können. Um ein Regressionsmodell abschließend zu bewerten, sollten stets auch die Residuen betrachtet werden. Die Residuen sind die Differenzen zwischen den geschätzten Werten des Regressionsmodells und den originalen Y-Werten. Sie stellen genau jene Differenzen zu den Y-Werten dar, welche durch das Regressionsmodell noch nicht erklärt werden. Damit repräsentieren Sie den Gesamtfehler des Regressionsmodells.

Die absolute Größe dieses Fehlers ist bereits dadurch minimiert worden, indem die Regressionsgleichung gemäß der Methode der kleinsten quadratischen Abstände optimiert wurde. Für die weitere Analyse ist jedoch auch die Streuung der Residuen interessant.

**Analyse der Residuen im Regressionsmodell**

Die Residuen eines Regressionsmodells stellen vereinfacht den Gesamtfehler des Regressionsmodells dar. Ein Fehler sollte stets zufällig streuen. Demzufolge sollten bei einem guten Regressionsmodell die Residuen auch zufällig um die die Schätzgerade streuen. Diese Bedingung kann dadurch gewährleistet werden, indem die Residuen normalverteilt sein sollten.

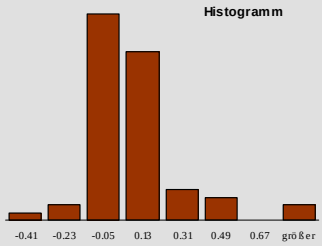



*Bild 6.6: Histogramm der Residuen aus Ubung_1.xls*

Bild 6.6 zeigt das Histogramm für die Residuen aus Uebung_1.xls. Deren Verlauf ist zudem in Bild 6.5 abgebildet. Das Histogramm widerspiegelt hinreichend eine Normalverteilung, es existieren aber einige sehr große Residuen. Besonders in diesen Fällen ist die Regressionsgerade schlechter als Modell geeignet.

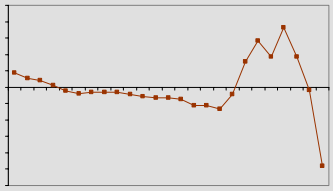


*Bild 6.7: Residuen für eine lineare Regression in Uebung_3.xls*

Interessant sind zudem Fälle, in welchen die Residuen zum Beispiel in einem S-förmigen Verlauf um die Regressionsgerade streuen. Bild 6.7 zeigt einen Verlauf von Residuen eines linearen Regressionsmodells (Datenbeispiel aus Uebung_3.xls). Dieser Verlauf zeigt eine systematische Abweichung der Residuen, offensichtlich repräsentiert das angenommene lineare Regressionsmodell den funktionalen Zusammenhang nicht korrekt. Den Daten aus Uebung_3.xls kann demzufolge auch ein logistisches Regressionsmodell besser angepasst werden. Neben der Linearen Regressionsanalyse werden deshalb in Kapitel 3 (Nichtlineare Regression) noch weitere Regressionsmodelle zu besprechen sein.

# 2 Systemsynthese: Erstellen von Szenarien mit Hilfe von regressionsanalytisch parametrisierten Modellen

Bislang haben wir uns intensiver damit auseinandergesetzt, wie gut eine lineare Funktion den Zusammenhang zwischen zwei verschiedenen Variablen wiedergeben kann. Anhand des Zusammenhanges zwischen dem Ackerflächenanteil und den Stickstofffrachten, haben wir festgestellt, dass sich ein Regressionsmodell für die zu untersuchenden Variablen recht leicht finden lässt. Allerdings muss ein solches Regressionsmodell auch anschließend hinsichtlich seiner Aussagefähigkeit bewertet werden. Ein Maß hierfür ist sicherlich das erhaltene Bestimmtheitsmaß, weitere Möglichkeiten haben wir kennen gelernt.

Wenn eine schlechte Anpassung vorliegt, kann dies ein Hinweis auf ein anderes als ein lineares Regressionsmodell sein. In Kapitel 3 werden deswegen noch weitere Regressionsmodelle vorgestellt. Mit diesen lassen sich auch nichtlineare Zusammenhänge anpassen.

Ein weiterer Grund für einen zunächst nur schlecht auffindbaren Zusammenhang können störende Einflüsse in den Daten selbst sein. Damit kann ein vorhandener Zusammenhang eventuell nur schwer erkannt werden, weil dieser durch zusätzliche unbekannte Prozesse verdeckt wird. In solchen Fällen kann es daher wieder sinnvoll sein, den Einfluss einiger

Variablen quasi-konstant zu halten. Damit wollen wir wieder, wie bereits in Kapitel IV mit dem partiellen Korrelationskoeffizienten, „experimentelle“ Bedingungen simulieren.

Unser Hauptaugenmerk lag bislang darauf, eventuelle Zusammenhänge zwischen verschiedenen Variablen und den Nährstofffrachten der Flüsse im Ostseeeinzugsgebiet zu erkennen.

Während der Clusteranalyse (Kapitel III) war ein Cluster mit zahlreichen talsperrenbeeinflussten Flüssen und geringen Nährstofffrachten aufgefallen. Der hohe Anteil an Talsperren drückte sich durch hohe Speicherausbaugrade (Anteil des Volumens der Talsperren am gesamten Jahresabfluss) aus. Hinter diesem eigentlich schwachen Zusammenhang verbirgt sich die Steuerung des Phosphortransportes durch Spitzenabflüsse. Phosphor wird in Flüssen vorwiegend über den sedimentären Pfad transportiert. Je stärker also die Abflussschwankungen, umso eher wird Phosphor in einem Fluss „fortgespült“.

Den Anteil von Abflussschwankungen am Jahresabfluss kann man über die Abflussvariabilität messen. Diese setzt die innerjährlichen Schwankungen des Abflusses zur Summe des Jahresabflusses ins Verhältnis. Da Talsperren durch ihre dämpfende Wirkung die Abflussvariabilität senken, wird deren Wirkung möglicherweise auch den Transport von Phosphor und damit die Phosphorfrachten beeinflussen.

Wir finden in unserem Datensatz Angaben zur Variabilität der Abflüsse vor und nach dem Bau von Talsperren. Läst sich für die Einzugsgebiete der Ostsee ein mathematischer Zusammenhang zwischen der Höhe der Phosphorfrachten und der Intensität der Talsperrenbewirtschaftung auffinden? Die Intensität der Talsperrenbewirtschaftung soll dabei über die Abflussvariabilität ausgedrückt werden.

Wenn dies gelingt, sollen mit einem solchen Modell die Phosphorfrachten im Falle eines nicht vorhandenen Einflusses von Talsperren geschätzt werden!

