UFZ-UMWELTFORSCHUNGSZENTRUM LEIPZIG-HALLE GMBH

UFZ-Bericht

UFZ-Bericht • UFZ-Bericht • UFZ-Bericht • UFZ-Bericht

Nr. 22/1999 Dissertation

Anwendung von Methoden der Mustererkennung zur Bewertung der Schadstoffverteilung im Einzugsbereich einer Mülldeponie

Mathias Rudolph

ISSN 0948-9452

Anwendung von Methoden der Mustererkennung zur Bewertung der Schadstoffverteilung im Einzugsbereich einer Mülldeponie

von der Fakultät für Elektrotechnik und Informationstechnik
der Technischen Universität Chemnitz

genehmigte
Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur

(Dr.-Ing.)

vorgelegt

von
Dipl.-Ing. Mathias Rudolph
geboren am 23.1.1968 in Leipzig
eingereicht am 1. November 1998

Gutachter: Prof. Dr. sc. techn. Steffen F. Bocklisch
Prof. Dr. rer. nat. habil. Werner Dilger
Prof. Dr. rer. nat. habil. Walter Gläßer
Prof. Dr. rer. nat. habil. Heinz W. Zwanziger

Tag der Verleihung: 24. Juni 1999

Bibliographische Beschreibung

Rudolph, Mathias
Anwendung von Methoden der Mustererkennung zur Bewertung der Schadstoffverteilung im Einzugsbereich einer Mülldeponie
Technische Universität Chemnitz, Diss.
194 S., 117 Lit., 60 Abb., 65 Tab.

Referat:

Richtung und Umfang des Austrages von Schadstoffen einer Mülldeponie in den Grundwasserleiter können wegen der komplexen stofflichen Matrizes, der Vielfalt möglicher analytischer Parameter und ihrer ökologischen Auswirkungen sowie der relativ großen Anzahl von Messpunkten (Sickerwasserpegel SWP und Grundwasserbeobachtungsbrunnen GWB) nur schwierig aus dem aktuellen, meist sehr umfangreichen und damit unübersichtlichen Datenmaterial abgeleitet werden.
An einem Modellstandort, einer Altdeponie bei Leipzig, wird mit Hilfe geeigneter datenanalytischer Methoden ermittelt, ob sich anhand der Verteilung von Mustern - hierunter wird die geordnete Gesamtheit der an einer Sicker- oder Grundwasserprobenahmestelle gemessenen schadstoffrelevanten Parameter verstanden - Rückschlüsse auf Kontaminationsquellen und Transportpfade des Sickerwassers ziehen lassen. Des Weiteren wird untersucht, welche analytischen Parameter den Sickerwassertransport am besten beschreiben und ob aufwendige Analyseverfahren durch einfachere Methoden bei vergleichbarem Informationsgehalt substituiert werden können.
Mit einfachen Verfahren der beschreibenden mathematischen Statistik erfolgt eine Erstauswertung des im Zeitraum von 1992 bis 1998 aufgenommenen Datenmaterials. Anschließend wird über die klassischen statistischen Auswertemethoden Korrelations- und Regressionsanalyse sowie Varianzanalyse zum Methodenspektrum der Mustererkennung (überwachte und automatische Klassifikation) übergegangen. Eine Zeitreihenanalyse wird im modellhaften Ansatz durchgeführt. Als Vertreter einer wissensbasierten Methode kommt ein hybrides Neuro-Fuzzy-Klassifikationssystem zur Anwendung.
Für den konkreten Anwendungsfall wird eine vergleichende Bewertung der angewandten Verfahren sowohl der überwachten als auch der automatischen Klassifikation bezüglich des mit ihnen erzielten Informationsgehaltes sowie ihrer Adaptionsfähigkeit an die gegebene Aufgabenstellung vorgenommen. Die diesbezüglichen Erkenntnisse können verallgemeinert werden und sind damit für ähnliche praktische Anwendungsfälle relevant.

Schlagwörter

Datenanalyse, Multivariate Analyse, Mustererkennung, Umweltanalyse, Deponie, Sickerwasser, Grundwasser

Inhalt

Abkürzungsverzeichnis

AAS	Atomabsorptions-Spektroskopie
AES	Atomemissionsspektrometrie
AKF	Autokorrelationsfunktion
BMBF	Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie
CA	Clusteranalyse
CSB	Chemischer Sauerstoffbedarf
DOC	Gelöster organischer Kohlenstoff (Dissolved Organic Carbon)
GW	Grundwasser
GWB	Grundwasserbeobachtungsbrunnen
IC	Ionenchromatographie
ICP	Induktiv gekoppeltes Plasma (Inductively Coupled Plasma)
IR	Infrarot
ISE	Ionensensitive Elektroden
KNN	k nächste Nachbarn (Methode)
LAGA	Länderarbeitsgemeinschaft Abfall
LAWA	Länderarbeitsgemeinschaft Wasser
LDA	Lineare Diskriminanzanalyse
NED	Nichtelementare Diskriminanzmerkmale
PAKF	Partielle Autokorrelationsfunktion
SW	Sickerwasser
SWA	Sickerwasseraustritt
SWP	Sickerwasserpegel
TA	Technische Anleitung
UBA	Umweltbundesamt

Danksagung

Die vorliegende Arbeit wurde am Umweltforschungszentrum Leipzig-Halle in der Sektion Analytik im Zeitraum von Oktober 1995 bis September 1998 angefertigt. Dabei möchte ich mich insbesondere bei Herrn Dr. J. Flachowsky, Leiter der Abteilung PPM der Sektion Analytik, für die Überlassung des Themas, die Ermöglichung seiner Bearbeitung und die stetige Unterstützung bedanken.

Mein ausdrücklicher Dank gilt Herrn Prof. Dr. St. F. Bocklisch, Herrn Prof. Dr. W. Dilger, Herrn Prof. Dr. W. Gläßer und Herrn Prof. Dr. H. W. Zwanziger für ihr Interesse an der Thematik und die Übernahme eines Gutachtens. Durch ihre Anregungen und helfende Kritik ermöglichten sie mir die Anfertigung der Arbeit.

Für die ständige Hilfsbereitschaft, die zahlreichen praktischen Hinweise und fachlichen Diskussionen möchte ich dem EDV-Beauftragten der Sektion Analytik, Herrn Dipl.-Math. L. Brüggemann, meinen nachdrücklichen Dank aussprechen.

Herr Dr. H. Borsdorf und Herr Dr. G. Schulte haben durch ihre hervorhebenswerte Unterstützung, insbesondere beim kritischen Durchsehen des Manuskripts, wesentlich zum Gelingen dieser Arbeit beigetragen - deshalb danke ich ihnen ganz besonders.

Sehr herzlich bedanken möchte ich mich bei Herrn Dipl.-Ing. A. Rämmler für die Durchführung eines großen Teils der Probenahmen sowie bei Frau Chem.-Ing. P. Fiedler, Frau Chem.-Ing. A. Lange und Frau W. Großmann für die Durchführung der zahlreichen Parameterbestimmungen.

Mein Dank gilt auch den weiteren namentlich nicht genannten Mitarbeitern der Abteilung PPM der Sektion Analytik für ihre ständige Hilfsbereitschaft und Unterstützung.

1 Einleitung

1.1 Themenstellung

In der Bundesrepublik Deutschland werden seit mehreren Jahren Untersuchungen zum Stoffhaushalt von Deponien durchgeführt (BUNDESANSTALT FÜR GEOWISSENSCHAFTEN UND ROHSTOFFE, 1995; UMWELTBUNDESAMT, 1995 und 1997). Dabei ist die Aufklärung von Zusammenhängen zwischen abgelagerten Stoffen und zu erwartenden Emissionen nur in begrenztem Umfang möglich. Deshalb behilft man sich u. a. in der Regel damit, im An- und Abstrom des unter dem Deponiekörper befindlichen Grundwasserleiters anthropogene Schadstoffe nachzuweisen (BAUMANN et al., 1993; FLACHOWSKY, 1996). Hierfür gibt es eine Reihe landeshoheitlich geregelter Szenarien der Probenahme, der Auswahl der Untersuchungsparameter und der Abschätzungsformalismen über festgelegte Listenwerte (Grenzwerte, Richtwerte); diese hat z. B. DER RAT VON SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN (1990 und 1995) zusammengestellt.

Die Übertragung der bisher gesammelten Erkenntnisse und Bewertungsstrategien auf alte Mülldeponien (nachfolgend auch als Deponien bezeichnet) in den neuen Bundesländern ist bedenklich, da diese sich sowohl in der Technik des Abfalleinbaus als auch in der Abfallzusammensetzung erheblich von den Deponien in den alten Bundesländern unterscheiden. Die Ausbreitung der Schadstoffe über den „Wasserpfad" ist bei Deponien der ehemaligen DDR durch die in der Regel fehlenden Abdeckungen, Abdichtungen und Sickerwassererfassungen begünstigt. Die Ablagerungen liegen größtenteils in ungeordneter bzw. gemischter Form vor. Charakteristisch sind weiterhin der hohe Braunkohlenascheanteil und die geringe Müllverdichtung bei vermindertem Anteil an organischem Kohlenstoff. Dadurch ist der aerobe Abbau zu Beginn der Ablagerung begünstigt. Die Aschefraktionen wirken dabei immobilisierend, so dass beispielsweise der Austrag an organischen Schadstoffen relativ gering bleibt. Die Salzfrachten bilden bei ostdeutschen Mülldeponien die wesentlichen Schadstoffausträge in das Grundwasser (DER RAT VON SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN, 1995).

Richtung und Umfang des Austrages in den Grundwasserleiter können wegen der komplexen stofflichen Matrizes, der Vielfalt möglicher analytischer Parameter und ihrer ökologischen Auswirkungen sowie der relativ großen Anzahl von Messpunkten nur schwierig aus dem aktuellen Datenmaterial einer Deponie abgeleitet werden. Die zunehmende Automatisierung der

Bestimmungsmethoden (BROEKAERT, 1992; GUHL und WERNER, 1997) - insbesondere im Bereich der mobilen Vor-Ort-Analytik (FLACHOWSKY, 1998) - erlaubt es zudem, an interessierenden Untersuchungsobjekten (hier: Messstellen) in immer kürzerer Zeit immer mehr Parameter simultan zu bestimmen. Die zur Verfügung stehenden Datenmengen stellen sich somit in der Regel als sehr umfangreich und damit unübersichtlich dar.

Darüber hinaus ist durch die Kombination der Eigenschaften eines Müllkörpers - Grobstruktur mit geringem Wasserrückhalt und gleichzeitig Humusstruktur mit hohem Wasserrückhalt - eine genaue Beschreibung der Wasserbewegung in einem Müllkörper praktisch nicht möglich (EHRIG, 1989), wodurch die Vorbestimmung des Schadstoffaustrages erschwert wird.

Aus dieser Problemstellung resultierend bestand das Ziel der vorliegenden Arbeit in der Modifizierung und Anwendung von Methoden zur Auswertung komplexer Datenmengen eines Modellstandortes, einer Altdeponie in Leipzig. Dabei sollte primär ermittelt werden, ob sich anhand der Verteilung von Mustern - hierunter wird die geordnete Gesamtheit der an einer Sicker- oder Grundwasserprobenahmestelle gemessenen schadstoffrelevanten Merkmale (Parameter) verstanden - Rückschlüsse auf Kontaminationsquellen und Transportpfade des Sickerwassers ziehen lassen, um somit eine Aussage darüber zu treffen, wie und in welchem Umfang Sickerwasser aus dem Deponiekörper in den Grundwasserleiter gelangt. Des Weiteren sollten Erkenntnisse darüber gewonnen werden

- zwischen welchen der gemessenen Merkmale signifikante Korrelationen bestehen bzw. welche analytischen Parameter den Sickerwassertransport am besten beschreiben, um hieraus gesicherte Aussagen über die Zulässigkeit der Substitution besonders zeit- und kostenaufwendig zu bestimmender Parameter abzuleiten und
- ob sich aufwendige Analyseverfahren durch einfachere Methoden bei vergleichbarem Informationsgehalt substituieren lassen.

Die Lösung der Aufgabenstellung, das durchzuführende data mining (GROTELÜSCHEN, 1997), erforderte die Konzipierung eines spezifischen Algorithmus zur Auswertung der Daten sowie eine entsprechende Adaption der Methoden. Mit einfachen Verfahren der **beschreibenden Statistik** wurde zunächst eine Erstauswertung des umfangreichen Datenmaterials vorgenommen und anschließend über die klassischen statistischen Auswertemethoden **Korrelations-** und **Regressionsanalyse** sowie **Varianzanalyse** zum **Methodenspektrum der Mustererkennung** (überwachte und automatische Klassifikation) übergegangen. Des Weiteren wurde im modellhaften Ansatz eine **Zeitreihenanalyse** durchgeführt. Die Aufgabe der Mustererken-

nung ist eine typische Domäne auch der Neuronalen Netze. Durch den Einsatz eines **Neuro-Fuzzy-Klassifikationsverfahrens** erhält man einen Klassifikator auf der Grundlage linguistischer Regeln. Zur Anwendung kam das NEFCLASS-System, welches auf dem generischen Modell eines dreischichtigen Fuzzy-Perceptrons aufbaut.

Ein weiteres Ziel der Arbeit bestand darin, für den konkreten Anwendungsfall eine individuelle und vergleichende Bewertung der Methoden bezüglich des mit ihnen erzielten Informationsgehaltes sowie ihrer Adaptionsfähigkeit an die gegebene Aufgabenstellung vorzunehmen. Insbesondere die diesbezüglichen Erkenntnisse können in hohem Maße verallgemeinert werden und sind damit auch für andere praktische Anwendungsfälle relevant.

Das Thema der Arbeit steht im Zusammenhang mit den innerhalb des BMBF-Verbundprojekts REGNAL („Regeneration und nachhaltige Landnutzung hochbelasteter Ökosysteme (Landschaften) - der Ballungsraum Leipzig-Halle-Bitterfeld als Modellregion"; Projektzeitraum vom 1.1.1993 - 30.6.1996; Projektnummer 0339419 K) durchgeführten Untersuchungen zur Gefährdung auenlandschaftlicher Ökosysteme durch Deponien. Die generelle Bedeutung des Themas sowie der hohe Applikationsbedarf auf andere Modellstandorte sind beispielsweise dadurch gegeben, dass die untersuchte Altdeponie zu den knapp Dreiviertel der deutschen Deponien zu zählen ist, die laut UBA nicht nach dem Stand der Technik - hierzu gehören Oberflächen- und Basisabdichtung, Regenwasserdrainage und Sickerwasserbehandlung - betrieben werden (RIESE, 1998).

Den gegenwärtigen Veröffentlichungen auf diesem Gebiet ist zu entnehmen, dass die im Rahmen dieser Arbeit vorgenommene Anwendung von Methoden der Mustererkennung zur Bewertung der Schadstoffverteilung im Einzugsbereich einer Deponie gegenüber den ansonsten gemeinhin verwendeten Verfahren (BUNDESANSTALT FÜR GEOWISSENSCHAFTEN UND ROHSTOFFE, 1995; UMWELTBUNDESAMT, 1995 und 1997) einen alternativen und neuartigen Ansatz zur Lösung dieses Problems darstellt.

1.2 Beschreibung des Modellstandortes

Der Modellstandort befindet sich im Nordwesten der Stadt Leipzig im Waldgebiet der Elster-Luppe-Aue. Von 1937 (Beginn der Müllverbringung) bis 1983 (Schließung der Deponie) kamen hier ca. 4,5 Mill. m^3 Müll in größtenteils ungeordneter und unverdichteter Form zur Ablagerung. Dieser setzt sich zum Großteil aus Hausmüll, Bauschutt, Industrie- und Gewerbe-

müll (75 %) sowie Braunkohlenasche (15 %) zusammen. Die Entwicklungsgeschichte der Deponie, die zur Ablagerung gekommenen bekannten Abfallarten und -mengen sowie ein anzunehmender unbekannter Anteil weisen auf ein beträchtliches, ernstzunehmendes Schadstoffpotential hin. Die Niederschläge versickern aufgrund der fehlenden Oberflächenabdichtung direkt im Deponiekörper. Durch das eindringende Regenwasser, im folgenden Sickerwasser genannt, werden Inhaltsstoffe aus dem abgelagerten Müll gelöst, welche bei weiterem Durchdringen des Deponiekörpers wegen der ebenfalls fehlenden Basisabdichtung bis in den Grundwasserleiter transportiert werden und damit eine erhebliche Gefährdung desselben verursachen können (DASSOW et al., 1992; BARKOWSKI et al., 1993).

Vor der Müllverbringung wurde auf dem Gelände Lehm für die Ziegelverarbeitung ausgestochen. Die dadurch entstandenen Auslehmungen stellen eine besondere Gefahr für den unter dem Lehm gelegenen Grundwasserleiter dar, weil man davon ausgehen muss, dass zuerst diese Gruben, die an der Basis durch minimal 20 cm mächtigen Auelehm vom oberen Grundwasserleiter getrennt sind, mit Klärschlamm und Müll verfüllt wurden. Die Basis der Deponie befindet sich im ca. 1,4 m mächtigen Grundwasserschwankungsbereich, so dass man mit hoher Wahrscheinlichkeit von einer direkten Wechselwirkung zwischen Grund- und Stau- bzw. Sickerwasser, besonders in den genannten Bereichen verminderter Auelehmmächtigkeit, den ehemaligen Lehmgruben, ausgehen muss (FISCHER, 1993).

Nach DASSOW et al. (1992, S. 67) muss das Sickerwasser „ ... als extrem belastet eingestuft werden." Sie stellen weiterhin fest, dass die erforderlichen natürlichen Basisabdichtungselemente in keiner Weise ausreichend und zudem in ihrer Wirkungsweise durch anthropogene Eingriffe (Lehmgruben) beeinträchtigt bzw. in weiten Teilen vollkommen abgebaut sind. Somit erfolgt ständig ein direkter Eintrag kontaminierten Sickerwassers aus den Ablagerungen in das Grundwasser. Dabei ist insbesondere nach Niederschlägen eine rasche Änderung des Grundwasserchemismus im Deponieunterstrom zu verzeichnen.

Sickerwasserbilanzierungen und -behandlungen lassen sich nur über Sickerwassererfassungen vornehmen; diese sind am Modellstandort nicht vorhanden. Somit ist i. Allg. die Menge des gebildeten Sickerwassers nicht bekannt, Abschätzungen über mittlere Niederschläge und Verdunstungsraten sind außerordentlich fehlerhaft. Eine Vorbestimmung des Sickerwassertransfers im Deponiekörper ist aufgrund der örtlich inhomogenen Verteilungen der Müllbestandteile, des unterschiedlichen Alters der Deponiebereiche und der differenzierten Wasserwegsamkeiten im Müllkörper nicht möglich.

2 Probenahme, Probenahmepunkte, analytische Parameter und deren Bestimmungsmethoden

2.1 Probenahme und Probenahmepunkte

2.1.1 Grundwasser, Bodeninfiltration, Sickerwasser

Grundwasser ist unterirdisches Wasser, das die Hohlräume der Erdrinde zusammenhängend ausfüllt und dessen Bewegung ausschließlich oder nahezu ausschließlich von der Schwerkraft und den durch die Bewegung selbst ausgelösten Reibungskräften bestimmt wird. Es fließt, wenn Gefälle vorhanden ist. Unter **Infiltration** wird der Zugang von Wasser in die Erdrinde verstanden. Die Grundwasseroberflächen liegen verschieden tief, meist nicht unmittelbar unter der Erdoberfläche. Zwischen Erdoberfläche und Grundwasseroberfläche befindet sich jener Bereich, der von den in den Boden infiltrierenden Niederschlagsanteilen durchsickert werden muss, bevor diese das Grundwasser erreichen. Der Raum zwischen Erdoberfläche und Grundwasseroberfläche wird, da er nicht gänzlich mit Wasser ausgefüllt ist, als wasserungesättigte Bodenzone bezeichnet. Das in dieser Zone enthaltene Wasser ist kein Grundwasser, es wird als Wasser der ungesättigten Bodenzone, als **Sickerwasser**, bezeichnet (HÖLTING, 1995). Die Ausbreitungspfade sind in der ungesättigten Zone - in einer solchen befindet sich eine Deponie in der Regel - anders als in der gesättigten Zone, also im Grundwasserbereich. Die Ausbildung des Bodens, hier vor allem die Korngrößenverteilung, die Porosität, das Gefüge und der Mineralbestand, ist neben der Zusammensetzung und Konzentration des Sickerwassers ein entscheidender Faktor. Im System Boden-Boden-Grundwasser, das im Zusammenhang mit Deponien und industriellen Altlasten betrachtet wird, muss prinzipiell mit einem konvektiven und diffusen Transportvorgang gerechnet werden (CZURDA, 1992). Die an Abfluss und Niederschlag gekoppelten Quellen führen somit zu Grundwasserzuflüssen, die - sofern sie messtechnisch nicht fassbar sind - als diffus bezeichnet werden (JANDEL, 1998).

2.1.2 Probenahme

Die Grund- und Sickerwasserproben werden aus Beobachtungsbrunnen mit Pumpen, in Einzelfällen auch mit Schöpfern, entnommen, wobei die Pumpe bis in eine Filterstrecke hinab-

gelassen wird, s. Abb. 2-1. Vor Abfüllen der Wasserproben wird eine Mindestabpumpzeit (Richtwert: 0,5 h) eingehalten, um nicht im Brunnen stehendes Wasser zu entnehmen. Das Verfahren der Probenahme aus derartigen Probenahmestellen ist nach DIN 38402 - A 13 (1989) genormt und wird beispielsweise durch RUMP (1998) ausführlich erläutert.

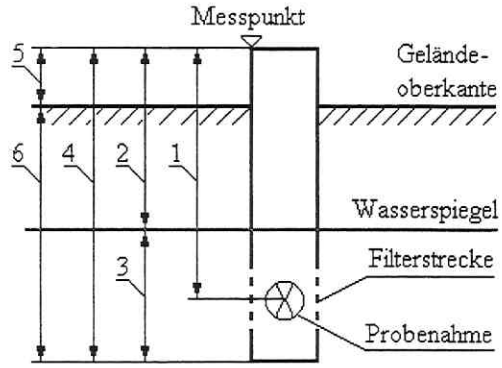

Abb. 2-1. Schema eines Beobachtungsbrunnens. - 1 Entnahmetiefe (m), 2 Wasserstand (m), 3 Wasserspiegelstand (m NN), 4 Rohroberkante (m NN), 5 Rohr über Geländeoberkante (m NN), 6 Geländeoberkante (m NN).

Die in diesem Zusammenhang nachfolgend ebenfalls als Probenahmestellen oder auch Messstellen bezeichneten Luppe und Nahle (beides Vorfluter) sowie SWA und SWA-Pegel sind genau genommen Fließgewässer; die Probenahme erfolgt hier entsprechend DIN 38402 - A 15 (1989).

2.1.3 Geographische Lage der Probenahmepunkte im Untersuchungsgebiet

Die Probenahme wird in der Regel an 11 Sickerwassermessstellen (SWP 1 - 5, 7, 9 - 11; SWA; SWA-Pegel) sowie an 14 Grundwassermessstellen (GWB 1 - 12; Nahle; Luppe) durchgeführt. Die geographische Lage der Probenahmepunkte im Untersuchungsgebiet ist Abb. 2-2 zu entnehmen.

Die Entwässerung des Untersuchungsgebietes erfolgt durch die Vorfluter Luppe und Nahle, die das Deponiegelände im Osten und Westen begrenzen. Nach DASSOW et al. (1992) trennt eine direkt unter dem Deponiekörper südost-nordwest verlaufende Wasserscheide die Einzugsbereiche beider Flüsse voneinander. Demzufolge sind im Deponiebereich zwei Grundwasserfließrichtungen zu verzeichnen. Östlich der Wasserscheide fließt das Grundwasser in die nördliche und westlich der Wasserscheide in die nordwestliche Richtung ab, wobei infolge des Untergrundaufbaus jedoch wechselnde Fließrichtungen auftreten können. Die GWB 7 - 9

liegen im Anstrombereich des Deponiekörpers. Die GWB 1 - 3 sind dem Abstrombereich der Luppe und die GWB 4 - 6 dem Abstrombereich der Nahle zuzuordnen.

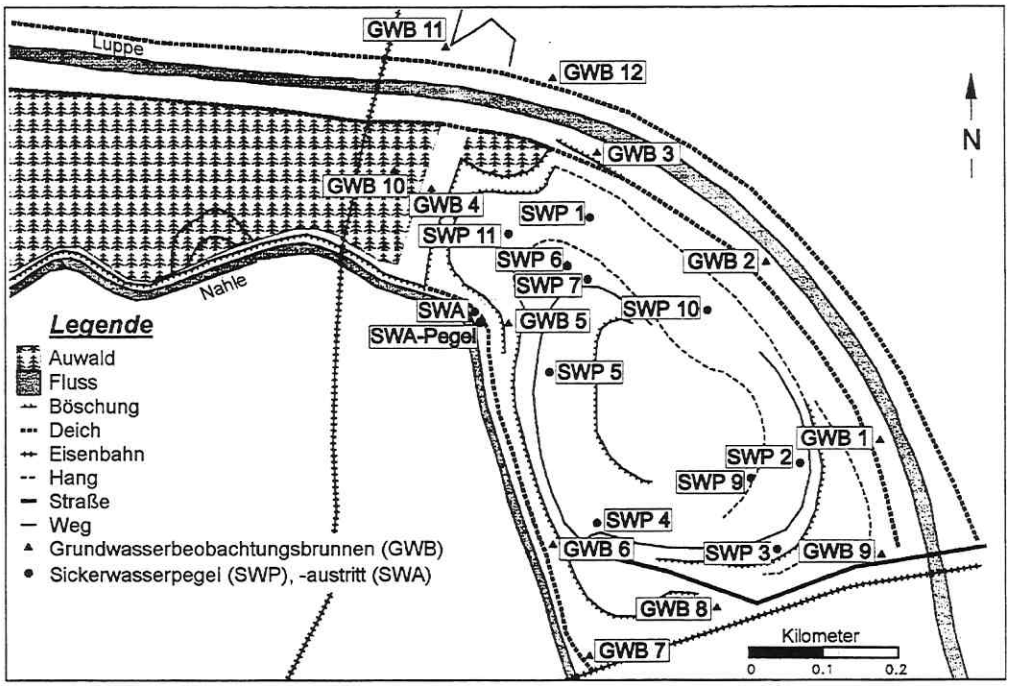

Abb. 2-2. Geographische Lage der Probenahmepunkte (MATTHES, 1995).

2.2 Analytische Parameter und deren Bestimmungsmethoden

2.2.1 Gesamtüberblick

Bei den im Zeitraum von Mai '92 bis April '98 an der Deponie durchgeführten Sicker- und Grundwassermessungen - eine genaue Übersicht hierüber gibt Tab. 2-2 - wurden jeweils die Werte von bis zu 56 Parametern aufgenommen. Für einige von diesen erfolgte eine Mehrfachbestimmung mit verschiedenen Analysetechniken - berücksichtigt man dies bei der Zählung, so erhöht sich die Gesamtanzahl auf 96. Die Messungen wurden in der Regel in Anlehnung an die jeweilige DIN-Vorschrift durchgeführt. Einen detaillierten Gesamtüberblick über sämtliche analytischen Parameter und deren Bestimmungsmethoden gibt RUDOLPH (1998a).

Bei den im Rahmen vorliegender Arbeit durchgeführten Untersuchungen wurde sich auf einige ausgewählte, zur Beschreibung der Schadstoffausbreitung als besonders relevant anzusehende (Erläuterungen hierzu s. nachfolgender Abschnitt), Parameter beschränkt, s. Tab. 2-1.

Tab. 2-1. Analytische Parameter und deren Bestimmungsmethoden.

Parameter			Bestimmungsmethode bzw. Gerät	DIN
Nr.	Name	Einheit		
1	Natrium	mg/l	AAS	DIN 38406 - E 3
2	Kalium	mg/l	AAS	DIN 38406 - E 3
3	Magnesium	mg/l	ICP-AES	DIN 38406 - E 22
4	Calcium	mg/l	ICP-AES	DIN 38406 - E 22
5	Chlorid	mg/l	IC	DIN 38405 - D 19
		mg/l	Titration, Titrator „TR 250" (Gerät der Fa. Schott)	DIN 38405 - D 1
		mg/l	Ionensensitive Elektroden	(keine)
6	Sulfat	mg/l	IC	DIN 38405 - D 19
		mg/l	Reaktionsküvetten, WTW-Photometer „MPM 1500"	(keine)
		mg/l	Säule mit anschließender Titration	DIN 38405 - D 5
		mg/l	gravimetrisch (fällen, absaugen, verglühen)	DIN 38405 - D 5
7	pH-Wert		Wasserchecker HORIBA	DIN 38404 - C 5
			Elektrode (pH-Meter „pH 196")	DIN 38404 - C 5
8	elektr. Leitfähigkeit	mS/cm	Wasserchecker HORIBA	DIN 38404 - C 8
		mS/cm	Elektrode (Konduktometer „LF 96")	DIN 38404 - C 8
9	Temperatur	°C	Wasserchecker HORIBA	DIN 38404 - C 4
		°C	Temperaturfühler	DIN 38404 - C 4
10	Wasserhärte	°dH	Titration (Titrator „TR 250")	DIN 38409 - H 6
11	CSB	mg/l	Titration (Titrator „TR 250")	DIN 38409 - H 41
		mg/l	Reaktionsküvetten, WTW-Photometer „MPM 1500"	(keine)
		mg/l	Küvettensatz der Dr. Bruno Lange GmbH	(keine)
12	DOC	mg/l	Gerät „liguiTOC"	DIN 38409 - H 3
13	Wasserspiegelstand	m	Lichtlot	DIN 38402 - A 13

Hierzu gehören vor allem die für Mülldeponien in den neuen Bundesländern typischen Salzfrachten, die primär durch die Einzelionen Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^- und SO_4^{2-} verkörpert werden (DER RAT VON SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN, 1995), aber auch die physikalisch-chemischen Kenngrößen pH-Wert, Temperatur und Leitfähigkeit, die gemeinhin bei Grundwasserprobenahmen von Interesse sind (SÄCHSISCHES LANDESAMT FÜR UMWELT UND GEOLOGIE, 1998). Zudem hat sich die Auswahl dieser sowie der

weiteren der in Tab. 2-1 aufgeführten Parameter bei Untersuchungen zur Schadstoffausbreitung an anderen Deponiestandorten (ANDREAS, 1993; BAUMANN et al., 1993; GADE, 1994) bereits als geeignet erwiesen.

Der pH-Wert, die elektrische Leitfähigkeit und die Temperatur wurden aufgrund ihrer Veränderlichkeit beim Transport direkt bei der Probenahme - in den meisten Fällen mittels Wasserchecker HORIBA - gemessen. Da es sich bei dieser mobilen Vor-Ort-Analytik genau genommen um die Anwendung von Feldmesstechnik (FLACHOWSKY, 1998) handelt, werden diese Größen nachfolgend häufig auch als Feldparameter bezeichnet. Die anderen Messgrößen wurden bis auf den Wasserspiegelstand (Lichtlot) ausschließlich laboranalytisch bestimmt.

2.2.2 Kurzbeschreibung der analytischen Parameter

Ein Forschungsschwerpunkt bei Langzeituntersuchungen an Deponien besteht darin herauszufinden, welche Emissionen auf welche abgelagerten Stoffe bzw. deren Reaktionen untereinander zurückgeführt werden können. Außerdem ist von Interesse, auf welchem Wege und durch welche „Trägersubstanzen" bestimmte Schadstoffe in die Deponie eingetragen werden, die dann durch Lösungs- und Auslaugungsprozesse mit dem Sickerwasser wieder austreten (UMWELTBUNDESAMT, 1995 und 1997). Die Auswahl der Parameter bzw. deren „Funktion" im Zusammenhang mit bestimmten (transportrelevanten) Vorgängen im Deponiekörper ist damit generell bei Untersuchungen zur Schadstoffausbreitung von besonderer Bedeutung - mit den nachfolgenden Ausführungen wird diesem Aspekt Rechnung getragen.

Die Salzfrachten im Sickerwasser sind, wie bereits erwähnt, als die wesentlichen Schadstoffausträge in das Grundwasser anzusehen. Die Gehalte an den betrachteten **Einzelionen** erklären sich überwiegend durch den Bauschuttanteil (Ca^{2+}, SO_4^{2-}) und durch den Ascheanteil (Mg^{2+}, Cl^-, in eingeschränktem Maße auch Na^+, K^+) des Müllkörpers (BARKOWSKI et al., 1993). Der Nachweis der Kationen erfolgte mittels Verfahren der AAS (Na^+, K^+) bzw. der AES (Mg^{2+}, Ca^{2+}) (HEIN und KUNZE, 1995) entsprechend den geltenden DIN-Vorschriften DIN 38406 - E 3 (1989) bzw. DIN 38406 - E 22 (1989). Die Konzentrationen der Anionen wurden größtenteils mittels Ionenchromatographie nach DIN 38405 - D 19 (1989) bestimmt. Um eine Vergleichbarkeit der ermittelten Stoffkonzentrationen zu gewährleisten, wurden für die Untersuchungen sämtliche der in mg/l gemessenen Ionenkonzentrationen in die äquimolare Größe mval/l umgerechnet (KOEHNE, 1948).

Der **pH-Wert** liegt in natürlichen Gewässern meistens zwischen 6,5 und 7,5. Abweichungen nach unten ergeben sich u. a. durch den Gehalt an freiem CO_2, das durch aerobe und anaerobe Prozesse entstehen kann (RUMP, 1998).

Temperaturerhöhungen im Sickerwasser können als Indikator für ablaufende biochemische Reaktionen im Deponiekörper angesehen werden. Beispielsweise laufen in Deponien ohne bzw. mit geringer Verdichtung sowie in frisch eingebautem Müll kurzzeitig aerobe Prozesse ab. Hierbei werden die leicht abbaubaren organischen Substanzen oxidiert, als Reaktionsprodukte entstehen im Wesentlichen Kohlendioxid (CO_2) und (Sicker-) Wasser. Dabei kann es zu Temperaturerhöhungen von bis zu 70°C im Deponiekörper kommen (ANDREAS, 1993).

Als Summenparameter ist die **elektrische Leitfähigkeit** ein Maß für die Gesamtheit der in einer Probe gelösten Salze (Ionen). Deutlich erhöhte Werte der Leitfähigkeit im Abstrombereich einer Altablagerung geben häufig bereits einen Hinweis auf eine Beeinflussung des Grundwassers durch Deponiewässer (BARKOWSKI et al., 1993).

Die **Härte eines Wassers** ist der Gehalt eines Wassers an Calcium-Ionen (Ca^{2+}) und Magnesium-Ionen (Mg^{2+}) (DIN 38409 - H 6, 1989). Als Summenparameter ist sie somit ein Vergleichsindikator für die beiden Einzelparameter.

Bezüglich der Belastung von Wasser oder Abwasser mit organischen Substanzen - hervorgerufen durch große Teile des Hausmülls sowie Garten- und Parkabfälle - sind die Summenparameter CSB (Chemischer Sauerstoffbedarf) und DOC (gelöster organischer Kohlenstoff) von Bedeutung. Der **CSB** gibt die Menge an Sauerstoff an, die in Form von Oxidationsmitteln für die Oxidation organischer Wasserinhaltsstoffe verbraucht wird (RUMP, 1998). Die Bestimmung erfolgte größtenteils mittels Titration, dem Normverfahren nach DIN 38409 - H 41 (1989). Bei den Messungen ab September '95 kamen verstärkt Küvettentests hinzu. Der **DOC** ist ebenfalls ein Maß für den Anteil gelöster organischer Verbindungen - das bei der Oxidation dieser Verbindungen entstehende CO_2 wird IR-spektroskopisch (DIN 38409 - H 3) analysiert. Beim Sickerwasser sind infolge der biologischen Abbauprozesse des Mülls hohe CSB- und DOC-Werte vorhanden, beim Grundwasser sind diese theoretisch niedrig. Eine Erhöhung der Werte beim Grundwasser kann somit als Indikator für einen Schadstoffeintrag in das Grundwasser angesehen werden. Im Vergleich der Normverfahren kann der DOC mit experimentell wesentlich geringerem Aufwand als der CSB bestimmt werden. Vor allem aus diesem Grund wurde in den letzten Jahren sowohl eine verstärkte Diskussion über das Für und Wider beider Summenparameter geführt (FUNK und KLEIN, 1994) als auch die Entwicklung be-

triebsanalytischer Messmethoden in Form von Küvetten-Tests für den CSB vorangebracht (PASS und SCHMIDT, 1998).

Es ist anzumerken, dass internationale und nationale Richtlinien, Empfehlungen und Richtwerte wesentliche Hilfsmittel bei der Auswertung und Beurteilung der Messergebnisse von Wasser- und Bodenanalysen sind. In diesem Zusammenhang sei insbesondere auf die Holländische Liste (LEIDRAAD BODEMSANERING, 1995) verwiesen. Die hierin enthaltenen Richtwerte für Boden- und Grundwasserkontaminationen werden häufig als Richtlinie bei der Gefährdungsabschätzung von Altlasten betrachtet (HEIN und KUNZE, 1995). Die landeshoheitlichen Richtlinien zu Prüf- und Maßnahmewerten sind nach den entsprechenden LAGA- und LAWA-Empfehlungen festgelegt. Speziell für den Deponiebetrieb und die Deponiekontrolle gelten die Vorschriften der TA Abfall (WAGNER, 1995) und TA Siedlungsabfall (BERGS, 1997).

2.3 Das auszuwertende Datenmaterial

Eine Aufstellung der Messkampagnen, aus denen das für die Untersuchungen verfügbare Datenmaterial hervorging, ist in Tab. 2-2 enthalten. Die Ergebnisse der 23 Sicker- und 27 Grundwassermesskampagnen wurden jeweils in einer (Roh-) Datenmatrix der Form

$$\underset{(n,p)}{\underline{X}} = \left(x_{ij}\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \text{Objektmuster } 1 \\ \text{Objektmuster } 2 \\ \\ \cdot \\ \cdot \\ \cdot \\ \text{Objektmuster } n \end{pmatrix} = \begin{pmatrix} \text{V} & \text{V} & & \text{V} \\ \text{a} & \text{a} & & \text{a} \\ \text{r} & \text{r} & & \text{r} \\ \text{i} & \text{i} & & \text{i} \\ \text{a} & \text{a} & \cdots & \text{a} \\ \text{b} & \text{b} & & \text{b} \\ \text{l} & \text{l} & & \text{l} \\ \text{e} & \text{e} & & \text{e} \\ 1 & 2 & & p \end{pmatrix}$$

zusammengestellt. In der Sickerwasser-Datenmatrix sind n = 253 Objekte (11 Messstellen·23 Messkampagnen) und in der Grundwasser-Datenmatrix n = 378 Objekte (14 Messstellen·27 Messkampagnen) enthalten.

Beide Matrizen bestehen aus p = 56 bzw. (bei Einbeziehung der teilweise verschiedenen Bestimmungsmethoden) aus p = 96 Parametern (nachfolgend auch als Variablen, Merkmale oder Messgrößen bezeichnet), wobei in die Untersuchungen vorliegender Arbeit nur die im vorhergehenden Abschnitt (s. Tab. 2-1) genannten Parameter einbezogen wurden. Existieren für diese mehrere Bestimmungsmethoden, so wurden - falls an entsprechender Stelle nicht anders

erwähnt - generell die mit der in Tab. 2-1 erstgenannten Methode ermittelten und in größter Anzahl vorliegenden (s. Tab. A-1 und A-2 des Anhangs) Werte verwendet.

In Abhängigkeit von der angewandten Auswertemethode erfolgte eine Vorbehandlung des Datensatzes, z. B. durch Standardisierungen, arithmetische Mittelwertbildungen u. a.. Darauf wird ebenfalls in den entsprechenden Abschnitten hingewiesen. Die bei einigen Untersuchungen aufgrund fehlender Einzelwerte vorab vorgenommene objektweise Streichung der Daten führte dazu, dass hier die Daten nicht aller der in Tab. 2-2 aufgeführten Messkampagnen berücksichtigt werden konnten.

Tab. 2-2. Aufstellung der Messkampagnen.

Monat	Sickerwassermesskampagne	Grundwassermesskampagne
Mai 1992	13.05.	13.05.
Oktober 1992 (I)	06.10.	06.10.
Oktober 1992 (II)	15.10.	15.10.
November 1992 (I)	(keine Messkampagne)	04.11.
November 1992 (II)	(keine Messkampagne)	17.11.
Dezember 1992	07.12.	07.12.
März 1993	24.03.	24.03.
April 1993 (I)	(keine Messkampagne)	05.04.
April 1993 (II)	(keine Messkampagne)	14.04.
Mai 1993	(keine Messkampagne)	19.05.
September 1993 (I)	16.09.	16.09.
September 1993 (II)	30.09.	30.09.
Oktober 1993	07.10.	07.10.
Januar 1994 (I)	10.01.	10.01.
Januar 1994 (II)	25.01.	25.01.
März 1994 (I)	02.03.	02.03.
März 1994 (II)	22.03.	22.03.
Juni 1994	02.06.	(keine Messkampagne)
August 1994	30.08.	29.08.
Oktober 1994	26.10.	26.10.
Dezember 1994	07.12.	06.12.
Februar 1995	09.02.	08.02.
März 1995	07.03.	08.03.
April 1995	04.04.	05.04.
Juni 1995	14.06.	14.06.
September 1995	05.09.	05.09.
November 1997	03.11.	03.11.
April 1998	20.04.	20.04.

3 Methoden zur Datenanalyse

3.1 Allgemeiner Überblick

Das für die Untersuchungen vorliegender Arbeit angewandte Methodenspektrum stellt sich als relativ breit und vielfältig dar. Zu dessen Beschreibung lassen sich eine Vielzahl von Begriffen wie z. B. univariate und multivariate Datenauswertung, Mustererkennung und Musterklassifikation sowie exploratives und konfirmatorisches Vorgehen verwenden. In der entsprechenden Literatur werden hierzu sowie zu einer möglichen orientierenden Klassifizierung datenanalytischer Auswertemethoden umfangreiche Hinweise gegeben - hier sei insbesondere auf HENRION et al. (1988), HENRION und HENRION (1994) sowie EINAX et al. (1997) verwiesen. Die nachfolgenden Ausführungen verschaffen zu dieser Thematik einen komprimierten Überblick.

Die Datenanalyse allgemein lässt sich in zwei Phasen einteilen - die **explorative** und die **konfirmatorische**. Die explorative (experimentelle) Datenanalyse als empirische Wissenschaftsdisziplin ist eine Entwicklung des Angloamerikaners J. W. Tukey (TUKEY, 1962). Sie umfasst die Suche nach unbekannten Strukturen und Auffälligkeiten in vorliegenden Daten mit dem Ziel, ein passendes Modell für die Daten zu finden und Hinweise zu einer kausalen Interpretation zu erhalten, sowie das Aufspüren von Hypothesen. Typische, auf mehrdimensionale Datensätze anwendbare, explorative Methoden sind beispielsweise die Clusteranalyse (Kapitel 8) und die Hauptkomponentenanalyse (Kapitel 9). Die konfirmatorische („erklärende") Datenanalyse beinhaltet die Sicherung reproduzierbarer Effekte oder Muster in den Daten. Wie in der mathematischen Statistik werden dabei Begriffe wie Irrtumswahrscheinlichkeit und Konfidenz verwendet, um z. B. die Wahrscheinlichkeit für eine mögliche Fehlinterpretation anzugeben. Zu dem Methodenspektrum hier sind beispielsweise die Korrelations- und die Regressionsanalyse (Kapitel 5) sowie die Varianzanalyse (Kapitel 6) zu zählen (NAGEL et al., 1988; FLEISCHER und NAGEL, 1989).

Weiterhin unterscheidet man prinzipiell zwischen einer **univariaten** und einer **multivariaten** Auswertung der Daten. Bei ersterer erfolgt eine Datenreduktion in dem Sinne, dass die (ggf. gruppiert vorliegenden) Werte eines messbaren Merkmals durch die zugehörigen statistischen Kenngrößen (Mittelwerte, Streumaße) und die Zahl der zugrunde liegenden Freiheitsgrade ersetzt werden. Daran anschließend kann beispielsweise die Frage beantwortet werden, ob sich

die a priori gegebenen Gruppen des betrachteten Merkmals voneinander unterscheiden oder ob die von Gruppe zu Gruppe beobachteten Unterschiede in den Messungen zufälligen Charakters sind. Derartige Untersuchungen werden im Kapitel 6 (Varianzanalyse) beschrieben. Die inhaltliche Abgrenzung dessen, was im Einzelnen unter multivariater Analyse zu subsummieren ist, erweist sich in der Literatur als keineswegs einheitlich. So werden beispielsweise die Methoden der multivariaten linearen Regression nur teilweise hierunter erfasst, insbesondere wegen der mit ihnen verfolgten praktisch nur merkmalsorientierten Auswertung (SCHULZE, 1998). Die in Kapitel 5 im Rahmen bivariater Häufigkeitsverteilungen betrachteten Korrelations- und Regressionsanalysen stellen den einfachsten Fall dieser multivariaten Verfahren dar. HENRION und HENRION (1994) heben hervor, dass sich die multivariate Datenanalyse mit der Systematisierung von Merkmalsmustern beschäftigt. Unter einem Muster wird dabei die geordnete Gesamtheit der Beobachtungs- bzw. Messergebnisse an einem Objekt für die verschiedenen Merkmale (Variablen, Parameter) verstanden. Dementsprechend werden die Verfahren der überwachten und der automatischen Klassifikation (Kapitel 7 und 8) sowie die Hauptkomponentenanalyse (Kapitel 9) dem Instrumentarium der **Mustererkennung** (pattern recognition) zugeordnet.

NAUCK et al. (1996a) nehmen im Zusammenhang mit den Lernparadigmen Neuronaler Netze und deren Einsatz im Bereich der Datenanalyse eine Unterscheidung zwischen **Musterklassifikation** (pattern classification) und **Mustererkennung** (pattern recognition) vor. Für die Musterklassifikation werden demnach Neuronale Netze (bzw. Neuronale Fuzzy-Systeme) eingesetzt, welche einen überwachten Lernalgorithmus (mit einer entsprechenden festen Lernaufgabe) verwenden und für die Mustererkennung solche, die auf der Basis eines nicht überwachten Lernalgorithmus eine freie Lernaufgabe erfüllen. Das in Kapitel 11 beschriebene NEFCLASS-System ist ein hybrides Neuro-Fuzzy-System. Es wird im Bereich der Datenanalyse zur Musterklassifikation eingesetzt und kann somit als (wissensbasierte) Methode den in Kapitel 7 betrachteten Verfahren der überwachten Klassifikation gegenübergestellt werden.

Eine weitere Möglichkeit, die Methoden der multivariaten Datenanalyse zu systematisieren, besteht darin, eine Unterteilung hinsichtlich ihrer Anwendung auf **homogene** oder **gruppierte Datensätze** vorzunehmen. Liegen erstere vor, so können beispielsweise die Verfahren der Clusteranalyse sowie die Hauptkomponentenanalyse eingesetzt werden. Ist hingegen von vornherein eine objektweise (Klassen von Objekten unterschiedlicher, aber jeweils bekannter Herkunft) oder auch variablenweise (unabhängig einstellbare bzw. davon abhängige, resultie-

rende Variablen) Gruppierung gegeben, dann kommen Methoden der überwachten Klassifikation bzw. der multivariaten Regression zur Anwendung (HENRION und HENRION, 1994). Eine Möglichkeit der orientierenden Klassifizierung der in die vorliegende Arbeit einbezogenen datenanalytischen Auswertemethoden ist in Tab. 3-1 zusammengefasst. Sie beruht auf einem Vorschlag von KNOBLOCH und ZWANZIGER (1995) und stellt eine vereinfachte Zusammenfassung der in diesem Abschnitt besprochenen Thematik dar.

Tab. 3-1. Orientierende Klassifizierung datenanalytischer Auswertemethoden.

Vorwissen	Vorgehen	Orientierung auf Objekte	Orientierung auf Merkmale
ohne	explorativ	z. B. Clusteranalyse; Hauptkomponentenanalyse	z. B. Korrelationsanalyse; Hauptkomponentenanalyse; Clusteranalyse
mit	konfirmatorisch	z. B. Diskriminanzanalyse; Zeitreihenanalyse	z. B. Varianzanalyse (hier: univariat); Regressionsanalyse (hier: bivariat); Zeitreihenanalyse

3.2 Realisierung der Methoden durch Software-Programme

Zur Realisierung der Methoden wurden im Rahmen der Untersuchungen verschiedene Software-Programme eingesetzt. Tab. 3-2 verschafft hierüber einen Gesamtüberblick. In den entsprechenden Kapiteln wird ggf. noch näher auf die Programme eingegangen.

Tab. 3-2. Realisierung der Methoden durch Software-Programme.

Kapitel	Methode	Software-Programm	Literaturhinweis
4	Beschreibende Statistik	STATISTICA, Version 5.1	STATSOFT, 1996
5	Korrelation, Regression	STATISTICA, Version 5.1	STATSOFT, 1996
6	Varianzanalyse	STATISTICA, Version 5.1	STATSOFT, 1996
7	k nächste Nachbarn	MULTIDAT	HENRION u. HENRION, 1994
7	Diskriminanzanalyse	STATISTICA, Version 5.1 MULTIDAT	STATSOFT, 1996 HENRION u. HENRION, 1994
8	Hierarchische CA	STATISTICA, Version 5.1 MULTIDAT	STATSOFT, 1996 HENRION u. HENRION, 1994
8	Nichthierarchische CA	MULTIDAT	HENRION u. HENRION, 1994
8	Fuzzy-Clusteranalyse	ECO-FUCS, Version 2.0	PAASCH, 1994
9	Hauptkomponentenanalyse	STATISTICA, Version 5.1	STATSOFT, 1996
10	Zeitreihenanalyse	STATISTICA, Version 5.1	STATSOFT, 1996
11	Neuro-Fuzzy-Datenanalyse	NEFCLASS-PC, Version 2.04	NAUCK et al., 1996a

4 Beschreibende mathematische Statistik

4.1 Vorbemerkungen

Nach dem Sammeln und Zusammenstellen der Daten wurde mit den Methoden der beschreibenden Statistik das sehr umfangreiche und damit unübersichtliche Datenmaterial zunächst geordnet und auf einzelne charakteristische Werte reduziert. Dies geschah in Form von Tabellen und graphischen Darstellungen sowie der Berechnung von bestimmten Kenngrößen.
Mit Hilfe solcher Kenngrößen, die man auch als statistische Maßzahlen bezeichnet, kann man eine aus n Messwerten $x_1, x_2, \ldots, x_n$ bestehende Folge (Messreihe) - genau genommen handelt es sich dabei um eine Stichprobe vom Umfang n aus der Grundgesamtheit X - durch einen einzigen Wert charakterisieren. Die Berechnung statistischer Kenngrößen bezieht sich immer nur auf ein messbares Merkmal X. Ausführliche Erläuterungen zu dieser Thematik nehmen beispielsweise OSE et al. (1974) sowie SCHÄFER und BUSSE (1978) vor.
In den nachfolgenden beiden Abschnitten werden die in den Untersuchungen berechneten statistischen Kenngrößen und verwendeten graphischen Darstellungen kurz erläutert und in Abschnitt 4.4 die diesbezüglichen Ergebnisse präsentiert.

4.2 Statistische Kenngrößen

Um für die ausgewählten analytischen Parameter eine Charakterisierung der empirischen Verteilung vorzunehmen, wurden das **arithmetische Mittel** $\bar{x}$ sowie der **Median** $\tilde{x}$ berechnet.
Während beim arithmetischen Mittel die einzelnen Messwerte $x_1, x_2, \ldots, x_n$ wertmäßig in die Berechnung eingehen, ist bei der Bestimmung des Medians - dieser wird in der Literatur häufig auch als Zentralwert oder 50 %-Perzentil bezeichnet - nur die Stellung, d. h. die Lage der einzelnen Messwerte zueinander, von Bedeutung. Der Median $\tilde{x}$ einer aus n Messwerten $x_1, x_2, \ldots, x_n$ bestehenden Folge ist derjenige Wert, der die nach der Größe der einzelnen Messwerte geordnete Folge $x_1^* \leq x_2^* \leq \ldots \leq x_n^*$ halbiert. Er wird vorzugsweise angewandt, wenn

a.) unter den Messwerten einige extreme Werte (sog. Ausreißer) auftreten, die das arithmetische Mittel stark beeinflussen und es zu einer fiktiven Größe machen würden,

b.) wertmäßige Veränderungen unterhalb und oberhalb des Mittelwertes sich nicht auf diesen auswirken sollen (OSE et al., 1974).

Die Charakterisierung empirischer Verteilungen durch die Mittelwerte ist i. Allg. nicht ausreichend, da zwei Folgen aus verschiedenen Beobachtungswerten z. B. ein und dasselbe arithmetische Mittel besitzen können, die Messwerte der Folgen sich jedoch durch ihre Lage zu dem betreffenden Mittelwert unterscheiden können; die Ausbreitung der Beobachtungswerte um einen feststehenden Wert bezeichnet man als empirische Streuung (OSE et al., 1974).
Im Rahmen der durchgeführten Untersuchungen wurden daher des Weiteren die **Variationsbreite R** (= Spannweite oder auch (engl.) range), die **mittlere quadratische Abweichung s** (= Standardabweichung), die **Varianz s^2** (= Dispersion) sowie der **Quartilsabstand** ermittelt. Letzterer berechnet sich aus der Differenz zwischen dem **unteren Quartil** (= 25 % -Perzentil) und dem **oberen Quartil** (= 75 %-Perzentil) einer aus n Messwerten $x_1, x_2, \ldots, x_n$ bestehenden Folge (Messreihe), in ihm liegt folglich die Hälfte der Messwerte. Als Maß für die Streuung ist er - analog dem Median - unempfindlich gegenüber Ausreißern (NAGEL et al., 1988).

4.3 Graphische Darstellungen

Die Kenngrößen Median, unteres Quartil, oberes Quartil und Quartilsabstand können mittels **Box-Plots** anschaulich dargestellt werden. Es werden exemplarisch die im gesamten Untersuchungszeitraum (Mai '92 bis April '98) an den Sickerwasserprobenahmestellen mittels Wasserchecker HORIBA gemessenen Werte der Leitfähigkeit herangezogen, s. Abb. 4-1.

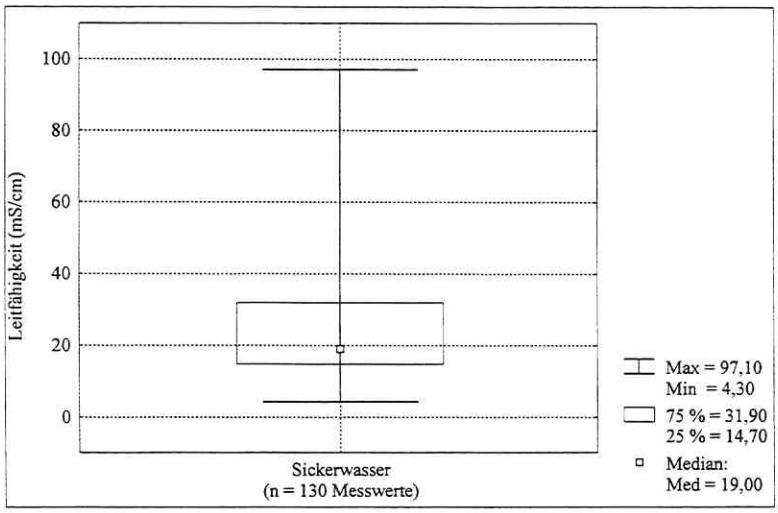

Abb. 4-1. Box-Plot.