🖫 GGA_Regression\Uebung_2xls

■ Versuchen Sie, ein lineares Regressionsmodell zwischen Abflussvariabilität (Q-Variabilität) und den Phosphorfrachten zu erstellen. Gelingt dies zufriedenstellend? Wenn nicht, warum schlägt der Versuch fehl? Welche Steuergrößen sind außer Acht gelassen worden, spielen aber dennoch eine Rolle? Wie kann man aber dennoch versuchen, einen Zusammenhang zwischen Phosphorfrachten und Abflussvariabilität zu ergründen?

Die zunächst schlechte Korrelation zwischen der Abflussvariabilität und den Phosphorfrachten ist ein Ergebnis des wesentlich stärkeren Einflusses der Ackerflächenanteile und der Bevölkerungsdichte auf die Höhe der Phosphorfrachten. Wir müssen also die Wirkung dieser beiden Einflüsse konstant halten. Damit kann deren störende Wirkung auf die eigentlich interessierende Fragestellung verringert werden. Mit einem solchen Ansatz kann versucht werden, „experimentelle“ Bedingungen nachzustellen.

## 🖫 GGA_Regression\Uebung_2xls

■ Sortieren Sie die beiden Variablen Ackerfläche und Besiedlungsdichte in aufsteigender Reihenfolge. Betrachten Sie die beiden Größen jeweils in einem Diagramm als ScreePlot (Bild 6.8)!

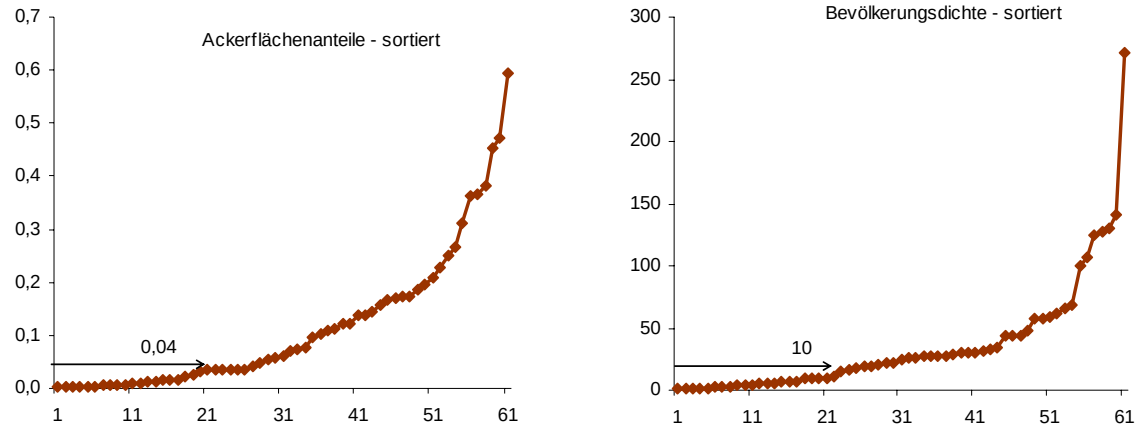



*Bild 6.8: Screeplots für die aufsteigend sortierten Variablen Bevölkerungsdichte und Ackerflächenanteil. Die Ackerflächenanteile können ab Werten < 0,04, die Bevölkerungsdichten ab Werten < 10 als annähernd konstant gelten.*

Können Sie durch Filtern einen Datensatz abtrennen, in welchem die beiden Steuergrößen Ackerflächenanteil und Bevölkerungsdichte annähernd konstant sind? Filtern Sie mit Hilfe der Autofilterfunktion von EXCEL alle Einzugsgebiete mit Ackerflächenanteilen < 0,04 und Bevölkerungsdichten < 10 ab! (Hinweise, wie Sie den Autofilter verwenden und Schwellenwerte abfiltern, finden Sie in Kapitel III.1)

Verwenden Sie für ein neuerliches Regressionsmodell schließlich nur jene Einzugsgebiete, welche Ackerflächenanteile und Bevölkerungsdichten unterhalb der gefundenen Trennwerte aufweisen! Da Sie den Einfluss unterschiedlicher Abflussvariabilitäten auf die Phosphorfrachten untersuchen wollen, könnten Sie in einem zweiten Schritt schließlich nur noch jene Einzugsgebiete verwenden, welche Talsperren aufweisen (Speicherausbaugrad größer 0)!

Zu welcher Aussage kommen Sie mit diesem Regressionsmodell?

Können Sie damit abschätzen, wie groß die Phosphorfrachten in die Ostsee aus diesen Einzugsgebieten ohne das Vorhandensein von Talsperren wären? Um wieviel Prozent würden sich die Nährstoffzuflüsse bei Phosphor erhöhen?

In 🖫 GGA_Regression\Uebung_2xls wurde mit Hilfe der Funktion RGP ein lineares Regressionsmodell entwickelt, welches allerdings auf nur 11 beobachteten Einzugsgebieten fußte (Bild 6.9). Wir mussten aber, um die Daten für die gesuchte Fragestellung adäquat zu extrahieren, zahlreiche der anderen Einzugsgebiete abfiltern. Entsprechend stellen die verbliebenen Fälle die für uns optimale Datenlage dar.

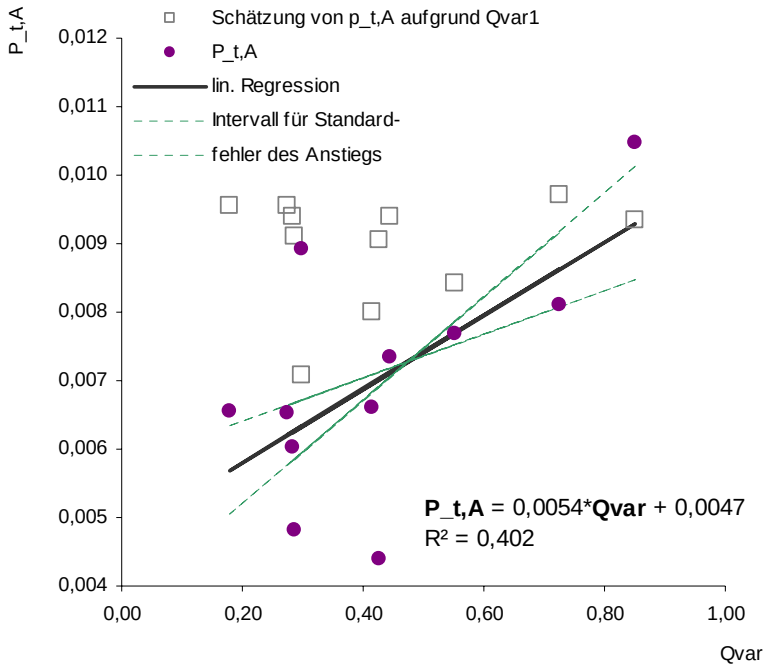



*Bild 6.9: Lineares Regressionsmodell für die Schätzung der Phosphorfrachten in Einzugsgebieten der Ostsee aufgrund der Abflussvariabilität der Flüsse*

Die geringe Zahl an Stichprobenelementen und das ebenso geringe Bestimmtheitsmaß weisen das Modell nicht als besonders gut aus. Nur 40% der Varianz der Phosphorfrachten können erklärt werden. Allerdings ist der Anstieg der Regressionsgleichung signifikant und es besteht ein inhaltlich bekannter Zusammenhang zwischen Phosphorfrachten und Abflussvariabilität. Dieser lässt die gefundene Beziehung zumindest in ihrer Richtung plausibel erscheinen. Demnach sollte es wie im Regressionsmodell einen Anstieg der Phosphorfrachten bei zunehmender Abflussvariabilität geben.