Durch diese Darstellungsform lässt sich u. a. die Verteilung der Leitfähigkeitsdaten gut erkennen: Diese ist rechtsschief, da der Abstand zwischen unterem Quartil und Median kleiner ist als der zwischen oberem Quartil und Median und zudem der vom oberen Quartil ausgehende sog. Schwanz (Entfernung bis zum Maximum) länger ist. In der Box, welche 50 % der Daten umfasst, ist der Median markiert. Die Höhe der Box, welche sich aus dem Quartilsabstand ergibt, ist ein Maß für die „mittlere Breite" der Datenmenge (NAGEL et al., 1988).
Zum Vergleich von zwei oder auch mehreren Datengruppen derselben physikalischen Einheit ist das **Multiple Box-Plot** geeignet. Als Beispiel werden die aus dem gesamten Untersuchungszeitraum hervorgegangenen Messwerte der Leitfähigkeit des Sickerwassers und des Grundwassers (Wasserchecker HORIBA) verwendet. Die Gegenüberstellung der statistischen Kenngrößen führt zu Abb. 4-2. Aus der gemeinsamen Darstellung der beiden Datenmengen sind Lage- und Streuungsunterschiede direkt ablesbar.

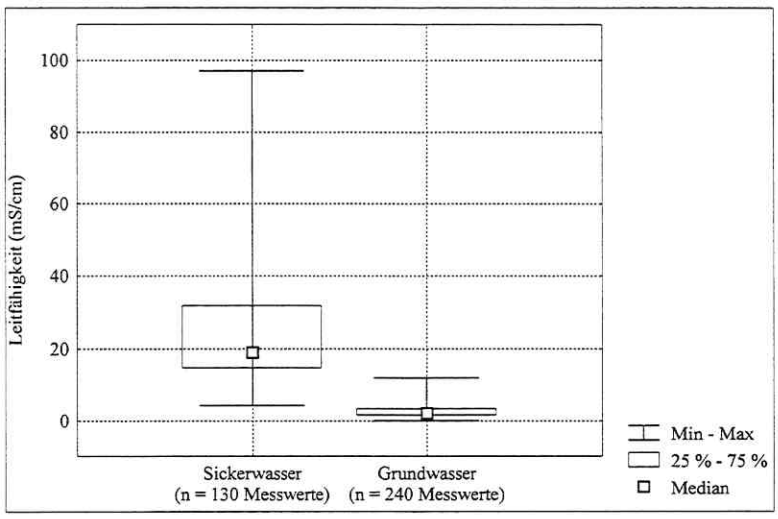

Abb. 4-2. Multiple Box-Plot.

Mit Hilfe eines **Histogramms** erhält man einen Überblick über die Häufigkeitsverteilung der Werte eines Merkmals und kann so Muster und Eigenschaften erkennen, wie z. B. symmetrische oder schiefe Verteilung der Daten, Ein- oder Mehrgipfligkeit der Daten, gehäuftes Auftreten von Werten, weit abgelegene Werte (Ausreißer) usw.. Das o. g. Beispiel (Leitfähigkeitsdaten des Sickerwassers im Untersuchungszeitraum) führt zur graphischen Darstellung gemäß Abb. 4-3.

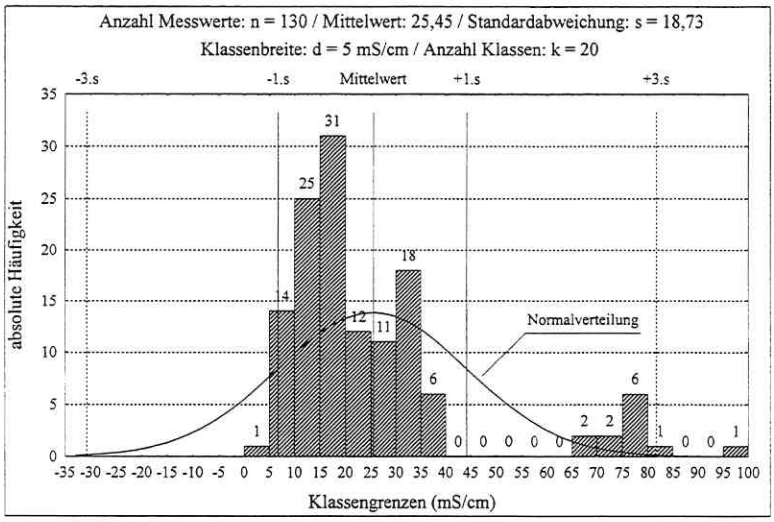

Abb. 4-3. Histogramm.

Weiterhin ist damit ein Vergleich der empirischen Daten mit einem Verteilungsmodell möglich. In Abb. 4-3 sind die Residuen - diese sind das Ergebnis einer Subtraktion der gemessenen Daten von einem an diese Daten angepassten Modell (NAGEL et al., 1988) - zwar nicht direkt ablesbar, der Unterschied zwischen dem Modell der (hier gewählten) GAUßverteilung und den empirischen Werten (Darstellung durch Häufigkeitssäulen) wird jedoch deutlich. Auf den Säulen sind die absoluten Häufigkeiten der Daten für die betreffende Klasse angegeben.

STORM (1995) schlägt vor, die für alle k Klassen konstant zu wählende Klassenbreite d so vorzunehmen, dass man nicht weniger als sechs und nicht mehr als 20 Klassen ($6 \leq k \leq 20$) erhält. Eine Veränderung der Klassenbreite führt zu Veränderungen der Häufigkeiten. Eine Veränderung der Klasseneinteilung bei gleichbleibender Klassenbreite bringt ebenfalls Veränderungen der Häufigkeiten (verschiedene Reduktionslagen) mit sich. Die graphische Darstellung der empirischen Verteilung wird dadurch jeweils beeinflusst (STORM, 1995).

Abschließend sei das **Ganglinienplot** erwähnt, welches - angeregt durch BAUMANN et al. (1993) - ebenfalls für die Untersuchungen genutzt wurde. Mit diesem lassen sich durch die gemeinsam vorgenommene räumliche und zeitliche Differenzierung sowohl Zusammenhänge zwischen den Messstellen als auch evtl. saisonale Tendenzen erkennen. Weiterhin kann durch den gezielten Vergleich eines Summenparameters mit den entsprechenden Einzelparametern der Grad der Korrelation visualisiert werden, um daraus Aussagen über die Zulässigkeit des Summenparameters zur Langzeitüberwachung abzuleiten.

4.4 Untersuchungen

4.4.1 Gegenstand und Zielstellung der Untersuchungen

Es wurden - jeweils für die Sicker- und Grundwasserdaten - die statistischen Maßzahlen der in Tab. 2-1 aufgeführten Parameter berechnet. Hierzu wurden die aus dem gesamten Untersuchungszeitraum (Mai '92 bis April '98) hervorgegangenen Messwerte zusammengefasst. Dadurch sollte zunächst ein komprimierter Überblick über das umfangreiche Datenmaterial, d. h. über die Verteilung der Grundgesamtheiten, gegeben werden. Die durch die graphischen Darstellungen generell verfolgten Zielstellungen wurden im vorhergehenden Abschnitt erläutert. Um den Seitenumfang vorliegender Arbeit in vertretbarem Rahmen zu halten, musste sich auf einige ausgewählte, für die Interpretation jedoch als relevant und repräsentativ anzusehende Graphiken beschränkt werden.

1. Box-Plot und Multiple Box-Plot: Die Darstellung statistischer Kenngrößen der Einzelionen sollte differenzierte Aussagen zu den Salzfrachten ausgewählter Probenahmestellen ermöglichen. Aus diesem Grund wurden die Werte der Parameter wie folgt gruppiert:
   - alle Sickerwassermessstellen
   - GWB 7, 8, 9 (Grundwasser-Anstrombereich der Deponie)
   - GWB 1, 2, 3 (Grundwasser-Abstrombereich der Luppe)
   - GWB 4, 5, 6 (Grundwasser-Abstrombereich der Nahle)

   Mit Hilfe von Multiple Box-Plots wurden jeweils von zwei Datengruppen (Sickerwasser, Grundwasser) der ausgewählten Messgrößen die Lage- und Streuungsunterschiede untersucht.
2. Histogramm: Es wurden die Daten der Einzelionen (Zusammenfassung der Werte jeweils aller Sicker- und Grundwasserprobenahmestellen sowie nur der von einzelnen Messstellen) untersucht mit dem Ziel, bestimmte Muster bzw. Eigenschaften zu erkennen und festzustellen, inwieweit sich die empirischen Daten der Normalverteilung annähern. Die Klassenbreite d wurde in Anlehnung an STORM (1995) so festgelegt, dass $6 \leq k \leq 20$ gilt.
3. Ganglinienplot: Es wurden - jeweils für das Sicker- und Grundwasser - folgende Summenparameter mit den entsprechenden spezifisch bestimmten Einzelparametern verglichen:
   - Leitfähigkeit $\leftrightarrow$ (berechnete) Summe der Na^+-, K^+-, Mg^{2+}-, Ca^{2+}-, Cl^-- und SO_4^{2-}-Ionen
   - Wasserhärte $\leftrightarrow$ (berechnete) Summe der Mg^{2+}- und Ca^{2+}-Ionen

4.4.2 Untersuchungsergebnisse

Die statistischen Maßzahlen sind in den Anlagen, Tab. A-1 (Sickerwasser) und A-2 (Grundwasser) enthalten. Die ausgewählten graphischen Darstellungen sind den folgenden Abbildungen zu entnehmen:

1. Darstellung statistischer Kenngrößen ausgewählter Messgrößen mittels Box-Plots
   - Na^+-Ionen (Sickerwassermessstellen; GWB 7 - 9; GWB 1 - 3; GWB 4 - 6): s. Abb. 4-4
   - vergleichende Darstellung (Sicker- und Grundwasser) statistischer Kenngrößen der Temperatur (Multiple Box-Plot): s. Abb. 4-5
2. Darstellung der Häufigkeitsverteilung der Na^+-Ionen mittels Histogramms
   - Sickerwasser (alle Messstellen): s. Abb. 4-6
   - Grundwasser (alle Messstellen): s. Abb. 4-7
   - Grundwasser (GWB 5): s. Abb. 4-8
3. Ganglinienplot
   - Leitfähigkeit und entsprechende Summe der Ionen (Grundwasser): s. Abb. 4-9
   - Wasserhärte und entsprechende Summe der Ionen (Sickerwasser): s. Abb. 4-10

4.4.3 Diskussion der Untersuchungsergebnisse

Durch die Darstellung statistischer Kenngrößen der Einzelionen mittels Box-Plots (hier gezeigt anhand der Na^+-Ionen, s. Abb. 4-4) werden drei Aspekte deutlich:

1. Die Salzfrachten des Sickerwassers sind deutlich höher als die des Grundwassers (s. auch Abb. 4-2, wo dies durch die Leitfähigkeit bestätigt wird), wobei die breite Streuung der Werte (range = 864,01 mval/l) jedoch ein Indiz für lokal unterschiedliche Belastungen ist.
2. Im Grundwasser-Anstrombereich des Deponiekörpers sind die Salzfrachten geringer als im Abstrombereich, was ein Hinweis auf den prinzipiellen Salzfrachteneintrag durch den Deponiekörper in das Grundwasser ist.
3. Im Abstrombereich der Nahle sind höhere Salzfrachten vorhanden als im Abstrombereich der Luppe. Die Ursachen hierfür sind erhöhte Schadstoffausträge durch diesen Bereich des Deponiegeländes sowie Wasserscheiden unter dem Müllkörper, welche zu erhöhten Wasserwegsamkeiten in Richtung Nahle führen. Dadurch ist in diesem Bereich der Deponie die Schadstoffbelastung für das Grundwasser offensichtlich am höchsten.

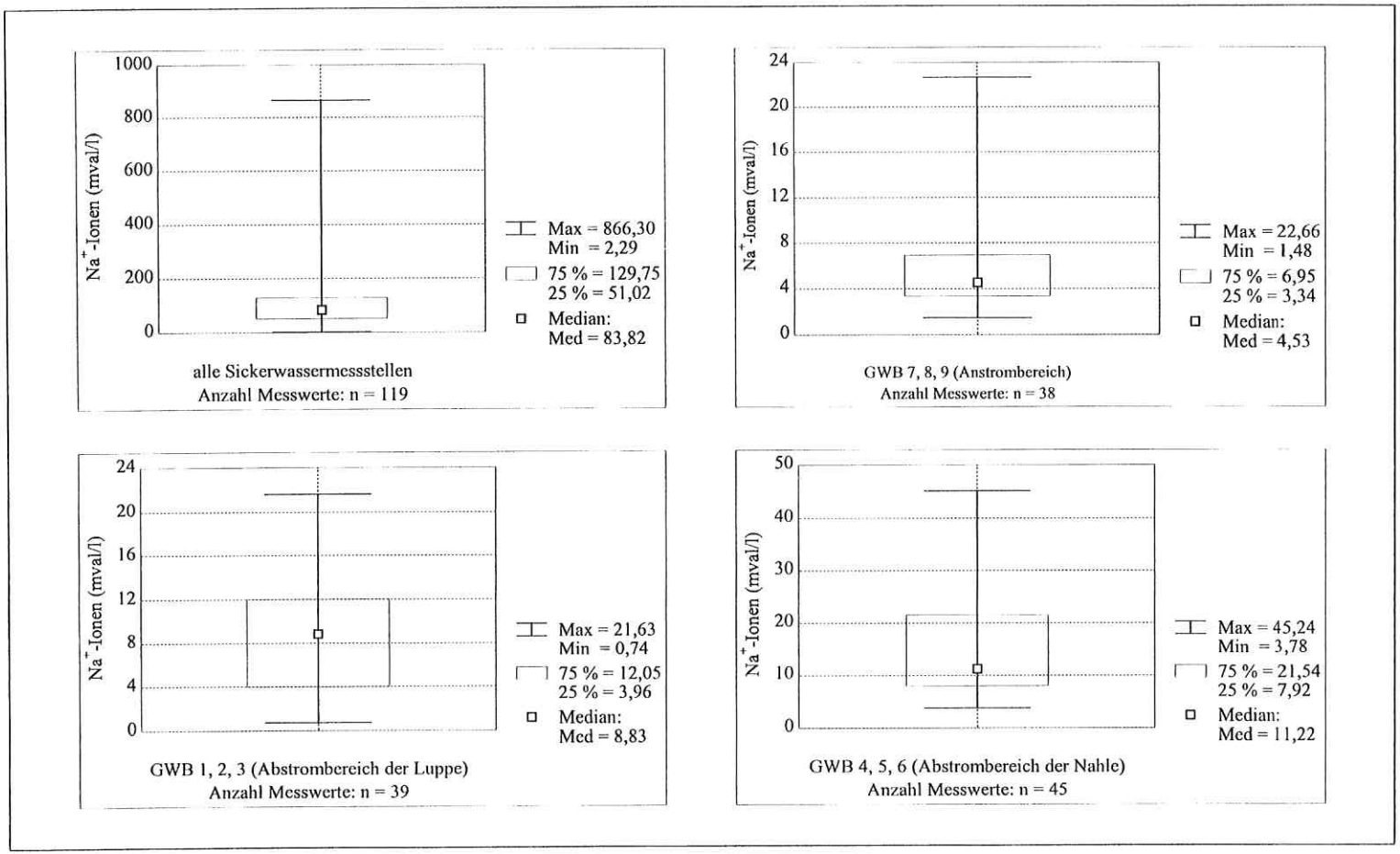

Abb. 4-4. Darstellung statistischer Kenngrößen der Na^+-Ionen.

Die vergleichende Darstellung (Sicker- und Grundwasser) der statistischen Kenngrößen der Variable Temperatur mittels Multiple Box-Plots zeigt, dass die Messwerte beim Sickerwasser deutlich höher liegen, s. Abb. 4-5. Dies ist ein Hinweis auf stattfindende biochemische Reaktionen (aerobe und anaerobe Prozesse) im Deponiekörper. Bei den Daten des Sickerwassers ist eine leichte Linksschiefe und bei denen des Grundwassers eine leichte Rechtsschiefe vorhanden, ansonsten ist die Verteilung der Werte annähernd gleich. Die vorhandenen Streuungen sind durch jahreszeitlich bedingte Temperaturschwankungen begründet.

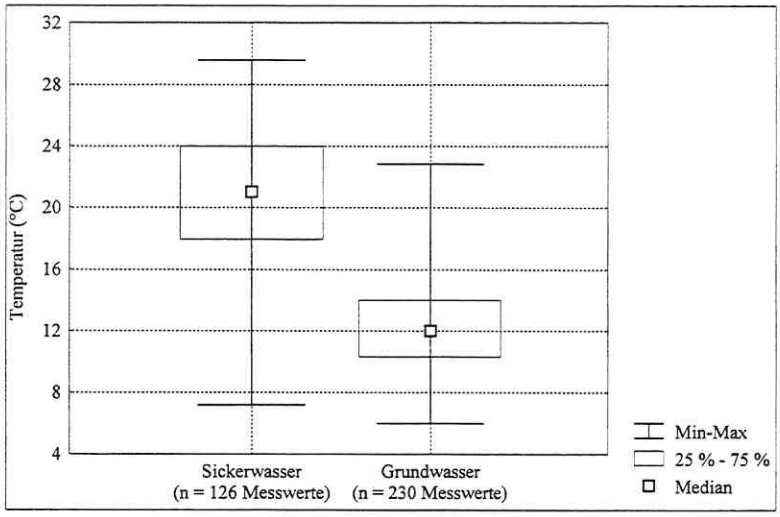

Abb. 4-5. Vergleichende Darstellung statistischer Kenngrößen der Temperatur.

Durch die graphische Darstellung der Häufigkeitsverteilung mittels Histogramms wird sowohl für die Sicker- als auch für die Grundwasserdaten der Na^+-Ionen eine rechtsschiefe Verteilung sichtbar. Die jeweils von allen Probenahmestellen zusammengefassten Werte weichen deutlich von der GAUßschen Normalverteilung ab, s. Abb. 4-6 und 4-7. Auch ein Vergleich der statistischen Kenngrößen Mittelwert und Median - s. Tab. A-1 (Sickerwasser) und A-2 (Grundwasser) - der Einzelionen zeigt, dass diese sich beachtlich voneinander unterscheiden. Man kann somit nicht davon ausgehen, dass die Daten den Voraussetzungen der GAUßverteilung genügen. Die entscheidende Ursache für das Abweichen könnte in der örtlich inhomogenen Verteilung der Müllbestandteile im Deponiekörper (stoffliche Zusammensetzung, Ablagerungsalter, verschiedene Korngrößenverteilungen) liegen. Eine Annäherung an die Normalverteilung wäre demnach bei einer räumlichen Differenzierung zu erwarten.

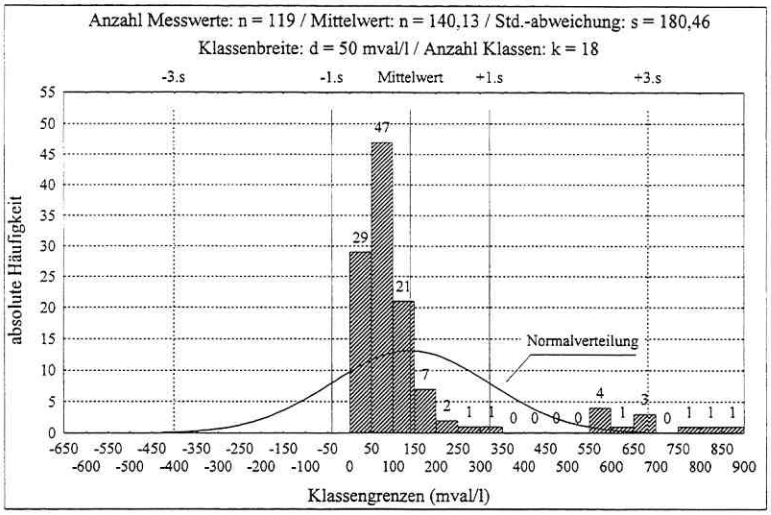

Abb. 4-6. Häufigkeitsverteilung der Na^+-Ionen (alle Sickerwassermessstellen).

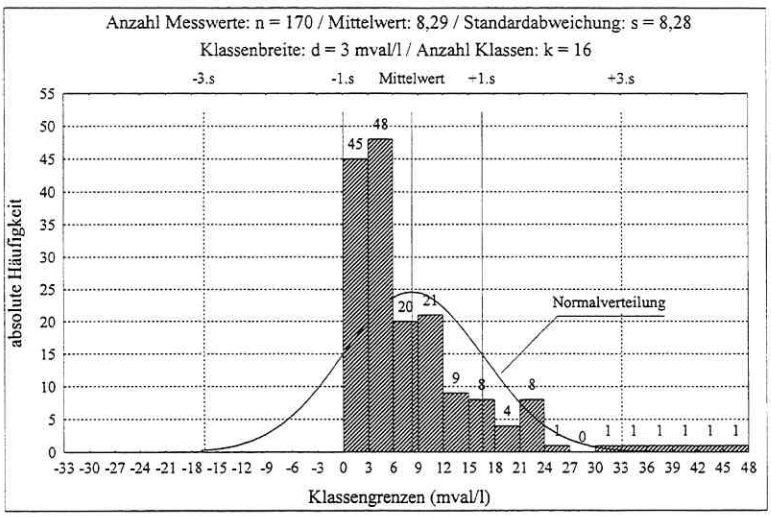

Abb. 4-7. Häufigkeitsverteilung der Na^+-Ionen (alle Grundwassermessstellen).

Dies wurde durch entsprechende Untersuchungen, bei denen die Werte jeweils nur einer Probenahmestelle betrachtet wurden, größtenteils bestätigt. Beispielhaft hierfür ist Abb. 4-8, welche die Häufigkeitsverteilung der ausschließlich an GWB 5 gemessenen Werte der Na^+-Ionen darstellt. Für diese stimmen zudem Mittelwert und Median (25,00 mval/l und 22,10 mval/l) weitestgehend überein, so dass man näherungsweise von einer Normalverteilung der Daten dieser Messstelle ausgehen kann.

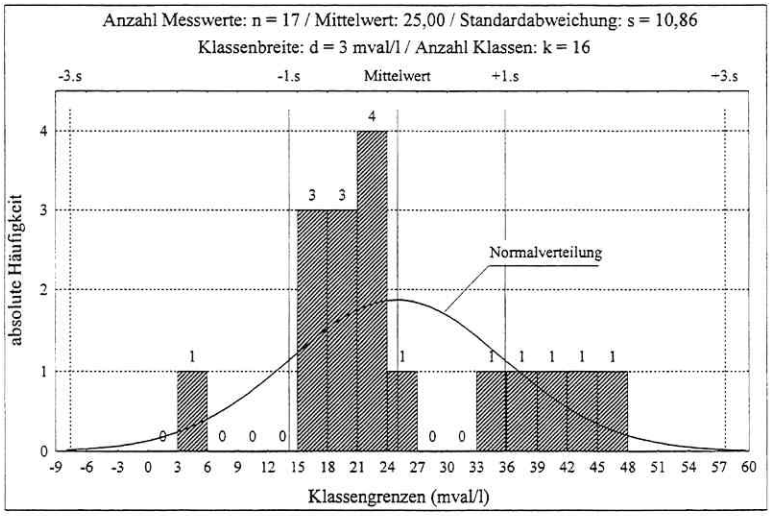

Abb. 4-8. Häufigkeitsverteilung der Na^+-Ionen (GWB 5).

Die Betrachtungen über die Normalverteilung sind insofern von Bedeutung, da bei einigen datenanalytischen Auswertemethoden hierüber bestimmte Annahmen gemacht werden. Beispielsweise sollte bei der univariaten Varianzanalyse (s. Kapitel 6) die abhängige Variable innerhalb der Gruppen annähernd normalverteilt sein (STORM, 1995) und für die lineare Diskriminanzanalyse (s. Kapitel 7) wird u. a. eine multivariate Normalverteilung der Objekte vorausgesetzt (AHRENS und LÄUTER, 1981). Darauf wird an entsprechender Stelle jeweils noch eingegangen.

Die Ganglinienplots - s. Abb. 4.9 und 4.10 - zeigen bei einer differenzierten Betrachtung der Probenahmestellen, dass außerordentlich hohe Salzfrachten am GWB 5 und am SWP 10 vorhanden sind. Beide Messstellen befinden sich im Abstrombereich der Nahle (nordwestlicher Teil des Deponiegeländes), so dass dies ebenfalls ein Hinweis darauf ist, dass in diesem Gebiet die Schadstoffbelastung für den umgebenden Aquifer offensichtlich besonders akut ist. Die mittels der Box-Plots gewonnene Erkenntnis, dass eine breite Streuung der Daten auf lokal unterschiedliche Belastungen hindeutet, wird durch die Schwankungen der Ganglinien bestätigt. Demhingegen werden signifikante saisonale Tendenzen sowohl für die Salzfrachten des Grundwassers als auch für die des Sickerwassers aus den Darstellungen nicht deutlich. Die Gleichartigkeit der Kurvenverläufe von Leitfähigkeit sowie Wasserhärte und den entsprechenden Summen der Einzelionen lässt die Möglichkeit einer schnellen Ionenbestimmung über die Feldmessung der Summenparameter Leitfähigkeit und Wasserhärte zu.

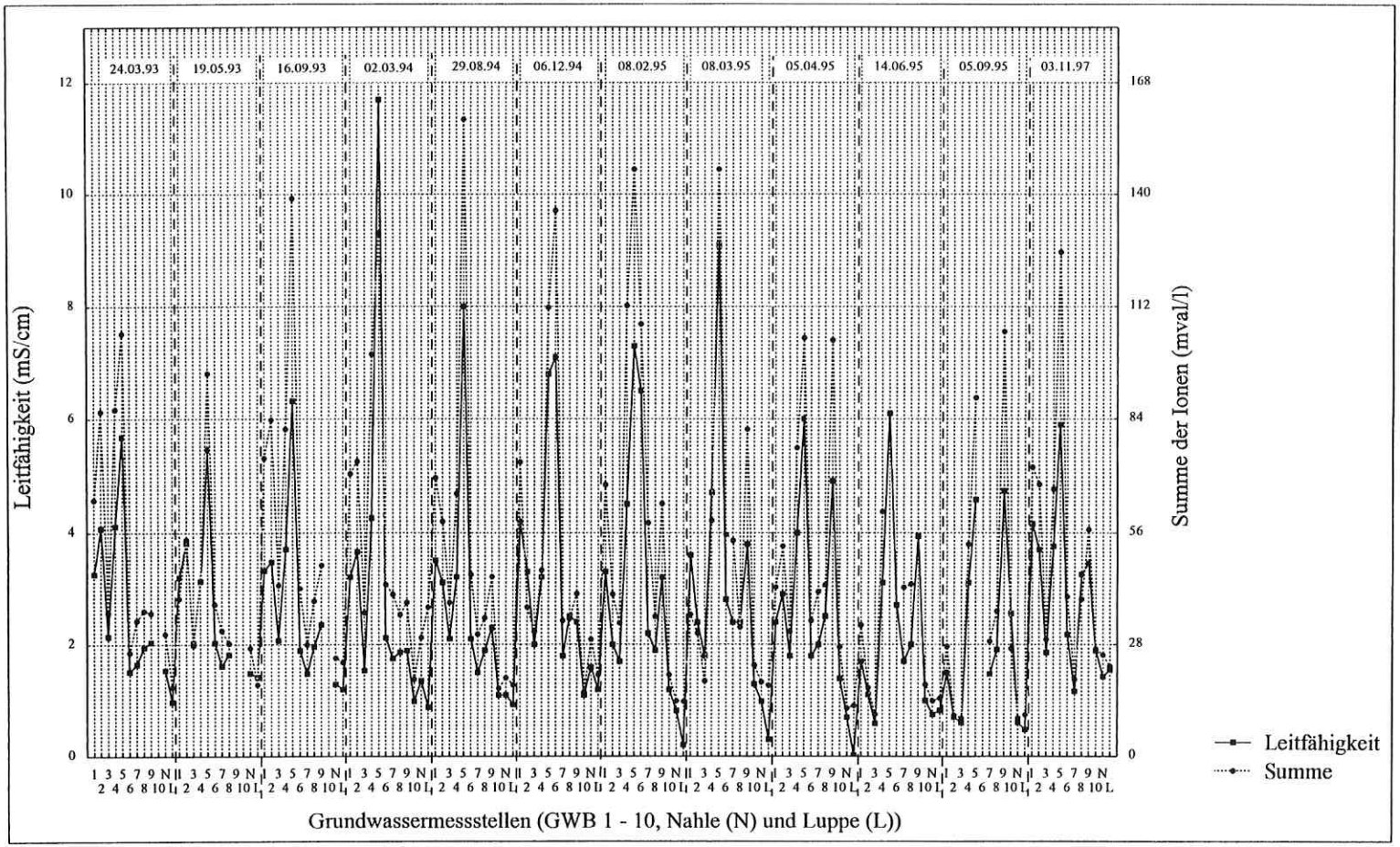

Abb. 4-9. Ganglinienplot der Leitfähigkeit und der Summe der Na^+-, K^+-, Mg^{2+}-, Ca^{2+}-, Cl^-- und SO_4^{2-}-Ionen (Grundwasser).

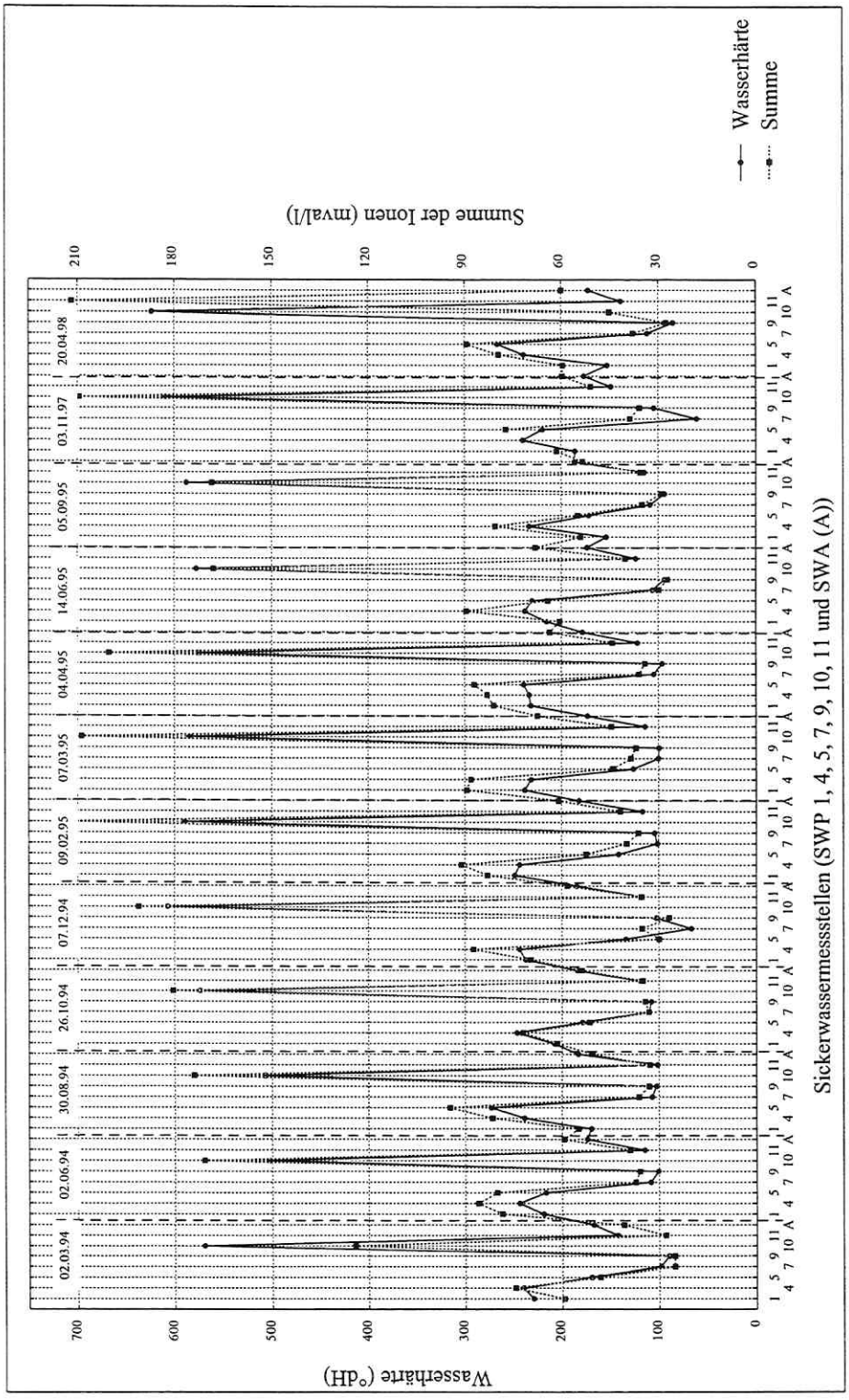

Abb. 4-10. Ganglinienplot der Wasserhärte und der Summe der Mg^{2+}- und Ca^{2+}-Ionen (Sickerwasser).

5 Lineare Korrelation und Regression

5.1 Vorbemerkungen

Die Begriffe „Korrelation“ und „Regression“ stammen aus dem Lateinischen und bedeuten soviel wie „Wechselbeziehung“ und „Rückbildung“.

Wurde in den Ausführungen von Kapitel 4 mit der Berechnung statistischer Maßzahlen jeweils nur ein messbares Merkmal X betrachtet, so soll nun die Abhängigkeit zwischen zwei Merkmalen X und Y untersucht werden. X und Y stellen zwei stetige Zufallsgrößen dar, die die Werte x_i und y_i ($i = 1, 2, \ldots, n$) annehmen können. Beide Folgen von Messwerten werden nicht als einzelne Stichproben betrachtet, sondern ihre Merkmalswerte x_i und y_i werden jeweils gleichzeitig an jedem Element einer Menge von zu untersuchenden Objekten gemessen. Es ergibt sich eine Stichprobe, welche aus Wertepaaren (x_i; y_i) ($i = 1, 2, \ldots, n$) besteht. Zwischen beiden Messwerten x_i und y_i besteht keine eindeutige funktionale Zuordnung, sondern ein korrelativer oder stochastischer Zusammenhang (OSE et al., 1974).

Bei den im Rahmen vorliegender Arbeit diesbezüglich durchgeführten Untersuchungen wurden die in den nachfolgenden beiden Abschnitten beschriebenen Verfahren der linearen Korrelation und Regression angewandt.

5.2 Lineare Korrelation

Mit Hilfe der Korrelationsanalyse kann man aus einer Folge geordneter Paare (x_i; y_i) der Messwerte (x_i) und (y_i) ($i = 1, 2, \ldots, n$) den Grad des Zusammenhangs zwischen zwei Merkmalen X und Y quantitativ bestimmen. Als Kenngröße wird der empirische Korrelationskoeffizient r_{XY} verwendet, dieser ist ein Maß für die Straffheit des linearen Zusammenhangs zwischen X und Y (OSE et al., 1974). Er ist in der allgemein (OSE et al., 1974; GÖHLER, 1989) verwendeten Schreibweise definiert als

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

d. h., er ergibt sich aus der Kovarianz s_{XY} der beiden Zufallsgrößen X und Y, normiert auf das Produkt der beiden Standardabweichungen s_X und s_Y. Somit gilt stets $-1 \leq r_{XY} \leq +1$.

Der berechnete Korrelationskoeffizient lässt sich wie folgt interpretieren:

1. $r_{XY} < 0$: negative (ungleichsinnige) Korrelation, d. h., zu großen Werten von X gehören kleine Werte von Y und umgekehrt
2. $r_{XY} > 0$: positive (gleichsinnige) Korrelation, d. h., zu großen Werten von X gehören große Werte von Y und umgekehrt
3. $r_{XY} = +1$ bzw. $r_{XY} = -1$: vollständige positive bzw. vollständige negative Korrelation
4. $r_{XY} = 0$: die Zufallsgrößen X und Y sind stochastisch unabhängig

EINAX et al. (1997) beschreiben den Test auf Signifikanz des Korrelationskoeffizienten und führen ihn auch exemplarisch vor. Eine Hypothese H_0 über die Unabhängigkeit der beiden Zufallsgrößen X und Y ist demnach dann abzulehnen, wenn für eine Testgröße

$$t_{exp} = \frac{r_{XY} \cdot \sqrt{n-2}}{1 - r_{XY}^2}$$

gilt

$$\left|t_{exp}\right| \geq t_{f;\,q}, \quad \text{mit} \quad f = n - 2 \text{ und } q = 1 - \alpha/2 \quad \text{(zweiseitiger t-Test)},$$

wobei $t_{f;\,q}$ das Quantil der Ordnung q einer t-Verteilung mit f Freiheitsgraden und α das festgelegte Signifikanzniveau (Irrtumswahrscheinlichkeit) sind. In diesem Fall sind die berechneten Korrelationswerte statistisch signifikant, d. h., die stochastische Abhängigkeit von X und Y kann als „statistisch sicher" zu dem (in den Untersuchungen auf 0,05) festgelegten Signifikanzniveau α angesehen werden. In den Korrelationsmatrizen der Untersuchungsergebnisse, s. Tab. A-3 und A-4 des Anhangs, sind diese Werte durch Fettdruck hervorgehoben.

5.3 Lineare Regression

Die (bivariate) Regressionsanalyse untersucht den korrelativen Zusammenhang zwischen zwei Merkmalen X und Y und beschreibt die Art des Zusammenhangs. Die Verteilung der einen Zufallsgröße Y wird als abhängig angenommen und für bestimmte Werte der anderen Zufallsgröße X untersucht. Über den Ansatz eines bestimmten Regressionsmodells wird versucht, diese Abhängigkeit funktional auszudrücken. Über die Verteilung der unabhängigen Variablen X werden keine Annahmen gemacht, für die abhängige Größe Y muss gelten, dass sie für jeden Wert x_i (i = 1, 2, ... , n) normalverteilt sein muss mit dem Erwartungswert E(Y) = Y(X) (theoretische Regressionsfunktion) und der konstanten Varianz $D^2(Y) = \sigma^2$. Im Falle des linearen Regressionsmodells ist der Erwartungswert ein Funktionswert der Form

$$E(Y) = Y(X; a_0, a_1) = a_0 + a_1 \cdot X$$

mit der entsprechenden empirischen Regressionsgeraden (Schätzgeraden)

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 \cdot X,$$

wobei $\hat{a}_0$ und $\hat{a}_1$ die Schätzwerte für die unbekannten Koeffizienten a_0 und a_1 sind und es Aufgabe der Regressionsanalyse ist, diese zu berechnen (OSE et al., 1974).

Nach Auswahl einer bestimmten Fehlerfunktion werden die Schätzwerte über deren Minimierung bestimmt. Grundlage der durchgeführten Berechnungen ist die GAUßsche Minimumbedingung, nach dieser gilt für die Berechnung der Schätzwerte $\hat{a}_0$ und $\hat{a}_1$ der Ansatz

$$S = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min.$$

Durch das Einsetzen der linearen Regressionsfunktion $\hat{y}_i = \hat{a}_0 + \hat{a}_1 \cdot x_i$, die Bildung der partiellen Ableitungen nach $\hat{a}_0$ und $\hat{a}_1$ sowie das Nullsetzen der beiden resultierenden Gleichungen lassen sich die gesuchten Regressionskoeffizienten $\hat{a}_0$ und $\hat{a}_1$ bestimmen. Die dadurch erhaltene lineare Regressionsgleichung (Schätzgerade) hat die Form

$$\hat{Y} = \bar{y} + \frac{\sum_{i=1}^{n} y_i \cdot x_i - \bar{y} \cdot \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \cdot \sum_{i=1}^{n} x_i} \cdot (X - \bar{x}).$$

Zur Schätzung der „Streuung" (= Varianz) σ^2 der Zufallsgröße Y wird nach OSE et al. (1974) die empirische „Streuung" um die Regressionsgerade

$$\hat{s}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

verwendet und mit

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

erhält man die zugehörige Standardabweichung.

Der in den Untersuchungsergebnissen angegebene Vertrauensbereich (95 % Konfidenz) für die Regressionsgerade $Y(X) = a_0 + a_1 \cdot X$ ergibt sich aus

$$\hat{y}_i - s_{\hat{y}_i} \cdot t_{f;q} < y(x_i) < \hat{y}_i + s_{\hat{y}_i} \cdot t_{f;q},$$

mit

$$s_{\hat{y}_i} = \hat{s} \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1) \cdot s_X^2}} \quad \text{und} \quad s_X^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{(EINAX et al., 1997).}$$

$t_{f;\,q}$ mit $f = n - 2$ und $q = 1 - \alpha$ ist ein im Zusammenhang mit der Korrelation bereits erwähntes Quantil der Ordnung q einer t-Verteilung mit f Freiheitsgraden. Das Signifikanzniveau (Irrtumswahrscheinlichkeit) α wurde in den Untersuchungen zur linearen Regression ebenfalls auf 0,05 festgelegt, die statistische Sicherheit (Konfidenzniveau) beträgt somit $1 - \alpha = 0{,}95$.

In den graphischen Darstellungen der Untersuchungsergebnisse ist weiterhin das Bestimmtheitsmaß $B_{XY} = r_{XY}^2$ ($0 \leq B_{XY} \leq 1$) angegeben. Dieses wird im Zusammenhang mit der linearen Regression in vielen praktischen Anwendungsfällen (ERTEL et al., 1997; WÜBBOLD et al., 1987) verwendet - seine Größe gibt Aufschluss darüber, wie gut sich die Punkte um die Regressionsgerade konzentrieren (STORM, 1995).

EINAX et al. (1997) geben mehrere Bedingungen für die lineare Regression der beschriebenen Art an. Sie heben dabei insbesondere hervor, dass

(1)der Fehler der unabhängigen (fest vorgegebenen) Variablen X vernachlässigbar ist und nur die abhängige Variable Y fehlerbehaftet ist,

(2)die Y-Werte unabhängig voneinander und normalverteilt sein müssen,

(3)die Varianzen der Y-Werte bei allen X-Werten vergleichbar groß bzw. homogen sein müssen und

(4)die Residuen damit unabhängig, normalverteilt und homogen sind.

Um zu prüfen, ob diese einzelnen Voraussetzungen vorliegen, schlagen sie die Durchführung statistischer Testverfahren vor, beispielsweise für (2) den DIXON-Ausreißertest und für (3) - die hier beschriebene Eigenschaft wird Homoskedastizität genannt - den COCHRAN-Test. Diese beiden Testverfahren werden auch in der DIN ISO 5725 (1988) empfohlen.

BRONSTEIN und SEMENDJAJEW (1983) weisen darauf hin, dass die Beantwortung der Frage, ob annähernd eine lineare Korrelation zwischen X und Y vorliegt, nur in seltenen Fällen theoretisch entschieden werden kann, z. B. dann, wenn bekannt ist, dass X und Y normalverteilt sind. Sie kommen zu der Einschätzung, dass sich eine Bewertung der stochastischen Abhängigkeit zwischen X und Y dann gut vornehmen lässt, wenn man die gesamte Stichprobe als Punkteschar in der X-Y-Ebene darstellt.

In die Untersuchungen wurden ebenfalls jeweils alle vorhandenen Messwertepaare (x_i, y_i) einbezogen, d. h., es wurde die gesamte Stichprobe in der X-Y-Ebene aufgetragen, um eine Bewertung vorzunehmen. Auf die Durchführung der erwähnten Testverfahren zum Prüfen der Voraussetzungen wurde daher ebenso verzichtet wie auf die (beispielsweise von EINAX et al. (1997) erläuterte) Durchführung des Tests zur Signifikanz der Regressionskoeffizienten.

5.4 Untersuchungen

5.4.1 Gegenstand und Zielstellung der Untersuchungen

- lineare Korrelation

  Es wurden die Variablen Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^-, SO_4^{2-}, pH-Wert, Leitfähigkeit, Wasserhärte, CSB, DOC und Wasserspiegelstand in die Untersuchungen einbezogen. Die Korrelationswerte wurden, getrennt nach Sicker- und Grundwasserdaten, durch den Vergleich der Messreihen von jeweils zwei Variablen berechnet. Die Anzahl der hierbei verfügbaren Wertepaare ist in Klammern unter dem jeweils berechneten Koeffizienten angegeben. Die resultierenden Korrelationsmatrizen sind vom Typ (12, 12), d. h., es sind 12-reihige quadratische Matrizen, wobei die Koeffizienten der Hauptdiagonalen den Wert eins besitzen (identische Messreihen). Da die Werte bezüglich der Hauptdiagonalen identisch sind, wurden in den Untersuchungsergebnissen die unterhalb gelegenen weggelassen.

  Das Ziel dieser Untersuchungen bestand vereinfacht ausgedrückt darin, signifikante Abhängigkeiten (Korrelationen) zwischen den verschiedenen Parametern zu ermitteln und diese in komprimierter Form darzustellen.

- lineare Regression

  Es wurden - jeweils für die Sicker- und Grundwasserdaten - Untersuchungen für ausgewählte Merkmalspaare durchgeführt, s. Tab. 5-1.

Tab. 5-1. Lineare Regression. - Gegenstand der Untersuchungen.

Nr.	Merkmalspaar		Anzahl der Messwertepaare	
	Merkmal X	Merkmal Y	Sickerwasser	Grundwasser
1	Na^+-Ionen	K^+-Ionen	118	170
2	Mg^{2+}-Ionen	Ca^{2+}-Ionen	119	170
3	Cl^--Ionen	SO_4^{2-}-Ionen	82	155
4	Summe der Kationen	Summe der Anionen	81	155
5	CSB	DOC	90	117
6	CSB (Titration)	CSB (Küvettensatz)	27	37

Die dabei verfolgten Zielstellungen erlaubten es, sich auf den (im vorhergehenden Abschnitt erläuterten) einfachen Fall der bivariaten Regressionsanalyse zu beschränken.

Durch die ersten drei Untersuchungen sollte ermittelt werden, inwiefern sich die a priori bekannte identische (Na^+- und K^+-Ionen: vorzugsweise Aschen) oder unterschiedliche (Mg^{2+}-Ionen: Aschen und Ca^{2+}-Ionen: Bauschutt bzw. Cl^--Ionen: Aschen und SO_4^{2-}-Ionen: Bauschutt) Herkunft der Salzfrachten im Deponiekörper durch analog „gute" bzw. „schlechte" Ergebnisse entsprechender Regressionsanalysen bestätigen lässt. Ähnliches galt für die 4. Untersuchung, bei der aufgrund des vermuteten chemischen Gleichgewichts sowohl im Deponiekörper als auch im Grundwasserbereich eine annähernd lineare Regression zwischen den Summen der Kationen (Na^+, K^+, Mg^{2+} und Ca^{2+}) und den Summen der Anionen (Cl^- und SO_4^{2-}) zu erwarten war. Bei der 5. Analyse sollte die erwartet gute lineare Regression des DOC zum CSB die Aussage über die Zulässigkeit der Substitution des zeit- und kostenaufwendig zu bestimmenden CSB unterstützen. Die letzte Untersuchung sollte ein Argument liefern zu der Diskussion darüber, ob sich die aufwendige Titrationsbestimmung des CSB bei vergleichbarem Informationsgehalt durch Küvettentests ersetzen lässt.
Die generelle praktische Bedeutung von Regressionsanalysen, mit Hilfe leicht messbarer („billiger") Variablen die Bestimmung schwer messbarer („teurer") Variablen zu umgehen, zeigt sich damit insbesondere bei den letzten beiden Untersuchungen.

5.4.2 Untersuchungsergebnisse

Die Korrelationsmatrizen der ausgewählten Messgrößen sind im Anhang, Tab. A-3 (Sickerwasser) und A-4 (Grundwasser), enthalten. Die hierin fett hervorgehobenen Werte sind statistisch signifikant, d. h., die stochastische Abhängigkeit der Variablen Y von der Variablen X kann als „statistisch sicher" angesehen werden bei einem festgelegten Signifikanzniveau (Irrtumswahrscheinlichkeit) von $\alpha = 0{,}05$.
Die graphischen Darstellungen der empirischen Regressionsgeraden - hier wurde sich auf einige ausgewählte beschränkt - sind den folgenden Abbildungen zu entnehmen:

- Na^+- und K^+-Ionen, Grundwasser: s. Abb. 5-1
- Mg^{2+}- und Ca^{2+}-Ionen, Sickerwasser: s. Abb. 5-2
- Cl^-- und SO_4^{2-}-Ionen, Sickerwasser: s. Abb. 5-3
- Kationen und Anionen, Grundwasser: s. Abb. 5-4
- CSB und DOC, Grundwasser: s. Abb. 5-5
- CSB (Titration) und CSB (Küvettensatz), Sickerwasser: s. Abb. 5-6

5.4.3 Diskussion der Untersuchungsergebnisse

Die Korrelationsmatrizen, s. Tab. A-3 und A-4 des Anhangs, verdeutlichen folgende Aspekte:

- Es besteht sowohl beim Grundwasser als auch beim Sickerwasser praktisch keine signifikante Korrelation zwischen dem Wasserspiegelstand und den analytischen Parametern.
- Eine erwartungsgemäß hohe Korrelation besteht zwischen folgenden Variablen:
  - Leitfähigkeit ↔ Einzelionen
  - Wasserhärte ↔ Mg^{2+}- und Ca^{2+}-Ionen
  - Anionen ↔ Kationen

Eine Ausnahme hierbei bildet die vor allem beim Sickerwasser geringe Korrelation der Ca^{2+}-Ionen mit den anderen Messgrößen. Eine mögliche Ursache hierfür ist die Schwerlöslichkeit des Calciumsulfats ($CaSO_4$) im Deponiekörper.

Die mögliche Annahme, dass eine Erhöhung der Salzfrachten zwangsläufig eine Erhöhung des Anteils transportierter organischer Stoffe zur Folge hat, wird durch die Untersuchungen bestätigt, da die Korrelationswerte zwischen dem CSB bzw. dem DOC und den Einzelionen - mit Ausnahme wiederum der Ca^{2+}-Ionen - statistisch signifikant sind.

Aus der sowohl beim Sickerwasser als auch beim Grundwasser nachgewiesenen signifikanten Korrelation zwischen DOC und CSB geht hervor, dass die Konzentrationsgehalte gelöster organischer Verbindungen hinreichend genau durch die Bestimmung des DOC erfasst bzw. quantifiziert werden können. Dies bestärkt die Annahme, dass eine Substitution des CSB durch den weniger aufwendig zu bestimmenden DOC zulässig ist, sichert sie jedoch keinesfalls ab (s. lineare Regression, Kommentar zu Abb. 5-5).

Aus den Untersuchungen zur linearen Regression lassen sich ergänzend die folgenden Aspekte hervorheben:

- Die Na^{+}- und K^{+}-Gehalte lassen sich vorzugsweise aus den Aschen ableiten, sind aber Bestandteile des gesamten abgelagerten Materials. Dadurch erklärt sich deren gute Regression bzw. Korrelation sowohl beim Sickerwasser als auch - s. Abb. 5-1 - beim Grundwasser.
- Die zwar statistisch signifikante aber dennoch relativ geringe Korrelation der Mg^{2+}- und Ca^{2+}-Ionen des Sickerwassers - s. Abb. 5-2 - lässt sich durch deren unterschiedliche Quellen im Deponiekörper begründen: Die Mg^{2+}-Ionen stammen primär aus den Aschen und die des Ca^{2+} aus dem Bauschutt. Auch für das Grundwasser wurde eine relativ hohe Streuung der Mg^{2+}- und Ca^{2+}-Werte ermittelt (B_{XY} = 0,2898), was ein Hinweis darauf ist, dass hier kein homogener Eintrag aus dem Deponiekörper in den Grundwasserleiter erfolgt.

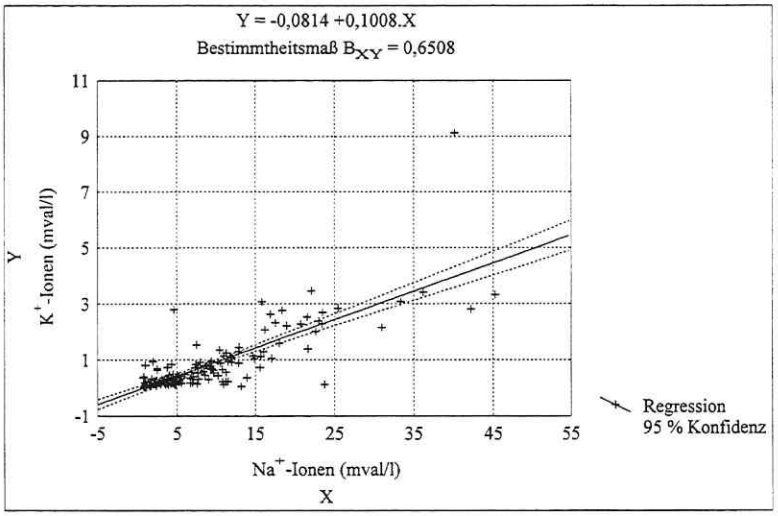

Abb. 5-1. Lineare Regression der Na^+- und K^+-Ionen des Grundwassers.

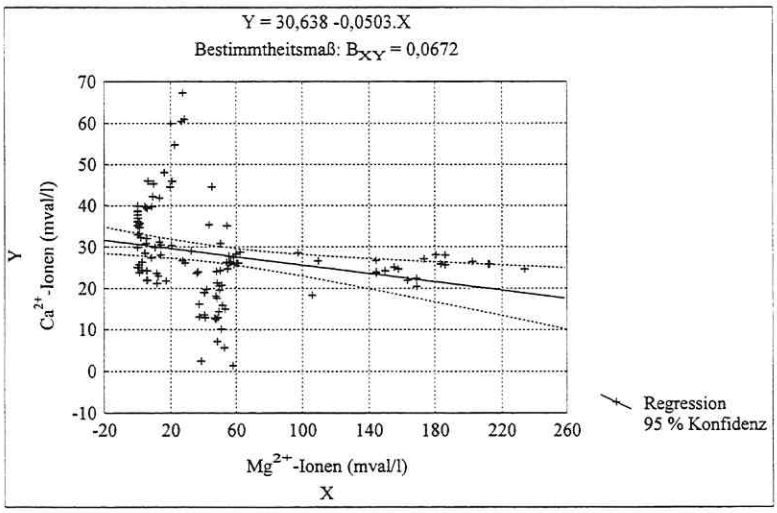

Abb. 5-2. Lineare Regression der Mg^{2+}- und Ca^{2+}-Ionen des Sickerwassers.

– Eine relativ geringe Streuung (B_{XY} = 0,7343) der Werte wird bei den Cl^-- und SO_4^{2-}-Ionen des Sickerwassers deutlich, s. Abb. 5-3. Dies steht im Widerspruch zu deren unterschiedlicher Herkunft im Deponiekörper: Der Ursprung der Cl^--Gehalte liegt in den Aschen und der der SO_4^{2-}-Gehalte im Bauschutt. Die hohe Korrelation (r_{XY} = 0,7837 bzw. B_{XY} = 0,6142) zwischen den Cl^-- und SO_4^{2-}-Ionen des Grundwassers hingegen erklärt sich zum einen mit der Durchmischung der Bestandteile des Sickerwassers beim Eintritt in den umgebenden

Aquifer und zum anderen mit der bereits vorhandenen Grundbelastung des anströmenden Grundwassers durch Cl^-- und SO_4^{2-}-Bestandteile.

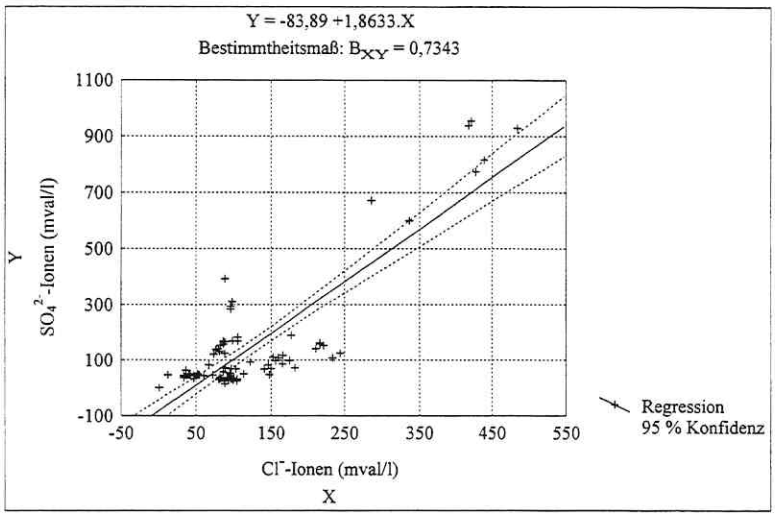

Abb. 5-3. Lineare Regression der Cl^-- und SO_4^{2-}-Ionen des Sickerwassers.