Wir wollen mit diesem Beispiel demonstrieren, dass die Regressionsanalyse nicht unbedingt beendet werden muss, wenn ein akzeptabler funktionaler Zusammenhang gefunden worden ist. Im Rahmen der Systemsynthese ist es jetzt sinnvoll, verschieden Szenarien zu berechnen. Anhand unseres Beispiels bietet es sich zum Abschluss zumindest an, die Phosphorfrachten zu schätzen, welche auftreten könnten, wenn keine Talsperren vorhanden wären. Hier wird übrigens auch deutlich, worin der Vorteil solcher Szenarien liegt. Ein reales Experiment könnte mit dieser Frage nicht durchgeführt werden.

Unter der Annahme des von uns gefunden linearen Zusammenhangs zwischen der Abflussvariabilität und den Phosphorfrachten können wir in die Regressionsgleichung die Werte der Abflussvariabilität der einzelnen Flüsse vor der Errichtung der Talsperren einsetzen. Die hierdurch geschätzten Phosphorfrachten liegen deutlich über den heute gemessenen und sind in Bild 6.9 dargestellt. Wir können festhalten, dass sich in diesem angenommenen Fall die Phosphorfrachten der untersuchten 11 Flüsse auf insgesamt 127% erhöhen würden. Diese Aussage gilt allerdings wegen des Standardfehlers des Regressionsmodells nur innerhalb eines Fehlerintervalls von +/- 20% und ist damit doch vergleichsweise unsicher.

# 3 Nichtlineare Regression

## 3.1 Welche Frage soll beantwortet werden?

Häufig kann, obwohl ein Zusammenhang zwischen zwei Variablen zu erwarten ist, keine Linearität vorausgesetzt werden. Bildet man die Residuen der linearen Regression für einen solchen Fall in einem Diagramm ab, so sind systematische Über- oder Unterschätzungen festzustellen (Bild 6.10).

Dies ist zum Beispiel bei der Abnahme des Luftdrucks mit zunehmender Höhe oder der Beobachtung des Bevölkerungswachstums im zeitlichen Bezug zu beobachten. Die Frage lautet:

Wie kann trotz nichtlinearem Zusammenhang eine funktionale Beschreibung mittels Regressionsgleichung gefunden werden?

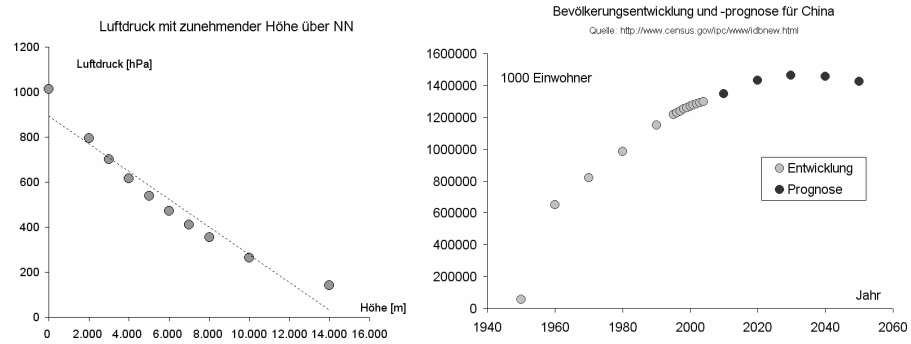



*Bild 6.10 Nichtlineare Zusammenhänge*

## 3.2 Welches ist das richtige Modell?

Soll ein nichtlineares Modell zur Beschreibung des funktionalen Zusammenhangs zweier Variablen verwendet werden, steht als erste Frage die Entscheidung über die Variante des Modells im Vordergrund.

Theoretisch müssten der Komplexität der Regressionsgleichung $y = f(x,a_1,a_2, \ldots)$ keine Grenzen gesetzt werden. Jede erdenkliche Funktion $f$ mit einer unabhängigen Variablen $x$ und einer endlichen Anzahl von Parametern $a_1,a_2,..$ wäre als Regressionszusammenhang vorstellbar. Das einzige Ziel wäre in solchem Fall, den Kurvenverlauf im Scatterplot (siehe Kapitel III) möglichst genau abzubilden.

In der Praxis existiert allerdings eine Reihe von Gründen, ein möglichst einfaches Modell zu wählen (siehe Box).

**Gründe für die Wahl eines einfachen Regressionsmodells**

- Je diffiziler ein Modell und je höher die Anzahl der verwendeten Parameter, umso aufwendiger die Parametrisierung. Irgendwann scheitern alle Methoden!
- Mit zunehmender Modellkomplexität steigt der Modellfehler, d.h. Prognosen und Interpolationen werden immer unsicherer.
- Für komplexe Modelle ist ihrer Kausalität nur schwer wissenschaftlich zu belegen. Wichtig ist jedoch, dass wir die Zusammenhänge im Modell auch begründen können!

Für einen Großteil der Probleme lässt sich die Auswahl der möglichen Regressionsmodelle auf eine handvoll gebräuchlicher Varianten einschränken. Diese sollen Gegenstand der nachfolgenden Betrachtungen sein.

### 3.2.1 Polynomiales Regressionsmodell

Unser erstes Regressionsmodell ist das am häufigsten verwendete bivariate Modell, oder besser die am häufigsten verwendete Modellgruppe. Denn es handelt bei den polynomialen Regressionsmodellen um unterschiedliche Modellvarianten, die sich durch den Grad des verwendeten Polynoms unterscheiden. Genau genommen besitzt sogar das im ersten Teil des Kapitels vorgestellte lineare Modell eine polynomiale Funktionsgleichung erster Ordnung und ist somit ein polynomiales Regressionsmodell.