– Das vorhandene chemische Gleichgewicht sowohl im Deponiekörper als auch im Grundwasserbereich spiegelt sich in einer annähernd linearen Regression zwischen den Summen der Kationen und den Summen der Anionen wider, s. exemplarisch Abb. 5-4.

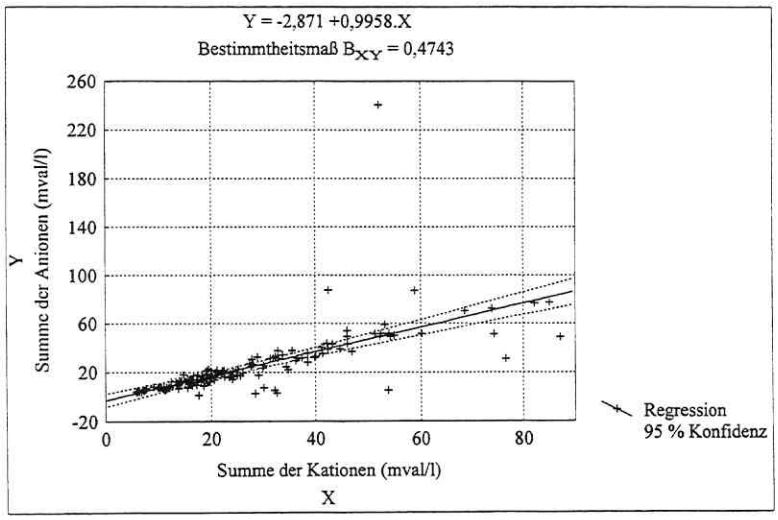

Abb. 5-4. Lineare Regression der Kationen und Anionen des Grundwassers.

– Die Annahmen über eine Substitution zum einen des CSB durch den DOC und zum anderen der aufwendigen Titrationsbestimmung des CSB durch Küvettentests werden durch die entsprechenden Untersuchungsergebnisse zur linearen Regression bestärkt, s. Abb. 5-5 und 5-6. Um diese Aussagen jedoch gänzlich abzusichern, sollten wesentlich mehr Messwertepaare vorliegen, und zudem müssten die am Ende von Abschnitt 5.3 erwähnten Berechnungen zur Signifikanz der Regressionskoeffizienten durchgeführt werden.

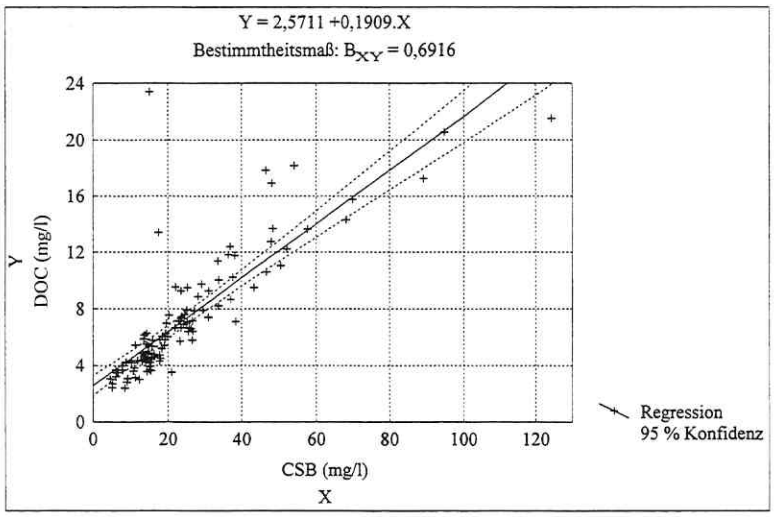

Abb. 5-5. Lineare Regression des CSB und DOC des Grundwassers.

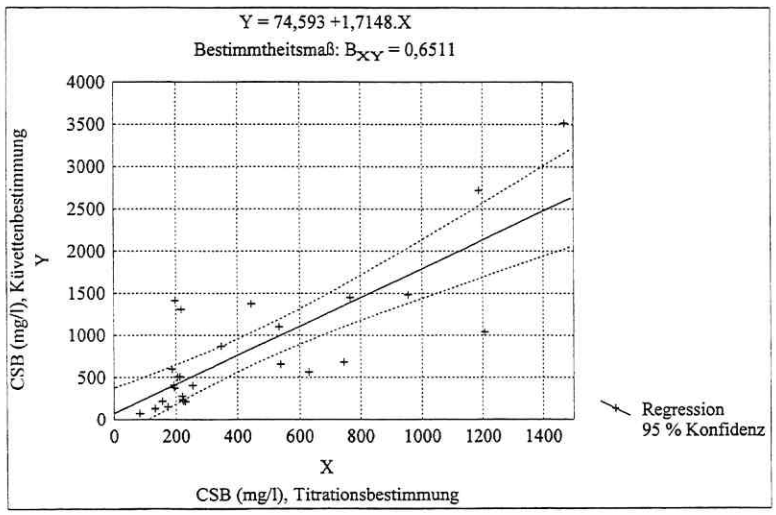

Abb. 5-6. Lineare Regression der CSB-Bestimmungsmethoden (Sickerwasser).

6 Varianzanalyse (univariat)

6.1 Vorbemerkungen

Die Varianzanalyse kommt - ebenso wie die in Kapitel 7 behandelten Verfahren der überwachten Klassifikation - als statistische Methode dann zur Anwendung, wenn von dem vorgegebenen Datensatz a priori eine Strukturierung in verschiedene Gruppen bekannt ist. Die Ausgangssituation unterscheidet sich damit grundlegend von derjenigen bei der Clusteranalyse (s. Kapitel 8), wo es darum geht, Strukturen im Datensatz überhaupt erst zu erkennen.

Die mit der Varianzanalyse zu behandelnden Probleme werden nach EISENHART (1947) in zwei grundsätzlich verschiedene Typen eingeteilt: Den Problemen 1. Art liegt das Modell mit festen Effekten (Modell I) und den Problemen 2. Art das Modell mit zufälligen Effekten (Modell II) zugrunde.

Bei den Problemen 1. Art - diesen sind die durchgeführten Untersuchungen zuzuordnen - ermittelt die Varianzanalyse quantitativ den Einfluss eines in k ($k \geq 2$) Stufen wirkenden Faktors auf die Versuchsergebnisse eines (daher univariate oder auch eindimensionale Varianzanalyse) messbaren Merkmals X, indem die Mittelwerte der Messergebnisse je Stufe untereinander verglichen werden und ihr Unterschied auf Signifikanz geprüft wird (multipler Mittelwertvergleich). Diese k Stufen sind hier fest vorgegeben, wohingegen sie bei den Problemen 2. Art als zufällige Stichproben vom Umfang k aus einer (gedachten unendlichen) Grundgesamtheit aufgefasst werden. Inhaltlich stellt die Varianzanalyse somit in gewissem Sinne eine Verallgemeinerung des doppelten t-Tests zum Vergleich zweier Mittelwerte aus zwei unabhängigen Stichproben dar. Auch die Wirkung mehrerer Faktoren - je nach Anzahl unterscheidet man zwischen einfacher, zweifacher usw. Varianzanalyse - auf die Messergebnisse kann untersucht sowie der Anteil der durch die einzelnen Faktoren hervorgerufenen Variabilität an der Gesamtvariabilität ermittelt werden (STORM, 1995; AHRENS und LÄUTER, 1981).

Im Rahmen der Untersuchungen wurden einfaktorielle sowie zweifaktorielle (einfache und mehrfache Besetzung) Varianzanalysen durchgeführt, welche jeweils univariat erfolgten und in den nachfolgenden Abschnitten zunächst erläutert werden. Die sich bei einer multivariaten Betrachtung ergebenden Problemstellungen (Bestimmung des Informationsgehaltes einzelner Merkmale bzw. Merkmalsgruppen, Aussonderung redundanter Merkmale) werden im Zusammenhang mit der Diskriminanzanalyse (s. Abschnitt 7.3) diskutiert.

6.2 Einfaktorielle Varianzanalyse

6.2.1 Problemstellung

Es lassen sich n Messwerte x_{ij} (i = 1, ... , k; j = 1, ... , n_i) eines messbaren Merkmals X in k ($k \geq 2$) Gruppen oder Klassen mit je n_1, ... , n_k Elementen $\left(\sum_{i=1}^{k} n_i = n\right)$ anordnen, d. h.: Ein Faktor A wirkt in k Stufen auf das Merkmal X.

Betrachtet man z. B. die n = 70 Grundwassermesswerte der Leitfähigkeit des Jahres 1995, so lässt sich der Einfluss des Faktors MESSSTELLE, welcher in k = 14 Stufen (GWB 1 - 12, Nahle und Luppe) mit jeweils $n_1 = ... = n_{14} = 5$ Messwerten (Monate Februar, März, April, Juni, September) auf das messbare Merkmal Leitfähigkeit wirkt, untersuchen.

Tab. 6-1 zeigt den sog. Versuchsplan in einer von STORM (1995) und AHRENS (1967) verwendeten Darstellungsweise.

Tab. 6-1. Versuchsplan der einfaktoriellen Varianzanalyse.

Anzahl der Messwerte je Gruppe	Gruppen (Stufen) des Faktors A				
	1	2	...	k	
1	x_{11}	x_{21}	...	x_{k1}	
2	x_{12}	x_{22}	...	x_{k2}	
...	...	...	...	...	
n_i	x_{1n_1}	x_{2n_2}	...	x_{kn_k}	
Summen S_i	S_1	S_2	...	S_k	S_{ges}
Mittelwerte $\bar{x}_i$	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$	$\bar{x}_{ges}$

Für die in Tab. 6-1 enthaltenen Kenngrößen gelten die folgenden Beziehungen:

$$S_i = \sum_{j=1}^{n_i} x_{ij}$$ ... Summe der i-ten Gruppe

$$\bar{x}_i = \frac{S_i}{n_i}$$ ... Mittelwert der i-ten Gruppe

$$S_{ges} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{k} S_i$$ ... Summe aller n Messwerte

$$\bar{x}_{ges} = \frac{1}{n}\cdot\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\cdot S_{ges}$$ ... Mittelwert gesamt

$$s_i^2 = \frac{1}{n_i - 1} \cdot \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2 \qquad \text{... Streuung der i-ten Gruppe}$$

$$s_{ges}^2 = \frac{1}{n - 1} \cdot \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{ges} \right)^2 \qquad \text{... Streuung gesamt}$$

Die Aufgabe der einfaktoriellen Varianzanalyse besteht darin, die Mittelwerte $\bar{x}_1, \ldots, \bar{x}_k$ der k Gruppen miteinander zu vergleichen und damit die Wirkung (den Effekt) des Faktors A auf das Merkmal X zu untersuchen. Sind die Umfänge n_i (i = 1, ... , k) der Gruppen verschieden, so ist der Versuchsplan unbalanziert (nichtorthogonal), anderenfalls balanziert (orthogonal). Zum Vergleich der Gruppenmittelwerte interpretiert man die k Gruppen als k unabhängige Stichproben mit den Umfängen $n_1, \ldots, n_k$ und setzt voraus, dass die i-te (i = 1, ... , k) Stichprobe aus der normalverteilten Grundgesamtheit mit dem Erwartungswert μ_i und der von i unabhängigen und damit für alle k Grundgesamtheiten gleichen, aber i. Allg. unbekannten Streuung σ^2 stammt. Zu prüfen ist die Hypothese

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$.

Sie besagt, dass alle Stichproben aus derselben Grundgesamtheit stammen, der Faktor A also keinen Einfluss auf das Merkmal X ausübt. Das vorliegende Beobachtungsmaterial ist folglich homogen. Aus der Ablehnung von H_0 folgt die Unterschiedlichkeit von mindestens zwei der μ_i (i = 1, ... , k) und somit die Aussage, dass der Faktor A einen Einfluss auf das Merkmal X ausübt (STORM, 1995).

6.2.2 Anwendung des F-Tests zur Prüfung der Hypothese H_0

Der F-Test dient zum Prüfen der Hypothese über die Gleichheit der Varianzen σ_X^2 und σ_Y^2 zweier unabhängiger normalverteilter Zufallsgrößen X und Y (STORM, 1995).

Um die Anwendbarkeit des F-Tests zur Prüfung der o. g. Hypothese H_0 zu verdeutlichen, sei zunächst auf die additive Zerlegbarkeit der Summe der Abweichungsquadrate hingewiesen. Für die einfache Klassifikation ist bei Einbeziehung von mehr als einem Merkmal die in Abschnitt 7.3.3.1 näher erläuterte sog. HUYGENSsche Dekompositionsformel $\underline{T} = \underline{W} + \underline{B}$ allgemeingültig (HENRION und HENRION, 1994). Im univariaten Fall vereinfacht sich diese zu

$SQ_T = SQ_I + SQ_z$.

AHRENS (1967) und auch STORM (1995) zeigen, wie man zu dieser Zerlegungsformel kommt. Hierin bedeuten:

$SQ_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_{ges}\right)^2$ ... Summe der Abweichungsquadrate insgesamt (total)

$SQ_I = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$ ... Summe der Abweichungsquadrate innerhalb der Gruppen

$SQ_Z = \sum_{i=1}^{k} n_i \cdot \left(\bar{x}_i - \bar{x}_{ges}\right)^2$ ... Summe der Abweichungsquadrate zwischen den Gruppen

Die entsprechenden mittleren Quadrate ergeben sich, indem man die Summen der Abweichungsquadrate durch die zugehörenden Freiheitsgrade dividiert:

$MQ_T = \frac{1}{n-1} \cdot SQ_T$ ... gesamte (totale) Varianz

$MQ_I = \frac{1}{n-k} \cdot SQ_I$ ... Varianz innerhalb der Gruppen

$MQ_Z = \frac{1}{k-1} \cdot SQ_Z$ ... Varianz zwischen den k Gruppen

Unter der Voraussetzung, dass die k Stichproben aus Grundgesamtheiten mit der gleichen, aber unbekannten Streuung σ^2 stammen, ist MQ_I eine erwartungstreue Schätzung für die Streuung σ^2 des Versuchsfehlers. Andererseits ist MQ_Z unter der genannten Voraussetzung eine i. Allg. nicht erwartungstreue Schätzung für σ^2, es gilt

$$E\left(MQ_Z\right) = \sigma^2 + \frac{1}{k-1} \cdot \sum_{i=1}^{k} n_i \cdot \left(\mu_i - \mu\right)^2 .$$

Ist H_0 richtig, d. h. $\mu_1 = \mu_2 = ... = \mu_k = \mu$, so ist MQ_Z ebenfalls erwartungstreu für σ^2. Aus dem Vergleich der beiden voneinander unabhängigen Streuungen MQ_Z und MQ_I lässt sich mit Hilfe des einseitigen F-Tests auf die Hypothese H_0 schließen. Es wird die Testgröße

$$F_{exp} = \frac{MQ_Z}{MQ_I}$$

gebildet. H_0 wird abgelehnt, falls

$F_{exp} \geq F_{crit} = F_{k-1;\, n-k;\, q}$ mit $q = 1 - \alpha$

gilt, wobei $F_{k-1;\, n-k;\, q}$, das Quantil der Ordnung q einer F-Verteilung mit k -1 und n -k Freiheitsgraden und α das festgelegte Signifikanzniveau (Irrtumswahrscheinlichkeit) sind. In diesem Fall ist die Varianz MQ_Z zwischen den Gruppen wesentlich größer als die Varianz MQ_I innerhalb der Gruppen, d. h., der Einfluss des Faktors A auf das Merkmal X kann als „statistisch sicher" zu dem (in den Untersuchungen auf 0,05) festgelegten Signifikanzniveau α gesehen werden (STORM, 1995; EINAX et al., 1997).

6.3 Zweifaktorielle Varianzanalyse mit einfacher Besetzung

Bei der zweifachen Klassifikation (= Kreuzklassifikation) wird der gleichzeitig und in mehreren Stufen wirkende Einfluss von zwei Faktoren A und B auf ein messbares Merkmal X quantitativ ermittelt.

Es lassen sich n Messwerte eines messbaren Merkmals X in $n = k \cdot s$ (k, $s \geq 2$) Zellen oder Unterklassen anordnen, d. h.: Ein Faktor A wirkt in k Stufen und ein Faktor B in s Stufen auf das Merkmal X. Für jede dieser Unterklassen erfolgt genau r = 1 Messung (einfache Besetzung), d. h., es liegt jeweils ein Element x_{ij} (i = 1, ... , k; j = 1, ... , s) vor.

Beispielhaft werden wiederum die n = 70 Grundwassermesswerte der Leitfähigkeit des Jahres 1995 betrachtet. Durch Anwendung der zweifaktoriellen Varianzanalyse (einfache Besetzung) lässt sich der simultane Einfluss der Faktoren MESSSTELLE und DATUM, welche in k = 14 Stufen (GWB 1 - 12, Nahle und Luppe) bzw. s = 5 Stufen (Monate Februar, März, April, Juni, September) auf das messbare Merkmal Leitfähigkeit wirken, untersuchen.

Das mit Hilfe des Verfahrens allgemein zu lösende Problem besteht darin, die $n = k \cdot s$ Messwerte miteinander zu vergleichen und damit die Wirkung (den Effekt) der Faktoren A und B auf das Merkmal X zu untersuchen. Zum Vergleich interpretiert man die n Messwerte als n unabhängige Stichproben vom Umfang 1 und setzt voraus, dass diese n Stichproben aus normalverteilten Grundgesamtheiten mit dem Erwartungswert μ_{ij} (i = 1, ... , k; j = 1, ... , s) und der von i und j unabhängigen, aber i. Allg. unbekannten Streuung σ^2 stammen. Eine weitere Voraussetzung besteht darin, dass die Wirkungen (Effekte) der Faktoren A und B additiv sind, d. h. keine Wechselwirkung zwischen den Effekten auftritt (STORM, 1995).

Es werden - analog der beschriebenen Vorgehensweise bei der einfachen Klassifikation - Hypothesen H_A und H_B aufgestellt, die besagen, dass alle n Stichproben aus derselben Grundgesamtheit stammen, die Faktoren A bzw. B also keinen Einfluss auf das Merkmal X ausüben. Durch Anwendung des einseitigen F-Tests werden diese Hypothesen überprüft, wobei die entsprechenden Testgrößen F_{exp} aus dem Verhältnis von der Varianz zwischen den A-Gruppen MQ_A bzw. B-Gruppen MQ_B zu einer Restvarianz MQ_R gebildet werden. Aus der Ablehnung von H_A bzw. H_B folgt die Aussage, dass der Einfluss der Faktoren A bzw. B auf das Merkmal X als „statistisch sicher" zu dem festgelegten Signifikanzniveau α angesehen werden kann.

STORM (1995) sowie AHRENS (1967) erläutern ausführlich die hier in komprimierter Form wiedergegebene Vorgehensweise und demonstrieren diese anhand eines Beispiels.

6.4 Zweifaktorielle Varianzanalyse mit mehrfacher Besetzung

Bei der zweifachen Klassifikation mit mehrfacher Besetzung wird neben dem Einfluss von zwei Faktoren A und B auf ein messbares Merkmal X auch eine evtl. vorhandene Wechselwirkung (Interaktion) zwischen diesen beiden Faktoren A und B quantitativ erfasst.

Es lassen sich n Messwerte eines messbaren Merkmals X in $k \cdot s$ (k, $s \geq 2$) Zellen oder Unterklassen anordnen, d. h.: Ein Faktor A wirkt in k Stufen und ein Faktor B in s Stufen auf das Merkmal X. Für jede dieser Unterklassen wird eine gleiche Anzahl r ($r > 1$) von Wiederholungsmessungen durchgeführt, d. h., es liegen jeweils mehrere Elemente x_{ijl} ($i = 1, \ldots, k$; $j = 1, \ldots, s$; $l = 1, \ldots, r$) vor. Die Gesamtanzahl der Messwerte beträgt somit $n = k \cdot s \cdot r$.

Exemplarisch werden auch hier die Grundwassermessungen des Parameters Leitfähigkeit im Jahr 1995 herangezogen. Wären für jede der $k \cdot s = 70$ Unterklassen $r = 3$ Wiederholungsmessungen durchgeführt worden (d. h., zu jedem Zeitpunkt hätten an jeder Probenahmestelle drei Messungen erfolgen müssen), so lägen für diesen Zeitraum insgesamt $n = 210$ Messwerte vor.

Die Aufgabe der zweifaktoriellen Varianzanalyse (mehrfache Besetzung) besteht darin, die $k \cdot s$ Gruppen miteinander zu vergleichen und damit die Wirkung (den Effekt) der Faktoren A und B auf das Merkmal X zu untersuchen. Zusätzlich soll eine evtl. vorhandene Wechselwirkung zwischen diesen beiden Faktoren ermittelt werden. Zum Vergleich interpretiert man die $k \cdot s$ Gruppen als $k \cdot s$ unabhängige Stichproben vom Umfang r und setzt voraus, dass diese $k \cdot s$ Stichproben aus normalverteilten Grundgesamtheiten mit dem Erwartungswert μ_{ij} ($i = 1, \ldots, k$; $j = 1, \ldots, s$) und der von i und j unabhängigen, aber i. Allg. unbekannten Streuung σ^2 stammen (STORM, 1995).

Es werden Hypothesen H_A und H_B aufgestellt, die besagen, dass die Faktoren A und B keinen Einfluss auf das Merkmal X ausüben sowie eine Hypothese H_{AB}, die besagt, dass zwischen beiden Faktoren keine Wechselwirkung besteht. Durch Anwendung des einseitigen F-Tests werden diese Hypothesen überprüft, wobei die entsprechenden Testgrößen aus dem Verhältnis von der Varianz zwischen den A-Gruppen MQ_A bzw. B-Gruppen MQ_B zu einer Restvarianz MQ_R sowie aus dem Verhältnis von der Varianz der Wechselwirkungen zwischen A und B MQ_{AB} zu einer (derselben) Restvarianz MQ_R gebildet werden. Aus der Ablehnung von H_A bzw. H_B folgt die Aussage, dass die Faktoren A bzw. B einen Einfluss auf das Merkmal X ausüben und die Ablehnung von H_{AB} besagt, dass zwischen beiden Faktoren eine Wechselwirkung besteht (STORM, 1995; AHRENS, 1967).

6.5 Anmerkungen zu den Voraussetzungen und zu möglichen Post hoc-Tests

In die Diskussion über Methodik und Anwendung der Varianzanalyse sollen abschließend noch zwei wesentliche Aspekte einbezogen werden.

1. (vor Durchführung einer Varianzanalyse): Erfüllt das vorliegende Beobachtungsmaterial die Voraussetzungen für eine Anwendung der Varianzanalyse?

   Für die univariate Varianzanalyse wird vorausgesetzt, dass

   - die abhängige Variable innerhalb der Gruppen normalverteilt ist und
   - die Varianzen in den verschiedenen Gruppen identisch sind (STORM, 1995).

   Die Überprüfung der Normalverteilung kann prinzipiell mit den entsprechenden Anpassungstests (z. B. χ^2-Anpassungstest oder KOLMOGOROV-Test) erfolgen. LINDMAN (1974) nimmt eine detaillierte Erörterung der Robustheit der F-Statistik vor und kommt u. a. zu dem Ergebnis, dass diese gegenüber geringen Abweichungen von der Normalverteilung verhältnismäßig unempfindlich ist und somit auch in solchen Fällen angewandt werden kann, in denen das Beobachtungsmaterial diese Voraussetzung nicht erfüllt. Im Rahmen vorliegender Arbeit wurde daher auf diesbezügliche Untersuchungen verzichtet.

   Die Prüfung der Hypothese über die Gleichheit von mehr als zwei Streuungen kann beispielsweise durch den LEVENE-Test oder den (älteren und weniger robusten) BARTLETT-Test erfolgen. Diesbezügliche Untersuchungen wurden durchgeführt mit dem Ergebnis, dass die Streuungen der jeweiligen Grundgesamtheiten gleich sind und somit diese Voraussetzung erfüllt war. LINDMANN (1974) zeigt jedoch auch hier, dass nur bei schwerwiegenden Verletzungen der Voraussetzung der Varianzhomogenität die Gültigkeit (Validität) der F-Statistik in Frage gestellt werden muss.

2. (nach Durchführung einer Varianzanalyse mit dem Ergebnis einer Ablehnung von H_0): Zwischen welchen der k Stichprobenmittelwerte bestehen signifikante Unterschiede?

   Die Beantwortung dieser Frage kann im Nachhinein (post hoc) beispielsweise durch den DUNCAN-Test erfolgen, dies ist ein nichtparametrischer Rangtest zum Vergleich von je zwei Mittelwerten aus einer Reihe von k Mittelwerten (STORM, 1995). Diesbezügliche Untersuchungen wurden jedoch nicht im Detail durchgeführt, da sie aufgrund der Vielzahl einbezogener analytischer Parameter sowie der relativ hohen Stufenwerte k sehr umfangreich und zudem für die Interpretation wenig relevant sind. In die Untersuchungsergebnisse wurden jedoch die graphischen Darstellungen der Gruppenmittelwerte einbezogen.

6.6 Untersuchungen

6.6.1 Gegenstand und Zielstellung der Untersuchungen

In die Untersuchungen zur univariaten Varianzanalyse wurden sämtliche der in Tab. 2-1 aufgeführten Parameter sowie - angeregt durch LESCHBER et al. (1993) - die Ionenverhältnisse Na^+/K^+, Mg^{2+}/Ca^{2+} und Cl^-/SO_4^{2-} einbezogen. Es wurden sowohl die Sicker- als auch die Grundwasserdaten dieser (abhängigen) Variablen mit der einfaktoriellen und der zweifaktoriellen (einfache und mehrfache Besetzung) Varianzanalyse getestet.

Bei den Untersuchungen zur **einfaktoriellen Varianzanalyse** wurden für die Variablen eine zeitliche Differenzierung (Haupteffekt: DATUM) und eine räumliche Differenzierung (Haupteffekt: MESSSTELLE) vorgenommen. Es erfolgte weiterhin eine Abstufung nach der Bestimmungsmethode (Haupteffekt: BESTIMMUNGSMETHODE) für die Variablen Cl^- (k = 3 Stufen: IC, Titration, ISE), SO_4^{2-} (k = 4: IC, Reaktionsküvetten, Säulenmessung, gravimetrische Messung) und CSB (k = 3: Titration, Reaktionsküvetten, Küvettensatz der Fa. Dr. Lange GmbH), s. Tab. 2-1. Die Umfänge n_i (i = 1, ... , k) in den k Gruppen waren bei einzelnen Messgrößen aufgrund fehlender Werte verschieden (unbalanzierter Versuchsplan).

Bei der **zweifachen Klassifikation mit einfacher Besetzung** wurde der simultane Einfluss der Faktoren DATUM und MESSSTELLE auf die genannten Merkmale X untersucht.

Bei der **zweifachen Klassifikation mit mehrfacher Besetzung** sollte neben dem gemeinsamen Einfluss der beiden Faktoren DATUM und MESSSTELLE auf die gemessenen Parameter mit Ausnahme des Wasserspiegelstandes auch eine eventuell vorhandene Wechselwirkung (Interaktion) zwischen diesen beiden Faktoren quantitativ erfasst werden. Die für jede der besprochenen Unterklassen notwendige Anzahl r (r > 1) von Wiederholungsmessungen wurde (zum jeweiligen Messdatum) an den Probenahmestellen durchgeführt.

Die im Rahmen von Kapitel 4 durchgeführten Untersuchungen führten u. a. zu der Erkenntnis, dass sowohl im Deponiekörper als auch im Grundwasserbereich die lokalen Belastungen sehr unterschiedlich sind. Indizien hierfür sind die breite Streuung der Daten (s. Abb. 4-4), deren Abweichung von der Normalverteilung (s. Abb. 4-6 und 4-7) sowie die Schwankungen der Ganglinien (s. Abb. 4-9 und 4-10). Mit Hilfe der einfaktoriellen Varianzanalysen zu den Haupteffekten DATUM und MESSSTELLE sollte ermittelt werden, ob die breite Streuung der Daten bzw. deren Abweichung von der Normalverteilung primär durch den Einflussfaktor

MESSSTELLE und weniger durch jahreszeitliche Effekte (Einflussfaktor DATUM) hervorgerufen wird. Die Zielstellung bestand des Weiteren darin, die lokale Unterschiedlichkeit statistisch abzusichern (zu quantifizieren) sowie durch den Vergleich der Gruppenmittelwerte schadstoffrelevanter Parameter örtlich signifikante Belastungen aufzuzeigen.
Die Differenzierung nach der Bestimmungsmethode (Haupteffekt BESTIMMUNGSMETHODE) für die Variablen Cl^-, SO_4^{2-} und CSB sollte aufzeigen, ob deren Wahl einen signifikanten Einfluss auf das resultierende Beobachtungsmaterial hat, um somit eine prinzipielle Erkenntnis darüber zu erhalten, ob sich aufwendige Analyseverfahren - z. B. die Titrationsbestimmung des CSB - durch einfachere Methoden (Küvettentests) bei vergleichbarem Informationsgehalt substituieren lassen.
Die zweifaktoriellen Varianzanalysen mit einfacher Besetzung boten die Möglichkeit, die mit Hilfe der einfaktoriellen Verfahren erzielten Untersuchungsergebnisse zu bestätigen. Dasselbe galt für die mit mehrfacher Besetzung, wobei die Resultate hier jedoch kritisch betrachtet werden müssen: In allen Fällen lagen nur relativ wenige Unterklassen - minimal sechs und maximal 88 - vor, die zudem jeweils nur mit r = 2 Wiederholungsmessungen bestückt waren.

6.6.2 Untersuchungsergebnisse

Die graphischen Darstellungen sind den folgenden Abbildungen zu entnehmen, wobei hier wiederum eine Auswahl getroffen werden musste:

- einfaktorielle Varianzanalyse
  - Zeitreihenplot der Mittelwerte der Ionen (Sickerwasser): s. Abb. 6-1
  - Zeitreihenplot der Mittelwerte der Ionen (Grundwasser): s. Abb. 6-2
  - Messstellenplot der Mittelwerte der Ionen (Sickerwasser): s. Abb. 6-3
  - Messstellenplot der Mittelwerte der Ionen (Grundwasser): s. Abb. 6-4
  - Messstellenplot der Mittelwerte des CSB und des DOC (Sickerwasser): s. Abb. 6-5
  - Messstellenplot der Mittelwerte des CSB und des DOC (Grundwasser): s. Abb. 6-6
- zweifaktorielle Varianzanalyse (mehrfache Besetzung)
  - Zeitreihenplot (Vergleich: Messstellen) der Leitfähigkeit (Sickerwasser): s. Abb. 6-7
  - Messstellenplot (Vergleich: Messdaten) der Leitfähigkeit (Sickerwasser): s. Abb. 6-8
  - Zeitreihenplot (Vergleich: Messstellen) der Leitfähigkeit (Grundwasser): s Abb. 6-9
  - Messstellenplot (Vergleich: Messdaten) der Leitfähigkeit (Grundwasser): s. Abb. 6-10

In den tabellarisch zusammengestellten Untersuchungsergebnissen sind für jede Variable die entsprechenden mittleren quadratischen Abweichungen, die sich aus diesen ergebende Testgröße F_{exp} sowie die berechnete Irrtumswahrscheinlichkeit α_{calc} angegeben. Die dabei durch Fettdruck hervorgehobenen Werte sind statistisch signifikant bei einer Irrtumswahrscheinlichkeit von $\alpha_{calc} < \alpha = 0{,}5$. Die Testgröße F_{exp} liegt im kritischen Bereich. Die entsprechenden Hypothesen werden abgelehnt, d. h., dass die betrachteten Faktoren in diesen Fällen signifikanten Einfluss haben bzw. zwischen ihnen eine Wechselwirkung besteht. Den im Anhang (Tab. A-5 - A-12) bzw. im nachfolgenden Abschnitt (Tab. 6-2 und 6-3) enthaltenen Tabellen können die folgenden Untersuchungsergebnisse entnommen werden:

- einfaktorielle Varianzanalyse
  - Haupteffekt DATUM (Sickerwasser): s. Tab. A-5
  - Haupteffekt DATUM (Grundwasser): s. Tab. A-6
  - Haupteffekt MESSSTELLE (Sickerwasser): s. Tab. A-7
  - Haupteffekt MESSSTELLE (Grundwasser): s. Tab. A-8
  - Haupteffekt BESTIMMUNGSMETHODE (Sickerwasser): s. Tab. 6-2
  - Haupteffekt BESTIMMUNGSMETHODE (Grundwasser): s. Tab. 6-3
- zweifaktorielle Varianzanalyse (einfache Besetzung)
  - Haupteffekte DATUM und MESSSTELLE (Sickerwasser): s. Abb. A-9
  - Haupteffekte DATUM und MESSSTELLE (Grundwasser): s. Abb. A-10
- zweifaktorielle Varianzanalyse (mehrfache Besetzung)
  - Haupteffekte DATUM und MESSSTELLE (Sickerwasser): s. Abb. A-11
  - Haupteffekte DATUM und MESSSTELLE (Grundwasser): s. Abb. A-12

6.6.3 Diskussion der Untersuchungsergebnisse

Beim Testen des zeitlichen Verhaltens der Parameter mit der einfaktoriellen Varianzanalyse wurde das DATUM als Ursache für deutliche Unterschiede zwischen den Messwerten der Temperatur (Sickerwasser und Grundwasser) und des Wasserspiegelstandes (Grundwasser) ermittelt, s. Tab. A-5 und A-6. Die Schwankungen der Temperatur und des Grundwasserspiegelstandes haben jahreszeitlich bedingte Ursachen, diese Größen unterliegen meteorologischen Einflüssen. Bei Einbeziehung aller durchgeführten Messkampagnen und Probenahmestellen ergibt sich für die Temperatur des Sickerwassers eine Variationsbreite von 22,4 °C und

für die des Grundwassers eine von 16,9 °C. Abb. 4-5 (Multiple-Box-Plot) veranschaulicht die Lage- und Streuungsunterschiede. Die Spannweite des Grundwasserspiegelstandes beträgt 3,03 m. Der Sickerwasserspiegelstand ist praktisch jahreszeitlich unabhängig (α_{calc} = 0,5420). Damit zeigt sich: Beim Eintrag der Niederschläge in den Deponiekörper hat der Sickerwasserspiegel eine Pufferwirkung - sein Stand wird durch die (räumlich abhängigen) Strömungswiderstände im Deponiekörper bestimmt, der Grundwasserspiegel jedoch steigt an.
Der zeitliche Verlauf der Mittelwerte der Ionen des Sickerwassers ist Abb. 6-1 zu entnehmen. Insbesondere die Kurven der Na^+- und SO_4^{2-}-Ionen weisen zwar deutliche Schwankungen auf, die Berechnungen ergaben jedoch, dass der Faktor DATUM weder hier noch auf die anderen gemessenen Merkmalswerte einen signifikanten Einfluss ausübt, s. Tab. A-5.

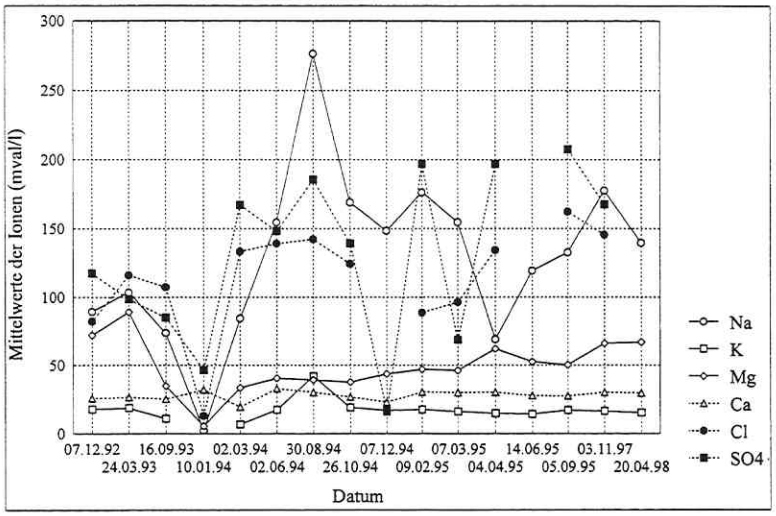

Abb. 6-1. Zeitreihenplot der Mittelwerte der Ionen (Sickerwasser).

Beim Grundwasser zeigt die Ganglinie der Na^+-Ionen in ihrem zeitlichen Verlauf ebenfalls deutliche Schwankungen, wohingegen die der K^+-Ionen eine relative Konstanz besitzt, s. Abb. 6-2. Der Einfluss des Faktors DATUM auf die Na^+-Ionen des Grundwassers spiegelt sich auch in den numerischen Ergebnissen wider, s. Tab. A-6: Für diesen Einzelparameter wurde nur eine geringe Überschreitungswahrscheinlichkeit (α_{calc} = 0,0549) in der Nähe von 0,05 berechnet, die deutlich unter der der K^+-Ionen (α_{calc} = 0,1872) liegt. Während die Na^+-Gehalte unbeeinflusst von der Menge des durch den Deponiekörper dringenden Wassers sind (allgemein schneller Durchgang durch das Kompartiment Boden), werden die K^+-Bestandteile in bo-

denähnlichen Strukturen zurückgehalten. Fließt mehr Sickerwasser, so kommt es zu einer erhöhten Verdrängung und somit zum Anstieg der K^+-Werte. Unter diesem Aspekt wäre für das Grundwasser mehr eine Signifikanz des Faktors DATUM auf die K^+-Ionen als auf die Na^+-Ionen zu erwarten gewesen und insofern ist das Untersuchungsergebnis widersprüchlich. Durch die Varianzanalysen zum Haupteffekt DATUM wurden des Weiteren für die Cl^-- und SO_4^{2-}-Ionen des Grundwassers deutliche Unterschiede zwischen den Werten ermittelt, so dass auch hier die Schadstoffbelastung offensichtlich von meteorologischen Einflussgrößen, insbesondere den jahreszeitlich bedingten Niederschlägen, abhängig ist.

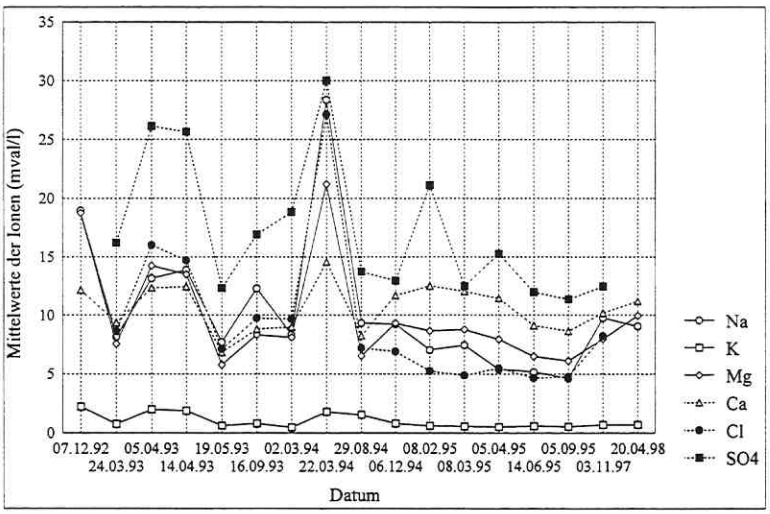

Abb. 6-2. Zeitreihenplot der Mittelwerte der Ionen (Grundwasser).

Sowohl für das Sickerwasser als auch für das Grundwasser wurde ein signifikanter Einfluss des Faktors MESSSTELLE auf alle untersuchten Parameter ermittelt, s. Tab. A-7 und A-8. Die Abb. 6-3 und 6-4 visualisieren die lokal unterschiedlichen Belastungen durch die Salzfrachten. Dabei sind insbesondere drei Aspekte hervorzuheben:

1. Die Schwankungen der Sickerwasserzusammensetzung sind durch die Inhomogenität des Deponiekörpers (stoffliche Zusammensetzung, Ablagerungsalter, verschiedene Korngrößenverteilungen) begründet.
2. Die räumliche Inhomogenität der Salzfrachten im Grundwasser begründet sich durch bevorzugte Austrittspfade des Sickerwassers sowie durch wechselnde Grundwasserfließrichtungen.

3. Die räumliche Differenzierung lässt erkennen, dass die Salzfrachten des Sickerwassers bei SWP 10 und die des Grundwassers bei GWB 5 (Abstrombereich der Nahle) am höchsten sind. Dasselbe gilt für die CSB- und DOC-Werte (s. Abb. 6-5 und 6-6). Dies unterstreicht nochmals die erhöhte Schadstoffbelastung durch den nordwestlichen Teil des Deponiegeländes und führt zu der Vermutung, dass zwischen SWP 10 und GWB 5 eine hydraulische Verbindung besteht.

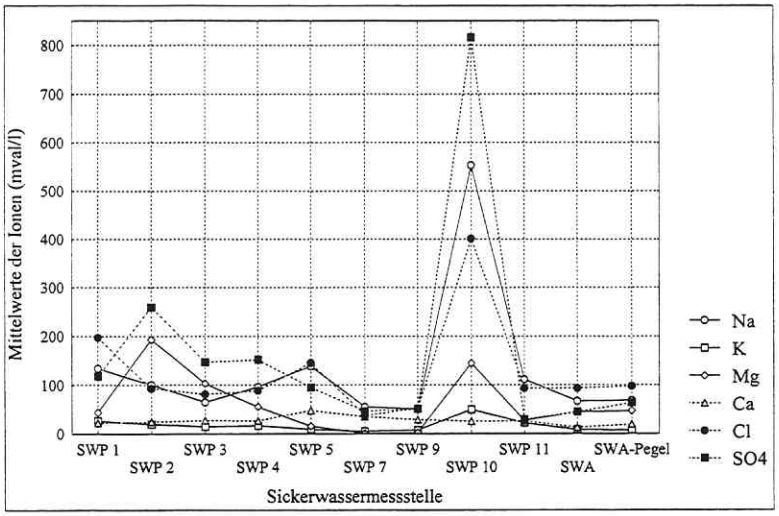

Abb. 6-3. Messstellenplot der Mittelwerte der Ionen (Sickerwasser).

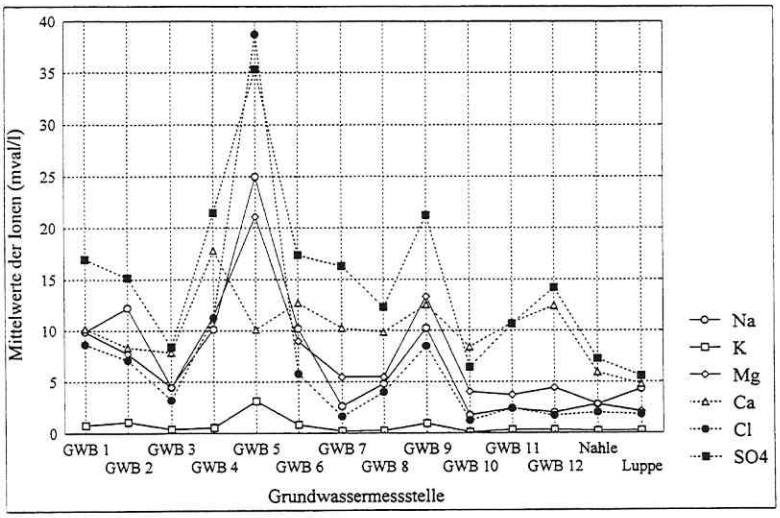

Abb. 6-4. Messstellenplot der Mittelwerte der Ionen (Grundwasser).

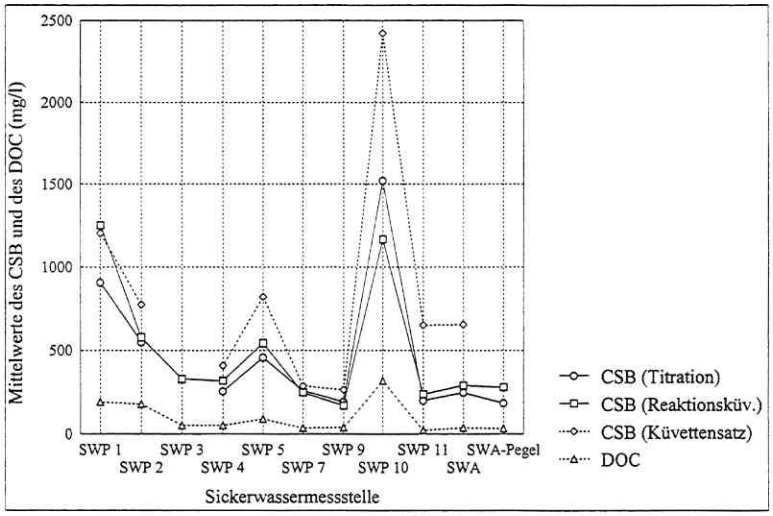

Abb. 6-5. Messstellenplot der Mittelwerte des CSB und des DOC (Sickerwasser).

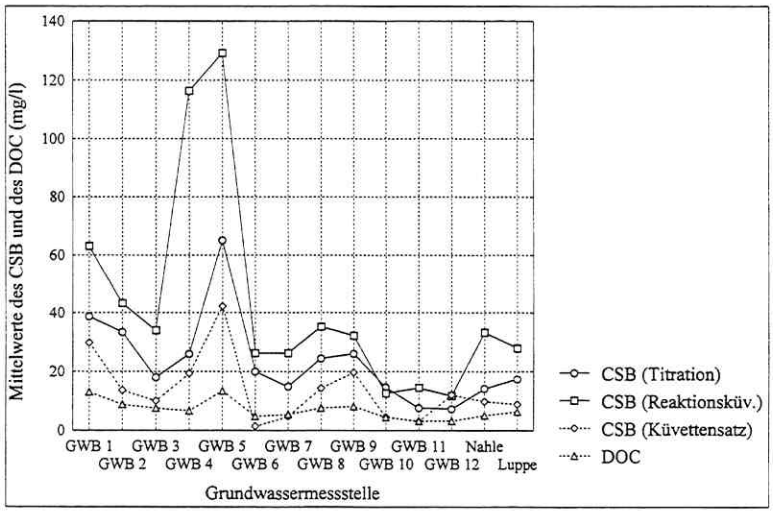

Abb. 6-6. Messstellenplot der Mittelwerte des CSB und des DOC (Grundwasser).

Die Untersuchungen führten weiterhin zu dem Ergebnis, dass die CSB-Werte offensichtlich sehr von der angewandten BESTIMMUNGSMETHODE (Titration, Reaktionsküvetten, Küvettensatz der Fa. Dr. Lange GmbH) abhängig sind - dieser Faktor übt sowohl auf die Werte des Sickerwassers als auch auf die des Grundwassers einen signifikanten Einfluss aus, s. Tab. 6-2 und 6-3. Damit wird die im Zusammenhang mit den Untersuchungsergebnissen der linearen Regression diskutierte Möglichkeit einer Substitution der aufwendigen Titrationsbestim-

mung des CSB durch Küvettentests in Frage gestellt, wobei jedoch berücksichtigt werden muss, dass beide Küvettenmethoden (WTW-Photometer „MPM 1500", Küvettensatz der Fa. Dr. Lange GmbH) als Stufen des Faktors wirkten und hier die Unterschiede in den Gruppenmittelwerten ebenfalls zur Signifikanz beigetragen haben. Demgegenüber scheinen die Cl^-- und SO_4^{2-}-Werte methodenunabhängig zu sein - der α_{calc}-level liegt hier z. T. sogar nahe eins.

Tab. 6-2. Einfaktorielle Varianzanalyse zum Haupteffekt BESTIMMUNGSMETHODE (Sickerwasser).

Merkmal	MQ_Z	MQ_I	F_{exp}	α_{calc}
Cl^- (mval/l)	717,4303	8782,7116	0,0817	0,92159560
SO_4^{2-} (mval/)]	18417,9161	46729,0154	0,3941	0,75739891
CSB (mg/l)	1340691,5353	325173,1750	4,1230	**0,01802962**

Tab. 6-3. Einfaktorielle Varianzanalyse zum Haupteffekt BESTIMMUNGSMETHODE (Grundwasser).

Merkmal	MQ_Z	MQ_I	F_{exp}	α_{calc}
Cl^- (mval/l)	38,6161	161,5145	0,2391	0,62527825
SO_4^{2-} (mval/l)	7,2976	103,7002	0,0704	0,97571995
CSB (mg/l)	24788,2550	3266,4062	7,5888	**0,00063044**

Die Untersuchungsergebnisse der zweifaktoriellen Varianzanalyse (einfache Besetzung) zu den Haupteffekten DATUM und MESSSTELLE zeigen, dass die Signifikanz des Faktors MESSSTELLE deutlich über der des Faktors DATUM liegt, was insbesondere für die Sickerwasserdaten gilt, s. Tab A-9 und A-10. Die Resultate der einfachen Klassifikation, bei der diese Faktoren getrennt hinsichtlich ihres Einflusses betrachtet wurden, erfahren damit im Wesentlichen ihre Bestätigung. Die für die Leitfähigkeit des Sickerwassers ermittelte Signifikanz beider Einflussfaktoren wird durch die Abb. 6-7 und 6-8 veranschaulicht, wobei jedoch im Messstellenplot auffälligere Schwankungen zu verzeichnen sind. Derartige lokal unterschiedliche Belastungen werden auch für den Grundwasserbereich deutlich, s. Abb. 6-10. Die Zeitreihen hingegen verlaufen auch hier relativ konstant, s. Abb. 6-9.

Die Untersuchungsergebnisse der zweifaktoriellen Varianzanalyse mit mehrfacher Besetzung weisen insbesondere bei den mittels Vor-Ort-Analytik bestimmten Variablen pH-Wert, Leitfähigkeit und Temperatur extrem niedrige Werte der Irrtumswahrscheinlichkeit auf. Somit kann

bei Einbeziehung der Wiederholungsmessungen bei diesen Feldparametern von einer deutlichen Signifikanz gesprochen werden, sowohl was den Einfluss der Faktoren DATUM und MESSSTELLE als auch deren Wechselwirkung (Interaktion) angeht. Letztere bedeutet, dass die örtliche Probenahme in jedem Fall zeitabhängig ist bzw. andersherum, für die zeitliche Abhängigkeit der mittels Feldanalytik bestimmten Größen der gewählte Standort der Probenahme als Parameter fungiert.

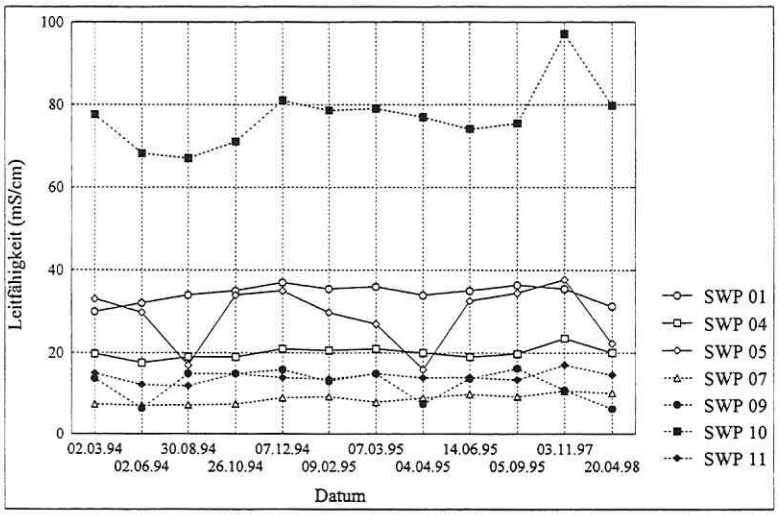

Abb. 6-7. Zeitreihenplot der Leitfähigkeit (Sickerwasser).

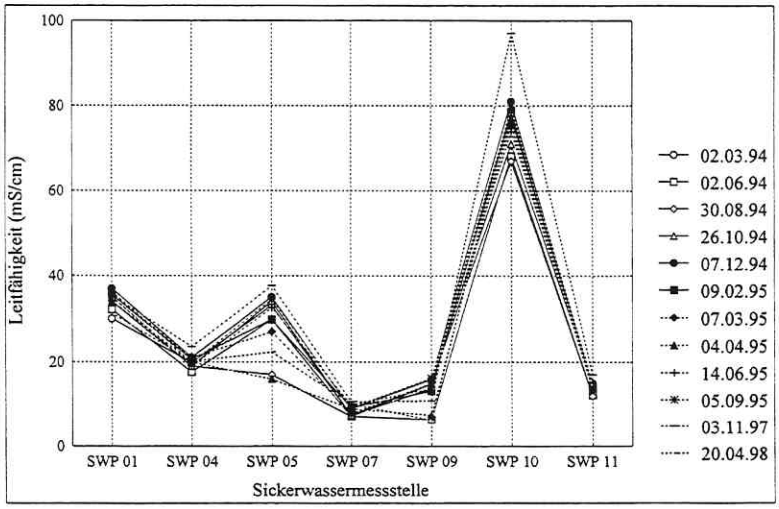

Abb. 6-8. Messstellenplot der Leitfähigkeit (Sickerwasser).

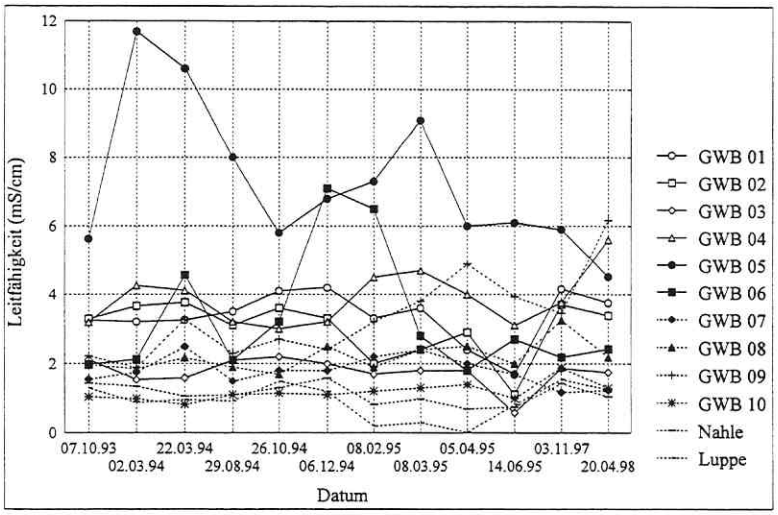

Abb. 6-9. Zeitreihenplot der Leitfähigkeit (Grundwasser).

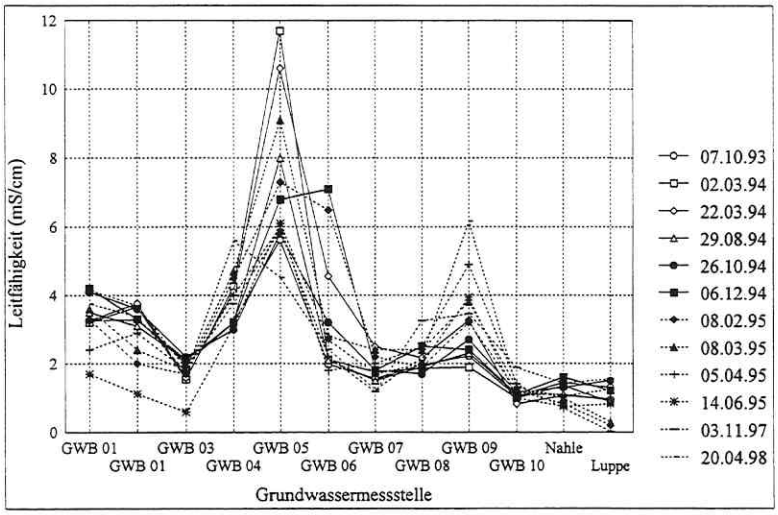

Abb. 6-10. Messstellenplot der Leitfähigkeit (Grundwasser).

Ein weiteres Ergebnis der Untersuchungen zur zweifachen Klassifikation mit mehrfacher Besetzung ist hervorzuheben: Eine Vielzahl der laboranalytisch bestimmten Einzelionen - insbesondere des Sickerwassers (Na^+, K^+, Mg^{2+}, SO_4^{2-}) - zeigt sich auch bei Einbeziehung der Wiederholungsmessungen unbeeinflusst vom Zeitpunkt der Messung, d. h. vom Haupteffekt DATUM. Damit wird nochmals deutlich, dass die Kontraste in den Daten der schadstoffrelevanten Parameter primär durch die lokal unterschiedlichen Belastungen hervorgerufen werden.

7 Überwachte Klassifikation

7.1 Vorbemerkungen

Der Begriff der Klassifikation wird im Zusammenhang mit der Datenanalyse üblicherweise in zwei verschiedenen Bedeutungen verwendet:

1. Klassen (= Gruppen) von Objekten bilden das Resultat einer Analyse
   ⇒ automatische Klassifikation (non supervised learning) bzw. Clusteranalyse (s. Kap. 8)
2. Klassen von Objekten sind im Datensatz (= Lerndatensatz) von vornherein vorhanden, neue Objekte unbekannter Herkunft (= Testdatensatz), anhand ihres Musters in denselben Variablen wie im ursprünglichen Datensatz, werden einer der Gruppen mit möglichst großer Sicherheit zugeordnet
   ⇒ überwachte Klassifikation (supervised learning)

Abb. 7-1 veranschaulicht - unter Bezugnahme auf HENRION und HENRION (1994) - die Datenstruktur bei Anwendung von Methoden der überwachten Klassifikation.

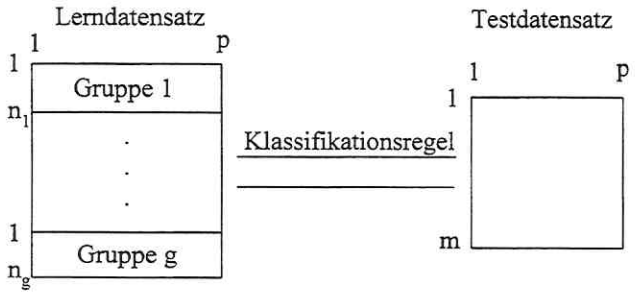

Abb. 7-1. Datenstruktur bei überwachter Klassifikation.

Jede Methode der überwachten Klassifikation wird in zwei Phasen unterteilt: Zunächst wird, ausgehend von den Variablenmustern einer bestimmten Menge von gruppierten (Lern-) Objekten, in der Lernphase das Klassifikationsmodell aufgestellt, anschließend erfolgt die Klassifizierung der Testobjekte (HENRION und HENRION, 1994).

Im Rahmen der Untersuchungen zur überwachten Klassifikation kamen die Methode der k nächsten Nachbarn (s. Abschnitt 7.2) sowie die lineare Diskriminanzanalyse (s. Abschnitt 7.3) zur Anwendung. Eine detaillierte Beschreibung weiterer Verfahren (BAYESsche Klassifikation, Mittelpunktregel u. a.) wird beispielsweise durch HENRION et al. (1988) sowie HENRION und HENRION (1994) vorgenommen. Erstere stellen zudem die mittels BAYESscher

Klassifikation und Methode der k nächsten Nachbarn erzielten Ergebnisse bei Anwendung auf den von FISHER (1936) publizierten Iris-Datensatz vor. Darüber hinaus hat ZWANZIGER (1988) in seiner Arbeit u. a. eine tabellarische Zusammenstellung der zahlreichen praktischen Anwendungen verschiedener Methoden - er bezeichnet sie gemeinhin als Diskriminationstechniken bzw. Diskriminationsmethoden - einschließlich Literaturübersicht vorgenommen.

7.2 Methode der k nächsten Nachbarn

7.2.1 Klassifizierung der Testobjekte

Die Methode der k nächsten Nachbarn (kurz: KNN-Methode) gehört zu den nichtparametrischen Klassifizierungsregeln, d. h., die Kenntnis bestimmter Parameter für die Wahrscheinlichkeitsverteilung der gegebenen Gruppen ist nicht erforderlich (HENRION et al., 1988).
Das von FIX und HODGES (1951) vorgeschlagene Verfahren ordnet ein Testobjekt derjenigen Gruppe zu, deren Lernobjekte unter den k nächsten Nachbarn (im Sinne des EUKLIDischen Abstands) des Testobjekts die Mehrheit bilden ($k \in N^+$). Der entscheidende Schritt besteht in der Festlegung des Parameters k. Ein zu kleiner Wert von k (z. B. $k = 1$) bedeutet zu wenig Information über die lokale Punktdichte („Mehrheitsverhältnisse“) in der Umgebung eines Testobjekts, die Klassifikation wäre aufgrund von Ausreißern sehr zufallsabhängig. Ein zu großer Wert von k (z. B. $k = n$) wiederum führt dazu, dass die Gruppe mit dem größten Stichprobenumfang die Mehrheit unter den sehr zahlreichen Nachbarn bildet. Die Klassifikation wäre damit zwar nicht mehr zufällig, dafür aber unabhängig vom Muster des Testobjekts von vornherein festgelegt und somit ebenfalls ohne Aussage. Die optimale Wahl von k erfolgt so, dass sich eine minimal zu erwartende Rate von Fehlklassifikationen ergibt (Schätzung mit der Leave-one-out-Methode, s. Abschnitt 7.2.2) (HENRION und HENRION, 1994).
Der Vorteil des Verfahrens besteht darin, dass es sehr leicht implementierbar ist und keine Verteilungsannahmen (z. B. multivariate Normalverteilung) vorausgesetzt werden. Dadurch ergeben sich zwischen Gruppen mit relativ unregelmäßiger Punktverteilung sehr flexible Trennkurven (bei mehr als zwei Variablen: Trennflächen). Zur Klassifizierung eines Testobjekts muss jedoch der gesamte Lerndatensatz zur Verfügung stehen, d. h. gespeichert sein, um die k nächsten Nachbarn in der Menge der Lernobjekte aufzufinden. Dies ist aufgrund des erhöhten Speicher- und Rechenzeitaufwandes von Nachteil (HENRION et al., 1988).

7.2.2 Schätzung der Klassifikationsfehlerrate

Die Einordnung von Testobjekten in vorgegebene Klassen erfolgt bei Anwendung der KNN-Methode dadurch, dass man zunächst bei variierendem Parameter k für jedes Objekt des Lerndatensatzes eine minimale Anzahl von Fehlklassifikationen ermittelt und anschließend mit Hilfe des optimalen k-Wertes die Einordnung vornimmt. Das bedeutet, dass zur eigentlichen Schätzung der Fehlerrate (= Anzahl Fehlzuordnungen / Gesamtanzahl Lernobjekte; auch: Irrtumswahrscheinlichkeit) nur Objekte bekannter Klassenherkunft (d. h. Lernobjekte) im Sinne von Pseudo-Testobjekten benutzt werden und die letztlich getroffene Entscheidung für einen (optimalen) k-Wert allein von diesen abhängt. Die Schätzung der Fehlerrate kann mittels Resubstitution oder Leave-one-out-Methode erfolgen, wobei die Anzahl der Fehlzuordnungen sich bei beiden Prozeduren aus der Mehrheitsbildung ergibt.

Bei der **Resubstitution** erfolgt die Anwendung der Klassifikationsregel für alle Lernobjekte, d. h. auch für das jeweilige Lernobjekt selbst (Resubstitution = Wiedereinsetzen). Unter den angegebenen k nächsten Nachbarn ($1 \leq k \leq n$) erscheint somit auch das jeweilige Lernobjekt selbst als zu sich nächster Nachbar mit dem EUKLIDischen Abstand null, daher ist die Mehrheitsbildung von vornherein zugunsten der korrekten Klasse verzerrt. Für $k = 1$ wird demnach stets eine korrekte Zuordnung (Fehlerrate 0 %) angezeigt, was bei der Anwendung auf echte Testobjekte aufgrund von Ausreißern u. ä. unrealistisch ist. Die Ursache der Verzerrtheit der Fehlerschätzung bei Anwendung der Resubstitution liegt somit darin, dass dieselben Objekte zur Anwendung gelangen, die schon an der Modellbildung beteiligt waren. Bei der **Leave-one-out-Methode** nach LACHENBRUCH (1967) wird jedes der n Lernobjekte genau einmal vom Datensatz isoliert. Die Klassifikationsregel wird auf das isolierte Objekt angewandt und ergibt sich aus den verbleibenden n -1 Objekten. Somit ist in jedem der n Klassifikationsläufe (jeweils mit $1 \leq k \leq n-1$), bestehend aus Modellbildung und Zuordnung, das zuzuordnende Objekt nicht an der Klassifikationsregel beteiligt, wodurch sich eine realistischere, i. Allg. höhere, Fehlerrate als bei der Resubstitution ergibt (HENRION und HENRION, 1994).

Bei der Resubstitution liegt das Minimum der Fehlerrate automatisch bei $k = 1$, diese Methode ist zur Ermittlung eines optimalen k-Wertes ungeeignet. Die Leave-one-out-Methode besitzt den Vorteil, dass sie fast unverzerrte Schätzungen liefert (LACHENBRUCH, 1967).

Als optimal ist daher der k-Wert (ggf. der kleinste unter gleich guten) mit minimaler Fehlerrate der Leave-one-out-Methode anzusehen.

7.2.3 Untersuchungen

7.2.3.1 Gegenstand und Zielstellung der Untersuchungen

Es wurden insgesamt drei Untersuchungen durchgeführt:

1. Bildung von sechs Lerndatensätzen, jeweils Zuordnung von einem Testdatensatz GWB 5
   Die sechs Lerndatensätze resultieren aus den jeweils zusammengefassten Werten gemeinsam durchgeführter Sicker- und Grundwassermesskampagnen von sechs verschiedenen Monaten, s. Tab. 7-1. Für jeden der Lerndatensätze wurde eine Unterteilung in die beiden Gruppen Sickerwasser und Grundwasser vorgenommen und die Zuordnung von einem Testdatensatz bzw. Testobjekt, gebildet aus den Werten, die im jeweils gleichen Monat am GWB 5 gemessen wurden, überprüft.
2. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5
   Der Lerndatensatz besteht aus den zusammengefassten Werten gemeinsam durchgeführter Sicker- und Grundwassermesskampagnen von insgesamt sechs Monaten, s. Tab. 7-1. Es wurde eine Unterteilung in 12 Gruppen (Sickerwasser und Grundwasser für jede Messkampagne) vorgenommen und die Zuordnung von sechs Testobjekten (GWB 5, sechs Monate) kontrolliert.
3. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7
   Die Struktur der Datensätze ist identisch mit der von 2. mit dem Unterschied, dass die sechs Testdatensätze aus den am SWP 7 gemessenen Werten rekrutierten.

In Tab. 7-1 sind die den Datensätzen zugrunde liegenden Messkampagnen aufgeführt, wobei sich in den Lerndatensätzen die Anzahl der Objekte in den jeweiligen Gruppen aufgrund der Herausnahme der Testobjekte (jeweils eines bei der 1. Untersuchung und insgesamt sechs bei der 2. und 3.) entsprechend reduziert.

Die an den Objekten (Messstellen) gemessenen Merkmale sind die Einzelionen Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^- und SO_4^{2-}.

Im betrachteten Untersuchungszeitraum ist der GWB 5 bezüglich der Salzfrachten (Vergleich der Summationswerte der Einzelionen) die höchstbelastete Grundwassermessstelle. Die Zielstellung der 1. bzw. 2. Untersuchung bestand darin festzustellen, inwiefern sich bei Herausnahme der an GWB 5 gemessenen Daten als Testobjekt diese hohe Belastung in der Klassifikation widerspiegelt, d. h., ob bei 1. in dem jeweiligen Monat eine (inkorrekte) Zuordnung zur

Gruppe Sickerwasser und bei 2. generell eine (inkorrekte) Zuordnung zu einer der Gruppen Sickerwasser erfolgt. Der SWP 7 enthält die geringsten Gehalte der betrachteten Einzelionen (Vergleich der Summationswerte). Die 3. Untersuchung sollte zeigen, ob die geringe Schadstoffbelastung ausreichend für eine (inkorrekte) Klassifikation in eine der Gruppen Grundwasser ist.

In den Untersuchungen wurde jeweils zunächst der optimale k-Wert ermittelt (Schätzung mit der in Abschnitt 7.2.2 besprochenen Leave-one-out-Methode).

Tab. 7-1. KNN-Methode. - Sicker- und Grundwasserdaten für die Untersuchungen.

Messkampagne (Datum)	Anzahl n der Objekte (Messstellen)		
	gesamt	Gruppe Sickerwasser	Gruppe Grundwasser
02. 03. 1994	20	8	12
30. 08. 1994	22	8	14
09. 02. 1995	20	6	14
04.04. 1995	24	10	14
05.09. 1995	22	9	13
03.11.1997	20	8	12

7.2.3.2 Untersuchungsergebnisse

Sämtliche Untersuchungsergebnisse (Abbildungen und Tabellen) sind im nachfolgenden Abschnitt enthalten. An dieser Stelle wird hierüber eine Übersicht gegeben.

1. Bildung von sechs Lerndatensätzen, Zuordnung von jeweils einem Testdatensatz GWB 5
   - Ermittlung der optimalen k-Werte: s. Abb. 7-2
   - Überprüfung der Gruppenzugehörigkeit: s. Tab. 7-2
2. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5
   - Ermittlung des optimalen k-Wertes: s. Abb. 7-3
   - Überprüfung der Gruppenzugehörigkeit: s. Tab. 7-3
3. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7
   - Ermittlung des optimalen k-Wertes: s. Abb. 7-4
   - Überprüfung der Gruppenzugehörigkeit: s. Tab. 7-4

Zu den Tabellen ist anzumerken, dass hier jeweils die (inkorrekten) Zuordnungen zu einer Sickerwassergruppe (Tab. 7-2 und 7-3) bzw. Grundwassergruppe (Tab. 7-4) durch Fettdruck und die absolut korrekten Zuordnungen durch Unterstreichung hervorgehoben sind.

7.2.3.3 Diskussion der Untersuchungsergebnisse

1. Bildung von sechs Lerndatensätzen, Zuordnung von jeweils einem Testdatensatz GWB 5
Trotz seiner relativ hohen Belastung an Salzgehalten erfolgt für den GWB 5 in allen sechs Fällen keine Zuordnung zur Gruppe der Sickerwassermessstellen, s. Tab. 7-2. Zu diesem Untersuchungsergebnis kommt man bei Anwendung sowohl des optimalen k-Wertes als auch der in dessen unmittelbarer Nähe gelegenen „erzwungenen" k-Werte, für welche die Fehleranzahl (und damit die Klassifikationsfehlerrate) ebenfalls noch relativ niedrig ist, s. Abb. 7-2.

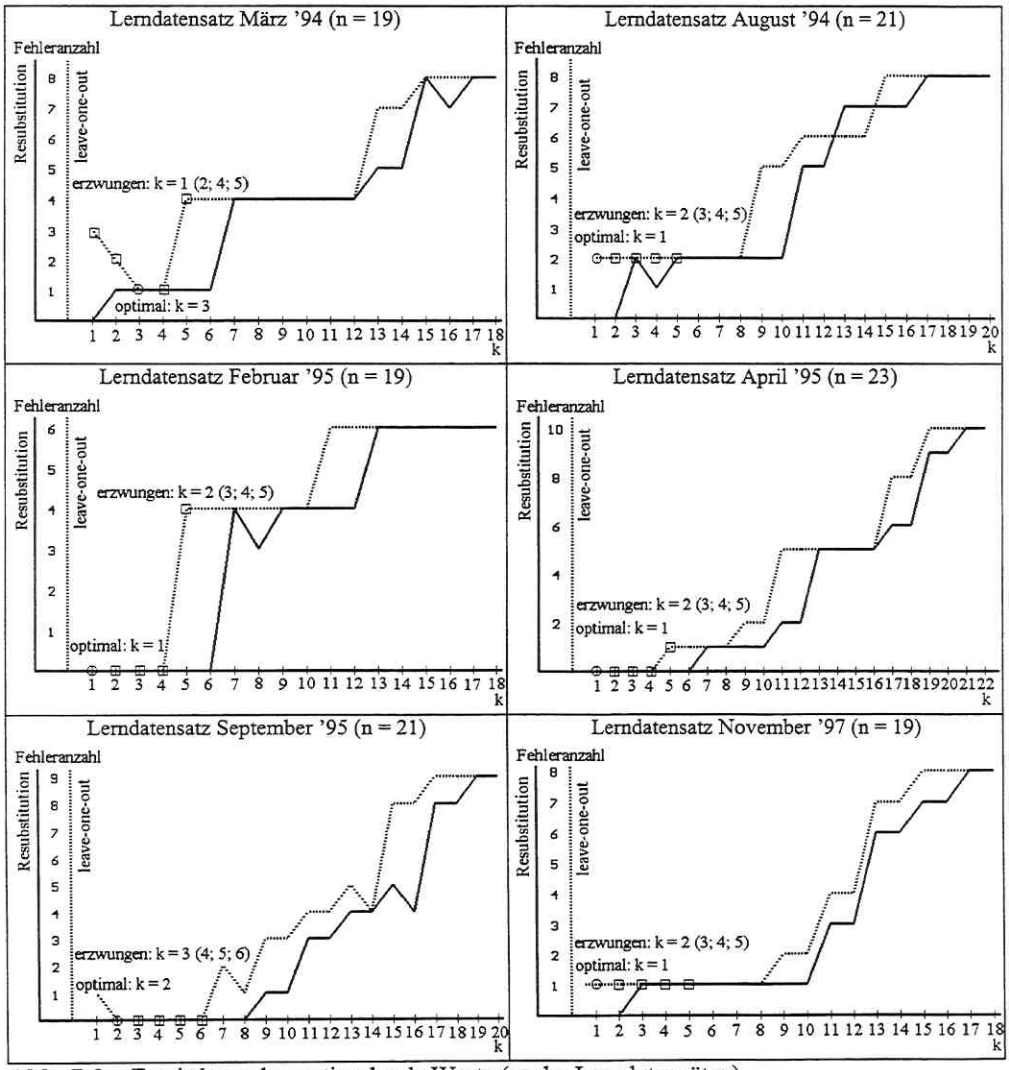

Abb. 7-2. Ermittlung der optimalen k-Werte (sechs Lerndatensätze).

Tab. 7-2. Klassifikationsergebnis (sechs Lerndatensätze, je ein Testdatensatz GWB 5).

Lern-, Test-datensatz	Gruppenzuordnung bei Variation von k									
	k (optimal)	Gruppe	k	Gruppe	k	Gruppe	k	Gruppe	k	Gruppe
März '94	3	GW	1	GW	2	GW	4	GW	5	GW
August '94	1	GW	2	GW	3	GW	4	GW	5	GW
Februar '95	1	GW	2	GW	3	GW	4	GW	5	GW
April '95	1	GW	2	GW	3	GW	4	GW	5	GW
September '95	2	GW	3	GW	4	GW	5	GW	6	GW
November '97	1	GW	2	GW	3	GW	4	GW	5	GW

Dies war in der Tat nicht zu erwarten und erscheint recht außergewöhnlich. Eine Ursache für dieses Untersuchungsergebnis könnte darin liegen, dass von einer sehr groben Partitionierung der Datensätze (jeweils nur zwei Gruppen) ausgegangen wurde. Somit kann lediglich vermutet werden, dass der korrekten Zuordnung zwar unsichere Mehrheiten zugrunde liegen, aber letztendlich das der hohen Belastung entsprechende Merkmalsmuster des GWB 5 eben doch nicht ausreichend für eine entsprechende Fehlklassifikation ist. Als eine weitere Ursache muss die Klassenhäufigkeit angesehen werden: In allen sechs Fällen liegt die Anzahl der Objekte in der „Grundwassergruppe" deutlich über der Anzahl derer in der „Sickerwassergruppe", s. Tab. 7-1, wodurch die Bildung der Mehrheitsverhältnisse insbesondere bei höheren Werten des Parameters k beeinflusst wird.

2. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5

Abb. 7-3 stellt den Verlauf der Fehleranzahl (Leave-one-out) des vergrößerten (Zusammenfassung der Werte der Messkampagnen) und entsprechend feiner partitionierten Lerndatensatzes dar. Die Klassifikationsfehlerrate liegt generell relativ hoch - selbst für den optimalen k-Wert beträgt sie 77,87 % (95 Fehler).

Das Testobjekt GWB 5 des Monats März '94 wird für k = 1 der dem gleichen Datum entsprechenden Sickerwassergruppe zugeordnet. Dies kann sowohl an den hohen Werten liegen, die in diesem Monat am GWB 5 gemessen wurden, als auch an den relativ geringen Werten, die zum gleichen Zeitpunkt an den Sickerwassermessstellen (SWP 1, SWP 4, SWP 7) registriert wurden, wie die Ergebnisse der Varianzanalyse (s. Abb. 6-7 - 6-10 mit der Leitfähigkeit als Summenparameter) deutlich belegen.

Ansonsten zeigen aber auch hier die Untersuchungsergebnisse, dass die hohe Belastung im Bereich vom GWB 5 nicht ausreichend für eine entsprechende (Fehl-) Klassifikation in eine der Sickerwassergruppen ist.

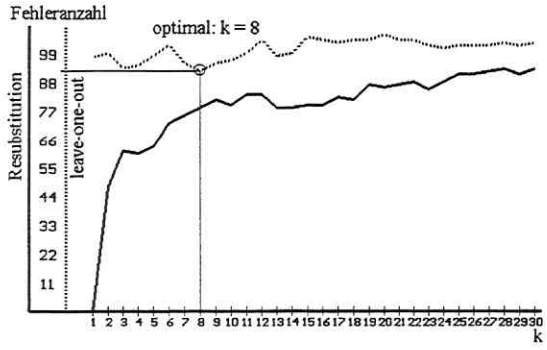

Abb. 7-3. Ermittlung des optimalen k-Wertes (ein Lerndatensatz, n = 122).

Tab. 7-3. Klassifikationsergebnis (ein Lerndatensatz, sechs Testdatensätze GWB 5).

Testdatensatz	Gruppenzuordnung bei Variation von k				
GWB 5	k = 8 (optimal)	k = 1	k = 2	k = 3	k = 4
März '94	GW Aug. '94	**SW März '94**	GW Nov. '97	GW Nov. '97	GW Aug. '94
August '94	GW Aug. '94	GW Nov. '97	GW Nov. '97	GW März '94	GW März '94
Februar '95	GW Aug. '94	GW Nov. '97	GW Febr. '95	GW Aug. '94	GW Febr. '95
April '95	GW Aug. '94	GW Nov. '97	GW Febr. '95	GW Nov. '97	GW Febr. '95
September '95	GW Aug. '94	GW Aug. '94	GW Aug. '94	GW Aug. '94	GW Nov. '97
November '97	GW Nov. '97	GW Aug. '94	GW Aug. '94	GW Aug. '94	GW März '94

3. Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7

Bei dieser Untersuchung wurde für den Lerndatensatz ebenfalls ein optimaler k-Wert von acht ermittelt, s. Abb. 7-4. Die hier dargestellten Kennlinien zur Schätzung der Klassifikationsfehlerrate ähneln denen von Abb. 7-3, da der Lerndatensatz bis auf die Herausnahme von SWP 7 (anstelle von GWB 5) als Testobjekt unverändert blieb.

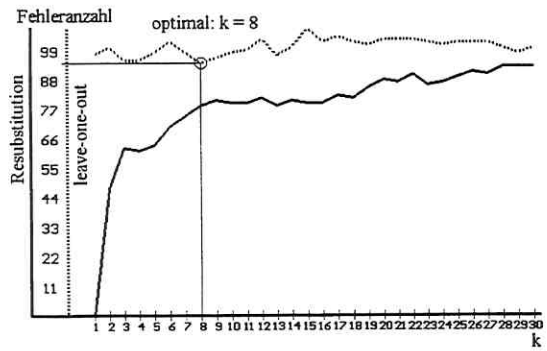

Abb. 7-4. Ermittlung des optimalen k-Wertes (ein Lerndatensatz, n = 122).

Eine (Fehl-) Klassifikation in eine der Grundwassergruppen wurde für den relativ gering belasteten SWP 7 nicht ermittelt, s. Tab. 7-4.

Tab. 7-4. Klassifikationsergebnis (ein Lerndatensatz, sechs Testdatensätze SWP 7).

Testdatensatz GWB 5	Gruppenzuordnung bei Variation von k				
	k = 8 (optimal)	k = 1	k = 2	k = 3	k = 4
März '94	SW April '95	SW Sept. '95	SW Aug. '94	SW Aug. '95	SW Aug. '94
August '94	SW Febr. '95	SW Aug. '94	SW Febr. '95	SW Febr. '95	SW Febr. '95
Februar '95	SW April '95	SW Febr. '95	SW April '95	SW April '95	SW April '95
April '95	SW April '95	SW April '95	SW April '95	SW April '95	SW April '95
September '95	SW April '95	SW April '95	SW April '95	SW April '95	SW April '95
November '97	SW April '95	SW April '95	SW April '95	SW April '95	SW April '95

Es wird jedoch deutlich, dass diese Probenahmestelle für k = 8 (optimaler Wert) fast ausschließlich der Sickerwassergruppe des Monats April '95 zugeordnet wird. Für diesen Zeitpunkt liegen die Schadstoffwerte der Sickerwassermessstellen im Vergleich zu den anderen Monaten relativ niedrig, s. Abb. 6-7 bzw. 6-8 des Kapitels Varianzanalyse. Somit spiegelt sich die geringe Belastung von SWP 7 zumindest in dieser Hinsicht in einer entsprechenden Zuordnung wider. Dies gilt umso mehr, da der Testdatensatz SWP 7 des Monats April '95 auch bei Variation von k seine absolut korrekte Zuweisung erfährt.

Bei den bis April '95 durchgeführten Messkampagnen erfolgt in drei von vier Fällen für k = 1 eine absolut korrekte Zuordnung des SWP 7 zu genau der Gruppe, welcher er entstammt, s. Tab. 7-4. Die aus den Salzfrachten induzierten Merkmalsmuster der Gruppen des Lerndatensatzes führen offensichtlich zu einer guten „Getrenntheit" der vorgegebenen Cluster und somit zu einer korrekten Identifikation des Testobjektes bei Wahl dieses kleinstmöglichen k-Wertes. Die Vermutung, dass sich die hohe (GWB 5) bzw. niedrige (SWP 7) Belastung in entsprechenden (Fehl-) Klassifikationen widerspiegelt, konnte somit bis auf eine Ausnahme (2. Untersuchung) nicht bestätigt werden. Ursachen hierfür können beispielsweise die bereits angesprochene zu grobe Partitionierung des Lerndatensatzes (1. Untersuchung) sowie die unterschiedliche Verteilung der Klassenhäufigkeiten (alle drei Untersuchungen) sein. Des Weiteren ist hervorzuheben, dass das Verfahren für Lerndatensätze geringeren Umfangs offensichtlich besser geeignet ist als für solche mit einer großen Anzahl von Objekten bzw. Gruppen. Die Fehleranzahlen bzw. Fehlerraten für den optimalen k-Wert liegen hier generell relativ niedrig (1. Untersuchung maximal 9,52 %, s. Abb. 7-2, 2. und 3. Untersuchung hingegen jeweils bei 77,87 %, s. Abb. 7-3 und 7-4). Damit ist in diesen Fällen eine höhere (realistischere) Klassifikationsleistung für echte Testobjekte zu erwarten, wie sich bei der 1. Untersuchung durch die jeweils korrekte Einordnung des Pseudo-Testobjekts (seine Gruppenzugehörigkeit war schließlich bekannt) GWB 5 in die Gruppe der Grundwassermessstellen gezeigt hat.

7.3 Lineare Diskriminanzanalyse

7.3.1 Vorbemerkungen

Bei der Diskriminanzanalyse erfolgt die Klassifizierung mit Hilfe sog. nichtelementarer Diskriminanzmerkmale (NED), wobei im Gegensatz zu den anderen Methoden der überwachten Klassifikation Aspekte wie Verschiedenheit der Gruppen und Variablenreduktion entscheidend sind. Die Zielstellung besteht zunächst darin, die verschiedenen Gruppen von Objekten bestmöglich zu trennen (zu diskriminieren). Die Unterschiede zwischen den Gruppen werden durch die Gesamtheit der unabhängigen Merkmale erklärt, sie sollen durch diese optimal separiert werden. Die Diskriminanzanalyse wird den dependenzanalytischen Verfahren zugeordnet: Die beobachteten Werte zweier oder mehrerer (unabhängiger) metrisch skalierter Merkmale werden durch die Werte eines abgeleiteten (abhängigen) nominal skalierten Merkmals ersetzt. Hierzu werden die Linearkombinationen der Merkmale benutzt, die verwendete Diskriminanzfunktion (= Trennfunktion) lautet:

$$Y = v_1 \cdot X_1 + v_2 \cdot X_2 + \ \ldots \ + v_p \cdot X_p ,$$

wobei Y das abhängige Diskriminanzmerkmal, X_j ein unabhängiges Merkmal j (j = 1, ... , p) und v_j die zu bestimmenden Diskriminanzkoeffizienten sind. Handelt es sich um einen Zwei-Gruppen-Fall, so spricht man von einer einfachen Diskriminanzanalyse (Y ist dichotom, es wird nur eine Diskriminanzfunktion berechnet), s. Abschnitt 7. 3. 2. Bei einer multiplen Diskriminanzanalyse (s. Abschnitt 7.3.3) sind mehr als zwei Gruppen vorgegeben (Y ist polytom, es werden mehrere Diskriminanzfunktionen ermittelt). Eine weitere Unterscheidung besteht zwischen linearen und (hier nicht betrachteten) nichtlinearen Diskriminanzfunktionen (SCHULZE, 1998; HENRION et al., 1988).

Nachdem die Gewichte v_1, ... , v_p so optimiert wurden, dass die gegebenen g Objektgruppen nach Transformation der Daten maximal getrennt sind, kann die Zuordnung von (bereits im Lerndatensatz enthaltenen) Pseudo-Testobjekten überprüft bzw. echte Testobjekte können unter Verwendung ebendieser Gewichte nachträglich klassifiziert werden. Die in den Untersuchungen zur Anwendung gekommene und in den nachfolgenden beiden Abschnitten zunächst erläuterte lineare Diskriminanzanalyse beruht auf dem von FISHER (1936) begründeten Distanzkonzept. Sie kann als eine Art Pendant der (interdependenzanalytischen) Hauptkomponentenanalyse (s. Kap. 9) für objekt-strukturierte Datensätze gesehen werden.

7.3.2 Lineare Diskriminanzanalyse im Zwei-Gruppen-Zwei-Merkmals-Fall

Es liegen zwei unabhängige Merkmale X_1 und X_2 vor mit einer Aufteilung der n Objekte in zwei Gruppen mit n_1 bzw. n_2 Objekten. Die Diskriminanzfunktion hat demnach die Form

$$Y = v_1 \cdot X_1 + v_2 \cdot X_2 .$$

Die Diskriminanzkoeffizienten sind so zu bestimmen, dass $\bar{y}_1$ (arithmetisches Mittel der Gruppe 1) möglichst weit von $\bar{y}_2$ (arithmetisches Mittel der Gruppe 2) entfernt ist (d. h., je größer der Abstand $d = |\bar{y}_1 - \bar{y}_2|$ bzw. $d^2 = (\bar{y}_1 - \bar{y}_2)^2$, umso größer wird die Unterschiedlichkeit zwischen beiden Gruppen sein) und gleichzeitig das Diskriminanzmerkmal Y in jeder der beiden Gruppen eine möglichst kleine Varianz s^2 aufweist. Dies bedeutet, den Ausdruck

$$A(v_1, v_2) = \frac{d^2}{s^2}$$

zu maximieren. Das Bilden der partiellen Ableitungen nach v_1 bzw. v_2 sowie das Einsetzen von d_1 bzw. d_2 (Differenz der Gruppenmittelwerte von Merkmal X_1 bzw. X_2) führt zu einem linearen Gleichungssystem, welches nach den gesuchten Koeffizienten v_1 und v_2 aufzulösen ist. Eine detaillierte Beschreibung der Vorgehensweise nimmt z. B. SCHULZE (1998) vor.

Im Falle standardisierter Ausgangsdaten (Diskussion s. Abschnitt 7.3.4) können die Diskriminanzkoeffizienten direkt als Maß für die Trenn- bzw. Separationsfähigkeit der Beobachtungsmerkmale angesehen werden. Durch entsprechende Normierung der Diskriminanzkoeffizienten lässt sich deren (prozentualer) Anteil am Gesamtbetrag der beiden Koeffizienten feststellen (HENRION et al., 1988; SCHULZE, 1998).

Mit Hilfe der Diskriminanzfunktion kann nun ermittelt werden, welcher Gruppe die Objekte zugeordnet werden müssen. Dies kann z. B. durch Vergleich der individuellen Diskriminanzwerte y_i mit einem zu berechnenden kritischen Diskriminanzwert y^* geschehen, die Zuordnungsvorschrift lautet: Ordne das Objekt i der Gruppe 1 zu, wenn $y_i < y^*$; ordne das Objekt i der Gruppe 2 zu, wenn $y_i > y^*$ (SCHULZE, 1998).

Zum Schluss lässt sich die Zuordnung der Objekte zu den Gruppen mit Hilfe einer Zuordnungs- bzw. Klassifikationsmatrix überprüfen.

Den dargelegten Algorithmus im Zwei-Gruppen-Zwei-Merkmals-Fall demonstriert RUDOLPH (1998a) anhand eines Beispieldatensatzes (Na^+- und Fe^{2+}- Werte der Wasserproben von 15 Messstellen (Objekten), die in zwei Gruppen (Sickerwasser und Grundwasser) aufgeteilt wurden).

7.3.3 Lineare Diskriminanzanalyse im Mehr-Gruppen-Mehr-Merkmals-Fall

7.3.3.1 Nichtelementare Diskriminanzmerkmale

Um die gesamte Information über die lineare Trennbarkeit zu erhalten, macht sich bei mehr als zwei Gruppen die Berechnung weiterer Diskriminanzmerkmale erforderlich. Das zunächst gesuchte „künstliche" Merkmal Y_1, das aus den Ausgangsmerkmalen $X_1, \ldots, X_p$ durch Linearkombination entsteht, muss gegenüber allen anderen Y_j ($j = 1, \ldots, t$) ein maximales Trennmaß besitzen. Daraus ergibt sich die gegenüber dem Zwei-Gruppen-Zwei-Merkmals-Fall nun allgemeingültige Forderung

$$\max_V \left\{ \frac{V^T \cdot \underline{B} \cdot V}{V^T \cdot \underline{W} \cdot V} / V \in R^p \right\},$$

d. h., der Begriff der Separation zwischen Objektgruppen ist als Verhältnis aus der Streuung zwischen den Gruppen (Streuung der Gruppenmittelpunkte) und der Streuung innerhalb der Gruppen zu definieren. HENRION und HENRION (1994) zeigen, dass die aus der HUYGENSschen Dekompositionsformel $\underline{T} = \underline{B} + \underline{W}$ resultierende äquivalente Forderung

$$\max_V \left\{ \frac{V^T \cdot \underline{B} \cdot V}{V^T \cdot \underline{T} \cdot V} / V \in R^p \right\}$$

auf die Lösung des Eigenwertproblems $\underline{T}^{-1} \cdot \underline{B} \cdot V = \lambda \cdot V$ führt. Hierbei sind $\underline{T}$, $\underline{B}$ und $\underline{W}$ quadratische Matrizen (p Zeilen und Spalten) von Abweichungsquadraten:

- $\underline{T}$ ... Matrix der Gesamtabweichungsquadrate („total"),
- $\underline{W}$ ... Matrix der Abweichungsquadrate innerhalb der Gruppen („within") und
- $\underline{B}$ ... Matrix der Abweichungsquadrate zwischen den Gruppen („between").

Ist der zum größten Eigenwert λ_1 gehörende und durch $V_1^T \cdot \underline{S} \cdot V_1 = 1$ ($\underline{S}$: Kovarianzmatrix) normierte Eigenvektor $V_1 = (v_{11} \quad v_{21} \quad \ldots \quad v_{p1})^T$ berechnet, so sind dessen Komponenten gerade die gesuchten Koeffizienten (optimalen Gewichte) für Y_1, d. h.

$$Y_1 = v_{11} \cdot X_1 + v_{21} \cdot X_2 + \ldots + v_{p1} \cdot X_p .$$

Dabei ist λ_1 das Trennmaß des Merkmals Y_1, es entspricht dem für die optimalen Gewichte $v_{11}, \ldots, v_{p1}$ erreichten maximalen Wert des Separationsquotienten. Anschließend wird aus dem „Rest" des Diskriminanzraumes, d. h. unter allen verbliebenen Vektoren V, erneut derjenige berechnet, der der bestmöglichen Linearkombination entspricht, nämlich V_2 mit dem zugehörigen zweitgrößten Eigenwert λ_2:

$$Y_2 = v_{12} \cdot X_1 + v_{22} \cdot X_2 + ... + v_{p2} \cdot X_p .$$

Y_1 und Y_2 bilden für die Separation der Objektgruppen das beste durch lineare Transformation aus den Ausgangsmerkmalen entstandene Merkmalspaar. Die hieraus resultierende Objektdarstellung in der Ebene wird als LDA-Display bezeichnet (HENRION et al., 1988).

Durch Fortführung der Berechnungen in der beschriebenen Art erhält man schließlich alle NED $Y_1, Y_2, ..., Y_t$ (mit dem gesamten Trennmaß $\lambda_1 + \lambda_2 + ... + \lambda_t$), welchen aufgrund der Polarisation des Trennmaßes eine sequentiell abnehmende Bedeutung zukommt. Die Anzahl t der NED ergibt sich aus

$$t = \min (g - 1; p),$$

d. h., es gibt i. Allg. weniger NED als Ausgangsmerkmale. HENRION et al. (1988) haben diesbezüglich eine Zusammenstellung vorgenommen, s. Tab. 7-5.

Obwohl $t \leq p$ ist, stimmt das multivariate Trennmaß T^2 aller NED mit dem aller Ausgangsmerkmale überein, d. h.

$$T^2 (Y_1, ..., Y_t) = T^2 (X_1, ..., X_p).$$

Das multivariate Trennmaß T^2 der Merkmalsmenge $X_1, ..., X_p$ ist dabei definiert als

$$T^2 (X_1, ..., X_p) = tr (\underline{B} \cdot \underline{T}^{-1}) .$$

Hierbei ist Spur tr („trace") die Summe aller Diagonalelemente der bezeichneten Matrix (HENRION et al., 1988; HENRION und HENRION, 1994).

Tab. 7-5. Anzahl nichtelementarer Diskriminanzmerkmale (NED).

Anzahl g der Gruppen	Anzahl p der Merkmale	Anzahl t der NED
2	beliebig	1
3	1	1
3	≥2	2
4	≥3	3

7.3.3.2 Klassifizierung echter Testobjekte

Während bisher die NED zur Informationsextraktion und Dimensionserniedrigung benutzt wurden, kann nun mit deren Hilfe die Klassifizierung echter Testobjekte erfolgen. Nachdem das neue Objekt mittels NED transformiert wurde (bei der Berechnung werden die bereits ermittelten Gewichte benutzt), gibt es nach HENRION et al. (1988) sowie HENRION und HENRION (1994) für dessen Zuordnung zwei Möglichkeiten:

1. Eindeutige Entscheidung

   Das Objekt wird derjenigen (l-ten) Gruppe zugeordnet, für die der Ausdruck

   $$\frac{n_l}{n_l + 1} \cdot d_l^2$$

   minimal ist. Hierbei sind d_l der EUKLIDische Abstand dieses Objektes zum Mittelpunkt der l-ten Gruppe und n_l deren Objektanzahl.

2. Mehrdeutige Entscheidung

   Das Objekt wird derjenigen Gruppe zugeteilt, in deren 95 %-Streubereich es fällt ($d_l \leq r_l$). Die Formel für die Berechnung von r_l (Streuradius der l-ten Gruppe) geben HENRION et al. (1988) an. Liegt es außerhalb sämtlicher Streubereiche, wird es als Ausreißer erkannt.

Die Schätzung des Diskriminationsfehlers (= Anzahl der Fehlklassifikationen / Gesamtanzahl der Objekte) kann wie bei dem KNN-Verfahren mittels Resubstitution oder der Methode nach LACHENBRUCH (1967) erfolgen, wobei hier jedoch zwei Kriterien eine Rolle spielen, die Signifikanz der NED und die Trenn- bzw. Separationsfähigkeit der Ausgangsmerkmale.

Bei Variation der Anzahl t von Diskriminanzmerkmalen durchläuft die Fehlerrate analog der des k-Wertes bei KNN ein Minimum, wodurch die optimale Reduktion ermittelt werden kann. Eine Verbesserung der Diskrimination, d. h. eine Verringerung des Diskriminationsfehlers, aber auch der Mehrfachzuordnungen, wird durch Benutzung ausschließlich signifikanter NED erreicht - exemplarisch zeigen dies HENRION et al. (1988) sowie EINAX et al. (1997). Ein analoges Verhalten tritt bei den Ausgangsvariablen selbst auf, denn auch hier beeinflussen nichtsignifikante (im Sinne einer nur geringen Separationsfähigkeit) die Fehlerrate negativ. HENRION und HENRION (1994) haben entsprechende Untersuchungen durchgeführt.

Bei der Schätzung mittels Leave-one-out-Methode erfolgt die Berechnung der NED ohne die echten Testobjekte, diese werden vorher jeweils aus dem Lerndatensatz herausgenommen und erst nachträglich wieder positioniert. Damit gestaltet sich die Anwendung hier nicht ebenso unkompliziert wie im KNN-Fall, da das der Diskrimination zugrunde liegende Modell mit jedem Objekt neu aufzustellen ist (HENRION et al., 1988).

Der in den Klassifikationsmatrizen der Untersuchungsergebnisse vorliegender Arbeit angegebene Diskriminationsfehler wurde prinzipiell mittels Resubstitution ermittelt, d. h., die Objekte des Lerndatensatzes selbst wurden als (Pseudo-) Testobjekte aufgefasst, um sie wieder zuzuordnen. Eine Schätzung im obigen Sinne erfolgte nicht, d. h., es wurden jeweils alle ermittelten NED unabhängig von ihrer Signifikanz sowie alle betrachteten Ausgangsmerkmale unabhängig von ihrer Separationsfähigkeit zur Klassifizierung herangezogen.

7.3.4 Anmerkungen zu den Voraussetzungen und zur Datenvorbehandlung

AHRENS und LÄUTER (1981) nennen vier Voraussetzungen, die für die Anwendung der Diskriminanzanalyse erfüllt sein müssen:

1. mehrdimensionale Normalverteilung
2. Gleichheit der Kovarianzmatrizen
3. stochastische Unabhängigkeit der Messwertvektoren (Variablen)
4. keine Messwertausfälle

Bezüglich der erstgenannten Bedingung weisen beispielsweise JOHNSON und WICHERN (1982) darauf hin, dass man nicht notwendigerweise für alle Populationen (Klassen) eine multivariate Normalverteilung annehmen muss, und auch gegenüber kleineren Abweichungen der Gleichheit der Kovarianzmatrizen zeigt sich das Verfahren robust (STATSOFT, 1996). Die Unabhängigkeit der Variablen betreffend, ist der Arbeit von ZWANZIGER (1988) zu entnehmen, dass die erwartete Diskriminationsfehlerrate mit der Dimension p der Beobachtungsvektoren sowie mit hoher Korrelation anwächst und sie abnimmt, wenn die Stichprobe größer wird. Messwertausfälle traten in den für die Untersuchungen verwendeten Datensätzen nicht auf, da vorher jeweils eine objektweise Streichung vorgenommen wurde.

Im Rahmen der diesbezüglich durchgeführten Untersuchungen wurde aufgrund der genannten Aspekte auf die Überprüfung der einzelnen Voraussetzungen verzichtet. Die Untersuchungsergebnisse sind daher bezüglich der ermittelten Fehlerraten differenziert zu betrachten.

Zum Abschluss soll der Effekt der Datenvorbehandlung erwähnt werden. Bei der linearen Diskriminanzanalyse wird die Größe der Diskriminanzkoeffizienten v_j durch die Maßeinheiten und die Standardabweichungen der Beobachtungsmerkmale X beeinflusst. Dies lässt sich durch eine Standardisierung der Merkmalsdaten eliminieren - die berechneten Koeffizienten v_j lassen sich in diesem Fall direkt als Maß für die Bedeutung der Ausgangsmerkmale interpretieren (HENRION et al., 1988; SCHULZE, 1998). In den Untersuchungen wurden daher die Spaltenvektoren der (Roh-) Datenmatrix vor Anwendung des Verfahrens entsprechend transformiert. Auf die Ergebnisse der Diskriminanzanalyse selber (d. h. die Berechnung der Koordinaten eines Objekts im Diskriminanzraum) hat die Datenstandardisierung jedoch - im Unterschied zur Hauptkomponentenanalyse (s. Kapitel 9) - keinen Einfluss (HENRION und HENRION, 1994). Weitergehende Ausführungen zu den Effekten der Datenvorbehandlung sind in der Arbeit von ZWANZIGER (1988) enthalten.

7.3.5 Untersuchungen

7.3.5.1 Gegenstand und Zielstellung der Untersuchungen

Das generelle Ziel von Untersuchungen mittels Diskriminanzanalyse ist auf die Beantwortung der folgenden Fragen ausgerichtet:

- Unterscheiden sich die a priori gegebenen Gruppen bezüglich der gemessenen Merkmale (hier: Na^+-, K^+-, Mg^{2+}-, Ca^{2+}-, Cl^--und SO_4^{2-}-Ionen) voneinander?
- Welche Merkmale sind bedeutsam bzw. redundant?
- Inwiefern ermöglichen die Merkmale eine korrekte Zuordnung von Objekten unbekannter Herkunft?

Die Untersuchungen vorliegender Arbeit unterteilten sich in zwei Abschnitte:

1. Lineare Diskriminanzanalyse mit dem primären Ziel der optimalen linearen Separation von gegebenen Objektgruppen (ohne Klassifizierung von echten Testobjekten)
   - Sieben-Gruppen-Sechs-Merkmals-Fall: Der Lerndatensatz wurde aus den Sickerwasserdaten von insgesamt 11 Messkampagnen gebildet. Die Unterteilung erfolgte in sieben Gruppen (SWP 1, 4, 5, 7, 9, 10 und 11), welche jeweils die zusammengefassten Werte einer Messstelle enthalten, s. Tab. 7-6.
   - Drei-Gruppen-Sechs-Merkmals-Fall: Der Lerndatensatz resultierte aus den Grundwasserdaten von insgesamt 17 Messkampagnen. Die a priori-Gruppierung erfolgte bezüglich dem Anstrombereich der Deponie (GWB 7 - 9) sowie den Abstrombereichen der Luppe (GWB 1 - 3) und der Nahle (GWB 4 - 6), wobei jeweils die Werte der drei Messstellen zusammengefasst wurden, s. Tab. 7-7.

   Durch diese Untersuchungen konnte ermittelt werden, inwiefern sich die Sickerwassermessstellen bezüglich der Salzfrachten voneinander unterscheiden und ob sich der resultierende Schadstoffeintrag in das Grundwasser in einer eindeutigen Abgrenzung zwischen dem Anstrombereich und den beiden Abstrombereichen des Deponiekörpers widerspiegelt. Der in den Klassifikationsmatrizen angegebene Diskriminationsfehler wurde mittels Resubstitution ermittelt, d. h., die Objekte des Lerndatensatzes selbst wurden als unbekannt aufgefasst, um sie wieder zuzuordnen.

   Weiterhin sollte festgestellt werden, welche der aus dem Bauschutt- bzw. Aschenanteil des Deponiekörpers resultierenden Salzfrachten für die vorgenommenen lokalen Differenzie-

rungen signifikant bzw. redundant sind. Aus diesem Grunde sind in den Ergebnissen auch die berechneten Diskriminanzkoeffizienten (Einbeziehung aller sechs Ausgangsmerkmale) aufgeführt. Die (prozentualen) Anteile der NED am Gesamttrennmaß sollen die Signifikanz von Y_1 und Y_2 verdeutlichen, da mittels dieser beiden „künstlichen" Merkmale die Objektdarstellung in der Ebene (LDA-Display) erfolgt.

Tab. 7-6. Sickerwasserdaten für den Sieben-Gruppen-Sechs-Merkmals-Fall.

Messkampagne (Datum)	Objekte gesamt	Anzahl n_i der Objekte je Gruppe l (g = 7)						
		SWP 1	SWP 4	SWP 5	SWP 7	SWP 9	SWP 10	SWP 11
07.12.92	3	1	1	1	0	0	0	0
24.03.93	3	1	1	1	0	0	0	0
16.09.93	2	0	1	1	0	0	0	0
02.03.94	7	1	1	1	1	1	1	1
02.06.94	7	1	1	1	1	1	1	1
30.08.94	7	1	1	1	1	1	1	1
26.10.94	7	1	1	1	1	1	1	1
09.02.95	5	0	1	1	1	1	0	1
04.04.95	7	1	1	1	1	1	1	1
05.09.95	7	1	1	1	1	1	1	1
03.11.97	6	1	0	1	1	1	1	1
gesamt	61	9	10	11	8	8	7	8

Tab. 7-7. Grundwasserdaten für den Drei-Gruppen-Sechs-Merkmals-Fall.

Messkampagne (Datum)	Objekte gesamt	Anzahl n_i Objekte je Gruppe l (g = 3)		
		GWB 1-3	GWB 4-6	GWB 7-9
07.12.92	1	0	1	0
24.03.93	9	3	3	3
05.04.93	2	0	2	0
14.04.93	2	0	2	0
19.05.93	8	3	3	2
16.09.93	9	3	3	3
02.03.94	9	3	3	3
22.03.94	2	0	2	0
02.06.94	2	0	2	0
29.08.94	9	3	3	3
06.12.94	9	3	3	3
08.02.95	9	3	3	3
08.03.95	9	3	3	3
05.04.95	9	3	3	3
14.06.95	6	3	1	2
05.09.95	8	3	2	3
03.11.97	9	3	3	3
gesamt	110	36	40	34

2. Lineare Diskriminanzanalyse mit dem primären Ziel der Klassifizierung von Testobjekten

   Der Gegenstand der Untersuchungen entsprach dem der KNN-Methode, d. h.:

   - Zwei-Gruppen-Sechs-Merkmals-Fall: Bildung von sechs Lerndatensätzen, Zuordnung von jeweils einem Testdatensatz GWB 5
   - Zwölf-Gruppen-Sechs-Merkmals-Fall: Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5
   - Zehn-Gruppen-Sechs-Merkmals-Fall: Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7

   Die berücksichtigten Messkampagnen sind in Tab. 7-1 des Abschnitts 7.2.3.1 aufgeführt.

   Das Ziel dieser Untersuchungen bestand analog denen der KNN-Methode im Wesentlichen darin festzustellen, inwiefern die Gehalte an Einzelionen ausreichend für eine Fehlklassifikation von besonders hoch- bzw. niedrigbelasteten Probenahmestellen sind.

Wie am Ende des vorhergehenden Abschnitts bereits erwähnt, wurde keine Schätzung des Diskriminationsfehlers durchgeführt, d. h., es wurden jeweils alle ermittelten NED unabhängig von ihrer Signifikanz sowie alle sechs Ausgangsmerkmale unabhängig von ihrer Separationsfähigkeit zur Klassifizierung herangezogen.

7.3.5.2 Untersuchungsergebnisse

Die im nachfolgenden Abschnitt enthaltenen Untersuchungsergebnisse beinhalten die folgenden Tabellen und Abbildungen:

- 1. Untersuchung
  - Sieben-Gruppen-Sechs-Merkmals-Fall (Lerndatensatz Sickerwassermessstellen)
    - Klassifikationsmatrix: s. Tab. 7-8
    - Fehlklassifikationen: s. Tab. 7-9
    - Diskriminanzkoeffizienten und Trennmaß der NED: s. Tab. 7-10
    - LDA-Display: s. Abb. 7-5
  - Drei-Gruppen-Sechs-Merkmals-Fall (Lerndatensatz Grundwassermessstellen)
    - Klassifikationsmatrix: s. Tab. 7-11
    - Fehlklassifikationen: s. Tab. 7-12
    - Diskriminanzkoeffizienten und Trennmaß der NED: s. Tab. 7-13
    - LDA-Display: s. Abb. 7-6

- 2. Untersuchung
  - Bildung von sechs Lerndatensätzen, Zuordnung v. jeweils einem Testdatensatz GWB 5
    - Klassifikationsergebnis: s. Tab. 7-14
    - eindimensionale Darstellung (Lern-, Testdatensatz August '94): s. Abb. 7-7
  - Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5
    - Klassifikationsergebnis: s. Tab. 7-15
  - Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7
    - Klassifikationsergebnis: s. Tab. 7-16
- 1. und 2. Untersuchung
  - Signifikanz der Merkmale für die Trennung der Gruppen: s. Tab. 7-17

Zu den Ergebnissen der beiden Untersuchungen ist folgendes anzumerken:

- Die in den beiden LDA-Display's (Abb. 7-5 und 7-6) für jede Objektgruppe vorgenommene Eintragung der Ellipse für den 95 % Konfidenzbereich ist in dem hier zur Anwendung gekommenen Programmpaket STATISTICA (Version 5.1) implementiert. Die Ellipse basiert auf der Annahme, dass die beiden Variablen einer bivariaten Normalverteilung genügen, ihre Ausrichtung wird durch das Vorzeichen des Korrelationskoeffizienten zwischen den beiden Variablen bestimmt (STATSOFT, 1996).
- Sämtliche Klassifikationsergebnisse (mit Ausnahme derer, die in Tab. 7-14 aufgeführt sind) beruhen prinzipiell auf einer eindeutigen Entscheidung für die Zuordnung der Objekte (s. vorhergehender Abschnitt), wobei ein Objekt in diejenige Gruppe klassifiziert wird, für die es die höchste sog. posteriori-Klassifikationswahrscheinlichkeit (STATSOFT, 1996) besitzt. Die teilweise angegebenen nachfolgenden Zuordnungen, s. Tab. 7-9, 7-15 und 7-16, beruhen auf den berechneten nächsthöheren Werten dieser Wahrscheinlichkeit.
- In den Tabellen 7-15 und 7-16 sind die (inkorrekten) Zuordnungen zu einer der Sicker- bzw. Grundwassergruppen durch Fettdruck und die absolut korrekten Zuordnungen durch Unterstreichung hervorgehoben.

7.3.5.3 Diskussion der Untersuchungsergebnisse

1. Untersuchung

Für den Lerndatensatz Sickerwassermessstellen wurde eine ausgezeichnete Separation der a priori gegebenen Gruppen bezüglich der gemessenen Merkmale (Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^-,

SO_4^{2-}) ermittelt, s. Tab. 7-8 (Klassifikationsmatrix) und Abb. 7-5 (LDA-Display). Es ergeben sich bei insgesamt 61 Objekten (Messstellen) lediglich drei Fehlklassifikationen (s. Tab. 7-9), wobei hier in der berechneten Rangfolge der neuen Zuordnung bereits an der 2. bzw. 3. Stelle wieder eine korrekte Klassifikation erfolgt.

Tab. 7-8. Klassifikationsmatrix für den Lerndatensatz Sickerwassermessstellen.