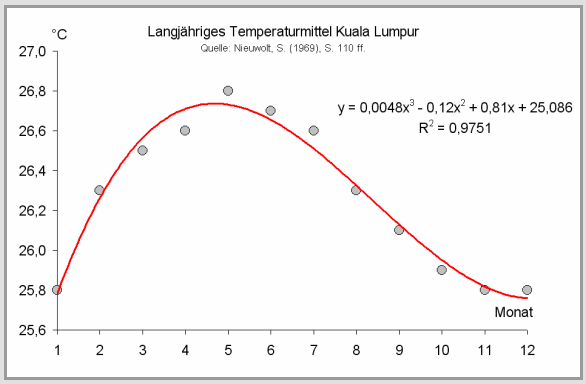
Da die Anzahl der Minima und Maxima abhängig vom Grad des Polynoms ist, lässt sich nahezu jeder Kurvenverlauf anpassen. Dieses Modell ist daher universell einsetzbar. Allerdings können nur Daten interpoliert, nie extrapolierte werden, da das Modell an den Rändern zum Ausschwingen neigt.

**Modellgleichung (allgemein):**

$$y = a_0 + a_1 x + a_2 x^2 + .... + a_n x^n$$

**Parameter:**

| | |
|---|---|
| $a_i$ | Koeffizienten |
| $n$ | Grad des Polynomes |
| $n-1$ | Extremstellen |
| $n-2$ | Wendestellen |




**Modellgleichung (SPSS):** $y = b_0 + b_1 \cdot x + b_2 \cdot x^2 + ...$

**Anwendungsbereich:**

1. Kurvenanpassung mit vorgegebener Anzahl von Extrem- und Wendestellen.
2. physikalische Probleme mit polynomialen Zusammenhängen wie Ausbreitung von Licht, Schall, Schwerkraft im Raum oder auf der Fläche.

**Problem:** Polynomiale Modelle hohen Grades neigen oft zum Ausschwingen an den Rändern des Anpassungsbereiches. (nicht begründbares Ansteigen oder Abfallen). Prognosen aus solchen Modellen besitzen keinerlei Aussagekraft.

### 3.2.2 Exponentielles Regressionsmodell

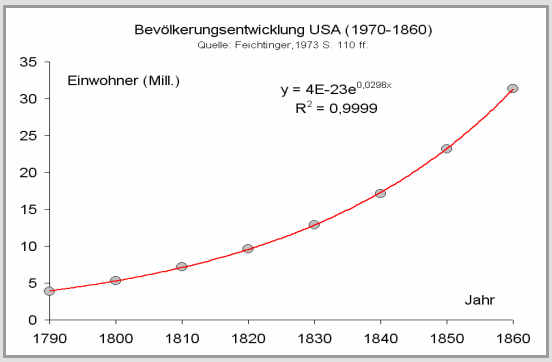
Das exponentielle Regressionsmodell bildet zum einen unbegrenztes exponentielles Wachstum (positiver Faktor *a*) oder unbegrenzten exponentiellen Zerfall (negativer Faktor *a*) ab. Zum anderen kann, bei negativem Exponenten *b*, auch die Annäherung an eine Sättigungsgrenze von unten (negativer Faktor *a*) oder von oben (positiver Faktor *a*) beschrieben werden. Um das Modell noch flexibler handhaben zu können, führen wir einen zusätzlichen Summanden *c* ein, der die Sättigungsgrenze festlegt.

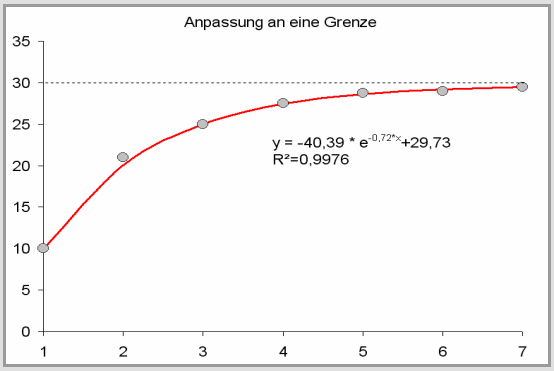
Modellgleichung (allgemein):

$$y = a \cdot e^{b \cdot x} \quad [+ c]$$

Parameter:

- $a$ <0 → Umkehrung der Richtung
- $b$ >0 → unbegrenztes Wachstum
  <0 → Annäherung an Null
- $c$ Verschiebung in y -Richtung




Modellgleichung (SPSS): $y = b_0 \cdot e^{b_1 \cdot x}$

Anwendungsbereich:

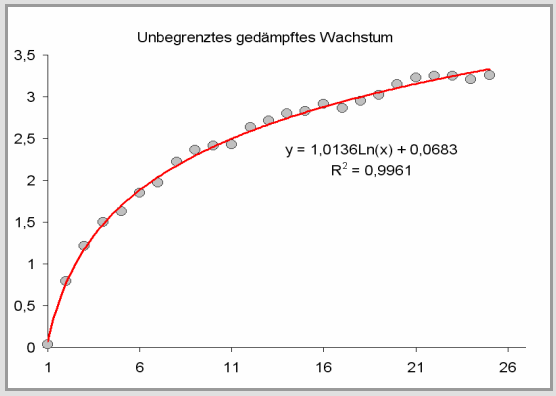
1. Kurvenanpassung von, mit stark zunehmendem Anstieg und ohne Dämpfung, verlaufende Punktmengen
2. **Oder:** Anpassung an einen vorgegebenen Grenzwert von oben oder von unten
3. Werte mit regelmäßiger Vervielfachung nach konstanten Zeitschritten (1; 2; 4; 8; 16; 32;…) oder (1; 0,5; 0,25; 0,125; 0,0625; …)
4. Beispiel: anfängliches, unbegrenztes Wachstum von Individuen; Kapitalwerte mit Verzinsung; Radioaktiver Zerfall

Problem: Das exponentielle Regressionsmodell mit positivem Exponenten geht von einem unbegrenzten Wachstum ohne Dämpfung aus. Nahezu alle Prozesse sind irgendwann begrenzt. Für Prognosen in die weite Zukunft ist Vorsicht geboten!

### 3.2.3 Logarithmisches Regressionsmodell

Dieses Regressionsmodell wird man in geographischen Anwendungen relativ selten in Benutzung finden. Wenn doch, dient es meist einer optimierten Kurvenapproximation, um Zwischenwerte bestimmen zu können.