			Korrigierte Gruppierung						
		% korrekt	01	04	05	07	09	10	11
Ur-	SWP 01	100,00	9	0	0	0	0	0	0
sprüng-	SWP 04	100,00	0	10	0	0	0	0	0
liche	SWP 05	90,91	0	0	10	1	0	0	0
Gruppie-	SWP 07	87,50	0	0	0	7	1	0	0
rung	SWP 09	87,50	0	0	0	1	7	0	0
	SWP 10	100,00	0	0	0	0	0	7	0
	SWP 11	100,00	0	0	0	0	0	0	8
	Σ	95,08	9	10	10	9	8	7	8

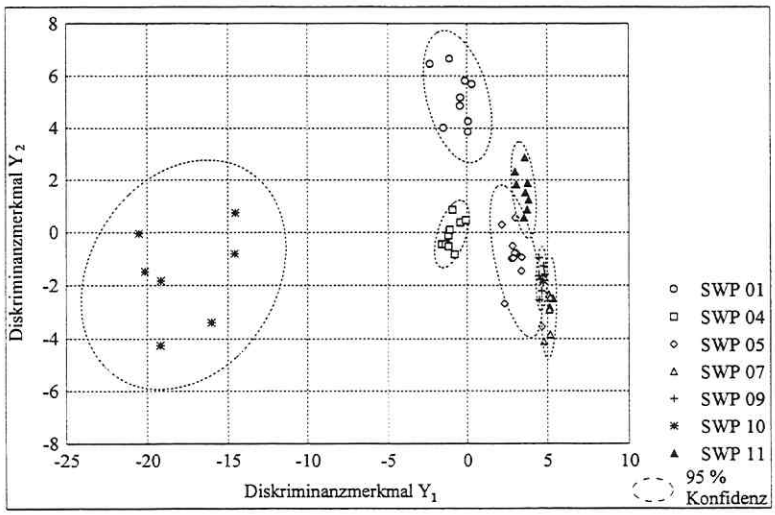

Abb. 7-5. LDA-Display für den Lerndatensatz Sickerwassermessstellen.

Tab. 7-9. Fehlklassifikationen des Lerndatensatzes Sickerwassermessstellen.

Lernobjekt	1. neue Zuordnung	2. neue Zuordnung	3. neue Zuordnung
SWP 05 (07.12.92)	SWP 07	SWP 09	SWP 05
SWP 07 (02.03.94)	SWP 09	SWP 07	SWP 11
SWP 09 (09.02.95)	SWP 07	SWP 09	SWP 05

Die in Tab. 7-10 aufgeführten Werte offenbaren, dass im ersten NED ein sehr hoher Anteil am Gesamttrennmaß enthalten ist (über 83 %), aber auch das zweite NED mit über 11 % noch einen relativ hohen Wert aufweist. Das LDA-Display reproduziert damit fast die Gesamtinformation zur Gruppentrennbarkeit und liefert diesbezüglich eine entsprechend realistische Darstellung. Die zum ersten bzw. zweiten NED gehörenden Diskriminanzkoeffizienten verdeutlichen, dass vor allem Mg^{2+} (im ersten NED) und SO_4^{2-} (im ersten und zweiten NED), aber auch Cl^- (zweithöchster Wert im zweiten NED) einen überragenden Beitrag zur Trennung der Gruppen liefern und demhingegen K^+ und Na^+ hier nur eine untergeordnete Rolle spielen.

Tab. 7-10. Diskriminanzkoeffizienten und Trennmaß der NED.

Variable	v_1	v_2	v_3	v_4	v_5	v_6
Na^+	-0,2551	-0,8823	-0,4685	0,6490	-0,1098	1,6542
K^+	0,1795	0,7911	0,2317	-0,1958	-1,0065	-1,2097
Mg^{2+}	-3,5937	1,9489	0,6970	-2,8317	-0,0943	0,4039
Ca^{2+}	0,4665	-1,2761	-1,1032	-0,8248	-0,2942	-0,2777
CL^-	-0,9165	4,4457	-2,2766	0,5626	0,4846	-0,0004
SO_4^{2-}	-2,3894	-6,3007	1,2030	1,7438	0,0920	-1,0049
λ_j	51,6681 ≙ 83,85 %	7,1025 ≙ 11,53 %	2,2863 ≙ 3,71 %	0,5181 ≙ 0,84 %	0,0424 ≙ 0,07 %	0,0028 ≙ 0,00 %

Für den Lerndatensatz Grundwassermessstellen stellt sich die Separation im LDA-Display weniger deutlich dar, s. Abb. 7-6.

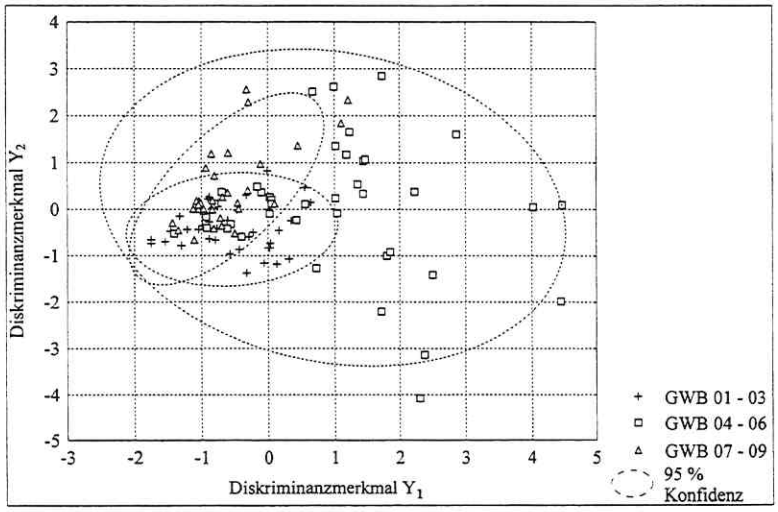

Abb. 7-6. LDA-Display für den Lerndatensatz Grundwassermessstellen.

Die Ursache hierfür liegt zunächst darin, dass für diesen Datensatz eine relativ grobe Partitionierung vorgegeben wurde - es bildet nicht jede Messstelle eine Gruppe für sich, sondern a priori werden hier lediglich drei Gruppen (Anstrombereich, Abstrombereiche der Nahle und der Luppe) vorausgesetzt. Die sich insbesondere für den Abstrombereich der Nahle (GWB 4 - 6) zeigende breite Streuung der Daten wird wohl vor allem durch die hier lokal sehr unterschiedlichen Belastungen hervorgerufen: Im nordwestlichen Deponiebereich (GWB 4 und 5) sind diese wesentlich höher als in dem vom GWB 6. Unter diesem Aspekt ist das Klassifikationsergebnis noch als relativ gut einzuschätzen: Bei 110 einbezogenen Objekten (Messstellen) werden nur 32 fehlklassifiziert, was dem prozentualen Anteil einer korrekten Zuordnung von 70,91 % entspricht, s. Tab. 7-11 und 7-12.

Tab. 7-11. Klassifikationsmatrix für den Lerndatensatz Grundwassermessstellen.

			Korrigierte Gruppierung		
		% korrekt	GWB 01 - 03	GWB 04 - 06	GWB 07 - 09
Ursprüngliche Gruppierung	GWB 01 - 03	77,78	28	3	5
	GWB 04 - 06	62,50	7	25	8
	GWB 07 - 09	73,53	7	2	25
	Σ	70,91	42	30	38

Tab. 7-12. Fehlklassifikationen des Lerndatensatzes Grundwassermessstellen.

Lernbjekt	1. neue Zuordnung	Lernobjekt	1. neue Zuordnung
GWB 02 (24.03.93)	GWB 04 - 06	GWB 08 (06.12.94)	GWB 01 - 03
GWB 06 (24.03.93)	GWB 01- 03	GWB 01 (08.02.95)	GWB 07 - 09
GWB 04 (05.04.93)	GWB 07 - 09	GWB 03 (08.02.95)	GWB 07 - 09
GWB 04 (14.04.93)	GWB 07- 09	GWB 05 (08.02.95)	GWB 07 - 09
GWB 04 (19.05.93)	GWB 01 - 03	GWB 06 (08.03.95)	GWB 07 - 09
GWB 06 (19.05.93)	GWB 01 - 03	GWB 01 (05.04.95)	GWB 07 - 09
GWB 08 (19.05.93)	GWB 01 - 03	GWB 03 (05.04.95)	GWB 07 - 09
GWB 06 (16.09.93)	GWB 01 - 03	GWB 06 (05.04.95)	GWB 07 - 09
GWB 06 (16.09.93)	GWB 01 - 03	GWB 09 (05.04.95)	GWB 04 - 06
GWB 08 (16.09.93)	GWB 01 - 03	GWB 04 (14.06.94)	GWB 07 - 09
GWB 09 (16.09.93)	GWB 01 - 03	GWB 04 (05.09.95)	GWB 07 - 09
GWB 01 (02.03.94)	GWB 07 - 09	GWB 09 (05.09.95)	GWB 04 - 06
GWB 06 (02.03.94)	GWB 01 - 03	GWB 01 (03.11.97)	GWB 04 - 06
GWB 06 (29.08.94)	GWB 01 - 03	GWB 06 (03.11.97)	GWB 07 - 09
GWB 01 (06.12.94)	GWB 04 - 06	GWB 07 (03.11.97)	GWB 01- 03
GWB 04 (06.12.94)	GWB 01 - 03	GWB 08 (03.11.97)	GWB 01- 03

Da lediglich g = 3 Gruppen vorliegen, werden nur t = 2 NED berechnet. Von diesen enthält das erste mit über 83 % den weitaus größeren Anteil am Gesamttrennmaß, s. Tab. 7-13. Die

Diskriminanzkoeffizienten verdeutlichen, dass diesmal Ca^{2+} (höchster Wert im ersten NED) einen überragenden Beitrag zur Trennung liefert, erneut Cl^- und SO_4^{2-} (in beiden NED) und diesmal auch Na^+ (höchster Wert im zweiten NED) diesbezüglich von Bedeutung sind.

Tab. 7-13. Diskriminanzkoeffizienten und Trennmaß der NED.

Variable	v_1	v_2
Na^+	0,2497	-1,2904
K^+	0,3891	0,1900
Mg^{2+}	0,1794	0,5603
Ca^{2+}	0,6876	0,2784
CL^-	0,5938	-0,5703
SO_4^{2-}	-0,2860	0,9089
λ_j	0,6000 ≙ 83,14 %	0,1217 ≙ 16,86 %

2. Untersuchung

- Bildung von sechs Lerndatensätzen, Zuordnung von jeweils einem Testdatensatz GWB 5
  Es wurde jeweils, da g = 2 Gruppen vorlagen, nur t = 1 NED berechnet. Die gemessenen Merkmale eignen sich hervorragend zur Trennung der Gruppen in den Lerndatensätzen (über ihre unterschiedliche Signifikanz hierbei wird am Ende des Abschnitts diskutiert), es wird lediglich in zwei Fällen jeweils ein Fehler ermittelt. Wie bei der KNN-Methode erfolgt bei allen Lerndatensätzen für das Testobjekt GWB 5 keine Zuordnung zur Gruppe der Sickerwassermessstellen, s. Tab. 7-14. In der eindimensionalen Darstellung wird aber zumindest deutlich, dass dieses Objekt fast immer deutlich vom „Kern" der anderen Grundwassermessstellen entfernt ist. Abb. 7-7 - hierin eingetragen sind die sich aus dem Quantil der Ordnung 0,95 einer F-Verteilung mit den entsprechenden Freiheitsgraden ($F_{0,95;\,1,\,19} = 4,38$) ergebenden beiden Streuradien - veranschaulicht dies (exemplarisch) für den Lern- bzw. Testdatensatz des Monats August '94.

Tab. 7-14. Klassifikationsergebnis (sechs Lerndatensätze, je ein Testdatensatz GWB 5).

Lern-, Testdatensatz	Fehlerrate des Lerndatensatzes	λ_1	Fehlzuordnungen des Lern- bzw. Testdatensatzes		
			eindeutig	mehrdeutig	Ausreißer
März '94	5,26 %	0,83	---	SWA	---
August '94	4,76 %	0,82	SWP 9	---	SWP 5
Februar '95	0,00 %	0,94	---	---	SWP 7
April '95	0,00 %	0,81	---	---	---
September '95	0,00 %	0,88	---	---	---
November '97	0,00 %	0,89	---	---	---

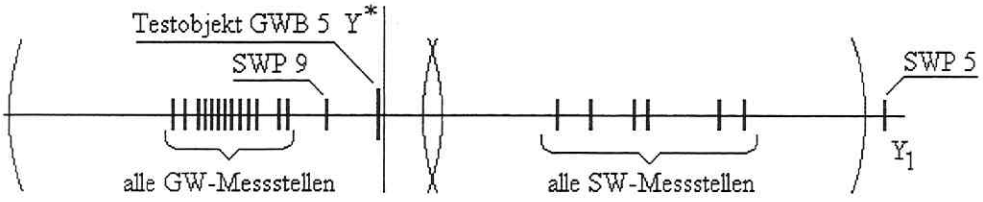

Abb. 7-7. Eindimensionale Darstellung (Lern-, Testdatensatz August '94).

Des Weiteren wird aus dieser Abbildung die nicht korrekte Zuordnung zweier im Lerndatensatz enthaltenen Objekte deutlich, die des SWP 5 (Ausreißer, der nicht in die Berechnung der Fehlerrate eingeht) und die des SWP 9 (Fehlzuordnung zur Gruppe der Grundwassermessstellen). Letztere ist insofern hervorhebenswert, da diese Probenahmestelle neben dem SWP 7 (dieser tritt im Datensatz des Monats Februar '95 als Ausreißer in Erscheinung) bezüglich der Salzfrachten gemeinhin relativ geringe Werte aufweist, s. Abb. 6-3 und 6-8, was sich hier in einem entsprechenden Klassifikationsergebnis widerspiegelt.

- Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen GWB 5

  Es wurden insgesamt sechs NED berechnet, von denen das erste mit 80,88 % einen hohen Anteil am multivariaten Trennmaß aller NED enthält, der des zweiten beträgt 5,21 %. Es ist auffällig, dass keine Fehlklassifikation erfolgt, im Gegenteil, der GWB 5 relativ oft genau der Gruppe zugeordnet wird, welcher er entstammt, s. Tab. 7-15. Letzteres wurde bei Anwendung der KNN-Methode ebenfalls festgestellt und führt zu der Vermutung, dass die Grundwassergruppen (nur diese, da die für den gesamten Lerndatensatz erzielte Fehlerrate relativ hoch ist) bezüglich der Merkmalswerte (Einzelionen) gut separiert werden.

Tab. 7-15. Klassifikationsergebnis (ein Lerndatensatz, sechs Testdatensätze GWB 5).

Testdatensatz GWB 5	Reihenfolge der Gruppenzuordnung des Testdatensatzes				
	1.	2.	3.	4.	5.
März '94	GW März '94	GW Nov. '97	GW Sept. '95	GW Aug. '94	GW April '95
August '94	GW Nov. '97	GW Aug. '94	GW April '95	GW Febr. '95	GW Sept. '95
Februar '95	GW Febr. '95	GW April '95	GW Nov. '97	GW März '94	GW Sept. '97
April '95	GW April '95	GW Nov. '97	GW Febr. '95	GW März '94	GW Sept. '95
September '95	GW Nov. '97	GW April '95	GW Febr. '95	GW Sept. '95	GW März '94
November '97	GW Nov. '97	GW April '95	GW Febr. '95	GW März '94	GW Sept. '95

- Bildung von einem Lerndatensatz, Zuordnung von sechs Testdatensätzen SWP 7

  Die ersten beiden NED enthalten 88,31 % bzw. 5,62 % des Gesamttrennmaßes. Für den SWP 7 wird hier zumindest einmal die Zuordnung zu einer der Grundwassergruppen er-

mittelt, s. Tab. 7-16, was ein Indiz dafür ist, dass sich seine relativ geringe Belastung offensichtlich in seinem durch die entsprechenden Merkmale induzierten Muster widerspiegelt. Dieser Nachweis konnte durch Anwendung der KNN-Methode nicht erfolgen.

Tab. 7-16. Klassifikationsergebnis (ein Lerndatensatz, sechs Testdatensätze SWP 7).

Testdatensatz	Reihenfolge der Gruppenzuordnung des Testdatensatzes				
GWB 5	1.	2.	3.	4.	5.
März '94	**GW Febr. '95**	**GW April '95**	**GW Nov. '97**	SW März '94	**GW Sept. '95**
August '94	SW Febr. '95	SW Aug. '94	**GW Febr. '95**	**GW April '95**	**GW Nov. '97**
Februar '95	SW Febr. '95	**GW Febr. '95**	**GW April '95**	**GW Nov. '97**	**GW März '94**
April '95	SW Febr. '95	**GW Febr. '95**	SW April '95	**GW April '95**	**GW Nov. '97**
September '95	SW Febr. '95	SW Sept. '95	**GW Febr. '95**	SW März '94	**GW April '95**
November '97	SW Febr. '95	SW April '95	SW Sept. '95	SW März '94	SW Aug. '94

Tab. 7-17 kann zusammenfassend entnommen werden, welche der Merkmale für die Trennung der Gruppen signifikant sind, wobei als Kriterium für die Reihenfolge die Beträge der in Klammern angegebenen Werte der Diskriminanzkoeffizienten des ersten NED angesehen wurden. Neben den diesbezüglich bereits untersuchten Lerndatensätzen der Sicker- und Grundwassermessstellen (s. Tab. 7-10 und 7-13) sind hierin auch die Ergebnisse für die kleineren Lerndatensätze der sechs Monate enthalten, die aufgrund der Unterteilung in die beiden Gruppen Sicker- und Grundwasser von besonderem Interesse sind. Die SO_4^{2-}- und Cl^--Ionen leisten hier einen überragenden Beitrag zur Trennung, so dass diese beiden analytischen Parameter als optimal zur Beschreibung des Sickerwassertransports angesehen werden können.

Tab. 7-17. Signifikanz der Merkmale für die Trennung der Gruppen.

Datensatz	Reihenfolge					
	1.	2.	3.	4.	5.	6.
SW-Messstellen	Mg^{2+} (-3,59)	SO_4^{2-} (-2,39)	Cl^- (-0,92)	Ca^{2+} (0,47)	Na^+ (-0,26)	K^+ (0,18)
GW-Messstellen	Ca^{2+} (0,69)	Cl^- (0,59)	K^+ (0,39)	SO_4^{2-} (-0,29)	Na^+ (0,25)	Mg^{2+} (0,18)
März '94	Cl^- (-2,64)	K^+ (-1,89)	Mg^{2+} (0,94)	SO_4^{2-} (0,92)	Ca^{2+} (-0,74)	Na^+ (0,63)
August '94	SO_4^{2-} (-3,39)	Na^+ (2,32)	Mg^{2+} (2,01)	Cl^- (1,83)	K^+ (-1,38)	Ca^{2+} (0,74)
Februar '95	Cl^- (-4,70)	Na^+ (4,40)	K^+ (-2,99)	Ca^{2+} (-0,94)	Mg^{2+} (0,56)	SO_4^{2-} (0,16)
April '95	SO_4^{2-} (-2,13)	K^+ (1,89)	Ca^{2+} (1,12)	Mg^{2+} (0,99)	Na^+ (0,56)	Cl^- (0,04)
September '95	SO_4^{2-}(-11,56)	Na^+ (10,00)	Cl^- (-5,27)	K^+ (4,85)	Mg^{2+} (3,29)	Ca^{2+} (2,23)
November '97	SO_4^{2-} (5,88)	Na^+ (-3,60)	K^+ (-2,75)	Mg^{2+} (-2,04)	Ca^{2+} (-1,78)	Cl^- (0,89)

Abschließend sei darauf hingewiesen, dass ein Vergleich der Klassifikationsleistung verschiedener Verfahren der überwachten Klassifikation in Abschnitt 12.2 vorgenommen wird.

8 Automatische Klassifikation (Clusteranalyse)

8.1 Vorbemerkungen

Ein allgemeines Problem jeder Datenanalyse besteht darin, aufgenommene Messreihen, hier allgemein als Objekte bezeichnet, in Strukturen derart zu überführen, dass Grundzusammenhänge zwischen diesen erkennbar sind. Die Menge von (endlich vielen) Objekten soll dabei systematisiert und eine vorhandene Ordnung aufgedeckt werden. In solchen Fällen werden Clusteranalysen sinnvoll angewandt. Mit diesen wird versucht, die Gesamtmenge in kleinere, möglichst homogene Gruppen (Cluster) zu zerlegen. In ein und derselben Gruppe befindliche Objekte sollen möglichst ähnlich, aus verschiedenen Gruppen stammende dagegen wenig ähnlich sein. Das Auffinden der günstigsten Zerlegung wird objektiviert und automatisiert, daher wird für Clusteranalyse gemeinhin auch die Bezeichnung automatische Klassifikation verwendet (HENRION et al., 1988).

Bei der Automatisierung der Suche nach homogenen Teilgruppen innerhalb der größeren Menge von Objekten müssen die Begriffe Ähnlichkeit bzw. Verwandtschaft quantifiziert werden. Setzen sich die Muster von Objekten aus Werten verschiedenster z. B. metrisch skalierter Variablen X_1, ... , X_p zusammen (Punktkoordinaten), so entspricht jedem Objekt ein Punkt. Die Punkte ähnlicher Objekte liegen dicht beisammen. Die Berechnung eines „Objektabstandes", der für Punkte im p-dimensionalen Raum definiert sein muss, ergibt ein indirektes Maß für die Ähnlichkeit zweier Objekte (große Ähnlichkeit entspricht geringem Abstand). Die Merkmale bzw. Variablen, welche ein Objekt „charakterisieren" bzw. „beschreiben", können sich prinzipiell auf verschiedenen Skalenniveaus befinden. Deren Unterscheidung ist notwendig, da von ihnen die jeweils passende Definition des Abstandes zweier Objekte abhängt. Man unterteilt i. Allg. zwischen nominal, ordinal und metrisch skalierten Merkmalen. Diese sind aufsteigend angeordnet, d. h., metrisch skalierte Merkmale besitzen den höchsten Informationsgehalt. Eine Niveauregression ist prinzipiell möglich (mit Informationsverlust), eine Niveauprogression nur, wenn Zusatzinformationen vorhanden sind (BOCKLISCH, 1987; HENRION et al., 1988).

Bei den im Rahmen vorliegender Arbeit durchgeführten Untersuchungen zur Clusteranalyse wurden zur Objektcharakterisierung ausschließlich metrische Merkmale benutzt. Bei diesen kann man das Objektmuster (Beobachtungsergebnisse für p festgelegte Merkmale im i-ten

Objekt x_{i1}, x_{i2}, ... , x_{ip}) als Koordinaten eines Objektpunktes im p-dimensionalen EUKLIDischen Raum $\Re^p$ interpretieren. Zu den gebräuchlichsten Abstandsfunktionen gehören:

- EUKLIDischer Abstand $$d_{AB} = \left[\sum_{j=1}^{p} \left(x_{Aj} - x_{Bj} \right)^2 \right]^{\frac{1}{2}}$$
- Quadratischer EUKLIDischer Abstand $$d_{AB} = \sum_{j=1}^{p} \left(x_{Aj} - x_{Bj} \right)^2$$
- Manhattan- oder City-Block-Abstand $$d_{AB} = \sum_{j=1}^{p} \left| x_{Aj} - x_{Bj} \right|$$
- MINKOWSKI- oder L_r-Metrik $$d_{AB} = \left[\sum_{j=1}^{p} \left| x_{Aj} - x_{Bj} \right|^q \right]^{\frac{1}{r}}$$

Die MINKOWSKI-Metrik mit den benutzerdefinierten Parametern q und r ist der allgemeine Fall, für den beim EUKLIDischen Abstand q = r = 2 und beim Manhattan-Abstand q = r = 1 ist. BOCKLISCH (1987) visualisiert die Nachbarschaftsbeziehungen für verschiedene Werte von (gleichgroßen) g und r der MINKOWSKI-Distanz im zweidimensionalen Merkmalsraum. Des Weiteren hat er den Einfluss der MINKOWSKI-Metrik (für gleichgroße q und r) und der Dimension p des Merkmalsraumes auf den Abstand eines Objektes vom Ursprung untersucht.
Die prinzipielle Gemeinsamkeit aller genannten Abstandsmaße besteht darin, dass deren Werte umso kleiner werden, je geringer die Differenzen zwischen den Objektpunktkoordinaten sind, d. h. je ähnlicher die Muster zweier Objekte sind. Die Anwendung bzw. der Wechsel des Abstandsmaßes sollte immer sachlogisch begründet sein. Der Manhattan-Abstand z. B. lässt sich dann vorteilhaft anwenden, wenn der „Übergang" zwischen zwei Objektpunkten durch sukzessive Änderung der einzelnen Parameter erfolgt. Für die durchgeführten Untersuchungen zur Clusteranalyse wurde der EUKLIDische Abstand angewandt. Dieser klassische „Luftlinien"-Abstand eignet sich gemeinhin am besten, da er eine gleichmäßige simultane Regelbarkeit (Veränderlichkeit) aller Parameter voraussetzt (HENRION und HENRION, 1994).
Abschließend sei erwähnt, dass für die Abstandsberechnung der Objekte die Messwertfolgen bei metrisch skalierten Merkmalen standardisiert vorliegen müssen, damit Werte verschiedener Einheiten bei der Berechnung des Abstandsmaßes auch vergleichbare Größen darstellen bzw. die Achsenskalierung des Abstandsmaßes im Dendrogramm (s. Abschnitt 8.2) vereinheitlicht ist (HENRION et al., 1988). Darüber hinaus weisen KNOBLOCH und ZWANZIGER (1995) darauf hin, dass das Nichtstandardisieren eigentlich ein „Kunstfehler" ist, welcher jedoch bei Vorliegen dimensionsgleicher Messgrößen bei der Clusteranalyse als wirksame Methode der „Ausreißererkennung" eingesetzt werden kann.

8.2 Hierarchisch agglomerative Methoden

8.2.1 Methodik allgemein

Das Ziel hierarchischer Verfahren besteht darin, die Objekte unter Verwendung des Abstandsmaßes miteinander so zu verbinden (aufzuteilen), dass aufeinanderfolgend größere (kleinere) Gruppen entstehen. Das Ergebnis hierarchischer Clusteranalyseverfahren ist ein Dendrogramm (= „hierarchischer Baum"). Entsprechend der Vorgehensweise unterscheidet man agglomerative Methoden, bei denen, beginnend mit den für sich isoliert stehenden Objekten, durch schrittweises Verschmelzen eine Agglomeration (= Anhäufung, Zusammenballung) zu einer Gesamtmenge vorgenommen wird und (in den Untersuchungen nicht zur Anwendung gekommene und daher nachfolgend nicht näher erläuterte) divisive Methoden, bei denen andersherum, von der Gesamtmenge ausgehend, eine systematische Aufspaltung in Splittergruppen erfolgt (HENRION et al., 1988).

Durch das Dendrogramm wird die Hierarchie „erklärt". Als Beispiel werden die Beobachtungsergebnisse von 17 Objekten (SWP 1, 4, 5, 7, 9 - 11; GWB 1 - 10) für vier Merkmale (Messwerte der Na^{+}-, K^{+}-, Mg^{2+}- und Ca^{2+}-Ionen des Monats April '98) gewählt. Betrachtet wird die horizontale Darstellungsform des Dendrogramms, welches sich aus der Clusteranalyse der Objekte (Messstellen) ergibt, s. Abb. 8-1.

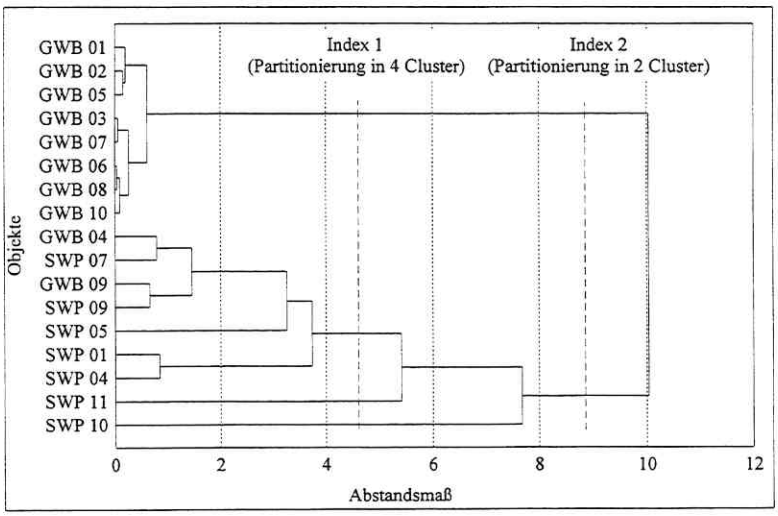

Abb. 8-1. Dendrogramm. - WARDs Methode (EUKLIDischer Abstand).

Die Abszisse stellt das Abstandsmaß dar, an der Ordinate sind die Objekte angetragen, welche, jedes noch eine Klasse für sich bildend, miteinander verglichen werden sollen. In kleinen aufeinanderfolgenden Schritten werden nun die Objekte entsprechend ihrer Zusammengehörigkeit zu größeren Gruppen miteinander verbunden. Das Abstandsmaß wird dabei zunehmend erhöht. Durch das Verbinden der Objekte entstehen immer größere Gruppen von zunehmend verschiedenen Elementen. Im letzten Schritt werden alle Objekte zusammengefügt. Für jeden Knoten am Graphen, d. h., an der Stelle, wo eine neue Gruppe gebildet wird, kann man an der horizontalen Achse das Abstandsmaß ablesen, mit welchem die betrachteten Elemente zu einer neuen einzelnen Gruppe verbunden werden. Bei Variation der Indexwerte lässt sich die Objektmenge in unterschiedliche Clusteranzahlen partitionieren. Dabei handelt es sich um ein Cluster im eigentlichen Sinn, wenn eine Objektgruppierung sich unter einem langen Zweig des Dendrogramms befindet, d. h. größeren Indexänderungen gegenüber stabil ist (HENRION et al., 1988; HENRION und HENRION, 1994).

Der allgemeine Agglomerationsalgorithmus lässt sich wie folgt angeben:

1. Berechnung des kleinsten vorkommenden Abstandes d_{AB}
2. Zusammenfassung der Objekte A und B zu einem (neuen) Objekt AB
3. Berechnung der Abstände aller verbliebenen alten Objekte zum neuen Objekt
4. zurück zu 1., falls noch nicht (n -1) Zusammenfassungen erfolgt sind, d. h. noch nicht alle Objekte zu einer Gesamtmenge agglomeriert sind (HENRION et al., 1988)

Die nachfolgend genannten verschiedenen Agglomerationsverfahren unterscheiden sich in der Art der Abstandsneuberechnung (3. Schritt):

- Single Linkage und Complete Linkage
- Average Linkage (ungewichtet) und Average Linkage (gewichtet)
- Median und Centroid Linkage
- WARDs Methode (HENRION und HENRION, 1994)

Ein Wechsel des Agglomerationsverfahrens führt zu einer Nuancierung des Ergebnisses einer Clusteranalyse, die sich im Dendrogramm widerspiegelt. Bei Single Linkage werden zwei Cluster als unrealistisch nah beieinander angesehen, bei Complete Linkage als zu weit voneinander entfernt. Single Linkage ist raumkontrahierend, d. h., es führt zu sog. Ketteneffekten, weil sich noch in keinem Cluster befindende Einzelobjekte eher dazu neigen, mit einem bereits gebildeten Cluster zu fusionieren, als selbst den Kern eines neuen Clusters zu bilden. Das Dendrogramm nimmt eine charakteristische Treppenform an. Complete Linkage hingegen ist

raumdilatierend, weil einzelne Objekte leichter Zentren für neue kleine Cluster bilden, als dass sie mit vorhandenen großen Clustern verschmelzen, es kommt zu charakteristischen Inselbildungen im Dendrogramm. Die Verfahren Average Linkage (gewichtet und ungewichtet), Median, Centroid Linkage und WARDs Methode werden als konservativ bezeichnet, da die Effekte Raumkontraktion bzw. Raumdilation, welche bei Single Linkage bzw. Complete Linkage auftreten, vermieden werden (HENRION et al., 1988).

Sämtliche der genannten Verfahren ergeben sich nach LANCE und WILLIAMS (1966) als Spezialfälle bei Verwendung der allgemeinen Rekursionsformel

$$d_{A,BC} = \alpha_B \cdot d_{AB} + \alpha_C \cdot d_{AC} + \beta \cdot d_{BC} + \gamma \cdot |d_{AB} - d_{AC}|$$

mit den Parametern α_B, α_C, β und γ. Unter den Restriktionen $\alpha_B + \alpha_C + \beta = 1$, $\alpha_B = \alpha_C$, $\beta < 1$ und $\gamma = 0$ lässt sich hieraus eine Untermenge extrahieren:

$$d_{A,BC} = \alpha_B \cdot (d_{AB} + d_{AC}) + (1 - 2 \cdot \alpha) \cdot d_{BC}.$$

Bei Variation von α ($\alpha > 0$) ist eine **flexible Strategie** möglich, man kann von kontrahierenden ($\alpha < 0{,}5$) über konservative (für $\alpha = 0{,}5$ ergibt sich exakt Average Linkage (ungewichtet)) bis hin zu dilatierenden Verfahren ($\alpha \rightarrow 1$) kontinuierlich „durchregeln", wobei $\alpha \approx 0{,}6$ - $0{,}7$ empfohlen wird (STEINHAUSEN und LANGER, 1977).

Die Leistungsfähigkeit hierarchisch agglomerativer Verfahren wird in der Literatur ausführlich vergleichend diskutiert. Allein HENRION et al. (1988) geben zu diesem Thema sieben Literaturhinweise, wobei die jeweiligen Autoren teilweise zu widersprüchlichen Bewertungen kommen. Bei pragmatischem Herangehen, d. h., man wendet verschiedene Methoden auf ein und denselben Datensatz an und bewertet sie nach der Interpretierbarkeit der Ergebnisse bzw. nach der Reproduktion von a priori bekannten Gruppen, kommt EVERITT (1979) zu dem Resultat, dass sich Single Linkage als durchgehend schlecht erweist, Average Linkage und WARDs Methode hingegen schneiden sehr gut ab. WARDs Methode ist ein in der Praxis relativ oft genutztes Agglomerationsverfahren (DAUS, 1995; EINAX et al., 1990). Es führt in der Regel zu sehr gut strukturierten Dendrogrammen mit homogenen Clustern. HENRION et al. (1988, S. 71) gelangen nach umfangreichen Untersuchungen zu der Einschätzung, dass WARDs Methode „ ... eines der besten Agglomerationsverfahren ist ... ", allerdings mit der Einschränkung, dass die Bewertung der Verfahren „ ... mehr oder weniger subjektiv ist ... " und vom konkreten Anwendungsfall abhängt. Sie empfehlen daher in jedem Fall zumindest einen Vergleich der Ergebnisse verschiedener Agglomerationsverfahren, um gewisse Unsicherheiten zu erkennen bzw. Fehlinterpretationen zu vermeiden.

8.2.2 Untersuchungen

8.2.2.1 Gegenstand und Zielstellung der Untersuchungen

Die Untersuchungen unterteilten sich in zwei Abschnitte. Den Resultaten sämtlicher Clusteranalysen liegt das EUKLIDische Abstandsmaß (standardisierte Merkmalswerte) zugrunde.

1. Untersuchung: Ermittlung des optimalen Agglomerationsverfahrens

Aus den Ausführungen des vorhergehenden Abschnitts wurde bereits deutlich, dass die Anwendung konservativer Verfahren i. Allg. vorzuziehen ist. Um zu überprüfen, ob diese auch für den hier vorliegenden konkreten Anwendungsfall (Clusteranalysen der Messstellen in Abhängigkeit von ausgewählten Ionenpaaren bzw. Ionenverhältnissen, s. zweiter Teil der Untersuchungen) geeignet waren, wurde zunächst eine Agglomeration der Objekte (Messstellen) nach der flexiblen Strategie (LANCE und WILLIAMS, 1966) vorgenommen, d. h., für die Werte (Muster) jedes der in Tab. 8-1 aufgeführten Ionenpaare bzw. Ionenverhältnisse erfolgte die „Durchregelung" von kontrahierenden ($\alpha = 0{,}3$) über konservative ($\alpha = 0{,}5$) bis hin zu dilatierenden ($\alpha = 0{,}7$) Verfahren (insgesamt 18 Clusteranalysen).

2. Untersuchung: Ermittlung der Ähnlichkeiten zwischen den Probenahmestellen

Es wurden 12 Clusteranalysen für jeweils 19 Objekte durchgeführt, s. Tab 8-1. Die 19 Objekte sind die Messstellen (GWB 1 - 10, Nahle, Luppe; SWP 1, 4, 5, 7, 9 - 11), die Variablen sind ausgewählte Ionenpaare bzw. Ionenverhältnisse (Begründung s. u.), wobei die maximale Anzahl der zur Verfügung stehenden Werte des gesamten Untersuchungszeitraumes (Mai '92 bis April '98) genutzt wurde, um die Eindeutigkeit der induzierten Merkmalsmuster abzusichern.

Wie bereits mehrfach erwähnt, sind die Salzfrachten des Sickerwassers die wesentlichen Schadstoffausträge in das Grundwasser, und eine Vorbestimmung des Sickerwassertransfers ist nicht möglich. Bei Messungen von Salzfrachtausträgen aus dem Deponiekörper als punktuelle Schadstoffquellen in den umgebenden Aquifer ist eine einfache Zuordnung der Ausbreitungsrichtung zu den Salzfrachtbestandteilen aufgrund der analytischen Ununterscheidbarkeit im An- und Abstrom ebenfalls nicht machbar. Induzieren die ausgewählten Ionenpaare der Sicker- und Grundwassermessstellen jedoch charakteristische Muster, so kann durch entsprechende Clusteranalysen versucht werden, über deren Ähnlichkeiten (Verteilungsstrukturen) einen lokalen Zusammenhang beim Ausbreitungsverhalten der Salzfrachten des Sickerwassers im Grundwasser zu finden. Das Ziel dieses Teils der Untersuchungen bestand somit darin,

bevorzugte Austrittsrichtungen des Sickerwassers in die aquatische Umgebung festzustellen. Die Unterschiede in den Löslichkeitsverhältnissen der Einzelionen führen jedoch dazu, dass diese nicht als gleichwertige Indikatoren für Fließrichtungen angesehen werden können, die Ergebnisse der Diskriminanzanalyse (s. Abschnitt 7.3.5.3) haben dies gezeigt. Aus diesem Grund sowie um differenzierte Aussagen über mögliche Transportrichtungen treffen zu können, wurde eine Unterscheidung der Muster nach ausgewählten Ionenpaaren vorgenommen. Darüber hinaus wurden als Variablen die entsprechenden Ionenverhältnisse einbezogen. Diese sind ein Maß für die Abweichung vom stöchiometrischen Verhältnis der beiden Einzelionen, und die durch sie induzierten Merkmalsmuster können ebenfalls als Hinweis auf einen anthropogenen Schadstoffaustrag in den Aquifer angesehen werden (LESCHBER et al., 1993).

Tab. 8-1. Clusteranalyse der Messstellen. - Gegenstand der Untersuchungen.

Nr.	Variablen	Anzahl Werte	Agglomeration	Nr.	Variablen	Anzahl Werte	Agglomeration
1	Na^+, K^+	16	WARDs Methode	7	Na^+/K^+	8	WARDs Methode
2	Na^+, K^+	16	Average Linkage	8	Na^+/K^+	8	Average Linkage
3	Mg^{2+}, Ca^{2+}	16	WARDs Methode	9	Mg^{2+}/Ca^{2+}	8	WARDs Methode
4	Mg^{2+}, Ca^{2+}	16	Average Linkage	10	Mg^{2+}/Ca^{2+}	8	Average Linkage
5	Cl^-, SO_4^{2-}	7	WARDs Methode	11	Cl^-/SO_4^{2-}	3	WARDs Methode
6	Cl^-, SO_4^{2-}	7	Average Linkage	12	Cl^-/SO_4^{2-}	3	Average Linkage

8.2.2.2 Untersuchungsergebnisse

Bezüglich des zweiten Teils der Untersuchungen wurde sich auf eine repräsentative Auswahl von Dendrogrammen beschränkt. Im Einzelnen sind im nachfolgenden Abschnitt die folgenden Untersuchungsergebnisse enthalten:

- Dendrogramme (Ionenpaare), flexible Strategie: Zusammenfassung in Abb. 8-2
- Dendrogramme (Ionenverhältnisse), flexible Strategie: Zusammenfassung in Abb. 8-3
- Dendrogramm (Variablen: Cl^-- und SO_4^{2-}-Ionen), WARDs Methode: s. Abb. 8-4
- Dendrogramm (Variablen: Mg^{2+}- und Ca^{2+}-Ionen), WARDs Methode: s. Abb. 8-5
- Dendrogramm (Variablen: Ionenverhältnisse Na^+/K^+), WARDs Methode: s. Abb. 8-6
- Dendrogramm (Variablen: Ionenverhältnisse Cl^-/SO_4^{2-}): s. Abb. 8-7

Die sich bei variierenden Indexwerten (Aufteilung in zehn, fünf und drei Gruppen) in den Dendrogrammen (12 Clusteranalysen der 2. Untersuchung) ergebenden Objektpartitionierungen sind in den Tab. A-13 (Ionenpaare) und A-14 (Ionenverhältnisse) des Anhangs zusammengefasst, wobei die fett hervorgehobenen Cluster stabil gegenüber zweifacher Indexänderung und die unterstrichen dargestellten stabil gegenüber einfacher Indexänderung sind.

8.2.2.3 Diskussion der Untersuchungsergebnisse

1. Untersuchung

Die Effekte kontrahierender, konservativer und dilatierender Agglomerationsverfahren spiegeln sich in den Dendrogrammen, s. Abb. 8-2 (Ionenpaare) und 8-3 (Ionenverhältnisse), gut wider (Mit „S" bzw. „G" werden in diesen die entsprechenden SWP bzw. GWB bezeichnet.).

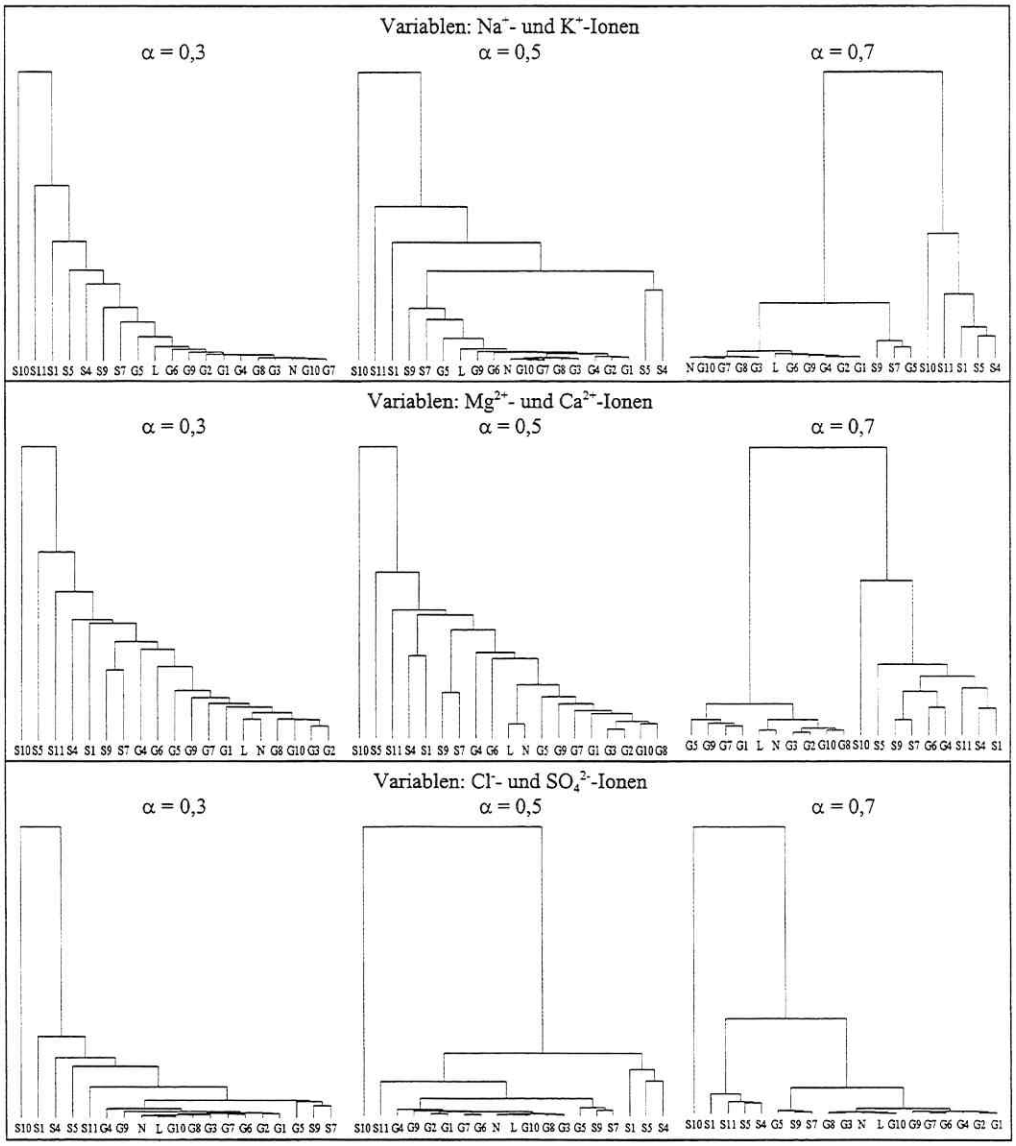

Abb. 8-2. Agglomeration nach der flexiblen Strategie (Variablen: Ionenpaare).

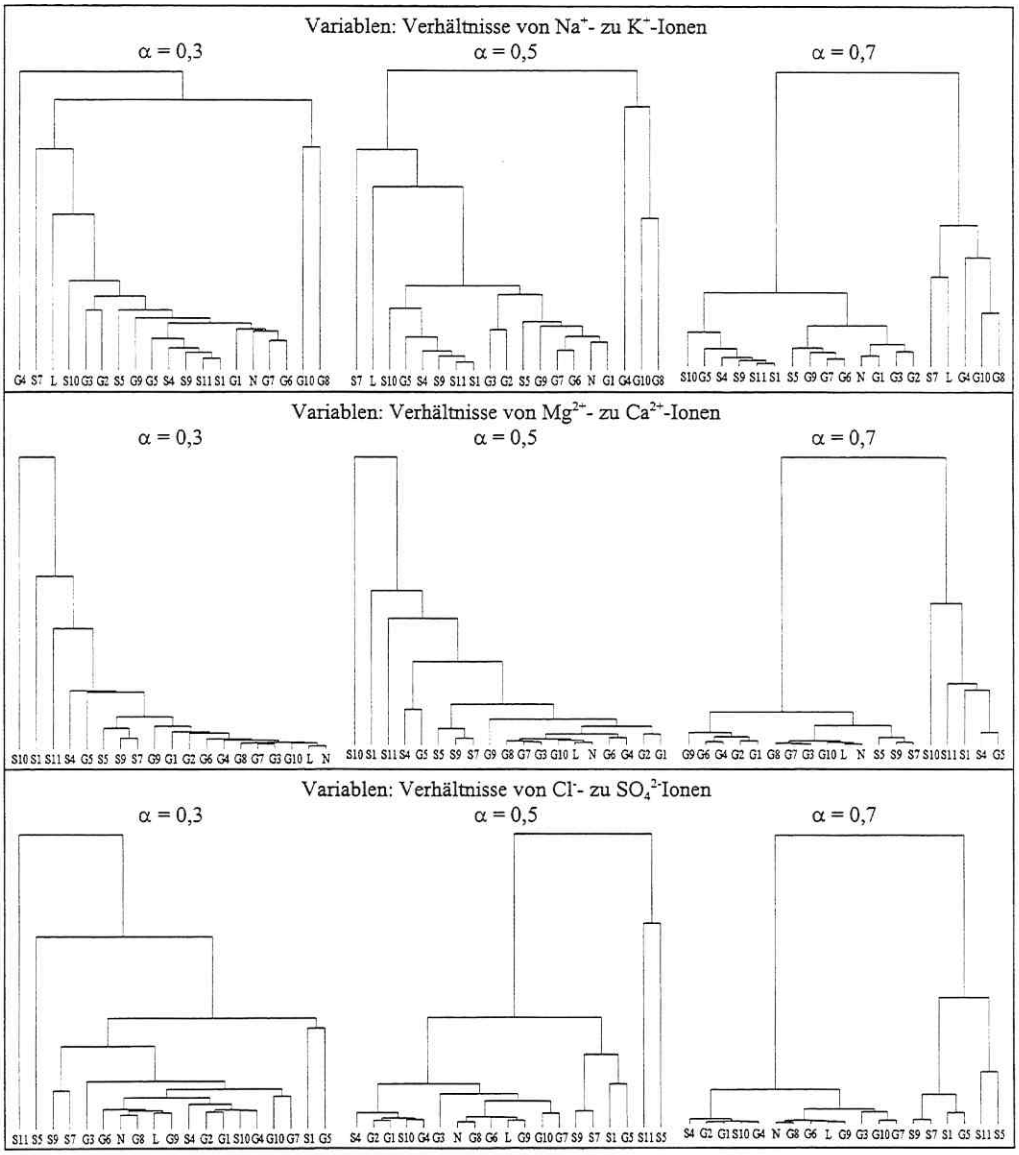

Abb. 8-3. Agglomeration nach der flexiblen Strategie (Variablen: Ionenverhältnisse).

Insbesondere bei den Ionenpaaren zeigen sich bei α = 0,3 die für kontrahierende Methoden typischen Ketteneffekte und bei α = 0,7 die für dilatierende Methoden charakteristischen Inselbildungen. Bei α = 0,5 wird hier ein besser abgestuftes (konservatives) Verhalten deutlich. Damit werden die von EVERITT (1979) sowie HENRION et al. (1988) getroffenen Empfehlungen bestätigt: Der Einsatz konservativer Verfahren wie WARDs Methode und Average Linkage war auch für diesen konkreten Anwendungsfall als am sinnvollsten anzusehen. Zum einen werden hier a priori-Strukturen korrekt wiedererkannt, z. B. bei den Mg^{2+}- und Ca^{2+}-Io-

nen die Ähnlichkeiten zwischen den GWB 2 und 3 (Abstrombereich der Luppe), den GWB 4 und 6 (Abstrombereich der Nahle) sowie den GWB 7 und 9 (beide Anstrombereich der Deponie). Zum anderen werden die hohen Belastungsunterschiede zwischen den Gruppen der Sicker- und Grundwassermessstellen hier am vorteilhaftesten nivelliert, so dass auch auf geringem Abstandsniveau bereits Ähnlichkeiten zwischen einzelnen „Vertretern" beider Gruppen erkennbar sind. Der Einsatz konservativer Verfahren wie die für den zweiten Teil der Untersuchungen verwendeten WARDs Methode und Average Linkage (hier gewichtet, um die unterschiedliche Anzahl von Sicker- und Grundwassermessstellen in die Abstandsberechnung einzubeziehen) ist damit auch für ähnliche praktische Anwendungsfälle zu empfehlen.

2. Untersuchung

Ein Wechsel des Agglomerationsverfahrens (WARDs Methode/Average Linkage (gewichtet)) führte zu einigen geringfügigen Veränderungen in den Dendrogrammen, weshalb sich die bei variierenden Indexwerten ergebenden Objektpartitionierungen teilweise unterscheiden, s. Tab. A-13 und A-14. WARDs Methode ist dabei als das leistungsfähigere Verfahren anzusehen (Diskussion s. Abschnitt 12.2). Die Unterschiede sind insgesamt gesehen jedoch relativ gering, die nachfolgend diskutierten Ähnlichkeiten zwischen bestimmten Objekten bzw. Objektgruppen wurden mit beiden Verfahren ermittelt.

- Clusteranalysen in Abhängigkeit von den Ionenpaaren (1. bis 6. Analyse)

  Die Ergebnisse der Clusteranalysen weisen generell eine jeweils dichte Gruppierung der Sicker- sowie der Grundwassermessstellen auf. Abb. 8-4 zeigt dies exemplarisch für die Clusterung in Abhängigkeit von den Cl^-- und SO_4^{2-}-Ionen.

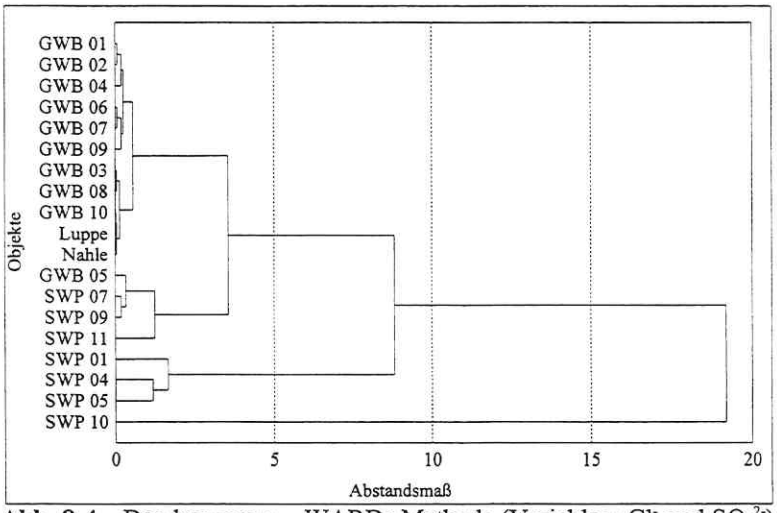

Abb. 8-4. Dendrogramm. - WARDs Methode (Variablen: Cl^- und SO_4^{2-}).

Die wesentlich höheren Salzfrachten im Sickerwasser gegenüber denen im Grundwasser führen zu dieser Differenziertheit bei der Clusterung. Die Grundwassermessstellen gruppieren sich bereits auf sehr niedrigem Abstandsniveau, die Sickerwassermessstellen fusionieren erst bei hohen Indexwerten, wobei durch die Standardisierung der Ausgangsdaten und die Anwendung konservativer Verfahren hier bereits eine Nivellierung erreicht wurde. Die Clusterung in Abhängigkeit von den Mg^{2+}- und Ca^{2+}-Ionen bildet hierbei eine Ausnahme, s. Abb. 8-5. Die SWP sind den GWB „ähnlicher", die entsprechenden Inselbildungen in den Dendrogrammen erfolgen auf etwa gleichem Niveau. Als Ursache für diese Ähnlichkeit kann die Schwerlöslichkeit des Calciumsulfats ($CaSO_4$) im Deponiekörper gesehen werden.

Der SWP 10 ist aufgrund seiner hohen Belastung von den anderen Objekten weit entfernt. Er zeigt sich bei allen sechs Clusteranalysen stabil gegenüber zweifacher Indexänderung, s. Tab. A-13. In den Graphiken wird er als Ausreißer deutlich, da die Kurve über diesem Objekt jeweils weit ausschlägt. Diese Aussage trifft auch auf den hoch belasteten GWB 5 in Bezug auf die Grundwassermessstellen zu, es wird bei fast allen Clusteranalysen ein signifikanter Abstand zu den anderen „Grundwasserobjekten" deutlich. Die Ergebnisse anderer Untersuchungen (z. B. die der Varianzanalyse, s. Abb. 6-3 und 6-4), welche zeigen, dass die Salzfrachten im Sickerwasser bei SWP 10 und im Grundwasser bei GWB 5 am höchsten sind, spiegeln sich somit gut wider.

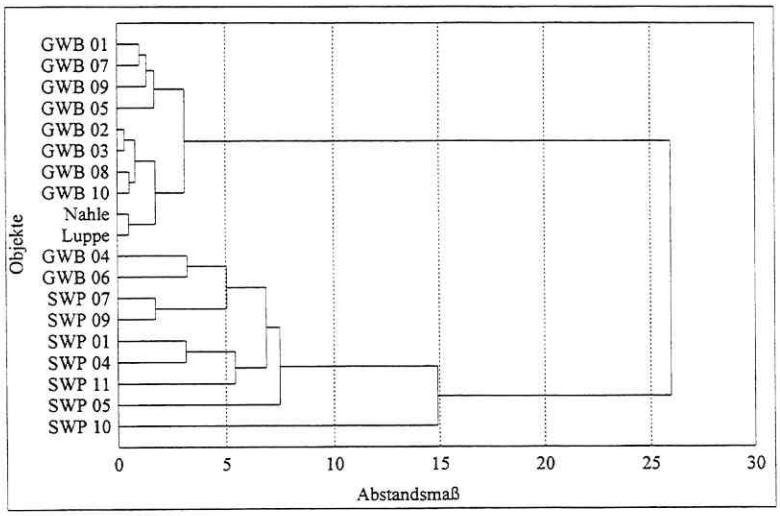

Abb. 8-5. Dendrogramm. - WARDs Methode (Variablen: Mg^{2+} und Ca^{2+}).