Soll der kausale Hintergrund auch eine Rolle spielen, müssen wir dieses Modell zur Simulation eines grenzenlosen, aber gedämpften Anstiegs oder Zerfalls parametrisieren.

| | |
|---|---|
| Modellgleichung (allgemein): $y = a \cdot \ln(x) + b$<br>Parameter:<br>$a$ <0 → Umkehrung der Richtung<br>$b$ Verschiebung in y –Richtung |  |
| Modellgleichung (SPSS): | $y = b_0 + b_1 \cdot \ln(x)$ |
| Anwendungsbereich: | 1. Kurvenanpassung für gedämpftes, aber unbegrenztes Wachsen oder Zerfallen<br>2. Der Anstieg verhält sich gleichmäßiger als beim exponentiellen Modell mit negativen Exponenten |
| Problem: | Das logarithmische Regressionsmodell mit positivem Exponenten geht von einem unbegrenzten Wachstum mit Dämpfung aus. Nahezu alle Prozesse sind irgendwann begrenzt. Für Prognosen in die weite Zukunft ist Vorsicht geboten! |

### 3.2.4 Logistisches Regressionsmodell

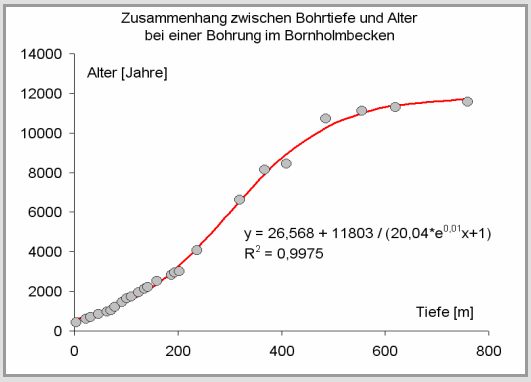
Das logistische Regressionsmodell findet in den Geowissenschaften eine Reihe von Anwendungsmöglichkeiten. Für Bevölkerungsmodelle oder die Simulation von Pflanzenwachstum besitzt es die gleiche Verwendungsberechtigung, wie für die Modellierung von Schichtdichten in Bohrkernen und alle Arten von Speicherfüllungen. Wir sollten es immer dann einsetzen, wenn Anfangs ein nahezu exponentieller Anstieg der abhängigen Größe auftritt, der jedoch später durch einen Sättigungswert begrenzt ist.

Modellgleichung (allgemein):

$$y = a + \frac{b}{1 + c \cdot e^{d \cdot x}}$$

Parameter:

| | |
|---|---|
| $a$ | untere Grenze der S-Kurve |
| $b$ | obere Grenze der S-Kurve |
| $c, d$ | steuern den Verlauf der S-Kurve |




Modellgleichung (SPSS):

$$y = \frac{1}{\frac{1}{o} + b_0 \cdot b_1^{\,x}}$$

Anwendungsbereich:

1. Kurvenanpassung an eine Punktmenge mit zuerst wachsendem, dann fallendem Anstieg
2. Eine Vielzahl von Wachstumsmodellen mit Anfangs günstigen Wachstumsbedingungen und späterer Begrenzung durch zunehmende Überbevölkerung
3. Wechsel zwischen zwei Zuständen (untere Grenze, obere Grenze)

## 3.3 Wie findet man die passenden Parameterwerte für ein Regressionsmodell?

Haben wir uns für das passende Modell entschieden, müssen wir im nächsten Schritt die Werte für die Parameter festlegen. Dies kann wie in der linearen Einfachregression mit Hilfe des Excel-Solvers und der Methode der kleinsten Quadrate erfolgen (siehe Abschnitt 2.1).

Mit dieser Methode kann theoretisch jedes beliebige Modell angepasst werden. Allerdings ist dazu notwendig, dass der Solver auch das globale Minimum des quadratischen Abstands zwischen gemessenen und approximierten Werten findet. Je mehr Parameter verwendet werden und je komplizierter das Modell ist, umso schwieriger wird für uns die Suche nach den Startwerten für einen erfolgreichen Iterationsprozess sein.

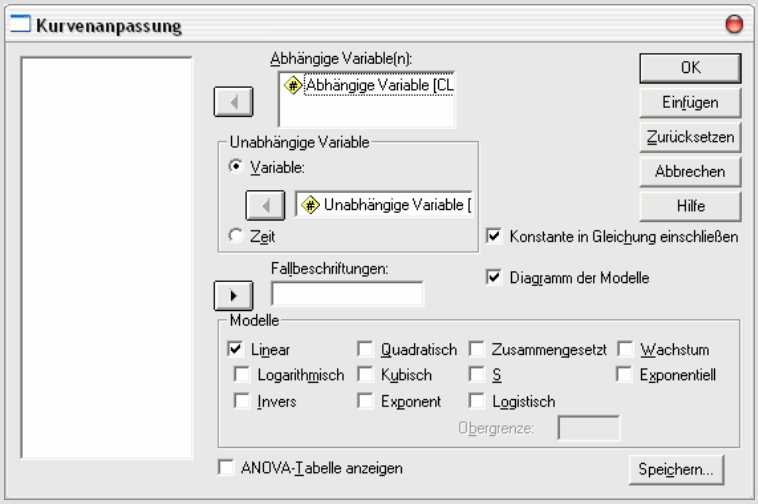
Eine weitere Möglichkeit ist die Verwendung der Trendlinienfunktion für Excel-Diagramme. Mit Hilfe eines Klicks der rechten Maustaste auf einen Diagrammdatensatz und dem Menübefehl „Trendlinie hinzufügen“ können schnell entsprechende Kurvenanpassungen erzeugt werden. Der Nachteil dieser Methode ist, dass mit dem Bestimmtheitsmaß nur ein Gütekriterium und nur eine eingeschränkte Auswahl möglicher Modelle zur Verfügung stehen.

Schnell gefunden sind die Parameterwerte auch mit Hilfe des Statistikprogramms SPSS (siehe Box). Von Vorteil ist hier, dass eine Bewertung der Modellgüte mit Hilfe verschiedener mitgelieferter Kriterien sehr einfach gemacht wird. Außerdem können unter einer Vielzahl von Modellen ein passendes ausgewählt, und sogar mehrere Modelle miteinander kombiniert werden.

Die Methoden der abschließenden Modellbewertung lassen sich von der linearen auf die nichtlineare Regression übertragen.

**Nichtlineare Regressionsanalyse mit SPSS**

1. Analysieren> Regression> Kurvenanpassung…




2. metrisch skalierte Variablen für abhängige und unabhängige Größen auswählen> mit „►“-Taste übertragen
3. Modell auswählen
4. OK>
5. SPSS gibt einen Scatterplot mit angepasstem Regressionsmodell aus. Leider werden nur Parameter, nicht aber gesamte Modellformel ausgegeben Deswegen weisen wir in den Modellboxen noch einmal auf die speziellen SPSS-Formeln hin.

# Nichtlineare Regression am Beispiel

GGA_Regression\Uebung_3.xls

■ In diesem Beispiel interessieren wir uns für die Modellierung eines Zusammenhangs zwischen dem Alter und der Bohrtiefe eines Bohrkerns aus der Ostssee.