Die Clusteranalysen in Abhängigkeit von den Na^+- und K^+-Ionen sowie von den Cl^-- und SO_4^{2-}-Ionen (s. Abb. 8-4) lassen weiterhin eine Ähnlichkeit zwischen GWB 5 und dem Cluster {SWP 7; SWP 9} erkennen. Besonders bei letztgenanntem Ionenpaar ist diese Ähnlichkeit in einem solchen Maße signifikant, dass sie, wenn man von einer Grobpartitionierung der Objektmenge in Sicker- und Grundwassermessstellen ausgeht, sogar zu einer Zuordnung des GWB 5 zum Cluster der Sickerwassermessstellen führt. Der Aussage einer entsprechenden differenzierten Wasserwegsamkeit würden auch aus hydrogeologischer Sicht keine Einwände entgegenstehen, da eine Grundwasserfließrichtung von Südost nach Nordwest (d. h. von SWP 9 in Richtung SWP 7) besteht und GWB 5 in unmittelbarer Nähe von SWP 7 im nordwestlichen Teil des Deponiegeländes liegt.

- Clusterung in Abhängigkeit von den Ionenverhältnissen (7. bis 12. Analyse)

  Diese Clusteranalysen zeigen - unabhängig vom Agglomerationsverfahren - folgendes Ergebnis: Die a priori gegebene Einteilung der Objekte in Sicker- und Grundwassermessstellen lässt sich nicht automatisch aus den gemessenen Parametern wiederfinden. Es tritt ein sog. Effekt unvollständiger Korrespondenz (HENRION und HENRION, 1994) ein: Rechnerisch erzeugte Cluster vereinigen mehrere Strukturklassen in sich oder umgekehrt wird eine Klasse in mehrere Cluster aufgespalten. Die Anwendung beider Verfahren auf die (dimensionslosen) Ionenverhältnisse führt offensichtlich zu einer Abnahme von deren Leistungsfähigkeit, weshalb die Untersuchungsergebnisse differenziert zu betrachten sind.

  Die Clusteranalysen in Abhängigkeit der Ionenverhältnisse Na^+/K^+ (s. Abb. 8-6) sowie Mg^{2+}/Ca^{2+} offenbaren eine Ähnlichkeit zwischen GWB 5 und SWP 4 bzw. SWP 10 und die Ergebnisse anderer Untersuchungen zeigen, dass die Salzfrachten des Sickerwassers bei SWP 10 und die des Grundwassers bei GWB 5 (Abstrombereich der Nahle) außerordentlich hoch sind. Ein entsprechendes (theoretisches) Transportmodell, wonach ein Salzfrachtenaustrag von SWP 10 in Richtung Nahle zu GWB 5 erfolgt, wäre somit ebenfalls möglich, zumal auch dieses in Übereinstimmung zur Grundwasserfließrichtung steht.

  Interessant ist in diesem Zusammenhang eine generelle Betrachtung der Grundwassermessstellen einerseits und des hochbelasteten SWP 10 andererseits, um bestimmte Ähnlichkeiten zu erkennen, welche auf örtlich signifikante Grundwasserbelastungen durch diesen Bereich des Deponiekörpers hindeuten können. Im Unterschied zur Anwendung auf die Ionenpaare zeigt sich hier nämlich in lediglich einem Fall (Mg^{2+}/Ca^{2+}) die Stabilität dieses Objekts gegenüber zweifacher Indexänderung, s. Tab. A-14. Die dem SWP 10

„nächstgelegenen" Grundwassermessstellen sind GWB 5 (Na^+/K^+, s. Abb. 8-6, und Mg^{2+}/Ca^{2+}) bzw. GWB 4 (Cl^-/SO_4^{2-}, s. Abb. 8-7). Auf eine entsprechende mögliche Transportrichtung der Schadstoffe von SWP 10 in Richtung Nahle zu GWB 5 wurde bereits hingewiesen und die geographische Lage von GWB 4, der ebenfalls auf diesem Wasserpfad liegt, unterstützt diese Aussage. Eine Ähnlichkeit besteht ferner zwischen SWP 10 und dem Cluster {GWB 1; GWB 2} (Cl^-/SO_4^{2-}, s. Abb. 8-7). Dies wiederum steht in Übereinstimmung dazu, dass SWP 10 sich geographisch in unmittelbarer Nähe zum Abstrombereich der Luppe befindet, zu welchem die GWB 1 und 2 gehören.

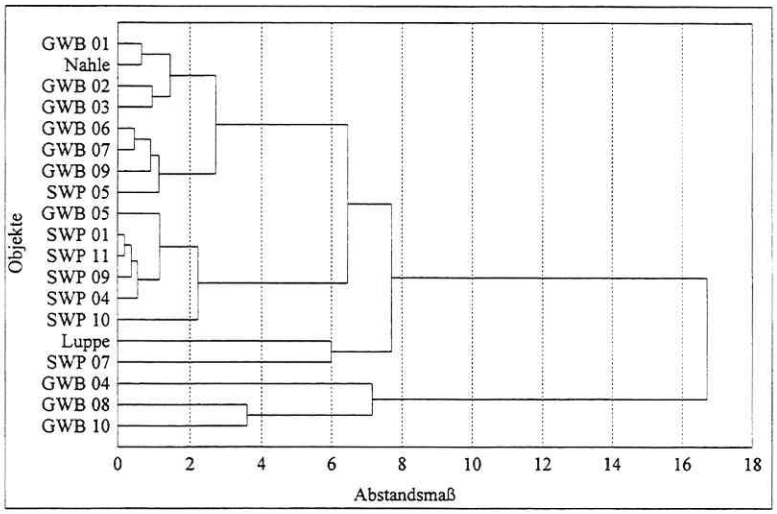

Abb. 8-6. Dendrogramm. - WARDs Methode (Variablen: Na^+/K^+).

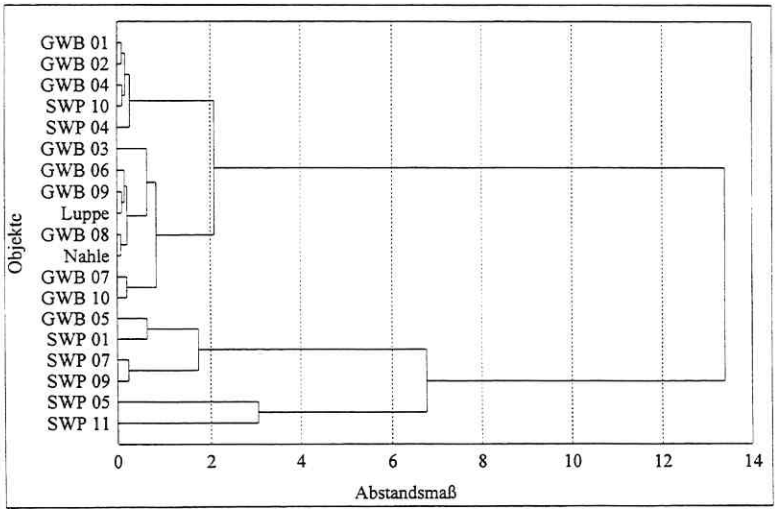

Abb. 8-7. Dendrogramm. - WARDs Methode (Variablen: Cl^-/SO_4^{2-}).

8.3 Nichthierarchische Methoden (Optimierende Clusterung)

8.3.1 Methodik allgemein

Nichthierarchische Verfahren liefern verschiedene Zerlegungen der Gesamtmenge der Objekte, die völlig unabhängig voneinander sind. Es wird eine bestimmte Gruppenanzahl g vorgegeben, und das Verfahren ermittelt eine weitgehend optimale Verteilung der Objekte auf die Gruppen. Die Zerlegungen der Objektmenge auf den verschiedenen Niveaus erfolgen unabhängig voneinander, d. h., bei Variation von g können bereits gebildete Gruppen sich wieder zerteilen. Dem Vorteil hierarchischer Methoden, ein eindeutiges Gesamtbild in Form eines Dendrogramms zu erhalten, steht somit der Nachteil gegenüber, dass eine Korrektur von auf niedrigem Niveau getroffenen (Fehl-) Entscheidungen auf höherem Niveau nicht mehr möglich ist. Insofern liefern nichthierarchische Verfahren eine realistischere Clusterbildung (HENRION et al., 1988).

Die Klassifikationsaufgabe besteht darin, bezüglich einer definierten Zielfunktion die Optimierung der Aufteilung der Objekte gegenüber einer gegebenen Anfangspartition vorzunehmen. Bei den betrachteten metrischen Daten ist die zu minimierende Zielfunktion das sog. Varianzkriterium (auch „Spur W-Kriterium"), die Zerlegung wird hierbei durch die Summe der quadrierten EUKLIDischen Abstände sämtlicher Objekte X_i (als Spaltenvektor geschriebenes i-tes Objekt des Datensatzes) zum Schwerpunkt $\overline{X}^l$ der jeweiligen Gruppe l bewertet, der sie angehören:

$$\operatorname{tr} \underline{W} = \sum_{j=1}^{p} w_{jj} = \sum_{l=1}^{g} \sum_{i \in g_l} \sum_{j=1}^{p} \left(x_{ij} - \overline{x}_j^l \right)^2 = \sum_{l=1}^{g} \sum_{i \in g_l} \left\| X_i - \overline{X}^l \right\|^2$$

... Spur (trace) einer „within scatter matrix", d. h. einer die Streuung innerhalb der Gruppen charakterisierenden Matrix.

Das Varianzkriterium ist nur für den Vergleich von Zerlegungen anwendbar, welche sich bei Variation der Gruppenanzahl g ergeben, wobei jeweils von einer Startpartition ausgegangen wird. Nur bezüglich dieser wird für jedes Niveau (Gruppenanzahl g) die beste Zerlegung extrahiert. Eine totale Enumeration, d. h. Berechnung und Vergleich der Zielfunktionswerte für alle möglichen Zerlegungen der n Objekte (bei jeweils gegebener Gruppenanzahl g), ist i. Allg. aufgrund des zu hohen Rechen- bzw. Zeitaufwandes nicht möglich. Die Anzahl der möglichen Partitionen, die sich bei einer Unterteilung der n Objekte in g Gruppen ergibt, lässt

sich für sehr große n näherungsweise nach der STIRLINGschen Formel (STEINHAUSEN und LANGER, 1977; BRONSTEIN, 1983) berechnen. Eine Auswahl kann Tab. 8-2 entnommen werden.

Tab. 8-2. Anzahl möglicher Startpartitionen bei Variation von n und g (Auswahl).

Objektanzahl n	Gruppenanzahl g	Anzahl Partitionen
6	2	41
50	10	$\approx 2{,}6 \cdot 10^{43}$
100	10	$\approx 2{,}8 \cdot 10^{93}$

In der Praxis werden daher heuristische Methoden angewandt: Ausgehend von einer festgelegten Startpartition werden iterativ durch Objektaustausch neue Partitionen bis hin zu einem lokalen Minimum des Varianzkriteriums erzeugt. Es lässt sich somit praktisch nur das lokale Optimum berechnen, die insgesamt beste Lösung (globales Optimum der Zielfunktion) wäre nur bei totaler Enumeration zu erhalten (HENRION und HENRION, 1994).

Zu den wichtigsten Vertretern der optimierenden Algorithmen gehören:

- das Minimaldistanzverfahren nach FORGY (1965)

  Der Algorithmus dieses Verfahrens lässt sich wie folgt angeben:

  1. Vorgabe einer Startpartition
  2. Berechnung der Gruppenschwerpunkte
  3. Überprüfung, ob alle n Objekte dem nächstliegenden Gruppenschwerpunkt (im Sinne des EUKLIDischen Abstands) zugeordnet sind, ggf. Verschiebung in die entsprechende Gruppe
  4. zurück zu 1., falls ein oder mehrere Objekte verschoben wurden

  STEINHAUSEN und LANGER (1977) zeigen, dass sich der Wert des Varianzkriteriums bei jeder Iteration verringert bzw. im ungünstigsten Fall gleich bleibt. HENRION und HENRION (1994) demonstrieren diesen Algorithmus anhand eines (fiktiven) Beispieldatensatzes.

- das k-means-Verfahren nach MACQUEEN (1967)

  Dieses Verfahren unterscheidet sich von dem erstgenannten dadurch, dass die Gruppenschwerpunkte nicht erst nach Verschiebung aller vorgemerkten Objekte neu berechnet werden, sondern unmittelbar nach Verschiebung eines ersten gemusterten Objekts. Die Itera-

tionsanzahl kann sich dadurch verringern, weil nach Verschiebung eines Objekts die unmittelbar aktualisierten Gruppenschwerpunkte noch in derselben Iteration die Verschiebung des danach erst gemusterten (weil im Ausgangsdatensatz als nächstes stehenden) Objekts erzwingen würden. Der Nachteil dieses Verfahrens besteht darin, dass die erhaltenen Partitionen von der Objektreihenfolge des Ausgangsdatensatzes abhängig sind.

- das Austauschverfahren nach RUBIN (1967)

  Die Objektverschiebung steht hier in unmittelbarem Zusammenhang mit dem Varianzkriterium. Die Auswahl der Gruppe, in welche ein als erstes gemustertes Objekt verschoben wird, erfolgt so, dass nach der unmittelbaren Aktualisierung der Gruppenschwerpunkte das Varianzkriterum größtmöglich abnimmt. Der Algorithmus ist ansonsten identisch mit dem des zweitgenannten Verfahrens, das Ergebnis der Partitionierung ist auch hier von der Objektreihenfolge abhängig.

Unter dem Aspekt, dass die optimalen Partitionen jeweils für verschiedene Clusteranzahlen g gebildet werden, stellt sich die alleinige Anwendung des Varianzkriteriums jedoch als unzureichend dar, da es mit wachsender Clusteranzahl für optimale Partitionen immer mehr abnimmt, bis es schließlich für die „feinste" Partition (jedes Einzelobjekt ist selbst ein Cluster) null wird. Die bereits mehrfach erwähnte HUYGENSsche Dekompositionsformel $\underline{T} = \underline{B} + \underline{W}$ lässt sich nach HENRION und HENRION (1994) in $\mathrm{tr}\,\underline{T} = \mathrm{tr}\,\underline{B} + \mathrm{tr}\,\underline{W}$ (unabhängig von der Partitionierung konstante Gesamtstreuung des Datensatzes) überführen, wobei $\mathrm{tr}\,\underline{B}$ die Spur (trace) einer „between scatter matrix", d. h. einer die Streuung zwischen den Gruppen charakterisierenden Matrix, darstellt. Die erhaltene Partition ist dann optimal, wenn die Objekte innerhalb ein und desselben Clusters möglichst ähnlich (kleine Streuung „innerhalb", d. h. „Homogenität") und solche aus verschiedenen Clustern wenig ähnlich (große Streuung „zwischen", d. h. „Separation") sind. Dies führt letztendlich zur Anwendung des sog. F-Kriteriums

$$F = \frac{\mathrm{tr}\,\underline{B}}{\mathrm{tr}\,\underline{W}} \cdot \frac{n-g}{g-1},$$

das eine unmittelbare Verallgemeinerung der im Zusammenhang mit der univariaten Varianzanalyse (s. Kapitel 5) diskutierten Streuungszerlegung auf den Fall von p Variablen darstellt. Je größer sein Wert bei variierenden Gruppenanzahlen g ist, umso besser ist die „Getrenntheit" der gefundenen Cluster bzw. der „Kontrast" des Clusterbildes (HENRION et al., 1988; HENRION und HENRION, 1994).

8.3.2 Untersuchungen

8.3.2.1 Gegenstand und Zielstellung der Untersuchungen

Es wurden insgesamt 18 nichthierarchisch optimierende Clusteranalysen für jeweils 19 Objekte durchgeführt. Die 19 Objekte sind die Messstellen (GWB 1 - 10, Nahle, Luppe; SWP 1, 4, 5, 7, 9 - 11), als Variablen wurden die bereits bei den hierarchischen Clusteranalysen betrachteten Ionenpaare bzw. Ionenverhältnisse (standardisierte Werte) herangezogen, die vorliegenden sechs Datensätze waren somit identisch. Es wurde eine Unterteilung in zehn, fünf und drei Gruppen vorgegeben. Einen Gesamtüberblick gibt Tab. 8-3.

Tab. 8-3. Clusteranalyse der Messstellen. - Gegenstand der Untersuchungen.

Nr.	Variablen	Anzahl Werte	Gruppenanzahl g	Nr.	Variablen	Anzahl Werte	Gruppenanzahl g
1	Na^+, K^+	16	10	10	Na^+/K^+	8	10
2	Na^+, K^+	16	5	11	Na^+/K^+	8	5
3	Na^+, K^+	16	3	12	Na^+/K^+	8	3
4	Mg^{2+}, Ca^{2+}	16	10	13	Mg^{2+}/Ca^{2+}	8	10
5	Mg^{2+}, Ca^{2+}	16	5	14	Mg^{2+}/Ca^{2+}	8	5
6	Mg^{2+}, Ca^{2+}	16	3	15	Mg^{2+}/Ca^{2+}	8	3
7	Cl^-, SO_4^{2-}	7	10	16	Cl^-/SO_4^{2-}	3	10
8	Cl^-, SO_4^{2-}	7	5	17	Cl^-/SO_4^{2-}	3	5
9	Cl^-, SO_4^{2-}	7	3	18	Cl^-/SO_4^{2-}	3	3

In allen Fällen wurde eine sog. Standardanfangspartition (HENRION et al., 1988) verwendet, bei dieser werden die Objekte (Messstellen) entsprechend der Reihenfolge ihres Auftretens im Datensatz (s. obige Klammerangabe) und der vorgegebenen Clusteranzahl zyklisch sukzessive durchnummeriert. Die Optimierung erfolgte nach dem im vorhergehenden Abschnitt besprochenen Minimaldistanzverfahren (FORGY, 1965).

Das Ziel der Untersuchungen bestand - analog denen zur hierarchisch agglomerativen Clusteranalyse (s. Abschnitt 8.2.2.1) - darin, über die Verteilungsstrukturen der Probenahmestellen bevorzugte Austrittsrichtungen des Sickerwassers in den umgebenden Aquifer festzustellen.

Anhand des F-Kriteriums wurde für alle Datensätze zunächst ermittelt, welche der a priori einheitlich vorgegebenen Gruppenanzahlen g näherungsweise ein Suboptimum darstellen.

8.3.2.2 Untersuchungsergebnisse

Im nachfolgenden Abschnitt bzw. im Anhang sind die folgenden Ergebnisse enthalten:

- Varianzkriterium und F-Kriterium bei variierenden Gruppenanzahlen g für alle sechs Datensätze: Zusammenfassung in Abb. 8-8
- Objektpartitionierungen bei variierenden Gruppenanzahlen g (Aufteilung in zehn, fünf und drei Gruppen): s. Tab. A-15

Die in Tab. A-15 fett hervorgehobenen Cluster sind stabil gegenüber zweifacher und die unterstrichen dargestellten stabil gegenüber einfacher Änderung der Gruppenanzahl g.

8.3.2.3 Diskussion der Untersuchungsergebnisse

Abb. 8-8 verdeutlicht, dass es problematisch ist, allein das F-Kriterium zum Auffinden einer optimalen Gruppenanzahl zu verwenden. Bei vier der sechs Datensätze erreicht dieses seinen Maximalwert erst bei einer Gruppenanzahl von 17 bzw. 18, was bei einer Gesamtanzahl von jeweils 19 Objekten keine aussagefähige Partitionierung ergibt.

In solchen Fällen ist es prinzipiell empfehlenswert, in die Bewertung das Varianzkriterium einzubeziehen. Die (gewählten) suboptimalen Clusteranzahlen der vier Datensätze ergaben sich dabei aus folgenden Überlegungen: Bei Erhöhung der Gruppenanzahl g zeigt sich zunächst eine starke, später nur noch geringe Abnahme des Varianzkriteriums tr $\underline{W}$. Eine geeignete, d. h. vorhandene Musterklassen gut widerspiegelnde Clusteranzahl lässt sich somit am „Knick" der die Punkte verbindenden Kurve ablesen, da vorher eine starke (d. h. eine weitere Verfeinerung der Zerlegung rechtfertigende) und dahinter nur noch eine sehr geringe (d. h. eine Erhöhung der Gruppenanzahl nicht mehr rechtfertigende) Abnahme von tr $\underline{W}$ erfolgt.

Generell von Nachteil für die Bewertung der Untersuchungsergebnisse nichthierarchischer Verfahren ist, dass diese sich nicht graphisch so anschaulich darstellen lassen, wie das für die der hierarchischen Clusteranalysen mittels Dendrogramms möglich ist. Dennoch lassen sich anhand von Tab. A-15 für die vorliegenden Ergebnisse einige interessante Aspekte feststellen.

- Clusterung in Abhängigkeit von den Ionenpaaren (1. bis 9. Analyse)
  Es wird deutlich, dass sich bei zweifacher Änderung der Gruppenanzahl der allein aus SWP 10 gebildete Cluster in zwei Datensätzen als stabil erweist. Die hohen Salzfrachten im Bereich dieser Messstelle spiegeln sich somit auch in diesen Ergebnissen gut wider.

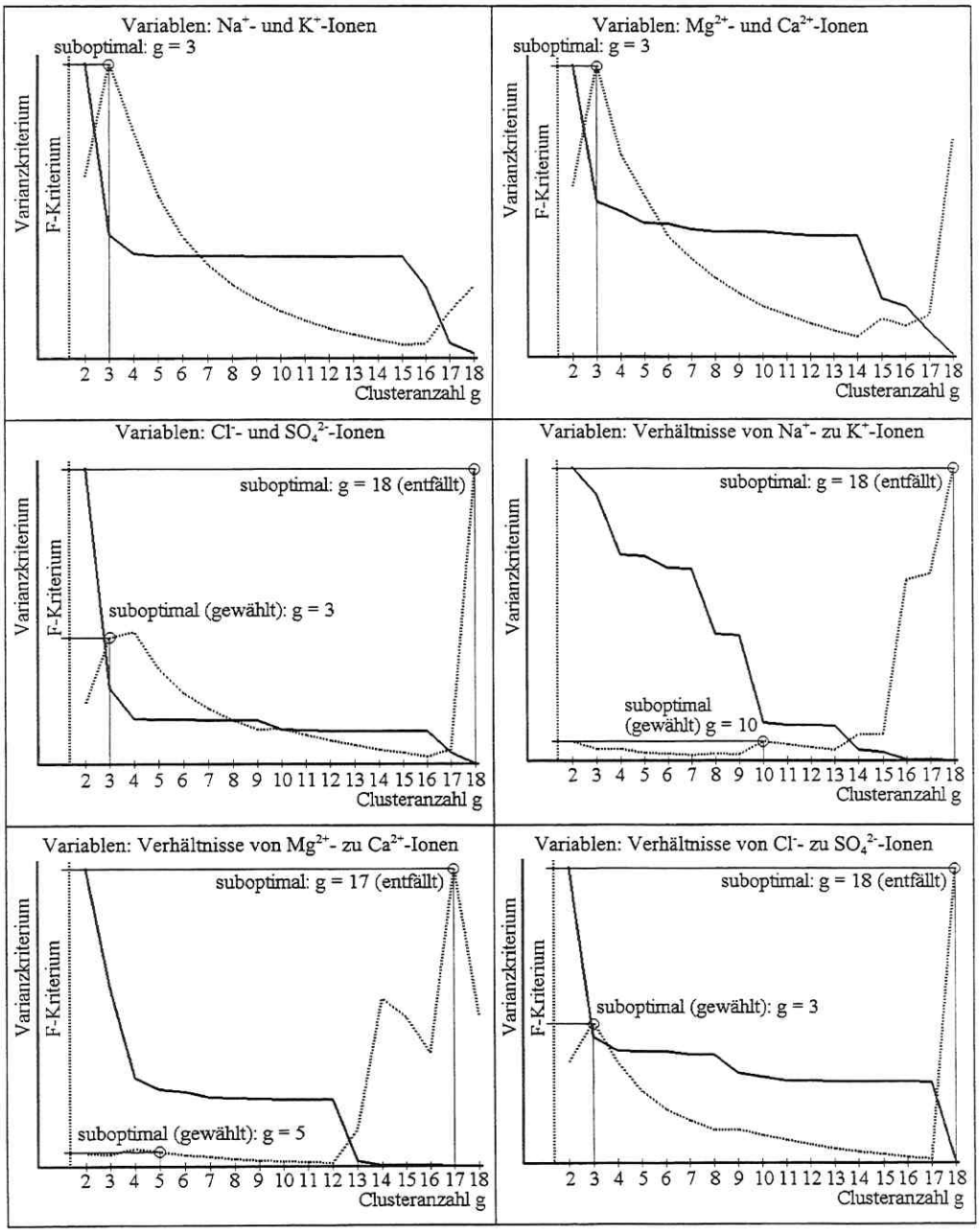

Abb. 8-8. Varianzkriterium und F-Kriterium bei variierenden Gruppenanzahlen g.

Die Vorgabe von drei Gruppen führt in allen drei Fällen dazu, dass - analog den Resultaten der hierarchischen Analysen - sämtliche Grundwassermessstellen aufgrund ihrer wesentlich geringeren Gehalte an Einzelionen jeweils demselben Cluster zugeordnet werden.

Bei der Objektpartitionierung in g = 3 Gruppen (suboptimale Lösung für alle drei Datensätze) fällt weiterhin auf, dass das Objekt SWP 7 (SWP 9) sich zweimal (einmal) in der Gruppe wiederfindet, in welcher sich sämtliche Grundwasserobjekte befinden (Na^+ und K^+, Cl^- und SO_4^{2-}), d. h., die Ähnlichkeit der Muster bezüglich der Salzfrachten könnte hier ein Indiz für entsprechende Austrittspfade sein. Die Ergebnisse hierarchischer Verfahren führten zu der Aussage, dass, von den Probenahmestellen SWP 7 und 9 ausgehend, ein Schadstoffaustrag zumindest in Richtung nordwestlicher Teil des Deponiegeländes (entsprechend der Grundwasserfließrichtung) zu den GWB 4 und 5 möglich ist. Eine Verfeinerung der Partitionierung auf fünf Gruppen zeigt dann auch, dass der SWP 7 erneut den Grundwassermessstellen zugeordnet wird (Na^+ und K^+) bzw. ein Cluster gebildet wird, in dem neben den SWP 7 und 9 auch SWP 11 - dieser liegt ebenfalls im nordwestlichen Deponiebereich - sowie GWB 5 enthalten sind (Cl^- und SO_4^{2-}). Auch bei der Clusterung in Abhängigkeit von den Mg^{2+}- und Ca^{2+}-Ionen finden sich bei g = 5 Gruppen die SWP 7, 9 und 5 (letzterer liegt ebenfalls auf diesem Wasserpfad) in einem gemeinsamen Cluster wieder. Um diesen Zusammenhang zu verdeutlichen, sei letztendlich noch auf die feinste Partitionierung (g = 10) hingewiesen: Es kommt hier u. a. zu den Clusterbildungen {SWP 7; GWB 5} (Na^+ und K^+), {SWP 7; SWP 9} (Mg^{2+} und Ca^2) sowie {SWP 9; GWB 5} (Cl^- und SO_4^{2-}).

Ansonsten lassen die Ergebnisse keine weiteren Interpretationen zu, da die gebildeten Gruppen entweder aus Probenahmestellen nur des Sickerwassers oder nur des Grundwassers rekrutieren. Als Ursache hierfür sind, wie mehrfach bereits erwähnt, die jeweils unterschiedlich hohen Salzfrachten zu sehen.

- Clusterung in Abhängigkeit von den Ionenverhältnissen (10. bis 18. Analyse)

  In Analogie zu den entsprechenden Untersuchungsergebnissen der hierarchischen Clusteranalyse werden eine Stabilität des Clusters {SWP 10} gegenüber zweifacher Änderung der Gruppenanzahl (Mg^{2+}/Ca^{2+}) sowie Ähnlichkeiten zwischen SWP 10 und GWB 5 (Na^+/K^+, g = 3) bzw. SWP 10 und GWB 4 (Cl^-/SO_4^{2-}, g = 3) ermittelt.

  Des Weiteren sei auf die Stabilität des Clusters {SWP 5; SWP 7; SWP 9} beim Übergang von g = 10 zu g = 5 (Mg^{2+}/Ca^{2+}) sowie die des Clusters {SWP 5; SWP 11} beim Übergang von g = 5 zu g = 3 (Cl^-/SO_4^{2-}) hingewiesen. Auch wenn daraus keine direkte Ähnlichkeit zwischen Sicker- und Grundwassermessstellen zu ersehen ist, so steht dies doch zumindest in Übereinstimmung mit der bisher getroffenen Aussage einer allgemein bevorzugten Sickerwasserfließrichtung von Südost nach Nordwest.

8.4 Fuzzy-Clusteranalyse

8.4.1 Vorbemerkungen

Die klassischen Clusteralgorithmen zielen auf sog. scharfe Partitionen ab, d. h., jedes Objekt wird in genau einen Cluster eingeordnet. Dies erweist sich bei guter Separation der Cluster als günstig, denn bei den Partitionen liegen somit keine Unsicherheiten vor. In den meisten praktischen Anwendungsfällen ist eine solche Situation jedoch nicht gegeben. Wie die bisherigen Untersuchungen zeigten, führen für ein und denselben Datensatz verschiedene hierarchische Verfahren mitunter zu unterschiedlichen Partitionen bei fixierter Clusteranzahl bzw. die mit Hilfe der nichthierarchischen Verfahren ermittelten Partitionen erweisen sich bei variierenden Gruppenanzahlen als wenig stabil. Ursachen hierfür können das Auftreten von hybriden Objekten (Objekte, die genau zwischen mehreren Clustern liegen) und von Ausreißern (Objekte, die weit entfernt von allen Clusterschwerpunkten liegen) sein. Deren Klassifizierung mit scharfen Clusteranalysealgorithmen muss als äußerst unsicher eingeschätzt werden, besonders unter dem Aspekt, dass keine der verschiedenen Methoden eine optimale Lösung gewährleistet (BORTZ, 1993).

Zusätzliche Probleme treten bei der Anwendung von scharfen Clusteranalysen speziell bei ökologischen Sachverhalten auf. Die Angabe der Daten ist hier oft mit einer Scheingenauigkeit versehen, weil die exakten Werte nicht ermittelt werden können. Dies äußert sich z. B. in einer subjektiven Schätzung der Werte („Die Temperatur liegt bei über 30 °C."), der Inter- bzw. Extrapolation von Werten oder der Angabe von Mittelwerten aus verschiedenen Messungen. Und in der Tat, die bisherige Analyse der Wasserproben (Messstellen) bezüglich ihres Musters in verschiedenen Salzfrachten zeigte, dass neben „eindeutigen" Wasserproben auch solche auftreten, die offenbar dem Einfluss mehrerer unabhängiger Emissionsquellen bzw. Wasserpfade ausgesetzt waren, deren mögliche Ursachen bereits diskutiert wurden. Die Zugehörigkeit solcher Objekte zu einem Cluster kann man somit nicht als eindeutig bezeichnen.

Durch die Anwendung von unscharfen bzw. Fuzzy-Clusteranalysen können solche Unsicherheiten berücksichtigt werden. Die Einordnung der Objekte in die verschiedenen Cluster erfolgt hier nicht scharf, sondern kann, entsprechend sog. Zugehörigkeitswerten, aufgeteilt werden. Darüber hinaus liefert die Verteilung der Zugehörigkeitswerte zwischen den verschiedenen Clustern eine Information über die Struktur der Datenmenge.

8.4.2 Methodik allgemein

Bei der Fuzzy-Clusterung wird, wie bei der nichthierarchisch optimierenden Clusterung, eine Anzahl g von (vermuteten) Clustern vorgegeben. Es werden die Koeffizienten u_{il} ermittelt, die angeben, mit welcher Wahrscheinlichkeit das Objekt i dem Cluster l angehört. Die Zugehörigkeit wird somit als Wahrscheinlichkeit interpretiert, weshalb für die zu berechnenden Koeffizienten gilt:

$$u_{il} \geq 0 \quad (1 \leq i \leq n;\ 1 \leq l \leq g) \quad \text{und} \quad \sum_{l=1}^{g} u_{il} = 1 \quad (1 \leq i \leq n).$$

Das Auffinden der optimalen Zugehörigkeitskoeffizienten erfolgt durch Minimierung einer geeigneten Zielfunktion. BEZDEK et al. (1984) geben diese mit

$$J(\underline{U}, \underline{V}) = \sum_{i=1}^{n} \sum_{l=1}^{g} \left(u_{il}\right)^m \left\| X_i - V_l \right\|^2$$

an, hierbei bedeuten

$\underline{U}$ ... (n, g)-Matrix der gesuchten Koeffizienten u_{il},

$\underline{V}$ ... (p, g)-Matrix von Cluster-Referenzpunkten (p: Anzahl der Variablen),

X_i ... als Spaltenvektor geschriebenes i-tes Objekt des Datensatzes,

V_l ... l-te Spalte der Matrix $\underline{V}$ (d. h. Referenzpunkt des l-ten Clusters) und

m ... Gewichtsexponent (wählbarer Parameter).

Bei einer scharfen Clusterung, wo die Zugehörigkeitskoeffizienten lediglich die Werte null oder eins annehmen können, sind im Sinne der Minimierung von J ($\underline{U}$, $\underline{V}$) als Cluster-Referenzpunkte unmittelbar die Clusterschwerpunkte zu wählen, d. h.

$$V_l = \overline{X}^l = \left(\overline{x}_1^l \ \ \overline{x}_2^l \ \ \ldots \ \ \overline{x}_p^l\right)^T \quad \text{mit} \quad \overline{x}_j^l = \frac{1}{n_l} \cdot \sum_{i \in g_l} x_{ij} \quad (j = 1, \ldots, p).$$

In diesem Fall ist J ($\underline{U}$, $\underline{V}$) identisch mit der Summation aller Abweichungsquadrate innerhalb der Cluster, d. h. mit dem bei der nichthierarchisch optimierenden Clusterung (s. Abschnitt 8.3) als zu minimierende Zielfunktion verwendeten Varianzkriterium tr $\underline{W}$:

$$J(\underline{U}, \underline{V}) = \sum_{i=1}^{n} \sum_{l=1}^{g} \left(u_{il}\right)^m \left\| X_i - V_l \right\|^2 = \sum_{l=1}^{g} \sum_{i \in g_l} \sum_{j=1}^{p} \left(x_{ij} - \overline{x}_j^l\right)^2 = \sum_{l=1}^{g} \sum_{i \in g_l} \left\| X_i - \overline{X}^l \right\|^2 = \operatorname{tr} \underline{W}.$$

tr $\underline{W}$ ist somit ein Spezialfall von J ($\underline{U}$, $\underline{V}$) bzw. umgekehrt J ($\underline{U}$, $\underline{V}$) die unscharfe Verallgemeinerung von tr $\underline{W}$. Die Aufstellung der Stationaritätsbedingungen für lokale Minima von J ($\underline{U}$, $\underline{V}$) (Nullsetzen der partiellen Ableitungen) führt nach BEZDEK et al. (1984) auf die beiden Gleichungen

$$V_l = \frac{\sum_{i=1}^{n} (u_{il})^m X_i}{\sum_{i=1}^{n} (u_{il})^m} \quad (1 \le l \le g) \quad \text{und} \quad u_{il} = \frac{1}{\sum_{l=1}^{g} \left(\frac{\|X_i - V_l\|}{\|\overline{X}^1 - V_l\|} \right)^{\frac{2}{m-1}}} \quad (1 \le i \le n;\ 1 \le l \le g;\ m > 1),$$

wobei die so definierten Koeffizienten u_{il} den o. g. Bedingungen genügen. Da die Matrix $\underline{V}$ (als Zusammenfassung der V_l) sich aus der Matrix $\underline{U}$ (als Zusammenfassung der u_{il}) und umgekehrt, die Matrix $\underline{U}$ sich aus der Matrix $\underline{V}$ berechnen lässt, besteht der Lösungsalgorithmus darin, ausgehend von Startkoeffizienten u_{il}, welche die o. g. Bedingungen erfüllen, durch alternierende Anwendung der beiden Gleichungen iterierte Matrizen $\underline{V}$ und $\underline{U}$ zu bestimmen. Die Konvergenz eines solchen sog. Fuzzy-c-Means-Algorithmus zeigt BEZDEK (1981).

Von entscheidender Bedeutung für die Ergebnisse der Fuzzy-Clusteranalyse ist die Wahl des Parameters m: Für m = 1 befindet man sich wieder in der Situation der klassischen scharfen Clusterung, darüber hinaus steigende Werte von m liefern zunehmend unscharfe Zerlegungen. Die beschriebene Vorgehensweise wird durch HENRION und HENRION (1994) exemplarisch anhand einer (fiktiven) Datenmatrix demonstriert.

8.4.3 Das Fuzzy-Clustering-System ECO-FUCS

Die Untersuchungen erfolgten mit dem Fuzzy-Clustering-System ECO-FUCS (PC-Version 2.0 für MS-DOS) (PAASCH, 1994). Die in ECO-FUCS verwendete Clustering Prozedur basiert auf dem im vorhergehenden Abschnitt erläuterten Fuzzy-c-Means-Algorithmus. Nach Vorgabe einer Startzerlegung werden deren Zugehörigkeiten mit jedem Schritt eines Iterationsprozesses in Abhängigkeit des Optimierungskriteriums modifiziert, bis ein Abbruchkriterium erfüllt ist. Ein Iterationsprozess beinhaltet dabei sowohl die Berechnung einer Startzerlegung (ausgehend von einer Standardanfangspartition entsprechend der vorgegebenen Clusteranzahl) als auch die in mehreren Iterationsschritten erfolgende Berechnung der Zugehörigkeitskoeffizienten und Cluster-Referenzpunkte.

Wie bereits erwähnt, garantiert kein Verfahren der nichthierarchischen Clusteranalyse das Auffinden von optimalen Lösungen, jeder Iterationsprozess konvergiert daher auch hier lediglich gegen ein lokales Minimum. Die Konvergenz ist abhängig von der Startzerlegung. Für deren Festlegung bietet ECO-FUCS sechs verschiedene Möglichkeiten, von denen die Wahl der (scharfen) Startzerlegung nach dem ISODATA-Algorithmus der Anwendung des in Abschnitt 8.3.1 erläuterten Minimaldistanzverfahrens nach FORGY (1965) entspricht.

Neben der Startzerlegung erfolgt vor jedem Iterationsprozess die Festlegung eines Abstandsmaßes (u. a. steht hier der EUKLIDische Abstand zur Verfügung) sowie eines geeigneten Wertes für den Gewichtsexponenten m. Letzterer beeinflusst die Unschärfe der zu findenden Zerlegung. Er kann in einem Bereich $1 \leq m \leq 3$ eingegeben werden, wobei Werte > 1 zunehmend unscharfe Zerlegungen liefern. Die Wahl m = 1 führt zu einer scharfen Clusteranalyse nach dem ISODATA-Algorithmus, d. h., in diesem Fall ist das Ergebnis der Berechnung einer Startzerlegung nach dem ISODATA-Algorithmus identisch mit dem Ergebnis der Endzerlegung, welches aus der (in diesem Fall scharfen) Fuzzy-Clusteranalyse resultiert.

Für die nach einem Iterationsprozess berechnete Partition erfolgt die Angabe der Zugehörigkeitskoeffizienten u_{il} (einschließlich Darstellung im sog. Fuzzy-Dendrogramm), der Cluster-Referenzpunkte V_l, verschiedener Gütekriterien sowie der sog. Cluster-Trennparameter.

Werden mehrere Iterationsprozesse in Abhängigkeit variierender (steigender) Clusteranzahlen g durchlaufen, so ist ein gemeinsamer Vergleich der Werte verschiedener Gütekriterien (Payoff, Zerlegungskoeffizient, Entropie, Non-Fuzziness-Index und Verhältnis Exponent) zum Auffinden der optimalen Partition unerlässlich.

Der **Payoff** entspricht J ($\underline{U}$, $\underline{V}$), also dem Optimierungskriterium eines Iterationsprozesses und sollte daher so klein wie möglich sein.

Der **Zerlegungskoeffizient** und die **Entropie** sind jeweils ein Ausdruck für die Schärfe der Zerlegung und liegen im Bereich zwischen null und eins. Ersterer ist genau null, wenn jedes Objekt ein Cluster bildet. Andererseits ist die Entropie null, wenn nur ein Cluster vorhanden ist. Die erhaltenen Partitionen beschreiben eine „gute" Zerlegung, wenn der Zerlegungskoeffizient möglichst groß und die Entropie möglichst gering ist. In diesem Fall weisen die Objekte in der Mehrzahl eine hohe Affinität (= Ähnlichkeit) zum jeweiligen Clusterschwerpunkt auf. Enthalten die Kurvenverläufe beider Kriterien in einem Intervall von möglichen Clusteranzahlen einen starken Anstieg bzw. einen starken Abfall, so ist dieser Bereich der Diskontinuität ein Hinweis auf die optimale Clusteranzahl (ROUBENS, 1982).

Der **Non-Fuzziness-Index** (ROUBENS, 1982) und der **Verhältnis-Exponent** (WINDHAM, 1981) sind weitere von der Clusteranzahl abhängige Gütekriterien. Numerische Berechnungen von ROUBENS (1982) zeigen, dass bei beiden Gütekriterien jeweils das Maximum gesucht werden muss, um die optimale Clusteranzahl zu finden.

Die nach jedem Iterationsprozess abrufbaren Cluster-Trennparameter sind ein Maß für die Trenngüte der Merkmale. Dabei haben nicht die absoluten Werte Einfluss auf die Trennung

der erhaltenen Cluster, sondern das Verhältnis der Werte untereinander. Höhere Verhältniswerte haben einen größeren Einfluss auf die Trennung (PAASCH, 1994).
Der praktische Einsatz des Fuzzy-Clustering-Systems ECO-FUCS wird beispielsweise in der Arbeit von FRIEDERICHS et al. (1996) anschaulich demonstriert.

8.4.4 Untersuchungen

8.4.4.1 Gegenstand und Zielstellung der Untersuchungen

Die Untersuchungen erfolgten analog denen zur (scharfen) nichthierarchisch optimierenden Clusteranalyse, s. Abschnitt 8.3.2.1, wobei hier jedoch für jeden der vorliegenden sechs Datensätze (mit jeweils 19 Objekten (Messstellen) und den Ionenpaaren bzw. Ionenverhältnissen als Variablen) eine Unterteilung in lediglich fünf Gruppen (somit insgesamt sechs Fuzzy-Clusteranalysen) vorgegeben wurde.
Die allgemeine Zielstellung entsprach im Wesentlichen derjenigen, welche mit Hilfe der scharfen Clusteranalysen verfolgt wurde. In den Abschnitten 8.2.2.1 und 8.3.2.1 wurde darauf eingegangen. Die sich darüber hinaus mit Hilfe der Fuzzy-Clusteranalyse zu lösenden spezifischen Problemstellungen lassen sich wie folgt zusammenfassen:

- Beurteilung der Separierbarkeit der Objektmenge bzw. der Schärfe der ermittelten Partitionen durch Betrachtung der Verteilung der Zugehörigkeitswerte
- Auffinden von Objekten (Messstellen), die repräsentativ für die Cluster sind
- Untersuchung des Einflusses der Merkmale (Ionen) auf die Trennung der erhaltenen Cluster

Um einen Vergleich mit den Resultaten der (scharfen) nichthierarchisch optimierenden Analysen (s. Tab. A-15) zu ermöglichen, wurde die Startzerlegung nach dem ISODATA-Algorithmus (= Minimaldistanzverfahren) ermittelt, d. h., bei einer Wahl von $m = 1$ für den Gewichtsexponenten wären die sich ergebenden (in diesem Fall scharfen) Partitionen identisch mit denen, die in Tab. A-15 für $g = 5$ aufgeführt sind.
In Anlehnung an die Untersuchungen von FRIEDERICHS et al. (1996) wurden für jeden der sechs Datensätze zwei Gütekriterienanalysen mit verschiedenen Gewichtskoeffizienten ($m = 1{,}3$ und $m = 1{,}6$) durchgeführt. Dadurch sollte ermittelt werden, für welchen der Koeffizienten die Gütekriterien bei $g = 5$ optimalere Werte annehmen, um anschließend mit diesem Koeffizienten die Analyse der sechs Datensätze vorzunehmen.

8.4.4.2 Untersuchungsergebnisse

Bei den graphischen Darstellungen (Gütekriterienanalysen, Fuzzy-Dendrogramme), musste sich wiederum auf eine Auswahl beschränkt werden. In Abschnitt 8.4.4.3 sind die folgenden Untersuchungsergebnisse enthalten:

- Gütekriterienanalyse (Variablen: Ionenverhältnisse Mg^{2+}/Ca^{2+}; m = 1,6): s. Abb. 8-9
- Fuzzy-Dendrogramm (Variablen: Cl^-- und SO_4^{2-}-Ionen): s. Abb. 8-10
- Fuzzy-Dendrogramm (Variablen: Ionenverhältnisse Cl^-/SO_4^{2-}): s. Abb. 8-11
- Trenngüte der Ionen (Cluster-Trennparameter): Zusammenfassung in Tab. 8-4

Die sich bei einer Gruppenanzahl von g = 5 ergebenden Objektpartitionierungen (Verteilung der Zugehörigkeitswerte) sind in den folgenden Tabellen des Anhangs aufgeführt, wobei hierin Zugehörigkeitswerte von ≥ 0,10 zu anderen Clustern unterstrichen dargestellt sind:

- Variablen: Na^+- und K^+-Ionen: s. Tab. A-16
- Variablen: Mg^{2+}- und Ca^{2+}-Ionen: s. Tab. A-17
- Variablen: Cl^-- und SO_4^{2-}-Ionen: s. Tab. A-18
- Variablen: Verhältnisse von Na^+- zu K^+-Ionen: s. Tab. A-19
- Variablen: Verhältnisse von Mg^{2+}- zu Ca^{2+}-Ionen: s. Tab. A-20
- Variablen: Verhältnisse von Cl^-- zu SO_4^{2-}-Ionen: s. Tab. A-21

8.4.4.3 Diskussion der Untersuchungsergebnisse

Ein Vergleich der Gütekriterienanalysen mit den beiden Gewichtskoeffizienten (m = 1,3 und m = 1,6) zeigte bei allen sechs Datensätzen, dass der höhere Wert zum Auffinden einer oder mehrerer (sub-) optimaler Clusteranzahlen prinzipiell besser geeignet ist bzw. die Gütekriterien hier für g = 5 optimalere Werte annehmen. Es wurde sich daher bei allen sechs Fuzzy-Clusteranalysen für den Gewichtskoeffizienten von 1,6 entschieden. Abb. 8-9 verdeutlicht dies exemplarisch anhand des Datensatzes mit den Variablen Mg^{2+}/Ca^{2+}, der hier dargestellte Verlauf der Gütekriterien bei einem m-Wert von 1,6 zeigt, dass die vorgegebene Clusteranzahl von g = 5 als eine (sub-) optimale Lösung angesehen werden kann.

Derartige Entscheidungen müssen generell bei einer Diskussion über die Schärfe der ermittelten Partitionen berücksichtigt werden, denn höhere Werte des Gewichtsexponenten liefern zunehmend unscharfe Zerlegungen.

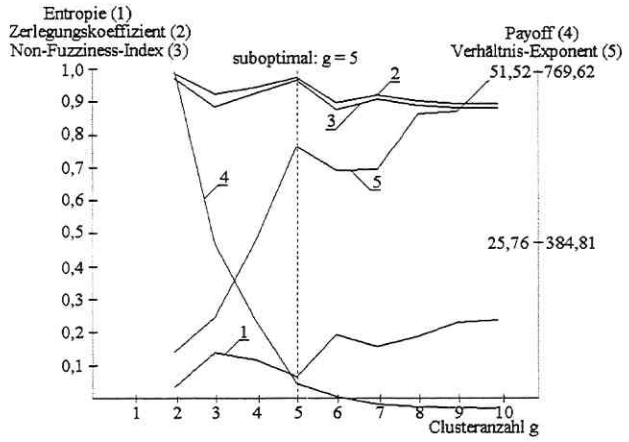

Abb. 8-9. Gütekriterienanalyse (Variablen: Mg^{2+}/Ca^{2+}; m = 1,6).

- Fuzzy-Clusterung in Abhängigkeit von den Ionenpaaren

  Es zeigt sich bei allen drei Datensätzen eine sehr gute Separierbarkeit der Objektmenge bei einer Vorgabe von fünf Gruppen, die Schärfe der resultierenden Partitionen ist relativ hoch, s. Tab. A-16, A-17 und A-18.

  Die Probenahmestellen des Grundwassers bilden bei den Mg^{2+}- und Ca^{2+}-Ionen (mit Ausnahme von GWB 4) sowie bei den Cl^-- und SO_4^{2-}-Ionen (mit Ausnahme von GWB 5) eine Gruppe für sich, wobei die Zuordnung bei den Mg^{2+}- und Ca^{2+}-Ionen weniger scharf erfolgt. Die hier generell bestehende größere Ähnlichkeit zwischen den Grund- und Sickerwassermessstellen wurde bereits mittels der scharfen Verfahren festgestellt und diskutiert.

  Im Ergebnis aller drei Fuzzy-Analysen bildet der SWP 10 ein separates Cluster mit dem größtmöglichen Zuordnungswert von eins. Damit werden die Resultate der scharfen Verfahren vollauf bestätigt, die das Objekt SWP 10 aufgrund seiner hohen Belastung als Ausreißer erkennen lassen.

  Bei den Cl^-- und SO_4^{2-}-Ionen besteht eine Übereinstimmung in den Merkmalsmustern zwischen GWB 5 einerseits sowie den SWP 7, 9 und 11 andererseits, s. Abb. 8-10. Damit wird die bereits im Zusammenhang mit den Resultaten der scharfen Clusteranalysen geäußerte Annahme über eine entsprechende Fließrichtung des Sickerwassers von SWP 9 über SWP 7 in Richtung nordwestlicher Teil des Deponiegeländes zu GWB 5 verstärkt, zumal bei den Mg^{2+}- und Ca^{2+}-Ionen eine Ähnlichkeit zwischen den beiden Sickerwassermessstellen SWP 7 und 9 sowie dem ebenfalls auf dem entsprechenden Wasserpfad gelegenen GWB 4 deutlich wird und bei den Na^+- und K^+-Ionen der sich geographisch in unmittelbarer Nähe von GWB 4 befindende SWP 11 ein separates Cluster bildet.

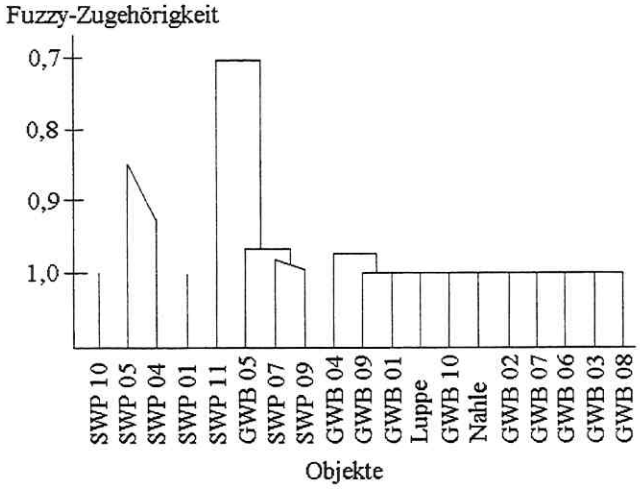

Abb. 8-10. Dendrogramm (Variablen: Cl^-, SO_4^{2-}).

Die Trenngüte der Merkmale (Einzelionen) kann in diesem Zusammenhang diskutiert werden, da ein direkter Vergleich der zu den jeweiligen Ionenpaaren gehörenden Ionen möglich ist, s. Tab. 8-4. Auffallend ist der hervorragende Einfluss der Ca^{2+}-Ionen auf die Trennung der erhaltenen Cluster bei Vergleich mit dem der Mg^{2+}-Ionen, bei sieben von acht Ionenpaaren liegen diese jeweils deutlich vorn. Dieselbe Aussage, wenn auch nur für insgesamt drei Ionenpaare, trifft auf die SO_4^{2-}-Ionen bei Vergleich mit den Cl^--Ionen zu.

Tab. 8-4. Trenngüte der Ionen (Cluster-Trennparameter).

Datensatz 1		Datensatz 2		Datensatz 3	
Merkmal (Messdatum)	Parameter	Merkmal (Messdatum)	Parameter	Merkmal (Messdatum)	Parameter
Na^+ (März '94)	20,574	Mg^{2+} (März '94)	6,753	Cl^- (März '94)	136,897
K^+ (März '94)	11,875	Ca^{2+} (März '94)	14,313	SO_4^{2-} (März '94)	302,815
Na^+ (Aug. '94)	46,689	Mg^{2+} (Aug. '94)	15,184	Cl^- (Aug. '94)	136,024
K^+ (Aug. '94)	23,040	Ca^{2+} (Aug. '94)	30,962	SO_4^{2-} (Aug. '94)	269,557
Na^+ (Dez. '94)	80,816	Mg^{2+} (Dez. '94)	5,022	SO_4^{2-} (Febr. '95)	266,321
K^+ (Dez. '94)	60,983	Ca^{2+} (Dez. '94)	32,018	Cl^- (April '95)	103,804
Na^+ (Febr. '95)	89,562	Mg^{2+} (Febr. '95)	11,833	SO_4^{2-} (April '95)	296,627
K^+ (Febr. '95)	56,902	Ca^{2+} (Febr. '95)	34,902		
Na^+ (März '95)	119,054	Mg^{2+} (März '95)	12,502		
K^+ (März '95)	73,290	Ca^{2+} (März '95)	33,954		
Na^+ (April '95)	97,769	Mg^{2+} (April '95)	13,732		
K^+ (April '95)	27,000	Ca^{2+} (April '95)	33,706		
Na^+ (Juni '95)	82,194	Mg^{2+} (Juni '95)	8,977		
K^+ (Juni '95)	200,682	Ca^{2+} (Juni '95)	34,794		
Na^+ (April '98)	52,197	Mg^{2+} (April '98)	12,341		
K^+ (April '98)	240,416	Ca^{2+} (April '98)	0,885		

- Fuzzy-Clusterung in Abhängigkeit von den Ionenverhältnissen

  Im Vergleich mit den Analysen in Abhängigkeit von den Ionenpaaren erfolgt die Zuordnung der Objekte zu den Gruppen hier weniger deutlich, s. Tab. A-19, A-20 und A-21.

  Bezüglich des SWP 10 wird keine signifikante Ähnlichkeit mit einzelnen Grundwassermessstellen deutlich, er findet sich entweder separat wieder (Mg^{2+}/Ca^{2+}) oder wird einer größeren Objektmenge zugeordnet (Na^{+}/K^{+} und Cl^{-}/SO_4^{2-}). In den beiden letzteren Fällen lässt sich jedoch feststellen, dass entweder der GWB 5 (Na^{+}/K^{+}) oder aber die GWB 1 und 2 (Cl/SO_4^{2-}, s. Abb. 8-11) ebenfalls in diesen Gruppen enthalten sind. Es besteht die Möglichkeit, dass diese Ähnlichkeiten als Indikatoren für eine entsprechende Transportrichtung der Salzfrachten (d. h. von SWP 10 in Richtung Nahle zu GWB 5 bzw. in Richtung Luppe zu GWB 1 und 2) angesehen werden können, da mit Hilfe der (scharfen) hierarchischen Verfahren analoge Ergebnisse erzielt wurden.

  Das außerdem mit Hilfe einiger scharfer Verfahren ermittelte Untersuchungsergebnis, dass ein Salzfrachteneintrag entsprechend der Grundwasserfließrichtung erfolgt und zu erhöhten Schadstoffgehalten im nordwestlichen Teil des Deponiegeländes führt, wird durch die unscharfe Clusteranalyse in Abhängigkeit der Verhältnisse von Cl^{-}- zu SO_4^{2-}-Ionen bestätigt: Die SWP 7 und 9 bilden - wenn auch mit nur relativ geringen Zugehörigkeitswerten - gemeinsam mit GWB 5 ein Cluster, zu welchem auch der in geographischer Nähe zu SWP 7 liegende SWP 1 gehört, s. Abb. 8-11.

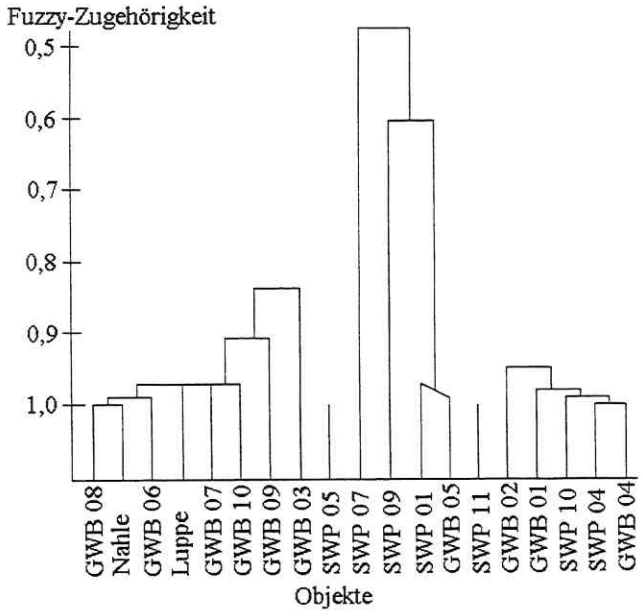

Abb. 8-11. Dendrogramm (Variablen: Cl^{-}/SO_4^{2-}).

9 Hauptkomponentenanalyse

9.1 Vorbemerkungen

Das Ziel der Hauptkomponentenanalyse besteht in der Verdichtung der in einem Datensatz enthaltenen Information (Informationsextraktion), indem viele voneinander abhängige messbare Merkmale (Variablen) auf wenige gemeinsame Ursachenkomplexe, sog. Hauptkomponenten, zurückgeführt werden (Dimensionserniedrigung). Dabei werden aus den ursprünglichen Variablen durch lineare Transformation neue gebildet, diese sind untereinander unkorreliert, ihre Anzahl ist geringer als die der ursprünglichen, und sie enthalten dennoch einen großen Teil der im Datensatz vorhandenen Information. Man erreicht die angestrebte Reduktion der Variablenanzahl ohne großen Informationsverlust (HENRION et al., 1988).

Die Hauptkomponentendarstellung ist ein unmittelbares Ergebnis der Hauptkomponentenanalyse. Mit ihr werden die Zusammenhänge aus dem p-dimensionalen Merkmalsraum in die Ebene der beiden ersten Hauptkomponenten übertragen, wodurch man einen unmittelbaren Überblick über die (Objekt-) Struktur des Datensatzes erhält. Dies ist insofern von Bedeutung, da man hierfür ansonsten bei einer vorliegenden (n, p)-Datenmatrix sämtliche n Objekte bezüglich ihrer Werte in zwei der p Variablen darstellen müsste. Ein solcher Aufwand ist praktisch nicht vertretbar, da bei p gemessenen Merkmalen insgesamt

$$\frac{p \cdot (p-1)}{2}$$

graphische Darstellungen (Kombinationen zur zweiten Klasse) möglich sind, d. h., für p = 6 beispielsweise beträgt deren Anzahl bereits 15.

Die Hauptkomponentendarstellung ermöglicht aber nicht nur (1) das Auftragen der Objekte, sondern auch (2) das der Variablen oder (3) gemeinsam das der Objekte und Variablen. Dementsprechend lassen sich drei Problemstellungen lösen:

(1) Gliedert sich der Datensatz in verschiedene Objektgruppen auf (Clusterbildung)?

(2) Welche Korrelationsbeziehungen existieren zwischen den Variablen?

(3) Durch welche Variablen wird die Lage bestimmter Objekte beeinflusst?

Eine umfassende Betrachtung aller drei Problemstellungen anhand von Beispielen nehmen HENRION et al. (1988) vor. KNOBLOCH und ZWANZIGER (1995) sowie DAUS (1996) favorisieren den Einsatz einer Hauptkomponentenanalyse bei Auftreten von Problemstellung (2)

und stellen in ihren Arbeiten diesbezügliche Untersuchungsergebnisse vor. Im Rahmen vorliegender Arbeit wurde sich auf Problemstellung (1) beschränkt, d. h., es wurde die Clusterbildung der Objekte (Probenahmestellen) in Abhängigkeit der aus den betrachteten p Ausgangsmerkmalen hervorgegangenen beiden ersten Hauptkomponenten untersucht. Die rechnerische Bestimmung der Hauptkomponenten wird im nachfolgenden Abschnitt kurz erläutert. Weitergehende Ausführungen zu den theoretischen Grundlagen dieses Verfahrens werden beispielsweise durch HENRION et al. (1988) und EINAX et al. (1997) vorgenommen.
HENRION und HENRION (1994) sowie HENRION (1998) stellen darüber hinaus die Dreiwege-Hauptkomponentenanalyse vor, eine Erweiterung der hier betrachteten klassischen Hauptkomponentenanalyse, die auf dreidimensionale Datenfelder (neben n Objekten und p Variablen Einbeziehung von q verschiedenen experimentellen Bedingungen bzw. Messzeiten) angewandt werden kann.

9.2 Berechnung der Hauptkomponenten

Werden an n Objekten p Variablen gemessen, so ergibt sich die entsprechende Datenmatrix

$$\underset{(n,p)}{\underline{X}} = \left(x_{ij}\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \text{Objektmuster 1} \\ \text{Objektmuster 2} \\ \\ \cdot \\ \cdot \\ \cdot \\ \\ \text{Objektmuster n} \end{pmatrix} = \begin{pmatrix} V & V & & V \\ a & a & & a \\ r & r & & r \\ i & i & & i \\ a & a & \cdots & a \\ b & b & & b \\ l & l & & l \\ e & e & & e \\ 1 & 2 & & p \end{pmatrix}.$$

Der Datensatz besteht somit aus einer Menge von n Objektpunkten im p-dimensionalen EUKLIDischen Raum $\Re^p$. HENRION und HENRION (1994) verweisen darauf, dass vor Anwendung des Verfahrens prinzipiell eine Datenvorbehandlung (Standardisierung der Merkmalswerte) erforderlich ist, und auch KNOBLOCH und ZWANZIGER (1995) bezeichnen - wie im Zusammenhang mit der Clusteranalyse bereits erwähnt - das Nichtstandardisieren als einen „Kunstfehler", welcher häufig zu Artefakten führt, die ihrerseits Folgefehler hervorrufen können. Daher werden zunächst die Spaltenvektoren des Datensatzes entsprechend transformiert.
Für die derart vorbehandelten Messwertfolgen der Variablen $X_1, \ldots, X_p$ werden die Korrelationskoeffizienten berechnet. Die zugehörige Korrelationsmatrix $\underline{R}$ ist vom Typ (p, p), die Koeffizienten der Hauptdiagonalen haben den Wert eins (identische Messwertfolgen), und die Werte in der Matrix sind symmetrisch bezüglich der Hauptdiagonalen. Die Ausgangsvariablen

$$X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \cdot \\ \cdot \\ \cdot \\ x_{n1} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \cdot \\ \cdot \\ \cdot \\ x_{n2} \end{pmatrix}, \quad \ldots \quad , X_p = \begin{pmatrix} x_{1p} \\ x_{2p} \\ \cdot \\ \cdot \\ \cdot \\ x_{np} \end{pmatrix}$$

werden linear transformiert. Die neuen Variablen

$$Y_1 = \begin{pmatrix} y_{11} \\ y_{21} \\ \cdot \\ \cdot \\ \cdot \\ y_{n1} \end{pmatrix}, \quad Y_2 = \begin{pmatrix} y_{12} \\ y_{22} \\ \cdot \\ \cdot \\ \cdot \\ y_{n2} \end{pmatrix}, \quad \ldots \quad , Y_p = \begin{pmatrix} y_{1p} \\ y_{2p} \\ \cdot \\ \cdot \\ \cdot \\ y_{np} \end{pmatrix}$$

berechnen sich wie folgt:

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \cdot \\ \cdot \\ \cdot \\ y_{n1} \end{pmatrix} = \begin{pmatrix} v_{11} \cdot x_{11} + v_{21} \cdot x_{12} + \ldots + v_{p1} \cdot x_{1p} \\ v_{11} \cdot x_{21} + v_{21} \cdot x_{22} + \ldots + v_{p1} \cdot x_{2p} \\ \cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ \cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ \cdot \quad\quad\quad\quad \cdot \quad\quad\quad\quad \cdot \\ v_{11} \cdot x_{n1} + v_{21} \cdot x_{n2} + \ldots + v_{p1} \cdot x_{np} \end{pmatrix},$$

d. h.

$$Y_1 = v_{11} \cdot X_1 + \ldots + v_{p1} \cdot X_p \quad \text{usw..}$$

Die Analogie dieser faktoriellen Methode zu ihrem Pendant für objektstrukturierte Datensätze, der Diskriminanzanalyse (s. Abschnitt 7.3), wird damit deutlich. Die Linearkombination der p Merkmale lässt sich zusammengefasst schreiben als

$$Y_j = \sum_{i=1}^{p} v_{ij} \cdot X_i \qquad (j = 1, \ldots, p).$$

Man bezeichnet Y_j (j = 1, ... , p) als j-te Hauptkomponente und die p^2 Koeffizienten v_{ij} als Faktorladungen. Zu diesen gelangt man durch die Lösung des Eigenwertproblems

$$\underline{R} \cdot V = \lambda \cdot V .$$

Die V sind hierin die Spaltenvektoren der Matrix

$$\underset{(p,p)}{\underline{V}} = \left(v_{ij}\right) = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{pmatrix}.$$

Die charakteristische Gleichung von $\underline{R}$ (mit $\underline{E}$ als Einheitsmatrix) lautet

$$\left|\underline{R} - \lambda \cdot \underline{E}\right| = 0 ,$$

deren Lösungen $\lambda_1, \lambda_2, \ldots, \lambda_p$ sind die Eigenwerte von $\underline{R}$.

Die zu den λ_j (j = 1, 2, ... , p) gehörenden nichttrivialen Lösungen V_j (j = 1, 2, ... , p) des Gleichungssystems

$$\left(\underline{R} - \lambda_j \cdot \underline{E}\right) \cdot V_j = 0$$

sind die Eigenvektoren, d. h.,

$$V_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \cdot \\ \cdot \\ \cdot \\ v_{p1} \end{pmatrix}$$

ist der zu λ_1 gehörende Eigenvektor usw.. Die Varianz ist das Informationsmaß des Datensatzes. Es sind nur solche Transformationen zulässig, bei denen die Gesamtvarianz der Variablen konstant bleibt, d. h., die Summe der Varianzen der X_j muss gleich der Summe der Varianzen der Y_j sein. Die entsprechende Normierungsbeschränkung

$$\sum_{i=1}^{p} v_{ij}^2 = 1 \quad (j = 1, 2, \ldots, p)$$

der Größen v_{ij} ist bei Lösung des o. g. Eigenwertproblems erfüllt. Die für die standardisierten Daten gültige Beziehung

$$\sum_{j=1}^{p} \lambda_j = p$$

verdeutlicht, dass nach erfolgter Transformation die Gesamtvarianz erhalten bleibt. Im Gegensatz zu den p standardisierten Ausgangsvariablen, welche die Varianz eins hatten, gleichen sich die Varianzen der neuen Variablen jedoch nicht mehr. Die Eigenwerte λ_j beinhalten (prozentual) den Anteil an der Gesamtvarianz. Der zum größten Eigenwert λ_1 von $\underline{R}$ gehörende Eigenvektor V_1 enthält die für die Bildung der ersten Hauptkomponente Y_1 gesuchten Koeffizienten, der zum zweitgrößten Eigenwert λ_2 gehörende Eigenvektor V_2 die für Y_2 usw.. Jeder Eigenwert liefert somit die Größe der Varianz, welche die zugehörende Hauptkomponente besitzt, wodurch diese nach fallender Bedeutung geordnet sind. Y_1 beinhaltet die größtmögliche Information aus dem Datensatz, Y_2 den größtmöglichen Teil der „Restinformation" usw.. Den letzten Hauptkomponenten kommt kaum noch Bedeutung zu. Sie werden weggelassen und nur die ersten q mit $q \leq p$ als relevant angesehen (HENRION et al., 1988; HENRION und HENRION, 1994; SCHÄFER, 1978).

Die rechnerische Bestimmung der Hauptkomponenten demonstrieren EINAX et al. (1997) anhand einer (5, 2)-Datenmatrix. HENRION et al. (1988) sowie HENRION und HENRION (1994) nehmen eine Diskussion über die Anzahl der zu extrahierenden Hauptkomponenten vor.

9.3 Untersuchungen

9.3.1 Gegenstand und Zielstellung der Untersuchungen

Es wurde für jeden der bereits mit Hilfe der Clusteranalysen untersuchten sechs Datensätze (jeweils mit 19 Objekten (Messstellen) und den Ionenpaaren bzw. Ionenverhältnissen als Variablen) eine Hauptkomponentenanalyse durchgeführt, s. Tab. 9-1.

Das unmittelbare Resultat einer Hauptkomponentenanalyse ist die Hauptkomponentendarstellung, in den beiden vorhergehenden Abschnitten wurde darauf eingegangen. Im Rahmen der Untersuchungen wurden jeweils die Objekte (Probenahmestellen) auf die Ebene der beiden ersten Hauptkomponenten aufgetragen, um deren Approximation zu verdeutlichen. Die Zielstellung entsprach damit im Wesentlichen derjenigen, die mit Hilfe der Clusteranalysen verfolgt wurde. Bei Betrachtung ausgewählter Ionenpaare bzw. Ionenverhältnisse der Sicker- und Grundwassermessstellen (Objekte) wurde mittels der Hauptkomponentenmethode versucht, einen lokalen Zusammenhang im Ausbreitungsverhalten der Salzfrachten zu ermitteln, d. h. bevorzugte Austrittsrichtungen des Sickerwassers in den Aquifer festzustellen. In diesem Sinne wurde die Hauptkomponentenanalyse zur Mustererkennung verwendet und kann bezüglich der Untersuchungsergebnisse mit den Methoden der Clusteranalyse verglichen werden.

Tab. 9-1. Hauptkomponentenanalyse. - Gegenstand der Untersuchungen.

Nr.	Variablen	Anzahl Werte	Nr.	Variablen	Anzahl Werte
1	Na^+, K^+	16	4	Na^+/K^+	8
2	Mg^{2+}, Ca^{2+}	16	5	Mg^{2+}/Ca^{2+}	8
3	Cl^-, SO_4^{2-}	7	6	Cl^-/SO_4^{2-}	3

9.3.2 Untersuchungsergebnisse

Es wurde sich auf die (repräsentative) Auswahl von vier Hauptkomponentendarstellungen beschränkt, diese sind im nachfolgenden Abschnitt enthalten:

- Hauptkomponentenanalyse der Mg^{2+}- und Ca^{2+}-Ionen: s. Abb. 9-1
- Hauptkomponentenanalyse der Na^+- und K^+-Ionen: s. Abb. 9-2
- Hauptkomponentenanalyse der Cl^-- und SO_4^{2-}-Ionen: s. Abb. 9-3
- Hauptkomponentenanalyse der Verhältnisse von Na^+- zu K^+-Ionen: s. Abb. 9-4

9.3.3 Diskussion der Untersuchungsergebnisse

In den Hauptkomponentendarstellungen spiegeln sich die Untersuchungsergebnisse der entsprechenden Clusteranalysen, insbesondere der hierarchischen, gut wider, die diesbezüglichen Interpretationen sind zu einem großen Teil übertragbar.

- Hauptkomponentenanalysen der Ionenpaare (1. bis 3. Analyse)

  Die Objektdarstellung in der Ebene der ersten beiden Hauptkomponenten zeigt in allen drei Fällen eine dichte Gruppierung jeweils der Probenahmestellen des Sickerwassers und des Grundwassers. Die unterschiedlich hohen Salzfrachten führen - analog den Ergebnissen der Clusteranalysen - zu dieser Differenziertheit bei der Approximation der Messstellen.

  Die Hauptkomponentenanalyse der Mg^{2+}- und Ca^{2+}-Ionen führt dabei zu einer vergleichsweise breiten Streuung der Objekte, s. Abb. 9-1, die Probenahmestellen des Sickerwassers zeigen eine größere Ähnlichkeit mit denen des Grundwassers. Damit wird die im Zusammenhang mit den Ergebnissen der Untersuchungen zur hierarchischen Clusteranalyse geäußerte Annahme, dass die Schwerlöslichkeit des Calciumsulfats ($CaSO_4$) zu einem veränderlichen bzw. differentiellen Austrag der entsprechenden Schadstoffe führt, bestätigt.

  Alle drei Hauptkomponentendarstellungen zeigen, dass der SWP 10 aufgrund seiner hohen Belastung von sämtlichen Objekten auffallend weit entfernt ist, was auch auf den hoch belasteten GWB 5 in Bezug auf die anderen Grundwassermessstellen zutrifft.

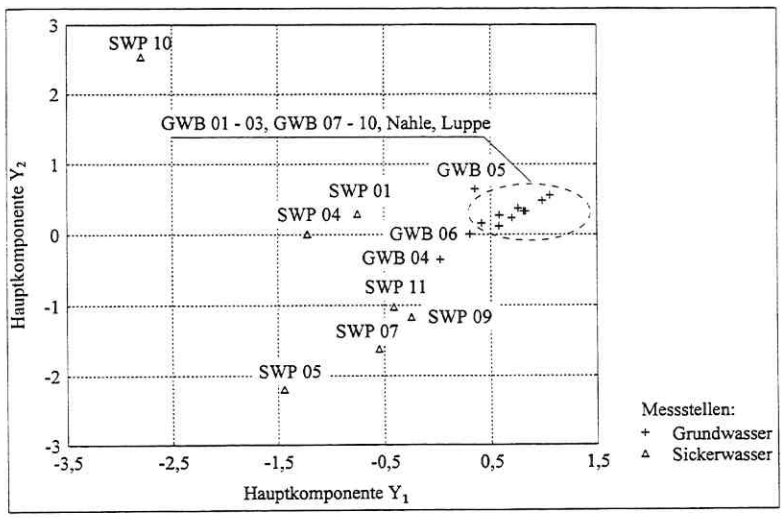

Abb. 9-1. Hauptkomponentendarstellung (Ausgangsvariablen: Mg^{2+}, Ca^{2+}).

In Bezug auf SWP 10 stellt es sich damit als problematisch dar, Ähnlichkeiten mit einzelnen Probenahmestellen des Grundwassers zu erkennen, um daraus Rückschlüsse auf differenzierte Transportrichtungen der Salzfrachten in diesem Deponiebereich zu ziehen. Aus der obigen Hauptkomponentendarstellung der Mg^{2+}- und Ca^{2+}-Ionen lässt sich aufgrund der relativ breiten Streuung der Objekte diesbezüglich noch am ehesten eine Aussage treffen. Der hier (relativ) geringe Abstand zwischen SWP 10 und GWB 5 wurde bei Anwendung der Clusteranalysen (dort allerdings bezüglich der Ionenverhältnisse) ebenfalls festgestellt und lässt eine entsprechende Fließrichtung des Sickerwassers von SWP 10 in Richtung Nahle zu GWB 5 vermuten.

Die Analysen führten zu einigen noch signifikanteren Approximationen der Objekte, so zeigt sich bei den Na^+- und K^+ und insbesondere bei den Cl^-- und SO_4^{2-}-Ionen ein nahes Beieinanderliegen von GWB 5 und SWP 7 bzw. SWP 9. In den entsprechenden Hauptkomponentendarstellungen - s. Abb. 9-2 und 9-3 - wird dies trotz der dichten Gruppierung jeweils der Sicker- und Grundwassermessstellen deutlich. Diese „Ähnlichkeit" zeigt sich auch in den Untersuchungsergebnissen der entsprechenden Clusteranalysen. Damit ist eine gesicherte Abschätzung entsprechender differenzierter Wasserwegsamkeiten möglich, d. h., der Schadstofftransport erfolgt entsprechend der Grundwasserfließrichtung von Südost (SWP 9) nach Nordwest (SWP 7), wobei es im nordwestlichen Deponiebereich (GWB 5) zu einem bevorzugten Austrag in das Grundwasser kommt.

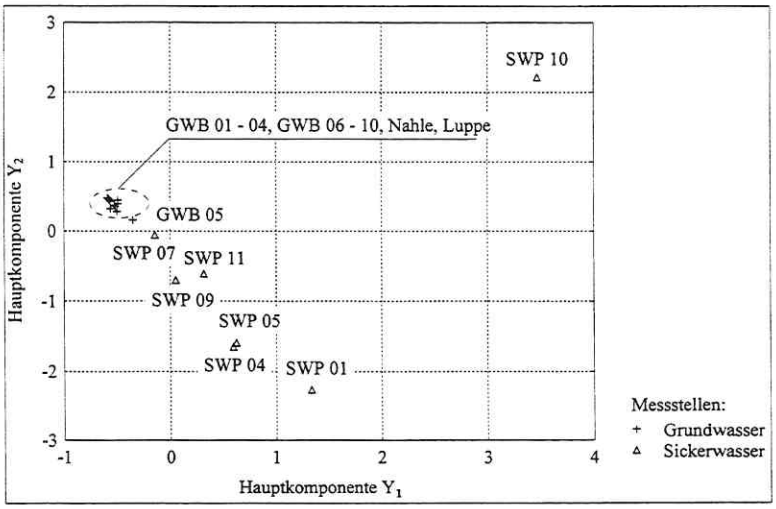

Abb. 9-2. Hauptkomponentendarstellung (Ausgangsvariablen: Na^+, K^+).

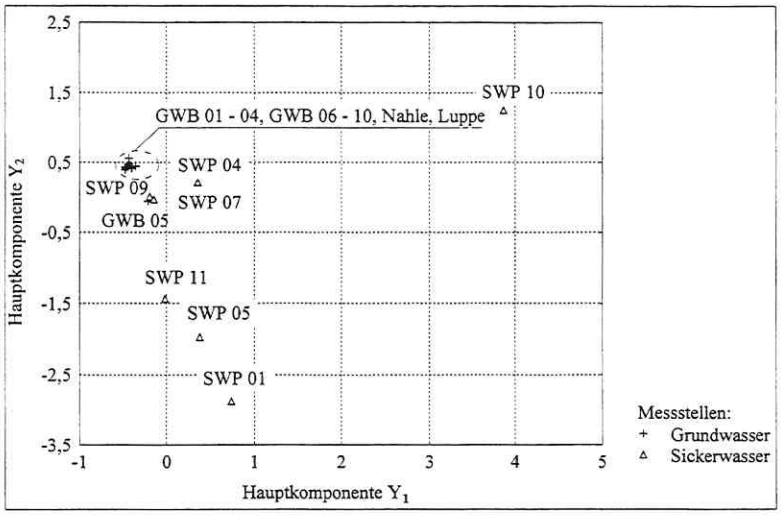

Abb. 9-3. Hauptkomponentendarstellung (Ausgangsvariablen: Cl^-, SO_4^{2-}).

- Hauptkomponentenanalysen der Ionenverhältnisse

  Der Kontrast zwischen den Gruppen der Sicker- und Grundwassermessstellen tritt hier bei weitem nicht so deutlich hervor. Abb. 9-4 zeigt dies exemplarisch für die Ionenverhältnisse Na^+/K^+. Des Weiteren zeigt sich in dieser Darstellung ein dichtes Beieinanderliegen von SWP 10 und GWB 5. Die bereits mehrfach geäußerte Annahme über eine entsprechende Transportrichtung des Sickerwasssers von SWP 10 in Richtung Nahle zu GWB 5 wird damit verstärkt. Weitere signifikante Approximationen gingen aus den Analysen nicht hervor.

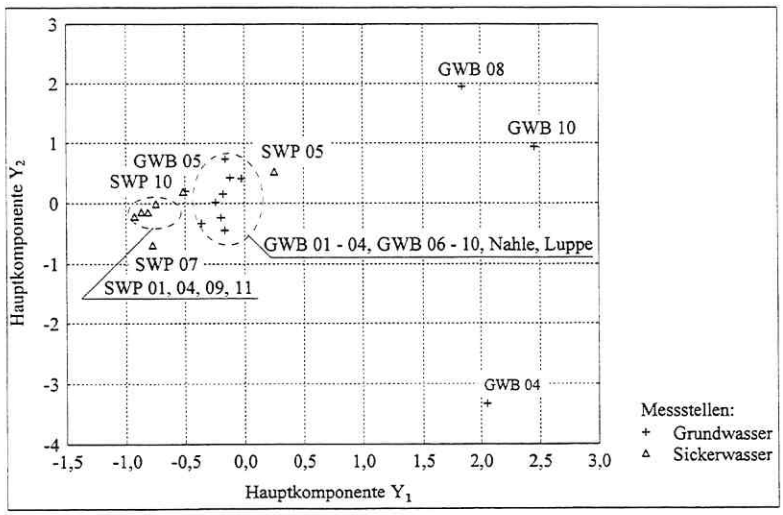

Abb. 9-4. Hauptkomponentendarstellung (Ausgangsvariablen: Na^+/K^+).

10 Zeitreihenanalyse

10.1 Vorbemerkungen

Bei den bisher durchgeführten Datenanalysen wurde davon ausgegangen, dass eine vorliegende (konkrete) Stichprobe $(x_1, \ldots, x_n)$ für ein Merkmal X als die Realisierung einer mathematischen Stichprobe $(X_1, X_2, \ldots, X_n)$ aus einer F-verteilten Grundgesamtheit X anzusehen ist, d. h., es wurde vorausgesetzt, dass die X_i $(i = 1, \ldots, n)$ voneinander unabhängig sind und alle die gleiche Verteilungsfunktion wie X besitzen. Die Betrachtungen werden in diesem Kapitel dahingehend erweitert, dass die zu untersuchende Stichprobe als eine zeitlich geordnete Folge von Beobachtungen eines quantitativen Merkmals angesehen wird. In diesem Fall bezeichnet man sie als eine **Zeitreihe** (x_t), $t \in T$. Es wird sich nachfolgend auf Zeitreihen mit diskreter Zeit beschränkt, wobei die Beobachtungszeitpunkte gleichabständig (äquidistant) sind. Für T gilt somit $T = \{1, 2, \ldots, n\}$, d. h., es liegt die Zeitreihe $(x_t) = (x_1, x_2, \ldots, x_n)$ vor. Diese ist als die endliche Realisierung eines **stochastischen Prozesses** anzusehen (STORM, 1995).

Ausgehend von einer Zerlegung der Zeitreihe in Komponenten unterscheidet man gemeinhin zwischen additiven und multiplikativen Zeitreihenmodellen. Erstere sind gegeben durch

$$x_t = m_t + k_t + s_t + e_t \quad \text{mit} \quad t \in T.$$

Hierin bedeuten:

- m_t ... Trend (langfristige systematische Änderung des mittleren Verlaufs der Zeitreihe)
- k_t ... Konjunkturkomponente (mehrjährige, oft wellenförmige Schwankung)
- s_t ... Saisonkomponente (regelmäßige zyklische Schwankung über die Saison, z. B. Jahr)
- e_t ... Restkomponente (kurzfristige, zufällig um null schwankende Einflüsse (Störungen))

Es wird zur Vereinfachung oftmals eine Zusammenfassung von Trend und Konjunkturkomponente zu einer glatten Komponente $g_t = m_t + k_t$ vorgenommen. Es existieren verschiedene Methoden zur Schätzung und Elimination der glatten und der saisonalen Komponente, zwei von diesen waren für die Untersuchungen relevant und werden nachfolgend vorgestellt. Die durch deren Beseitigung verbleibende Restkomponente e_t kann in vielen Fällen als die Realisierung eines **stationären Prozesses** angesehen werden. Durch nachfolgende Anpassung an spezielle (schwach) stationäre Prozesse lassen sich zum einen bestimmte Gesetzmäßigkeiten in der Zeitreihe erkennen und zum anderen können Vorhersagen über das zukünftige zeitliche Verhalten (= Prognosen) getroffen werden (STORM, 1995; EINAX et al., 1997).

Liegt eine Zeitreihe ohne Saisonkomponente vor, d. h. gilt $s_t \equiv 0$ und somit $x_t = g_t + e_t$, so besteht eine Möglichkeit der Trendbereinigung darin, zunächst mit Hilfe von **gleitenden Durchschnitten** (moving averages) eine stückweise lokale **Glättung** (smoothing) der Zeitreihe durchzuführen. Diese erfolgt durch Anwendung eines linearen Filters auf die Zeitreihe (x_t):

$$x_t^* = \sum_{j=-m}^{r} w_j \cdot x_{t+j} \quad \text{mit} \quad m, r \in N$$

Die reellen Konstanten w_j heißen Gewichte, die (Ursprungs-) Reihe (x_t) wird als Input und die gefilterte Reihe x_t^* als Output des Filters bezeichnet. Für einen einfachen gleitenden Durchschnitt (der ungeraden Ordnung $2 \cdot m + 1$, $m \in N$, $m < n$) gilt

$$r = m, \quad w_j = w_{-j} = \frac{1}{2 \cdot m + 1} \quad \text{für} \quad j = -m, \ldots, m \quad \text{und} \quad \sum_{j=-m}^{m} w_j = 1,$$

d. h., die Transformation erfolgt hier in

$$x_t^* = \frac{1}{2 \cdot m + 1} \cdot \sum_{j=-m}^{m} x_{t+j} \quad \text{für} \quad t = m+1, \ldots, n-m.$$

In diesem Fall handelt es sich um einen symmetrischen Filter, für den alle Gewichte gleich sind. Die geglättete Zeitreihe $x_{m+1}^*, \ldots, x_{n-m}^*$ besitzt nur noch $n - 2 \cdot m$ Werte, da an den Rändern ihres Wertebereichs jeweils m Werte wegfallen. Die Trendbereinigung (-elimination) wird anschließend durch entsprechende Subtraktion $x_t - x_t^*$ erreicht (STORM, 1995).

Enthält die Zeitreihe zusätzlich eine saisonale Komponente s_t, d. h. gilt $x_t = g_t + s_t + e_t$, so kann die Methode der gleitenden Durchschnitte in entsprechend modifizierter Form auch zu deren Elimination eingesetzt und damit zur lokalen Approximation der glatten Komponente g_t genutzt werden. Dabei wird vorausgesetzt, dass die Saisonkomponente s_t eine konstante und auf die Summe null normierte Saisonfigur mit bekannter Periodendauer s besitzt. Bildet man in diesem Fall **gleitende Durchschnitte der Ordnung s** (oder allgemeiner der Ordnung $k \cdot s$ mit $k \in N$), so bewirken diese, dass die Saisonkomponente verschwindet. Werden beispielsweise Quartalsdaten mit $s = 4$ verwendet, so ergibt sich

$$x_t^* = \frac{1}{4} \cdot \left(\frac{1}{2} \cdot x_{t-2} + x_{t-1} + x_t + x_{t+1} + \frac{1}{2} \cdot x_{t+2} \right),$$

wobei an den Rändern jeweils zwei Werte wegfallen. Die glatte Komponente wird dann durch x_t^* approximiert (STORM, 1995; SCHULZE, 1998).

Des Weiteren kann man die **Differenzenmethode** (STORM, 1995) zur Elimination der Saisonkomponente aus einer Zeitreihe anwenden, man bildet saisonale Differenzen 1. Ordnung:

$$x_t^* = \Delta x_t = x_t - x_{t-s} \quad \text{für} \quad t = s+1, \ldots, n \quad \text{(d. h. n -s Daten nach der Transformation)}$$

10.2 Modellierung einer Zeitreihe durch einen stochastischen Prozess

Wie im vorhergehenden Abschnitt bereits erwähnt, ist eine Zeitreihe (x_t) als die endliche Realisierung eines stochastischen Prozesses bzw. bei Beseitigung der glatten und der saisonalen Komponente als die Realisierung eines stationären Prozesses aufzufassen. Von derart unterschiedlichen Voraussetzungen ausgehend, kann deren Modellierung durch verschiedene stochastische Prozesse erfolgen, s. Tab. 10-1. STORM (1995) sowie EINAX et al. (1997) nehmen ausführliche Erläuterungen der hier genannten Prozessklassen einschließlich der ihnen zugrunde liegenden mathematischen Modellansätze vor.

Tab. 10-1. Modellierung von Zeitreihen durch parametrische Prozessklassen.

Zeitreihe enthält		Stationarität	Modellierung durch
m_t	s_t		
nein	nein	(schwach) stationär	MA (q)-Prozess
nein	nein	(schwach) stationär	AR (p)-Prozess
nein	nein	(schwach) stationär	ARMA (p, q)-Prozess
ja	nein	nichtstationär	ARIMA (p, d, q)-Prozess
nein	ja	nichtstationär	saisonaler ARMA (p, q)-Prozess mit Periode s
ja	ja	nichtstationär	multiplikativer saisonaler ARIMA-Prozess der Ordnung (p, d, q) × (P, D, Q)

Der multiplikative saisonale ARIMA (= autoregressive integrated moving average)-Prozess stellt die allgemeinste Form der genannten parametrischen Prozessklassen dar. Bei den im Rahmen vorliegender Arbeit durchgeführten Untersuchungen wurde eine optimale Anpassung von vier vorliegenden Zeitreihen an multiplikative saisonale (Periode s = 3) ARIMA-Prozesse der Ordnung (0, 0, 0) × (2, 1, 0) bzw. (0, 0, 0) × (3, 1, 0) ermittelt. Bei diesen Prozessklassen werden jeweils zur Saisonbereinigung die saisonalen Differenzen 1. Ordnung gebildet (D = 1) und für die so transformierte Reihe werden die im Modellansatz enthaltenen unbekannten Parameter berechnet bzw. geschätzt. Für einen Prozess der Ordnung (0, 0, 0) × (3, 1, 0) mit s = 3 bedeutet dies z. B.

$$X_t^* = C + \phi_1 \cdot X_{t-3}^* + \phi_2 \cdot X_{t-6}^* + \phi_3 \cdot X_{t-9}^* + \varepsilon_t^* \quad \text{mit} \quad X_t^* = X_t - X_{t-3},$$

d. h., hier erfolgt die Schätzung für C, ϕ_1, ϕ_2, und ϕ_3 und man erhält

$$\hat{x}_t^* = \hat{C} + \hat{\phi}_1 \cdot x_{t-3}^* + \hat{\phi}_2 \cdot x_{t-6}^* + \hat{\phi}_3 \cdot x_{t-9}^* .$$

Dabei ist (ε_t^*) ein reiner Zufallsprozess (= weißes Rauschen), zur Modellüberprüfung wird die Autokorrelationsfunktion der Residuen $\hat{e}_t^* = x_t^* - \hat{x}_t^*$ berechnet (s. Abschnitt 10.3.2).

10.3 Statistische Analyse stationärer Prozesse

10.3.1 Schätzung von Mittelwert- und Autokorrelationsfunktion

Unter der Voraussetzung, dass der stationäre Prozess mittelwert- und kovarianzergodisch ist, können für eine vorliegende Zeitreihe (x_t) erwartungstreu und konsistent geschätzt werden:

- der Mittelwert μ durch das zeitliche arithmetische Mittel

$$\bar{x} = \frac{1}{n} \cdot \sum_{t=1}^{n} x_t$$

- die Autokovarianzfunktion γ_k durch die empirische Autokovarianzfunktion

$$c_k = \frac{1}{n-1} \cdot \sum_{t=1}^{n-k} (x_t - \bar{x}) \cdot (x_{t+k} - \bar{x}) = c_{-k}$$

- die Autokorrelationsfunktion ρ_k durch die empirische Autokorrelationsfunktion

$$r_k = \frac{c_k}{c_0} = r_{-k} \qquad \text{für} \quad k = 0, \ldots, K \quad (K \text{ sollte im Bereich } k \leq K \leq \frac{n}{4} \text{ liegen})$$

- die von t unabhängige Varianz $\sigma^2 = \gamma_0$ durch die empirische Varianz

$$c_0 = \frac{1}{n-1} \cdot \sum_{t=1}^{n} (x_t - \bar{x})^2 \quad \text{(STORM, 1995)}$$

Im Rahmen der Untersuchungen erfolgte die Berechnung ($\bar{x}$, c_0) und graphische Darstellung (Korrelogramm bzw. Autokorrelationsfunktion $r_k = f(k)$) der charakteristischen Größen des jeweils betrachteten Prozesses. Zusätzlich zu der oben definierten (allgemeinen) Autokorrelationsfunktion (AKF) wurde jeweils auch die partielle Autokorrelationsfunktion (PAKF) dargestellt. Dies sind die partiellen Korrelationen einer Reihe mit sich selbst, d. h. mit ihren um eine bestimmte Anzahl von Beobachtungen (lag) zeitverschobenen Werten, wobei der Einfluss aller Korrelationen von lags niedrigerer Ordnung ausgeschaltet wird. Mit Hilfe eines Korrelogramms bzw. des in ihm eingetragenen asymptotischen Konfidenzintervalls kann überprüft werden, ob sich die (allgemeinen bzw. partiellen) Autokorrelationskoeffizienten r_k für k = 1, 2, ... signifikant von null unterscheiden, d. h. eine wesentliche Korrelation zwischen den Zufallsgrößen im zeitlichen Abstand k besteht (STORM, 1995; STATSOFT, 1996).

Des Weiteren gibt ein Vergleich mit den Autokorrelationsfunktionen insbesondere der einfachen MA (q)- oder AR (p)-Prozesse einen Hinweis darauf, welcher Prozess der vorliegenden Zeitreihe angepasst werden kann. EINAX et al. (1997) stellen die für einige Prozesse typischen Autokorrelationsfunktionen und partiellen Autokorrelationsfunktionen graphisch dar.

10.3.2 Anpassung eines multiplikativen saisonalen ARIMA-Prozesses

BOX und JENKINS (1976) schlagen eine Modellanpassung der Zeitreihe in drei Phasen vor:

1. Modellidentifikation

   In dieser Phase sind die (vorläufigen) Ordnungen p und q geeignet zu wählen. Der Vergleich des Verlaufs der empirischen Autokorrelationsfunktion bzw. der partiellen empirischen Autokorrelationsfunktion mit den entsprechenden Funktionen bestimmter Prozessklassen kann dabei bei einfachen MA (q)- oder AR (p)-Prozessen zur Orientierung genutzt werden. Insbesondere bei gemischten ARMA (p, q)-Prozessen ist diese Art der Wahl von p und q jedoch problematisch, weshalb die Modellselektion oftmals semiautomatisch vorgenommen wird: Aus einer Vielzahl von Modellen mit unterschiedlichen Ordnungen p und q erfolgt die Auswahl anhand bestimmter Gütekriterien, die sämtlichst auf der Minimierung der geschätzten Residuenvarianz $\hat{\sigma}_\varepsilon^2$ des Prozesses (ε_t) bezüglich p und q beruhen.

2. Modellschätzung

   Zur Schätzung der Modellparameter des ausgewählten Prozesses kam in den Untersuchungen das exakte Maximum-LIKELIHOOD-Verfahren nach MELARD (1984) zur Anwendung, wobei zur Minimierung der Summen der Fehlerquadrate das Quasi-NEWTON-Verfahren (FLETCHER und POWELL, 1963; FLETCHER, 1969) verwendet wurde.

3. Modellüberprüfung

   Ausgehend von den Parameterschätzungen ist die Güte der Modellanpassung an die Zeitreihe zu beurteilen. Hierzu werden die Residuen $\hat{e}_t$ berechnet bzw. deren Autokorrelationsfunktionen r_k für k = 1, 2, ... dargestellt. Das Modell ist dann geeignet, wenn die Residuenreihe $(\hat{e}_t)$ als die Realisierung eines reinen Zufallsprozesses (= weißes Rauschen) (ε_t) angesehen werden kann. In diesem Fall unterscheiden sich die r_k nicht wesentlich von null, sie liegen alle innerhalb des eingetragenen asymptotischen Konfidenzintervalls. Zur Überprüfung der entsprechenden Hypothese H_0 („Die Residuenreihe ist die Realisierung eines weißen Rauschens.“) wird ein von BOX und PIERCE (1970) angegebener Test verwendet. Der in den entsprechenden graphischen Darstellungen der Untersuchungsergebnisse angegebene Wert der Testgröße Q_k berechnet sich nach einer BOX-LJUNG-Q-Statistik (STATSOFT, 1996). Zu einem (in den Untersuchungen auf 0,05 festgelegten) Signifikanzniveau α wird zudem die Wahrscheinlichkeit p_0 dafür berechnet, mindestens Q_k zu erhalten, wenn H_0 richtig ist. Im Falle von $p_0 > \alpha$ wird H_0 angenommen, gilt $p_0 \leq \alpha$, so wird H_0 abgelehnt.

10.4 Prognose

10.4.1 Prognose durch exponentielle Glättung

Im Zusammenhang mit der Elimination des Trends und der Saisonkomponente wurde bereits über die stückweise lokale Glättung von Zeitreihen diskutiert, die sich mit Hilfe von gleitenden Durchschnitten durchführen lässt. Durch eine exponentielle (= geometrische) Glättung können bei trend- und saisonbereinigten Zeitreihen Prognosen getroffen werden, diese erreicht man durch Anwendung eines asymmetrischen Filters der Form

$$x_t^* = \sum_{j=0}^{\infty} w_j \cdot x_{t-j} \quad \text{mit} \quad w_j = \alpha \cdot (1-\alpha)^j \quad \text{für} \quad j = 0, 1, 2, \ldots \quad .$$

Der transformierte (geglättete) Wert

$$x_t^* = \alpha \cdot x_t + \alpha \cdot (1-\alpha) \cdot x_{t-1} + \alpha \cdot (1-\alpha)^2 \cdot x_{t-2} + \ldots$$

berechnet sich aus der letzten Beobachtung x_t und den vorangegangenen Werten x_{t-1}, x_{t-2} usw.. Der Glättungsfaktor α bestimmt den Einfluss zurückliegender Werte auf x_t^*. Je kleiner er gewählt wird, umso mehr Werte werden zur Glättung herangezogen. Ausgehend von einer Zeitreihe (x_t) mit $t = 1, \ldots, n$ ergibt sich z. B. eine 1-Schritt-Prognose $\hat{x}_{n,1}$ für den Wert x_{n+1} nach

$$\hat{x}_{n,1} = \alpha \cdot x_n + \alpha \cdot (1-\alpha) \cdot x_{n-1} + \ldots + \alpha \cdot (1-\alpha)^{n-1} \cdot x_1 \quad \text{(STORM, 1995).}$$

Zum Erhalten einer genauesten Prognose ist ein optimales α aus den Daten zu schätzen (GARDNER, 1985). In den Untersuchungen erfolgte eine automatische Suche nach dem besten α-Parameter über das Quasi-NEWTONsche Verfahren (FLETCHER und POWELL, 1963; FLETCHER, 1969) zur Funktionsminimierung. Bei diesem wird der mittlere quadratische Fehler als sog. Lack-of-Fit-Indikator (Indikator für Fehlanpassung) verwendet, der während des Schätzprozesses für den Parameter (als Startwert wurde $\alpha = 0{,}1$ gewählt) minimiert wird.

10.4.2 Prognose unter Verwendung eines angepassten ARIMA-Prozesses

Liegt der Zeitreihe ein multiplikativer saisonaler ARIMA-Prozess zugrunde, so kann man eine optimale lineare h-Schritt-Prognose $\hat{x}_{n,h}$ ($h \geq 1$) rekursiv mit Hilfe des Modellansatzes ermitteln. Dabei werden vor der Berechnung der Prognosewerte die (eventuell durchgeführten) Transformationen (Differenzenbildungen) "rückgängig" gemacht. Die Prognosewerte können dann im selben Maßstab wie die Werte der untransformierten Reihe interpretiert werden.

Wurde einer durch Differenzenbildung 1. Ordnung trendbereinigten Zeitreihe als Modell z. B. der AR (1)-Prozess angepasst (dies entspricht genau genommen einer Anpassung der unbereinigten Zeitreihe an einen ARIMA (1, 1, 0)-Prozess), so gilt nach der Rücktransformation

$$\hat{x}_t = a_0 + a_1 \cdot x_{t-1} + a_2 \cdot x_{t-2} \qquad \text{für} \quad t = 1, \ldots, n.$$

Die Prognosewerte $\hat{x}_{n,h}$ für x_{n+h}, $h \geq 1$, berechnen sich dann rekursiv wie folgt:

$$\hat{x}_{n,1} = a_0 + a_1 \cdot x_n + a_2 \cdot x_{n-1}$$

$$\hat{x}_{n,2} = a_0 + a_1 \cdot \hat{x}_{n,1} + a_2 \cdot x_n \quad \text{usw.}$$

Für $h \geq 2$ lässt sich somit allgemein schreiben

$$\hat{x}_{n,h} = a_0 + a_1 \cdot \hat{x}_{n,h-1} + a_2 \cdot \hat{x}_{n,h-2}.$$

Die Parameter eines angepassten multiplikativen saisonalen ARIMA-Prozesses sind i. Allg. unbekannt und müssen aus der Zeitreihe geschätzt werden. Dies hat zur Folge, dass die Prognosen nicht mehr optimal sind. Dass sich jedoch zumindest approximativ gute Vorhersagen treffen lassen, zeigen EINAX et al. (1997) anhand eines Beispieldatensatzes. Über die Prognosewerte hinaus lassen sich Prognoseintervalle zu einer (in den Untersuchungen auf 0,9 festgelegten) statistischen Sicherheit 1 -α berechnen (STATSOFT, 1996).

10.4.3 Auswertung von Prognoseergebnissen

Unter der Voraussetzung, dass zusätzlich zu den prognostizierten Werten x_{prog} einer Variablen X auch die tatsächlich gemessenen Werte x_{tat} zur Verfügung stehen, besteht eine relativ einfache und auch häufig in praktischen Anwendungsfällen (KAISER, 1995; KOBER, 1997; RUDOLPH, 1998b) genutzte Möglichkeit der Auswertung von Prognoseergebnissen in der Berechnung der Prognosegüte sowie weiterer charakteristischer Kenngrößen:

- absoluter Prognosefehler: $\varepsilon_{prog} = x_{tat} - x_{prog}$
- relativer Prognosefehler in Prozent: $\varepsilon_r = \frac{\varepsilon_{prog}}{x_{tat}} \cdot 100\% = \frac{x_{tat} - x_{prog}}{x_{tat}} \cdot 100\%$
- durchschnittlicher Prognosefehler in Prozent: $\bar{\varepsilon}_r = \frac{1}{n} \cdot \sum_{i=1}^{n} \varepsilon_{r,i}$
- mittlerer Prognosefehlerbetrag: $\left|\bar{\varepsilon}_{prog}\right| = \frac{1}{n} \cdot \sum_{i=1}^{n} \left|\varepsilon_{prog,i}\right|$
- mittlerer quadratischer Prognosefehler: $\tilde{\sigma} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\bar{\varepsilon}_{prog} - \varepsilon_{prog,i}\right)^2}$
- prozentuale Prognosegüte: $\lambda = \left(1 - \frac{3 \cdot \tilde{\sigma}}{x_{tat,max}}\right) \cdot 100\%$

10.5 Untersuchungen

10.5.1 Gegenstand und Zielstellung der Untersuchungen

Aufgrund des geringen Umfangs der zur Verfügung stehenden Daten jeweils einer Probenahmestelle sowie der unregelmäßigen (d. h. nicht äquidistanten) Zeitabstände zwischen den Probenahmen, s. Tab. 2-2, wurde der Gegenstand der Untersuchungen wie folgt festgelegt:

- Beschränkung auf die gemessenen Werte der Variablen pH-Wert, Leitfähigkeit, Temperatur und Wasserhärte des Grundwassers
- quartalsweise Zusammenfassung (arithmetische Mittelwertbildung) jeweils der Daten von GWB 1 - 3 (Abstrombereich der Luppe), GWB 4 - 6 (Abstrombereich der Nahle) und GWB 7 - 9 (Anstrombereich), wodurch für jedes Quartal quasi drei Werte simuliert wurden
- Beschränkung auf den Zeitraum vom 1. Quartal 1993 bis zum 3. Quartal 1995
- Ersetzung der in einer so entstandenen Zeitreihe fehlenden Werte durch arithmetische Mittelwertbildung der drei benachbarten Werte

Das aus diesen Festlegungen resultierende Datenmaterial ist in Tab. 10-2 zusammengefasst, die durch die Mittelwertbildung ersetzten Daten sind hierin unterstrichen dargestellt. Für alle vier Zeitreihen wurde eine Bereinigung von der (real so nicht vorhandenen, da durch die Unterteilung in drei Deponiebereiche entstandenen) Saisonkomponente (s = 3) durch Bildung saisonaler Differenzen 1. Ordnung vorgenommen.

Das Ziel bestand neben der statistischen Analyse der Zeitreihen darin zu ermitteln, inwieweit die vorgestellten Prognoseverfahren geeignet sind, Aussagen über das zukünftige zeitliche Verhalten der Merkmale zu treffen. Methodisch wurde dabei wie folgt vorgegangen: Die Zeitreihen enden mit dem 3. Quartal des Jahres 1995, s. Tab. 10-2. Die am 3. 11. 1997 an den GWB 1 - 9 gemessenen Werte wurden als repräsentativ für das 4. Quartal 1997 angesehen und durch Mittelwertbildung ebenfalls in die drei Bereiche GWB 1 - 3, GWB 4 - 6 und GWB 7 - 9 aufgeteilt. Damit war ein Vergleich der zu diesem Zeitpunkt beobachteten Werte mit den prognostizierten Werten möglich. Die Auswertung erfolgte über die Berechnung der besprochenen Prognosekenngrößen.

Aufgrund der vorgenommenen Datenmanipulierungen muss der überwiegend demonstrative Charakter der Untersuchungen hervorgehoben werden, sie sind als Versuchsmodell zu einer Zeitreihenanalyse der an der Deponie gemessenen Daten anzusehen.

Tab. 10-2. Zeitreihenanalyse. - Gegenstand der Untersuchungen.

Quartalszeitraum	pH-Wert	Leitf. (mS/cm)	Temperatur (°C)	Wasserhärte (°dH)
1/93 (GWB 1 - 3)	7,0883	3,1600	9,9333	59,6027
1/93 (GWB 4 - 6)	7,1167	3,7700	12,3333	74,0320
1/93 (GWB 7 - 9)	7,1267	1,8783	11,1500	43,6427
2/93 (GWB 1 - 3)	6,6594	3,0994	13,2778	61,5409
2/93 (GWB 4 - 6)	6,7300	3,5006	15,3444	61,5409
2/93 (GWB 7 - 9)	6,5019	1,9406	12,2313	61,5409
3/93 (GWB 1 - 3)	6,6850	3,1008	14,5833	60,4800
3/93 (GWB 4 - 6)	6,6167	4,1422	13,5167	83,8507
3/93 (GWB 7 - 9)	6,5300	1,9783	11,9500	47,6373
4/93 (GWB 1 - 3)	6,7444	2,8811	16,4000	65,6770
4/93 (GWB 4 - 6)	6,6467	3,5983	14,7667	65,6770
4/93 (GWB 7 - 9)	6,5083	1,9117	13,5000	65,6770
1/94 (GWB 1 - 3)	6,5622	2,8444	9,8222	53,8636
1/94 (GWB 4 - 6)	6,6889	6,2889	11,8556	91,9387
1/94 (GWB 7 - 9)	6,3956	2,3356	10,4222	56,2917
2/94 (GWB 1 - 3)	6,5744	3,4504	12,2389	62,8881
2/94 (GWB 4 - 6)	6,5744	3,4504	12,2389	62,8881
2/94 (GWB 7 - 9)	6,5744	3,4504	12,2389	62,8881
3/94 (GWB 1 - 3)	6,6000	2,9000	15,3333	53,2055
3/94 (GWB 4 - 6)	6,6667	4,4333	14,0000	76,8741
3/94 (GWB 7 - 9)	6,5333	1,9000	12,0000	45,1550
4/94 (GWB 1 - 3)	6,6333	3,2333	15,1667	59,2100
4/94 (GWB 4 - 6)	6,6833	4,8500	14,1667	100,6350
4/94 (GWB 7 - 9)	6,6333	2,1500	12,0000	52,7883
1/95 (GWB 1 - 3)	6,5167	2,4667	10,1667	48,7240
1/95 (GWB 4 - 6)	6,6250	5,8167	12,3333	100,5092
1/95 (GWB 7 - 9)	6,4417	2,6500	10,8333	63,2426
2/95 (GWB 1 - 3)	6,6667	1,7500	11,8833	32,0439
2/95 (GWB 4 - 6)	6,6167	3,9500	14,1667	69,7752
2/95 (GWB 7 - 9)	6,3950	2,8400	11,6833	64,7693
3/95 (GWB 1 - 3)	6,8033	0,9333	14,1000	19,1105
3/95 (GWB 4 - 6)	6,4750	3,8400	15,1000	70,0563
3/95 (GWB 7 - 9)	6,6133	2,7067	12,7667	65,0486

10.5.2 Untersuchungsergebnisse

Bei den Abbildungen wurde sich auf folgende exemplarische Auswahl beschränkt:

- Zeitreihenplot des pH-Wertes: s. Abb. 10-1
- saisonbereinigte Zeitreihe des pH-Wertes: s. Abb. 10-2
- AKF der saisonbereinigten Zeitreihe des pH-Wertes: s. Abb. 10-3
- PAKF der saisonbereinigten Zeitreihe des pH-Wertes: s. Abb. 10-4
- AKF der Residuen bei Anpassung eines ARIMA-Prozesses (pH-Wert): s. Abb. 10-5
- Anpassung eines ARIMA-Prozesses und Prognose (pH-Wert): s. Abb. 10-6

Die numerischen Untersuchungsergebnisse sind in den folgenden Tabellen zusammengefasst:

- Berechnung statistischer Kenngrößen: s. Tab. 10-3
- Vergleich der beobachteten mit den prognostizierten Werten (4. Quartal 1997): s. Tab. 10-4
- Berechnung der Prognosekenngrößen: s. Tab. 10-5

10.5.3 Diskussion der Untersuchungsergebnisse

Die graphische Darstellung der ursprünglichen Zeitreihen, s. Abb. 10-1 mit der Variable pH-Wert als Beispiel, lässt keine Trendkomponente erkennen bzw. durch die Untersuchungen konnte nicht ermittelt werden, dass bei einer Schätzung derselben sich diese sinnvoll interpretieren lässt. Eine Trendbereinigung wurde daher nicht durchgeführt. Eine Saisonbereinigung mit s = 3 hingegen war aufgrund der manipulierten Datenreihen erforderlich.

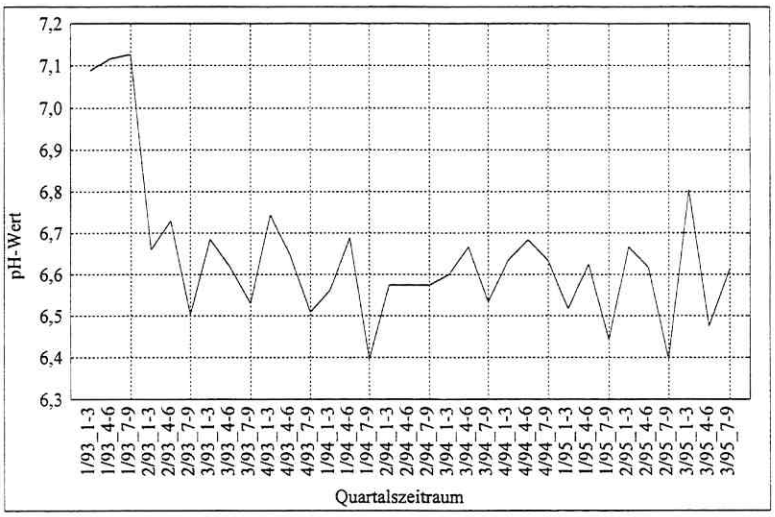

Abb. 10-1. Zeitreihenplot des pH-Wertes.