Ein für eine lineare Regression angefertigtes Residuendiagramm zeigt einen S-förmigen und somit systematischen Verlauf an (siehe Box Residuenanalyse). Die Anpassung eines linearen Regressionsmodells ist demnach ungeeignet, wir müssen auf ein nichtlineares Modell zurückgreifen.

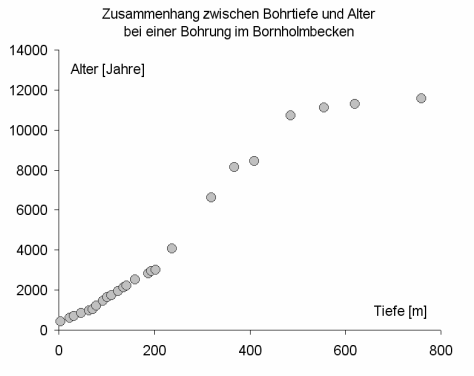



*Bild 6.11 Scatterplot*

Anhand des Verlaufs der Punktwolke im Scatterplot (Bild 6.11), können wir eine Aussage über den Typ des am besten beschreibenden Regressionsmodells wagen.

Scheinbar stimmt diese Anordnung sehr gut mit der Form eines logistischen Regressionsmodells überein. Wir sollten diese Annahme am Ende jedoch noch einmal überprüfen!

Im nächsten Schritt werden wir die Werte für die vier notwendigen Parameter mit der Methode der kleinsten Quadratsumme und dem Excel-Solver bestimmen.

In Spalte C berechnen wir dazu die Werte für das Alter nach dem Regressionsmodells in Abhängigkeit von der Bohrtiefe aus Spalte A. Es reicht aus, wenn wir dazu die dem Modell entsprechende Excelformel `=Gu+Go/(a*EXP(b*x)+1)` in die erste Datenzeile der Tabelle eintragen und die Formel auf die darunter liegenden Zellen übertragen. (*x, a, b, Gu, Go* wurden in diesem Excelblatt bereits definiert).

In der Spalte D werden als nächstes die quadratischen Differenzen `=(y-r?)^2` zwischen Original- und Schätzwerten kalkuliert. Schließlich tragen wir in der Zelle H8 die Summe der quadratischen Differenzen ein. Nach der Methode der kleinsten Quadratsumme soll dieser Zellwert möglichst klein werden!

Nachdem wir erste Startwerte für unsere Regressionsparameter *x, b, Gu* und *Go* festgelegt haben, starten wir den Solver (siehe Seite 89). Wir minimieren den Betrag der Zielzellen `H8`. Veränderbare Zellen sind `H3:H6`.

Wenn wir nicht sehr viel Glück haben, wird in diesem ersten Schritt keine oder nur eine schlechte Anpassung zustande kommen. Dem Solver fällt es schwer, das globale Minimum zu finden, da unser Modell recht komplex ist. Wir müssen also bessere Startwerte festlegen.

*Gu* = 400 und *Go* = 12000, als untere und obere Grenze der Wachstumskurve können wir noch vom Datensatz ablesen. Die Parameterstartwerte für *a* = 10 und *b* = -0,005 erhalten wir am besten durch mehrfaches Probieren.

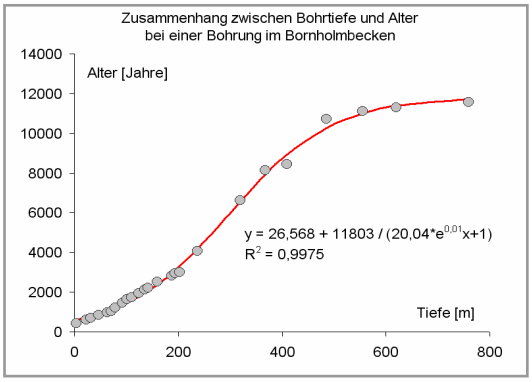



*Bild 6.12 Scatterplot mit angepasster Regression*

Nach einem nächsten Optimierungsgang durch den Solver können wir die Summe der quadratischen Abweichungen auf unter 900.000 minimieren. Wir haben offensichtlich eine sehr gute Anpassung gefunden. Dies bestätigt auch die optische Begutachtung im Diagramm und die Berechnung des Bestimmtheitsmaßes.

Wie gut ist aber das Bestimmtheitsmaß im Vergleich zu anderen Regressionsmodellen. Die Tabelle 6.1 gibt darüber Auskunft. Nur das polynomische Modell hohen Grades hat ein besseres Bestimmtheitsmaß. Im Allgemeinen kann ein polynomisches Modell, ist nur sein Grad hoch genug gewählt, immer das beste Bestimmtheitsmaß erreichen. Häufig lässt sich seine Verwendung aber nicht begründen.

Offensichtlich haben wir mit dem logistischen Modell eine gute Anpassung gefunden. Sie hilft uns auch Vorhersagen für nicht analysierte Bohrtiefen zu treffen. So wird das Alter der Probe aus 500 Meter Tiefe etwa 10760 Jahre, in 900 Meter Tiefe wird es laut Regressionsmodell etwa 11800 Jahre betragen. Die Schätzung innerhalb des gemessenen Bereichs können wir bejahen (Sie ist aber immer von der Güte des Regressionsmodells abhängig!). Vorsicht ist allerdings bei der Extrapolation dieses Datensatzes geboten. Denn mit dem nur etwas tiefer gelegenen Übergang zum Grundgebirge verändern sich alle Modellannahmen rapide, die vorhergesagten Werte sind nicht mehr zu verwenden.

| Modell | Bestimmtheitsmaß |
|---|---|
| Logistisch | 0,9975 |
| Linear | 0,9564 |
| Logarithmisch | 0,6613 |
| Polynomisch (Zweiten Grades) | 0,9676 |
| Polynomisch (Fünften Grades) | 0,9981 |
| Exponentiell | 0,8532 |

*Tabelle 6.1 Modellvergleich*

Eine Frage sollte man sich als Geograph immer stellen: Kann man die Verwendung eines Regressionsmodells auch wissenschaftlich rechtfertigen?

Scheinbar messen wir in dem Bohrkern eine Überlagerung zweier Effekte, nämlich die Veränderung der Sedimentverdichtung und der Sedimentationsrate mit der Tiefe.

Mit zunehmender Tiefe nimmt die Sedimentverdichtung der erbohrten Schichten und somit die Alterszunahme pro Tiefenmeter bis zur einer bestimmten Grenze exponentiell zu.

Anfangs ist die Sedimentationsrate im Verhältnis gering. Je tiefer er eindringt, umso mehr nähert sich der Bohrer den Schichten mit hohen pleistozänen Sedimentationsraten. Die Alterung pro Tiefenmeter verringert sich wieder.

# 4 SCHLUSSWORT

**Systemanalyse – Systemidentifikation – Systemsynthese**

Mit dem Überblick über die Methode der Regressionsanalyse endet dieser Leitfaden. Es war unser Ziel, Verfahren zur Analyse geographischer Daten kennen zu lernen. Diese Verfahren sollten jedoch stets nur Mittel zum Zweck bleiben. Nimmt ein Fachwissenschaftler, hier Sie als Geowissenschaftler, die Mühe auf sich, einen komplexen Datensatz zu analysieren, erwartet er hierdurch Antworten auf neue Fragen oder möchte neue, interessante Fragestellungen erkennen.

Als wir als Autoren das Konzept dieses Bandes erarbeiteten, wollten wir genau einen solchen Ansatz vermitteln: durch logisch aufeinander aufbauende Fragestellungen zu neuen Antworten und Erkenntnissen zu gelangen. Gleichzeitig sollten Sie als Lernende einen Leitfaden in der Hand halten, welcher Ihnen neben Grundkonzepten verschiedener Verfahren der Datenanalyse ebenso Schritt-für-Schritt–Anleitungen zum Durchführen der besprochenen Methoden vermittelt. Hierzu finden Sie die zahlreichen Übungsdateien auf der beigelegten CD-ROM. Etliche dieser Übungsdateien sind auch als Arbeitsblätter zu verwenden, mit denen Sie ihre Daten effizient auswerten können, die Sie dabei unterstützen, die unterschiedlichen Methoden zu handhaben.

Ein drittes Ziel war es natürlich auch, stets darauf hinzuweisen, dass man mit den erhaltenen Ergebnissen kritisch umgehen muss. Gerade heute lassen sich mit zahlreicher Software schnell und einfach verschiedenste, eben auch statistische Berechnungen durchführen. Wenn wir aber mit deren Hilfe fachspezifische Fragen beantworten wollen, so werden statistische Ergebnisse schnell und leicht in deterministische Zusammenhänge umformuliert. Mit statistischen Auswertemethoden gewinnt man aber stets nur Indizien, Schätzungen mit allerdings einer definierten Wahrscheinlichkeit. Erinnern Sie sich noch an den statistischen Zusammenhang zwischen Störchen und Geburtenrate in Kap. 4, welcher unzweifelhaft kein deterministischer war?

Innerhalb dieser Bandbreite von Ansprüchen entstand dieser Band aus einer weiterführenden Lehrveranstaltung zu statistischen Methoden in der Geographie. Wir hoffen, dass wir Ihnen hiermit einen Leitfaden in die Hand geben, mit welchem es Ihnen gelingt, stets auf die Kernfrage der Geographie zu fokussieren: Wie lässt sich anhand räumlicher Strukturen und Muster auf das Verhalten eines Systems schließen?

## Geographische Namen im verwendeten Datensatz

Die Namen der verwendeten Flusseinzugsgebiete sind wegen der Verwendung in den Programmen SPSS und EXCEL, hier wegen der Verknüpfung mit dem Kartenhintergrund, vereinfacht oder verändert worden. Nachfolgend eine Liste der verwendeten Flusseinzugsgebiete und deren korrekter geographischer Bezeichnung:

| *Verwendung im Datensatz* | *korrekte Schreibweise* |
|---|---|
| Aa | Aa |
| Abyalven, Byskealven | Abyälven, Byskeälven |
| Alan, Rosan | Alån, Rosån |
| Angermanalven | Angermanälven |
| Aurajoki | Aurajoki |
| Braknean | Brakneån |
| Dalalven | Dalälven |
| Danish Straits - Kavlingean, Saxan | Kavlingeån, Saxån |
| Daugava | Daugava |
| Eman | Emån |
| Gadean | Gadeån |
| Gavlean | Gavleån |
| Gidealven, Moalven, Orealven | Gideälven, Moälven, Öreälven |
| Helge a | Helge å |
| Iijoki | Iijoki |
| Indalsalven | Indalsälven |
| Kalixalven | Kalixälven |
| Kattegat – Atran | Atrån |
| Kattegat – Kungsbackean | Kungsbackeån |
| Kattegat - Lagan, Ronne a | Lagån, Ronne å |
| Kattegat – Viskan | Viskån |
| Kemijoki | Kemijoki |
| Kiiminginijoki | Kiiminginijoki |
| Klaralven, Gotaalven | Klarälven, Götaälven |
| Kokemaenjoki | Kokemaenjoki |

| | |
|---|---|
| Kymijoki | Kymijoki |
| Ljungan | Ljungan |
| Ljungbyan, Alsteran | Ljungbyån, Alsterån |
| Ljusnan | Ljusnan |
| Luga | Luga |
| Lulealven | Luleälven |
| Malaren | Mälaren |
| Morrumsan | Morrumsån |
| Motala strom | Motala ström |
| Narva, Pljussa | Narva, Pljussa |
| Neman | Neman |
| Neva | Neva |
| Odra | Odra |
| Oulujoki | Oulujoki |
| Pernu, Salaca | Pernu, Salaca |
| Pitealven | Piteälv |
| Polish Coast | Polish Coast |
| Porvoonjoki | Porvoonjoki |
| Pregola | Pregola |
| Ranealven | Raneälven |
| Ricklean, Kalabodaan | Rickleån. Kalabodaån |
| Ronnebyan | Ronnebyån |
| Sangisalven | Sangisälv |
| Selangersan | Selangersån |
| Siika-, Kala-, Lapuan-, Kyronjoki | Siikajoki, Kalajoki, Lapuanjoki, Kyronjoki |
| Simojoki | Simojoki |
| Skellefteälven | Skellefteälven |
| Skrabean | Skrabeån |
| Stockholm | Stockholm |
| Svagan, Harmangersan | Svagån, Harmangersån |
| Svartaan, Nykopingsan, Kilaan | Svartaån, Nyköpingsån, Kilaån |
| Tamnaran | Tamnarån |
| Tornealven | Torneälven |
| Umealven | Umeälven |
| Vindan, Storan | Vindån, Storån |
| Wisla | Wisla |

# Literatur


AURADA, K.-D. (1982): Zur Anwendung des systemtheoretischen Kalküls in der Geographie. Petermanns Geogr. Mitt., **126,** 241-249

AURADA, K.-D. (1993): Logik und Logistik der naturwissenschaftlichen Geographie. In: Beiträge des 9. Kolloquiums für Theorie und Quantitative Methoden in der Geographie. Klagenfurter Geogr. Schriften, **11**, 63-79.

AURADA, K.-D. (2003): Co-evolvierende + co-respondierende Systeme = co-operierendes System. Erdkunde, **57**, 309-330.

BACKHAUS, K., B. ERICHSON & W. PLINKE (2000): Multivariate Analysemethoden. Berlin.

BAHRENBERG, G., E. GIESE & J. NIPPER (1985): Statistische Methoden in der Geographie, Band 1 Univariate und bivariate Statistik. Stuttgart.

BAHRENBERG, G., E. GIESE & J. NIPPER (1991): Statistische Methoden in der Geographie, Band 2 Multivariate Statistik. Stuttgart.

BRYDSTEN, L. ET. AL. (1990): Element transport in regulated rivers and non-regulated rivers in Northern Sweden. Regulated rivers: research and management, Vol. 5 p. 167-176.

FEICHTINGER, G. (1973): Bevölkerungsstatistik. Berlin.

KLUG, H. & R. LANG (1983): Einführung in die Geosystemlehre. Darmstadt.

NIEUWOLT, S. (1969): Klimageographie der malaiischen Halbinsel. Mainz, Geograph. Inst. d. Johannes-Gutenberg-Univ.

SCHÖNWIESE, C.-D. [HRSG.] (1983): Statistische Methoden in der Klimatologie. Meteorologische Fortbildung, **13**, H. 1/2, Offenbach am Main.

SCHÖNWIESE, C.-D. (2000): Praktische Statistik. Berlin, Stuttgart.

STOYAN, D., H. STOYAN & U. JANSEN (1997): Umweltstatistik. Stuttgart, Leipzig.

MONKA, M. & W. VOß (2002): Statistik am PC: Lösungen mit Excel. München, Wien.

ZWERENZ, K. (2001): Statistik: Datenanalyse mit EXCEL und SPSS. München, Wien.

*http://www.grida.no/baltic/index.htm (eingesehen am 15. 11. 2004)*

**Greifswalder Geographische Arbeiten**

Band 14 BILLWITZ, K. (Red.): Mecklenburg-Vorpommern: Grundzüge der Naturraumausstattung, -erkundung und –bewertung. Fachsitzung 1 des 25. Deutschen Schulgeographentages vom 07.10.-11.10.1996 in Greifswald. 1997, 162 S.

Band 15 AURADA, K. D. & J. NEWIG (Red.): Die Ostsee und ihr Einzugsgebiet – Wandel des Natur- und Kulturraumes. Fachsitzung 3 des 25. Deutschen Schulgeographentages vom 07.10.-11.10.1996 in Greifswald. 1997, 131 S.

Band 16 LAMPE, R. (Red.): Greifswalder Bodden und Oder-Ästuar – Austauschprozesse (GOAP): Synthesebericht des Verbundprojektes. 1998, 490 S.

Sonder- ASMUS, I., H. T. PORADA & D. SCHLEINERT (Red.): Geographische und historische Beiträge zur Landeskunde Pommerns:

band Eginhard Wegner zum 80. Geburtstag. – Schwerin: Thomas Helms Verlag, 1998, 334 S.

Band 17 HELBIG, H.: Die spätglaziale und holozäne Überprägung der Grundmoränenplatten in Vorpommern. Diss. 1999, 110 S., 82 S. + Anhang und 14 Fototafeln

Band 18 RÖDEL, R.: Die Auswirkungen des historischen Talsperrenbaus auf die Zuflussverhältnisse der Ostsee. Diss. 2001, 118 S.

Band 19 UNVERZAGT, S.: Räumliche und zeitliche Veränderung der Gebiete mit Sauerstoffmangel und Schwefelwasserstoff im Tiefenwasser der Ostsee. Diss. 2001, 122 S. + Anhang

Band 20 HILBIG, A.: Kleinräumige Differenzierung der Bevölkerungsdynamik in Mecklenburg-Vorpommern. Diss. 2001, 99. S. + Anhang

Band 21 PAULSON, C.: Die Karstmoore in der Kreidelandschaft des Nationalparks Jasmund auf der Insel Rügen. Diss. 2001, 296 S.

Band 22 ZÖLITZ-MÖLLER, R. (Hrsg.): Historische Geographie und Kulturlandschaftsforschung. Beiträge zum Gedenkkolloquium für Dr. Eginhard Wegner am 4. Mai 2001 in Greifswald. 109 S.

Band 23 BILLWITZ, K. (Hrsg.): Geoökologische und landschaftsgeschichtliche Studien in Mecklenburg-Vorpommern. 2001, 296 S.

Band 24 KAISER, K.: Die spätpleistozäne bis frühholozäne Beckenentwicklung in Mecklenburg-Vorpommern – Untersuchungen zur Stratigraphie, Geomorphologie und Geoarchäologie. Diss. 2001, 208 S. + Anhang

Band 25 BILLWITZ, K. & P. KÜHN (unter Mitarbeit von H. BARTH, A. BAUMGART, S. HELMS, F. HOFMEISTER, K. KAISER, J. LUCKERT, W. OEHMICHEN, H. ROTHER & M. WIRNER): Der Bodenlehrpfad Jägerhof in Vorpommern. 2002, 57 S. + Anhang

Band 26 KAISER, K. (Hrsg.): die jungquartäre Fluss- und Seegenese in Norddeutschland. Beiträge zur Tagung in Hohenzieritz (Mecklenburg) vom 26.-28. Februar 2002. 2002, 243 S.

Band 27 LAMPE, R. (Ed.): Holocene Evolution of the South-Western Baltic Coast – Geological, Archaeological and Palaeo-environmental Aspects. Field meeting of INQUA Subcommission V: Sea-level Changes and Coastal Evolution. Western Europe, September 22.-27. 2002, 2002, 224 S.

Band 28 KÜHN, P.: Spätglaziale und holozäne Lessivégenese auf jungweichselzeitlichen Sedimenten Deutschlands. Dissertation 2003. 164 S. + Anhang

Band 29 BILLWITZ, K.: Bodenkundliche und landschaftsgenetische Studien in Mecklenburg-Vorpommern. 2003, 247 S.

Band 30 SUCCOW, M. & K. BILLWITZ: Landschaftsökologische Exkursionen in die Greifswalder Umgebung. 2003.

Band 31 TIMMERMANN, T., W. WICHTMANN, M. SUCCOW & K. BILLWITZ.: Alternative Nutzungsformen für Moorstandorte in Mecklenburg-Vorpommern. Beiträge einer Tagung in Greifswald vom 23. November 2002. 2003.

Band 32 DECKERS, B.: Die raumstrukturelle Wirkung von Transformation und EU-Osterweiterung. Zur Rolle der ortsansässigen Bevölkerung bei der Regionalisierung im nördlichen deutsch-polnischen Grenzraum. 2004, 179 S.

Band 33 HOFFMANN, T. & R. RÖDEL: Leitfaden für die statistische Auswertung geographischer Daten. 2004, 114 S.