Für die durch Differenzenbildung 1. Ordnung saisonbereinigten Zeitreihen zeigte sich dann auch im Wesentlichen eine jahreszeitliche Konstanz nahe dem Nullwert, Abb. 10-2 verdeutlicht dies exemplarisch für den pH-Wert. Die berechneten arithmetischen Mittelwerte der saisonbereinigten Zeitreihen (x_t^*) (mit der reduzierten Anzahl von n = 30 Werten) bestätigen dies, sie liegen bei allen vier Variablen nahe null. Zum Vergleich sind in Tab. 10-3 auch die Kenngrößen der originalen Zeitreihen (x_t) (n = 33 Werte) eingetragen.

Tab. 10-3. Berechnung statistischer Kenngrößen.

	pH-Wert		Leitfähigkeit (mS/cm)		Temperatur (°C)		Wasserhärte (°dH)	
	(x_t)	(x_t^*)	(x_t)	(x_t^*)	(x_t)	(x_t^*)	(x_t)	(x_t^*)
$\bar{x}$	6,6433	-0,0480	3,1273	-0,0443	12,8334	0,2850	62,5092	-0,7687
c_0	0,0313	0,0329	1,3093	0,9791	2,9625	5,4613	278,4770	227,5187

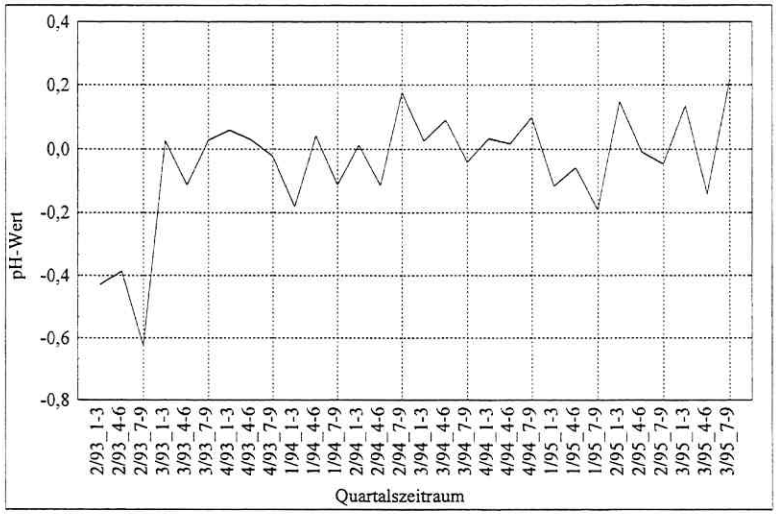

Abb. 10-2. Saisonbereinigte Zeitreihe des pH-Wertes.

Die berechneten Autokorrelationsfunktionen der saisonbereinigten Zeitreihen gestatteten es nicht, durch Vergleich mit den entsprechenden Funktionen bestimmter einfacher AR (p)- oder MA (q)-Prozesse die Ordnungen p und q geeignet zu wählen. Lediglich für den pH-Wert weisen die Bilder zumindest annähernd auf einen für den AR (2)-Prozess typischen Verlauf hin: abklingende AKF und fast verschwindende PAKF ab $k \geq 2$, s. Abb. 10-3 und 10-4. Die prototypische Darstellung hierfür kann EINAX et al. (1997) entnommen werden. Da zudem für den pH-Wert in der Phase der Modellidentifikation für $p = 2$ bei Vergleich mit den sich unmittelbar anschließenden p-Werten drei, vier und fünf die mit Abstand geringste Residuenvarianz ermittelt wurde, erfolgte die Wahl einer entsprechenden Anpassung. Bei den drei anderen Messgrößen erforderte die Modellidentifikation ein ähnlich heuristisches Vorgehen. In Abhängigkeit von verschiedenen p, q-Werten ($p, q \leq 5$) wurde sich für diejenige Kombination entschieden, für die die Residuenvarianz am geringsten war. Dies führte bei allen drei Variablen zu der Entscheidung für eine Anpassung an den AR (3)-Prozess. Die getroffene Wahl muss damit jedoch in allen Fällen als lediglich suboptimale Lösung angesehen werden.

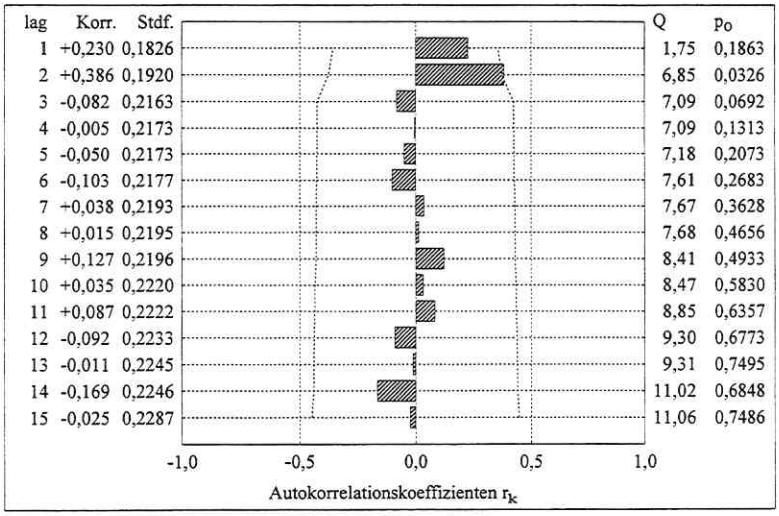

Abb. 10-3. AKF der saisonbereinigten Zeitreihe des pH-Wertes.

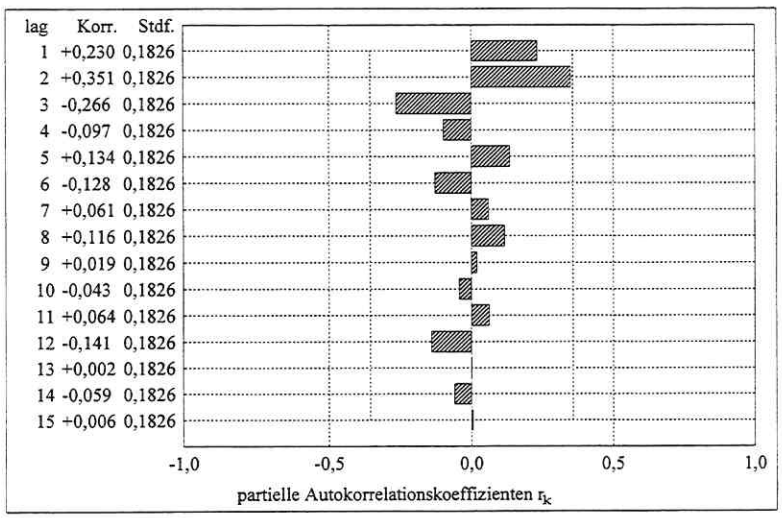

Abb. 10-4. PAKF der saisonbereinigten Zeitreihe des pH-Wertes.

Zur Überprüfung der Güte der (Prozess-) Modellanpassung wurden die Autokorrelationsfunktionen der Residuen herangezogen. Für die Variable pH-Wert wird deutlich, dass die Anpassung nicht optimal erfolgte, s. Abb. 10-5. Die r_k-Werte liegen hier zwar bis auf einen Ausreißer innerhalb des eingetragenen asymptotischen Konfidenzintervalls, jedoch muss die Hypothese dafür, dass die Residuenreihe als die Realisierung eines weißen Rauschens anzusehen ist, abgelehnt werden - die berechneten Wahrscheinlichkeiten p_0 liegen unter dem Signifikanz-

niveau von $\alpha = 0{,}05$. Bei den drei anderen Variablen hingegen hatte sich das heuristische Vorgehen als offensichtlich erfolgreich erwiesen. Die berechneten Wahrscheinlichkeiten p_0 liegen hier für alle k Verschiebungen ($1 \leq k \leq 15$) bei über 0,05, weshalb jede der Residuenreihen als die Realisierung eines weißen Rauschens angesehen werden kann.

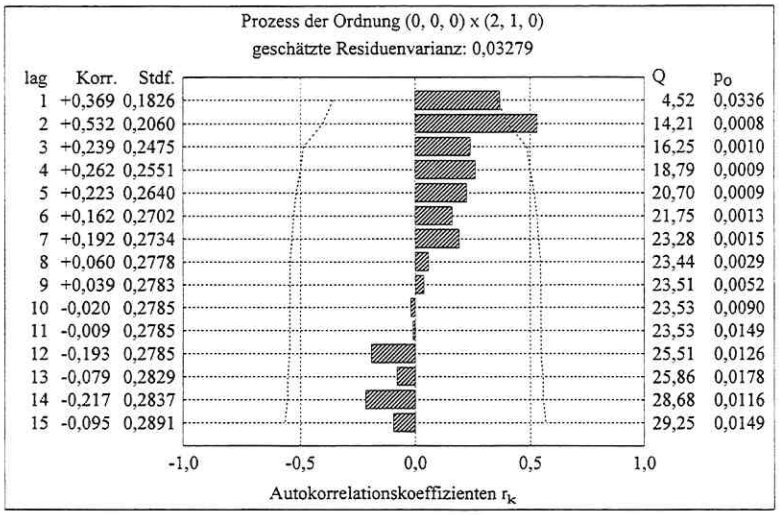

Abb. 10-5. AKF der Residuen bei Anpassung eines ARIMA-Prozesses (pH-Wert).

Die ermittelte Güte der Modellanpassung steht damit im Widerspruch zu den Prognosen, die unter Verwendung der angepassten Prozesse getroffen wurden. In Tab. 10-4 sind die beobachteten und prognostizierten Werte eingetragen und Tab. 10-5 kann entnommen werden, dass für den pH-Wert bei diesem Verfahren sogar die insgesamt beste Prognosegüte von 98,06 % vorliegt.

Tab. 10-4. Vergleich der beobachteten mit den prognostizierten Werten (4. Quartal 1997).

Bereich	pH-Wert			Leitfähigkeit (mS/cm)			Temperatur (°C)			Wasserhärte (°dH)		
	Beob.	Prognose		Beob.	Prognose		Beob.	Prognose		Beob.	Prognose	
		Glättung	Prozess		Glättung	Prozess		Glättung	Prozess		Glättung	Prozess
GWB 1 - 3	6,49	6,64	6,38	3,24	2,88	1,79	14,47	13,78	13,54	58,94	56,40	29,29
GWB 4 - 6	6,40	6,64	6,21	3,95	3,94	4,37	13,20	13,99	13,51	72,37	73,59	74,66
GWB 7 - 9	6,32	6,54	6,19	2,62	2,52	2,66	10,70	12,09	11,40	48,36	63,15	58,44

Tab. 10-5. Berechnung der Prognosekenngrößen.

Variable	Methode	$\bar{\varepsilon}_r$ (%)	$\lvert\bar{\varepsilon}_{prog}\rvert$	$\tilde{\sigma}$	λ (%)
pH-Wert	exponentielle Glättung	-3,2288	0,2063	0,0504	97,6703
	Prozessanpassung	2,2380	0,1432	0,0420	98,0586
Leitfähigkeit (mS/cm)	exponentielle Glättung	5,0692	0,1570	0,1843	85,9907
	Prozessanpassung	10,9464	0,6363	0,9894	24,7922
Temperatur (°C)	exponentielle Glättung	-4,7402	0,9534	1,0664	77,8857
	Prozessanpassung	-0,8616	0,6467	0,8490	82,3940
Wasserhärte (°dH)	exponentielle Glättung	-9,3219	6,1814	9,1162	62,2098
	Prozessanpassung	8,7622	14,0085	21,0543	12,7218

Die Güte der Modellanpassung bei Verwendung der exponentiellen Glättung ist prinzipiell schlechter gegenüber der, die bei Anpassung an eine der Prozessklassen erzielt wurde. Lediglich für den pH-Wert liegen die berechneten p_o über dem vorgegebenen Signifikanzniveau von $\alpha = 0,05$, die Modellanpassung erfolgte hier relativ gut. Diese Aussage bestätigt sich bei der Anwendung des Verfahrens zur Prognose - für den pH-Wert wurde hier bei Vergleich mit den drei anderen Messgrößen die beste Prognosegüte (97,67 %) erzielt.

Unter dem Aspekt, dass mit beiden Verfahren relativ langfristige Prognosen getroffen wurden, kann deren Leistungsfähigkeit als gut eingeschätzt werden. Bis auf zwei Ausreißer (Prognose für die Variablen Leitfähigkeit und Wasserhärte bei Prozessanpassung) liegt die ermittelte Prognosegüte bei über 60 %, für den pH-Wert bei beiden Methoden sogar bei ca. 98 %, s. Tab. 10-5 sowie Abb. 10-6. Der pH-Wert, der den Gehalt an freiem CO_2 (hier im Grundwasser) widerspiegelt, ist aus Anwendersicht für eine Prognose in jedem Fall von Interesse und sowohl das Verfahren der exponentiellen Glättung als auch das der Prozessanpassung lassen sich für diesen Parameter besonders gut anwenden.

Ein Vergleich der Prognosegüte beider Verfahren zeigt, dass diese für die Variablen pH-Wert und Temperatur annähernd gleich gut sind. Davon ausgehend sollte im praktischen Anwendungsfall die exponentielle Glättung vorgezogen werden. Bei dieser ist in der Phase der Modellidentifikation ein wesentlich geringerer Aufwand vonnöten als bei der Prozessanpassung, wo erst eine geeignete Ordnung der Parameter p und q heuristisch (und somit von Hand) ermittelt werden muss.

Es ist problematisch, eine generelle Aussage darüber zu treffen, inwieweit die beiden Verfahren für derartige langfristige Prognosen geeignet sind bzw. welches eine höhere Prognosegenauigkeit besitzt. Die Gründe hierfür liegen darin, dass zum einen das bereits erwähnte Problem des geringen Umfangs der zur Verfügung stehenden und zudem manipulierten Daten be-

stand und zum anderen die Berechnung der Prognosekenngrößen auf der Grundlage von lediglich drei zur Verfügung stehenden Wertepaaren erfolgte. Aufgrund von letzterem ist bereits ein relativ großer Wert des absoluten Prognosefehlers verantwortlich für eine extreme Verschlechterung der Prognosegüte, wie sich für die Variablen Leitfähigkeit und Wasserhärte bei Anwendung des Verfahrens der Prozessanpassung zeigt, s. Tab. 10-4 und 10-5.

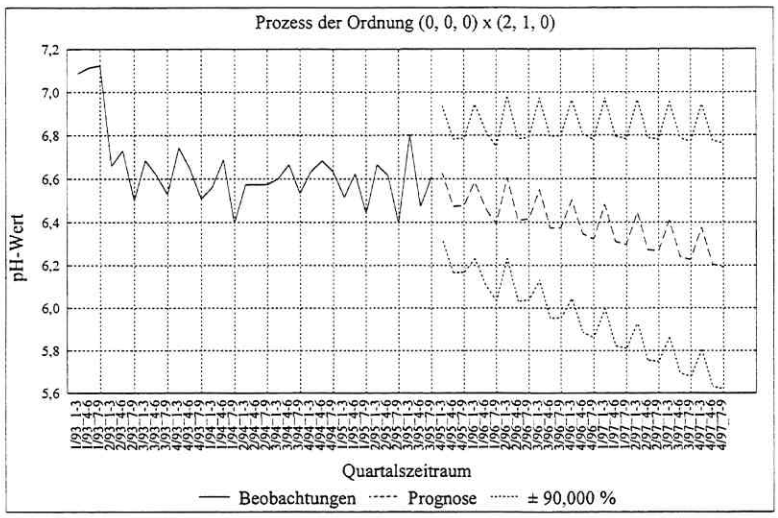

Abb. 10-6. Anpassung eines ARIMA-Prozesses und Prognose (pH-Wert).

Abschließend sei erwähnt, dass die beiden angewandten klassischen Prognoseverfahren nur einen sehr kleinen Ausschnitt des auf diesem Gebiet zur Verfügung stehenden Methodenspektrums darstellen. In den letzten Jahren wurden dabei zunehmend Verfahren entwickelt, die auf dem Konzept der Künstlichen Neuronalen Netze beruhen (s. hierzu auch die Ausführungen des nachfolgenden Kapitels). Diese lassen sich insbesondere bei Vorliegen umfangreicher Datenmengen (hohe Anzahl von Eingangsvariablen und große Tiefe des Datensatzes) hervorragend anwenden und werden daher beispielsweise sehr häufig von Energieversorgungsunternehmen zur Lastprognose eingesetzt (NEUMANN und ZIELONKA, 1992; HEINRICH, 1995). Weiterhin lassen sich mit ihnen als Werkzeug für Klimaprognosen globale Klimaänderungen wesentlich schneller simulieren, als mit herkömmlichen Modellen (SCHIELE-TRAUTH, 1997). Diese und zahlreiche weitere Anwendungsbeispiele belegen: Unter der Voraussetzung einer zeitkontinuierlichen Messung der Parameter am Deponiekörper lassen derartige Verfahren auch hier einen erfolgreichen Einsatz erwarten.

11 Datenanalyse durch Anwendung einer wissensbasierten Methode

11.1 Künstliche Intelligenz

Die bisher betrachteten klassischen Methoden der algorithmischen Datenverarbeitung stoßen dort an ihre Grenzen, wo Intelligenz oder komplexes Wissen für die Problemlösung unverzichtbar sind. Die systematische Repräsentation und Verarbeitung von Wissen aller Art, auch von informellem Alltagswissen, gehört zu den zentralen Themen eines Teilgebiets der Informatik, welches man als **Künstliche Intelligenz** bezeichnet (HERING et al., 1995).

Mit dem Begriff der Künstlichen Intelligenz lassen sich eine Vielzahl praktischer Anwendungsfälle assoziieren, deren Spektrum relativ weitreichend ist (LUTHER, 1997; LÖHN und SCHAEFER, 1998; MARTIN, 1998). Daher soll er zunächst eingegrenzt werden. Eine präzise Definition gibt DILGER (1997, S. 5) an: „Künstliche Intelligenz ist der Versuch, ein Artefakt zu konstruieren, das den TURING-Test sicher besteht." Unter einem Artefakt ist dabei ein System physikalischer Symbole zu verstehen, d. h. eine abstrakte Maschine, die Symbole (z. B. Bits, Buchstaben des Alphabets) manipulieren kann. Der TURING-Test geht auf einen Vorschlag des Mathematikers Alan M. Turing zurück, wonach einem künstlichen System dann Intelligenz zugesprochen werden kann, wenn eine Testperson es in einem „Gespräch" über Tastatur und Bildschirm nicht von einem Menschen unterscheiden kann (DENGEL, 1994).

Den Gegenstand der Künstlichen Intelligenz bilden drei Dinge: Suche, Wissensrepräsentation und Wissensverarbeitung. **Suche** dient dazu, in einem großen Suchraum (darstellbar als Suchgraph), in dem es an jedem Punkt der Problemlösung in der Regel mehrere Lösungen für den nächsten Lösungsschritt gibt, einen Weg von der Problemstellung zu einer Lösung zu finden. Da die Größe des Suchraums oft exponentiell mit der Größe des Problems wächst (sog. „kombinatorische Explosion"), ist es i. Allg. nicht möglich, die Suchräume vollständig zu erkunden. Durch Verwendung heuristischer Suchverfahren wird der Suchraum in einer optimalen oder nahezu optimalen Reihenfolge durchlaufen. Der Zweck der **Wissensrepräsentation** ist es, Fachwissen menschlicher Experten im Rechner darstellbar und verarbeitbar zu machen. Eine gewissermaßen universelle Wissensrepräsentationssprache stellt die Logik dar. Die **Wissensverarbeitung** beschäftigt sich mit der Anwendung der Suchmethoden und der Wissensrepräsentation zur Lösung von Problemen, die man als intelligent bezeichnet, z. B. Bilderkennen, Sprachverstehen u. a. (DILGER, 1997).

11.2 Systeme der Künstlichen Intelligenz

11.2.1 Überblick

Unter „Systemen der Künstlichen Intelligenz" sind an dieser Stelle genau genommen Typen von Systemen zu verstehen, nicht einzelne konkrete Systeme. Nachfolgend werden exemplarisch einige der bekanntesten Systemtypen aufgeführt.

Der Zweck von **Expertensystemen** besteht in der Übertragung bestimmter Fähigkeiten menschlicher Experten auf den Computer. Hierzu sind zwei Dinge erforderlich, die Darstellung von Expertenwissen im Computer und das Verarbeiten des Expertenwissens durch den Computer, d. h. insbesondere das Ziehen von Schlüssen, wie es der menschliche Experte macht (DILGER, 1997).

Ein weiterer Systemtyp sind die **Multiagentensysteme**. Agenten sind nach Definition der FIPA (Foundation for Intelligent Physical Agents) autonome, proaktive und intelligente Software-Einheiten, die im Auftrag ihrer Benutzer bestimmte Aufgaben vollführen. Dabei bedeutet Autonomie, dass sie sich von anderen Agenten Wissen und Ressourcen verschaffen und selbsttätig planen, auf welche Weise sie die von ihnen gesteckten Ziele erreichen können. Im typischen Fall sind Agenten Teil eines umfassenden Multi-Agenten-Systems, mit dessen anderen Angehörigen sie fortlaufend in Wechselwirkung treten (SCHMIDT, 1997).

Des Weiteren seien ohne zusätzlichen Kommentar **Planungssysteme** und **Verstehen natürlicher Sprache** genannt (DILGER, 1997).

Zwei Systemtypen lassen sich (u. a.) im Bereich der Datenanalyse für die Musterklassifikation (= überwachte Klassifikation) bzw. Mustererkennung (= automatische Klassifikation) hervorragend einsetzen. Die **Fuzzy-Systeme** besitzen dabei sowohl den Vorteil der Interpretierbarkeit, da sie auf linguistischen WENN-DANN-Regeln beruhen, als auch den der Initialisierbarkeit (mit a priori-Wissen). Bei den **Neuronalen Netzen** ist deren Lernfähigkeit hervorzuheben. Die Vorteile des einen Modells sind gerade die Nachteile des anderen und umgekehrt. Durch Kopplungen von Neuronalen Netzen mit Fuzzy-Systemen können die Nachteile der beiden Modelle aufgehoben werden. Sie führen zu interpretierbaren, lernfähigen Systemen, die zudem mit problemspezifischem Vorwissen initialisiert werden können, den **Neuronalen Fuzzy-Systemen**. Das in Abschnitt 11.3 vorgestellte NEFCLASS-Modell ist ein hybrides Neuro-Fuzzy-System zur Musterklassifikation und kam in den Untersuchungen zum Einsatz.

11.2.2 Neuronale Netze

Neuronale Netze bilden streng genommen keinen eigenen Systemtyp der Künstlichen Intelligenz, sondern werden den **lernenden Systemen** zugeordnet. Bei diesen wird unterschieden zwischen denen, die auf der symbolischen Ebene lernen und denen, die auf der subsymbolischen Ebene lernen. Letztere sind die (Künstlichen) Neuronalen Netze (DILGER, 1997).
Mit Neuronalen Netzen wird versucht, den Aufbau und die Funktionsweise des menschlichen Gehirns nachzuahmen. Ihr allgemeiner Aufbau (Netzwerkstruktur) entspricht dem eines gerichteten Graphen. Dessen Knoten sind die Prozessoreinheiten, die kleine informationsverarbeitende Einheiten sind. Jede Prozessoreinheit kann mehrere Ein- und Ausgänge, d. h. eingehende und ausgehende Kanten, haben. Die Prozessoreinheiten sind innerhalb eines Netzes in mehreren Schichten angeordnet, wobei es eine ausgezeichnete Eingabe- und Ausgabeschicht (input and output layer) gibt. Die Kanten verbinden meistens Prozessoreinheiten benachbarter Schichten (sog. vorwärtsbetriebene Netze), es kann aber auch andere Verbindungen oder Schleifen geben (DILGER, 1997).
Die Netzwerkstruktur ist während einer sog. Lern- oder Konditionierungsphase veränderbar. In dieser wird mit Hilfe eines Lernalgorithmus (= Lernregel) versucht, die Netzwerkstruktur derart zu bestimmen, dass das System auf bestimmte Anfangszustände oder Eingaben mit bestimmten Endzuständen oder Ausgaben reagiert. Für die Durchführung eines Lernvorganges ist neben einem geeigneten Lernalgorithmus die Vorgabe einer Lernaufgabe erforderlich, hierbei unterscheidet man zwischen festen und freien Lernaufgaben und dementsprechend zwischen überwachten und nicht überwachten Lernalgorithmen. Um für die verschiedenen Netztypen eine einheitliche Notation zu erreichen, führen NAUCK et al. (1996a) ein generisches Modell ein, das den formalen Rahmen für Neuronale Netze festschreibt.
Neuronale Netze besitzen zwei entscheidende Vorteile, zum einen ihre bereits erwähnte Lernfähigkeit und zum anderen ihre Fehlertoleranz, d. h., sie können auf gestörte oder unvollständige Muster reagieren. Der Nachteil ist, dass das in einem trainierten Netz enthaltene Wissen verteilt gespeichert ist und nicht in expliziter Form extrahiert werden kann (sog. „Black-Box-Verhalten"). Aus diesem Grund können sie auch nicht mit eventuell vorhandenem a priori-Wissen initialisiert werden. Ein weiteres Problem besteht darin, dass Parameter wie die Anzahl der Neuronen, ihre Verbindungen untereinander usw. zumeist nur experimentell oder auf der Basis von „Faustregeln" ermittelt werden können (NAUCK et al., 1996a).

11.2.3 Fuzzy-Systeme

Die Fuzzy-Logik basiert auf Theorien, die von ZADEH (1965) veröffentlicht wurden. Die Grundidee besteht darin, die klassische zweiwertige Modellierung von Konzepten und Eigenschaften (Prädikaten) wie z. B. hoch, niedrig, heiß, kalt im Sinne gradueller Erfüllung zu erweitern. Eine Temperatur z. B. kann nicht mehr nur als heiß oder nicht heiß angesehen werden, sondern das Prädikat heiß zu einem gewissen Grad zwischen null und eins erfüllen.
Fuzzy-Logik ist somit eine unscharfe Logik in dem Sinne, dass mit ihrer Hilfe unscharfe Eingangsbedingungen zu einer exakten Schlussfolgerung führen. Ihre drei wichtigsten Einsatzbereiche sind Fuzzy-Control (Regelungs-Systeme), Fuzzy-Logic (Expertensysteme) und Fuzzy-Analysis (Datenanalyse) (HERING et al., 1995).
Die Fuzzy-Set Theorie beinhaltet die Grundlagen zur Fuzzy-Logik, in ihr werden Begriffe wie linguistische Variablen, Terme, Basisvariablen, Zugehörigkeitsgrad und Zugehörigkeitsfunktion sowie Inferenz verwendet. Eine linguistische Variable (z. B. Temperatur) wird durch Terme (z. B. niedrig, hoch) beschrieben, mit Hilfe der Basisvariablen (Wertebereich der Variablen im abgeschlossenen Intervall) werden die Terme inhaltlich definiert, indem sie auf dieser abgebildet und gewichtet werden. Die geschlossene Kurve der so entstehenden Zugehörigkeitsgrade ergibt die Zugehörigkeitsfunktion. Auf diese Weise kann eine Eingangsgröße auf unscharfe bzw. Fuzzy-Mengen abgebildet werden. Unter Inferenz oder Ableitung versteht man den Übergang von einem Eingangszustand zu einem Endzustand aufgrund geltender Bedingungen. In Fuzzy-Systemen werden hierfür WENN-DANN- (if-then-) Regeln verwendet, diese entsprechen formal einer logischen Implikation $A \rightarrow B$. In den Prämissen sind die mit den linguistischen Termen assoziierten Fuzzy-Mengen der Eingangsgrößen enthalten, diese können ggf. durch die logischen Operatoren UND bzw. ODER miteinander verknüpft sein. Die Konklusion gibt einen Ausgabewert an. Bei regelungstechnischen Anwendungen wird die Ausgangsgröße (= Stellgröße) ebenfalls auf Fuzzy-Mengen abgebildet, mittels eines Defuzzyfizierungsverfahrens wird hier ein exakter Wert abgeleitet (HERING et al., 1995).
Fuzzy-Systeme haben die Vorteile, dass sie aufgrund der verwendeten linguistischen WENN-DANN-Regeln interpretierbar sind und dass sie mit a priori-Wissen über die Problemstellung initialisiert werden können. Problematisch hingegen ist die Festlegung konkreter Werte zwischen null und eins als Zugehörigkeitsgrade, die angeben, inwieweit ein Objekt (bzw. Element einer Grundmenge, z. B. ein Messwert) ein Konzept erfüllt (NAUCK et al., 1996a).

11.2.4 Neuronale Fuzzy-Systeme

Durch Kopplungen von Neuronalen Netzen mit Fuzzy-Systemen werden die Nachteile der beiden Modelle, das Black-Box-Verhalten Neuronaler Netze und die Schwierigkeiten der Festlegung konkreter Zugehörigkeitswerte bei Fuzzy-Systemen, aufgehoben. Sie führen, wie eingangs bereits erwähnt, zu Neuronalen Fuzzy-Systemen, lernfähigen, interpretierbaren Modellen, die mit unsicheren Informationen arbeiten und in die zudem problemspezifisches Vorwissen integriert werden kann. Die Formen der Kombination werden grob in zwei Klassen gegliedert. Bei **kooperativen** Neuro-Fuzzy-Systemen arbeiten Neuronale Netze und Fuzzy-Systeme grundsätzlich unabhängig voneinander. Die Kopplung besteht darin, dass das Neuronale Netz einige Parameter des Fuzzy-Systems im Sinne einer Optimierung erlernt. Kooperative Systeme können nur eines erlernen, entweder die unscharfen Mengen bei festgelegten Regeln oder die Regeln bei festgelegten unscharfen Mengen. Die **hybriden** Neuro-Fuzzy-Systeme hingegen basieren auf einer einheitlichen Struktur, die sich sowohl als Neuronales Netz als auch als Fuzzy-System interpretieren lässt. Sie können beides zugleich erlernen, sowohl die Regeln als auch die Mengen (NAUCK et al., 1996a; KRUSE et al., 1997).

11.3 Das NEFCLASS-Modell

11.3.1 Das formale Modell

Das NEFCLASS-Modell (NEuro Fuzzy CLASSification) ist ein hybrides Neuro-Fuzzy-System. Es beruht auf dem generischen Modell eines dreischichtigen Fuzzy-Perceptrons, welches die Architektur eines üblichen Multilayer-Perceptrons besitzt, und dient zur Bestimmung der Klasse eines gegebenen Eingabemusters. Dabei wird eine feste Lernaufgabe verwendet, d. h., die Ausgabe zu einem bestimmten Eingabemuster ist bekannt. Die zugehörigen überwachten Lernalgorithmen erfüllen die Aufgabe der Musterklassifikation (NAUCK et al., 1996a).
Ein dreischichtiges Fuzzy-Perceptron ist ein vorwärtsbetriebenes Neuronales Netz mit mehreren definierten Spezifikationen. Es kann als eingeschränkte „Fuzzyfizierung" eines herkömmlichen Multilayer-Perceptrons interpretiert werden, da lediglich die Gewichte der Verbindungen sowie die Netzeingabe und die Aktivierung der Ausgabeeinheiten mit Fuzzy-Mengen modelliert werden. NAUCK et al. (1996a und 1996b) geben die formale Definition eines

NEFCLASS-Modells an. Abb. 11-1 zeigt exemplarisch ein NEFCLASS-System mit zwei Eingaben, fünf (linguistischen) Regeln und zwei Klassen, wobei für die beiden Eingangsmerkmale jeweils eine Fuzzy-Partitionierung auf drei Fuzzy-Mengen verwendet wird.

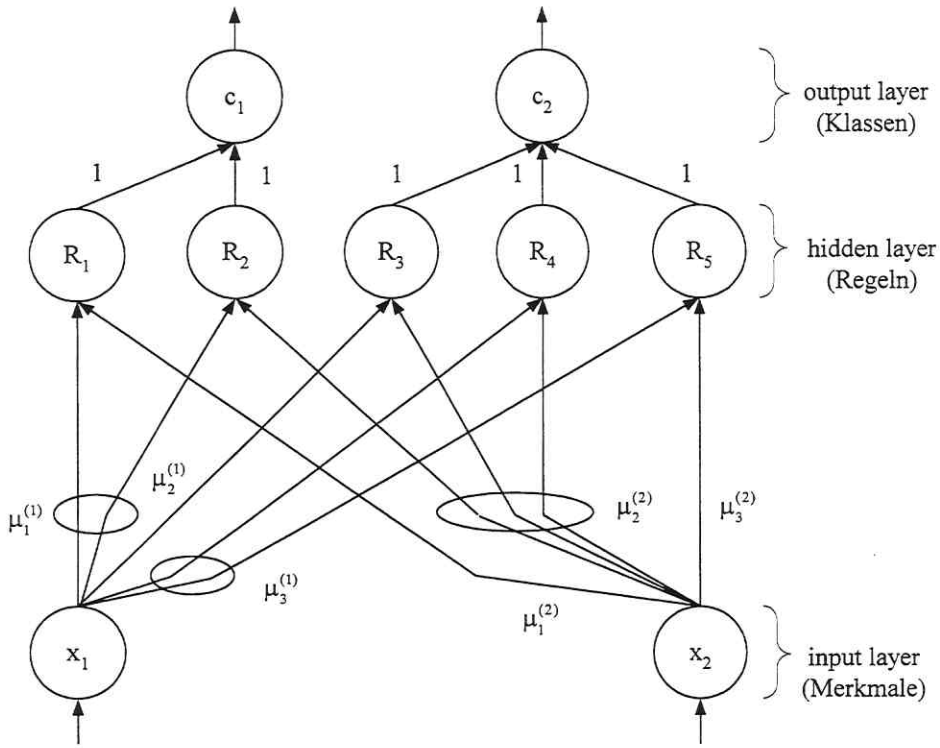

Abb. 11-1. Beispiel für ein NEFCLASS-System.

Für die Musterklassifikation können sowohl klassische statistische Verfahren (s. Kapitel 7) als auch Neuronale Netze eingesetzt werden. Die Verwendung eines Neuro-Fuzzy-Ansatzes wie NEFCLASS hat gegenüber diesen zwei entscheidende Vorteile:

- die Interpretierbarkeit des Klassifikators durch die Verwendung sog. Fuzzy-Regeln
- die Möglichkeit der Initialisierung des Systems mit (vollständigem oder partiellem) a priori-Wissen in Form von Fuzzy-Regeln

Die in einem NEFCLASS-System verwendeten Regeln haben die Form

R_r: **if** x_1 is $A^{(1)}_{j_1,r}$ and ... and x_n is $A^{(n)}_{j_n,r}$ **then** $(x_1 \ ... \ x_n) \in C_l$.

Hierin bedeuten:

$X = (x_1 \ ... \ x_n) \in R^n$ ... Eingabemuster, bestehend aus x_i (i = 1, ..., n) Komponenten

$C_l \subseteq R^n$ ... Menge von Mustern, die zur Klasse l gehören

$A^{(1)}_{j_1,r}, \ldots, A^{(n)}_{j_n,r}$ ... linguistische Terme (z. B. small, large), die durch Fuzzy-Mengen $\mu^{(1)}_{j_1,r}, \ldots, \mu^{(n)}_{j_n,r}$ repräsentiert werden

Die Fuzzy-Mengen werden dadurch erzeugt, dass der Wertebereich eines jeden Merkmals X_i ($i = 1, \ldots, n$) durch $j_i = 1, \ldots, q_i$ Fuzzy-Mengen $\mu^{(i)}_1, \ldots, \mu^{(i)}_{q_i}$ partitioniert wird. Die gesamte Regelbasis des Systems besteht aus $r = 1, \ldots, k$ solchen Fuzzy-Regeln $R_1, \ldots, R_k$.

Das NEFCLASS-System kann vollständig durch ein Lernverfahren erzeugt werden. Zur Initialisierung sind für jede Eingabegröße eine initiale Fuzzy-Partitionierung vorzugeben und die maximale Anzahl k_{max} von Regeleinheiten in der inneren Schicht festzulegen. In einem ersten Schritt werden zunächst durch Analyse der Musterdaten die $k \leq k_{max}$ Regeleinheiten erlernt (= Regellernalgorithmus) und anschließend wird deren Klassifikationsleistung dadurch verbessert, dass die Parameter der den linguistischen Termen zugeordneten Fuzzy-Mengen optimiert werden (= Lernalgorithmus für Fuzzy-Mengen). Es besteht jedoch auch die Möglichkeit, das System durch Vorgabe einer partiellen (welche durch das System dann ergänzt wird) oder vollständigen Regelbasis aufzubauen (Initialisierung mit a priori-Wissen) und anschließend wiederum durch Nachtrainieren der Fuzzy-Mengen in den Regelprämissen zu verbessern (Adaption des a priori-Wissens) (NAUCK et al., 1996a).

11.3.2 Das Propagationsverfahren

Zur Auswertung jeder Regel R_r der Regelbasis wird zunächst der Erfüllungs- oder Akzeptanzgrad α_r bestimmt, zu dem die Prämisse bei dem betrachteten Eingabemuster $X_0 = \left(x^{(0)}_1 \ \ldots \ x^{(0)}_n\right)$ erfüllt ist. Zu diesem Zweck wird für $i = 1, \ldots, n$ der Wert $\mu^{(i)}_{j_i,r}\left(x^{(0)}_i\right)$ berechnet, der den Zugehörigkeitsgrad der Merkmalsausprägung $x^{(0)}_i$ zu der mit dem linguistischen Term $A^{(i)}_{j_i,r}$ assoziierten Fuzzy-Menge $\mu^{(i)}_{j_i,r}$ angibt. In der Prämisse der Regel R_r wird verlangt, dass jede Merkmalsausprägung $x^{(0)}_1, \ldots, x^{(0)}_n$ den zugehörigen linguistischen Term $A^{(1)}_{j_1,r}, \ldots, A^{(n)}_{j_n,r}$ erfüllen muss, daher müssen die Werte $\mu^{(i)}_{j_i,r}\left(x^{(0)}_i\right)$ ($i = 1, \ldots, n$) in geeigneter Weise konjunktiv verknüpft werden. Da eine Schlussfolgerung nicht stärker sein kann als ihre Voraussetzungen, wird als Erfüllungsgrad der Regel R_r das Minimum der Erfüllungsgrade der Voraussetzungen verwendet. Bezüglich eines Eingabemusters X_0 berechnet er sich somit zu

$$\alpha_r = \min\left\{\mu^{(1)}_{j_1,r}\left(x^{(0)}_1\right), \ \ldots \ , \mu^{(n)}_{j_n,r}\left(x^{(0)}_n\right)\right\} \ (r = 1, \ldots, k).$$

Der Erfüllungsgrad einer Regel wird durch die Aktivierung der zugehörigen Regeleinheit repräsentiert, welche wiederum direkt an die Ausgabe weitergeleitet wird. Mit den so bestimmten Werten kann die Netzeingabe der Ausgabeeinheiten auf zwei verschiedene Arten (gewichtete Summe, Maximum) berechnet werden, womit gleichzeitig die Ausgabe des Systems unter Verwendung einer sog. „Winner-Take-All"-Interpretation bestimmt ist (NAUCK et al., 1996a und 1996b).
Eine exemplarische Erläuterung des beschriebenen Propagationsvorgangs nehmen sowohl HOFERICHTER (1996) als auch RUDOLPH (1998a) vor.

11.3.3 Der Regellernalgorithmus

Gibt man dem System eine partielle oder auch vollständige Regelbasis vor, so müssen die auf diese Weise spezifizierten Regeln nicht alle Eingangsgrößen (Merkmale) in den Prämissen enthalten. Bei den vom System gelernten Regeln hingegen treten immer alle Merkmale in den Prämissen auf.
Das System führt ein inkrementelles Regellernverfahren durch, d. h., die Regeln werden sukzessive der Regelbasis hinzugefügt. Die Musterpaare (d. h. Eingabe- und Ausgabemuster) werden nacheinander propagiert. Die Bildung einer Regel erfolgt so, dass zunächst für jedes Merkmal eines gegebenen Eingabemusters X_0 die Fuzzy-Menge $\mu_{max}^{(i)}$ bestimmt wird, die für den aktuellen Wert des Merkmals den größten Zugehörigkeitsgrad ergibt, d. h.

$$\mu_{max}^{(i)} = \mu_{j_i}^{(i)} \quad \text{mit} \quad \mu_{j_i}^{(i)}(x_i) = \max_{j_i \in \{1,\dots,q_i\}} \left\{ \mu_{j_i}^{(i)}(x_i) \right\} \quad (i = 1, \dots, n).$$

Die Konjunktion der mit diesen Fuzzy-Mengen $\mu_{max}^{(i)}$ (i = 1, ... , n) assoziierten linguistischen Terme bildet die Prämisse der gesuchten Fuzzy-Regel, die Konklusion wird durch die in der Lernaufgabe angegebene Klasse des Musters festgelegt. Ist eine Regel mit einer solchen Prämisse noch nicht in der Regelbasis des Systems enthalten und die (festgelegte) maximale Regelanzahl noch nicht erreicht, so wird sie hinzugefügt. Es ist somit prinzipiell möglich, in einem einzigen Durchlauf durch die Lernaufgabe (= Epoche) eine initiale Regelbasis aufzubauen. Ein solches **einfaches Regellernen** ist jedoch von der Reihenfolge der propagierten Muster abhängig, daher wird die erlernte Regelbasis nur in einfachen Fällen zufriedenstellend klassifizieren. Die Muster müssten dazu bezüglich ihrer Klassen gleichverteilt sein und idealerweise klassenweise abwechselnd propagiert werden. Ansonsten kann es sein, dass die maximale Regelanzahl bereits erreicht ist, bevor alle Klassen propagiert wurden, und somit eine

oder mehrere Klassen nicht erkannt werden. Es besteht außerdem die Möglichkeit, dass ein „Ausreißer" eine Regel erzeugt, deren Prämisse zur Klassifikation von Mustern einer anderen Klasse benötigt wird - dies ist umso wahrscheinlicher, je mehr sich die Klassen überlappen. Um solche Anomalien zu vermeiden, ist es notwendig, die Regeln in zwei weiteren Durchgängen durch die Lernaufgabe bezüglich ihrer Klassifikationsleistung zu bewerten, die Konklusion einzelner Regeln daraufhin zu ändern und schließlich nur die besten (im Sinne einer definierten größtmöglichen Bewertungszahl V_R) auszuwählen, während die übrigen aus dem NEFCLASS-System entfernt werden. Eine solche Vorgehensweise wird als **optimales** bzw. **klassenoptimales Regellernen** bezeichnet. Im Gegensatz zum einfachen Regellernen werden bei diesen Varianten im ersten Durchlauf durch die Lernaufgabe so viele Regeln wie möglich erzeugt. In zwei weiteren Epochen erfolgt dann die Reduktion auf die definierte Maximalanzahl k_{max} durch die sich an die Bewertung der Regeln anschließende Auswahl der $k \leq k_{max}$ besten Regeln. NAUCK et al. (1996a) geben die Definition des beschriebenen Regellernalgorithmus an. Aufgrund der genannten Probleme beim einfachen Regellernen ist eine Entscheidung für das optimale oder das klassenoptimale Regellernen zu treffen. Ersteres ist dann vorzuziehen, wenn vermutet wird, dass für das erfolgreiche Erkennen einer oder mehrerer Klassen eine größere Anzahl von Regeln erforderlich ist als für die anderen Klassen. Dabei ist darauf zu achten, dass die Anzahl k_{max} der zulässigen Regeln so hoch gewählt wird, dass bei Auswahl der besten k Regeln auch alle Klassen abgedeckt werden. Eine Entscheidung für das klassenoptimale Regellernen ist dann sinnvoll, wenn man annimmt, dass die Muster sich je Klasse auf eine etwa gleich große Anzahl von Clustern verteilen. Da jeder Cluster durch eine Fuzzy-Regel repräsentiert wird, sollte in diesem Fall auch je Klasse die gleiche Anzahl von Regeln eingesetzt werden. Die Anzahl möglicher Regeln ist nach oben beschränkt durch:

- die Anzahl s an Musterpaaren, d. h., im Extremfall kann jedes von insgesamt s Musterpaaren die Bildung einer Regel veranlassen
- die Kombinationsmöglichkeiten der Fuzzy-Mengen jeder Eingangsgröße; ist der Wertebereich jedes Eingabemerkmals x_i durch jeweils q_i Fuzzy-Mengen partitioniert, so lassen sich genau $\prod_{i=1}^{n} q_i$ verschiedene Kombinationsmöglichkeiten bilden

Die Regelbasis eines NEFCLASS-Systems kann somit im ungünstigsten Fall

$$k = \min\left\{s;\ \prod_{i=1}^{n} q_i\right\}$$

Regeln enthalten. Weisen die Muster eine „gutartige" Verteilung auf, so bilden sich in der Praxis jedoch viel weniger Regeln heraus (NAUCK et al., 1996b; HOFERICHTER, 1996).

11.3.4 Der Lernalgorithmus für Fuzzy-Mengen

Nachdem das System die Regelbasis erzeugt hat bzw. sie ihm vorgegeben wurde, werden die Fuzzy-Mengen bzw. deren Zugehörigkeitsfunktionen in mehreren Epochen angepasst.

Die Fuzzy-Mengen $\mu_{j_i}^{(i)}$ ($i = 1, \ldots, n; j = 1, \ldots, q_i$) werden durch die Partitionierung des Wertebereichs eines jeden Eingabemerkmals X_i durch q_i Fuzzy-Mengen erzeugt. NEFCLASS verwendet für diese dreiecksförmige (an den Rändern des Wertebereiches geschulterte) Zugehörigkeitsfunktionen, die zunächst gleichmäßig über dem Wertebereich verteilt sind. Es wird die Minimumsbildung (Durchschnitt zweier Fuzzy-Mengen) als t-Norm verwendet, um den Erfüllungsgrad α_r einer Regel zu bestimmen (s. Abschnitt 11.3.2). Der Lernalgorithmus für Fuzzy-Mengen ist eine einfache Heuristik. Nach der Propagation eines Musterpaares der Lernaufgabe wird für jede aktivierte Regeleinheit der Regelbasis anhand der korrekten Ausgabe ermittelt, ob deren Aktivierung für das aktuelle Muster verstärkt oder abgeschwächt werden muss. Hierzu wird die Fuzzy-Menge in der Prämisse der Regel bestimmt, die aufgrund ihres minimalen Zugehörigkeitsgrades (für das betreffende Merkmal X_i) für die Höhe der Regelaktivierung verantwortlich ist. Die Parameter dieser Fuzzy-Menge werden dann derart geändert, dass der Zugehörigkeitsgrad des entsprechenden Merkmals des Musterpaares bei der nächsten Propagation entsprechend höher (niedriger) ist, so dass die Aktivierung der betrachteten Regel dann auch höher (niedriger) ausfallen wird. Die Änderung der Parameter der Fuzzy-Menge verursacht eine Verschiebung und eine Verbreiterung bzw. Verengung der entsprechenden Dreiecksfunktion. Der beschriebene Algorithmus wird durch NAUCK et al. (1996a) definiert.

11.3.5 Das Datenanalysetool NEFCLASS-PC

Eine Implementierung des NEFCLASS-Modells ist das Datenanalysetool NEFCLASS-PC, Version 2.04 (NAUCK et al., 1996a), welches für die Untersuchungen verwendet wurde. HOFERICHTER (1996) und RUDOLPH (1998a) demonstrieren anhand von Beispieldatensätzen dessen Arbeitsweise. Die hier durchgeführten Versuche belegen, dass ein NEFCLASS-System nicht nach nur einem Lernvorgang aus den Daten eine optimale Regelbasis erzeugt, sondern die Analyse in mehreren Schritten zu vollziehen ist, um eine Verbesserung der Struktur des zu spezifizierenden Netzwerkes (Elimination von Variablen, mehr Regeln usw.) zu erreichen. In diesem Sinne ist NEFCLASS-PC als interaktives Werkzeug zur Datenanalyse zu verstehen.

11.4 Untersuchungen

11.4.1 Gegenstand und Zielstellung der Untersuchungen

Der Gegenstand der Untersuchungen und deren Zielstellung waren analog denen, die im Rahmen der überwachten Klassifikation (s. Kapitel 7) durchgeführt wurden. Da das NEFCLASS-Modell auf dieselben Datensätze angewandt wurde, die mit Hilfe der KNN-Methode bzw. der Diskriminanzanalyse ausgewertet wurden, war somit zudem ein unmittelbarer Vergleich der Klassifikationsergebnisse des (wissensbasierten) Neuro-Fuzzy-Ansatzes mit denen der konventionellen statistischen Verfahren möglich.

Die Untersuchungen unterteilten sich dementsprechend in zwei Abschnitte:

1. Neuro-Fuzzy-Datenanalyse mit dem Ziel der Separation von gegebenen Objektgruppen ohne Klassifizierung von echten Testobjekten

   Das NEFCLASS-System wurde hier auf die Lerndatensätze der Sicker- und Grundwassermessstellen (Variablen: Ionen) angewandt, s. Tab. 7-6 und 7-7 des Abschnitts 7.3.5.1.

2. Neuro-Fuzzy-Datenanalyse mit dem Ziel der Separation von gegebenen Objektgruppen und der Klassifizierung eines echten Testobjekts

   Bei diesen Untersuchungen wurden sechs Lerndatensätze - die an den Sicker- und Grundwassermessstellen aufgenommenen Werte der Ionen jeweils einer Messkampagne - gebildet und die Zuordnung von jeweils einem Testobjekt GWB 5 überprüft. Die berücksichtigten Datensätze (Messkampagnen) sind in Tab. 7-1 des Abschnitts 7.2.3.1 aufgeführt.

Den genannten Tabellen ist zu entnehmen, dass in allen Datensätzen die Muster bezüglich ihrer Klassen ungleichmäßig verteilt sind. Die Datensätze wurden dem NEFCLASS-System in Form einer festen Lernaufgabe vorgegeben. Um die Untersuchungen trotz ihres Umfangs übersichtlich zu gestalten, wurden die folgenden beiden Einschränkungen vorgenommen:

- Das NEFCLASS-System wurde für die verschiedenen Datensätze jeweils vollständig durch das Lernverfahren erzeugt, d. h., auf dessen Initialisierung mit Vorwissen wurde verzichtet, da dieses dem Autor nicht derart zur Verfügung stand, dass eine entsprechende Interpretation in Form linguistischer WENN-DANN-Regeln möglich war. Die zur Klassifikation verwendeten Fuzzy-Regeln wurden somit als „a priori-Wissen" durch das NEFCLASS-System selbst initialisiert - im übertragenen Sinne versteht sich. Hierfür wurde in allen Fällen das in Abschnitt 11.3.3 besprochene klassenoptimale Regellernen angewandt.

- Das zur Klassifikation der Daten jeweils verwendete System wurde nicht explizit auf seine Leistungsfähigkeit hin getestet, d. h., das letztendlich als optimal ermittelte (s. u.) wurde in einem einmaligen Lernvorgang erzeugt und anschließend das Klassifikationsergebnis bzw. der gefundene Klassifikator interpretiert.

  Demgegenüber wird in einigen anderen Anwendungsfällen die Leistungsfähigkeit des Klassifikationssystems anhand bereits klassifizierter Muster zunächst getestet. KANNE (1996) beispielsweise verwendet nur die Hälfte der verfügbaren Beispieldaten zur Bestimmung des Klassifikators, die andere Hälfte benutzt er, da die korrekte Klassifikation der Beispieldaten ja bekannt ist, zum Testen des erhaltenen Klassifikators. Die Güte des Klassifikationssystems wird dabei wesentlich durch die Aufteilung der Datenmenge beeinflusst, sowohl in der Trainings- als auch in der Testdatenmenge müssen repräsentative Muster enthalten sein (HOFERICHTER, 1996).

Die für jeden Datensatz durch das System erlernte Regelbasis wurde in mehreren Schritten ermittelt. In diesen wurde versucht, das System derart zu modifizieren (z. B. durch Elimination von Variablen, mehr Fuzzy-Mengen pro Variable usw.), dass es eine weitestgehend optimale (im Sinne einer niedrigen Fehlerrate sowie einer möglichst geringen Anzahl von Größen in den Prämissen) Regelbasis erzeugt. Die einzelnen Schritte werden nicht kommentiert, sondern es wird nur der letztendlich als optimal ermittelte Klassifikator zur Diskussion herangezogen.

11.4.2 Untersuchungsergebnisse

Im nachfolgenden Abschnitt sind die folgenden Untersuchungsergebnisse enthalten:

- 1. Untersuchung
  - Fehlklassifikationen des Lerndatensatzes Sickerwassermessstellen: s. Tab. 11-1
  - Klassifikationsregeln für den Lerndatensatz Sickerwassermessstellen: s. Tab. 11-2
  - Fehlklassifikationen des Lerndatensatzes Grundwassermessstellen: s. Tab. 11-3
  - Klassifikationsregeln für den Lerndatensatz Grundwassermessstellen: s. Tab. 11-4
- 2. Untersuchung
  - Klassifikationsregeln (sechs Lerndatensätze): s. Tab. 11-5
  - Klassifikationsergebnis (sechs Lerndatensätze, je ein Testdatensatz GWB 5): s. Tab. 11-6

Auf die Angabe der konkreten Parameterwerte der in den Regelprämissen mit den linguistischen Termen assoziierten Fuzzy-Mengen wird verzichtet. Diese resultieren aus dem im vor-

hergehenden Abschnitt beschriebenen Lernalgorithmus für Fuzzy-Mengen. Da bei diesem von einer zunächst gleichmäßigen Partitionierung des Wertebereichs jeder Eingabevariablen (Ionen) ausgegangen wird, können für die Datensätze der Sicker- und Grundwassermessstellen (1. Untersuchung) zur Orientierung die entsprechenden Maximal- und Minimalwerte der Ionen (s. Tab. A-1 bzw. A-2) herangezogen werden. Für den Datensatz der Messstellen des Monats November '97 (2. Untersuchung) werden exemplarisch die erlernten Fuzzy-Mengen der Variablen Cl^- und SO_4^{2-} in der entsprechenden zweidimensionalen Datenprojektion dargestellt, s. Abb. 11-2. Darüber hinaus ist das für diesen Datensatz zur Klassifikation verwendete System in Abb. 11-3 aufgezeichnet.

11.4.3 Diskussion der Untersuchungsergebnisse

1. Untersuchung

Die Anwendung des als optimal ermittelten Systems auf die Daten der Sickerwassermessstellen (61 Muster) führt zu einer Klassifikationsfehlerrate von 21,31 % (13 Fehler), s. Tab. 11-1.

Tab. 11-1. Fehlklassifikationen des Lerndatensatzes Sickerwassermessstellen.

gesamt (Fehlerrate)	SWP 1	SWP 4	SWP 5	SWP 7	SWP 9	SWP 10	SWP 11
13 (21,31 %)	0	0	1	2	1	1	8

Es zeigt sich, dass bei einer Spezifizierung von Na^+ und K^+ als „don't care"-Variablen (d. h. deren Nichtberücksichtigung für die Klassifikation der Muster) eine sehr gute Separation der a priori gegebenen Gruppen (Messstellen) bezüglich der verbliebenen Merkmale (Mg^{2+}, Ca^{2+}, Cl^-, SO_4^{2-} mit jeweils einer Partitionierung auf die vier Fuzzy Mengen *v_small*, *small*, *large*, *v_large*) möglich ist. Damit wird das mit der Diskriminanzanalyse erzielte Untersuchungsergebnis zwar nicht ganz erreicht - hier wurden für die 61 Objekte lediglich drei Fehlklassifikationen ermittelt - die dort getroffenen Aussagen können jedoch prinzipiell bestätigt werden. Demnach sind die Na^+- und K^+-Ionen für die Trennung der Gruppen weniger signifikant - in den Untersuchungsergebnissen der Diskriminanzanalyse spiegelt sich das durch deren geringe Diskriminanzkoeffizientenwerte im ersten und zweiten NED wider, bei Verwendung des NEFCLASS-Systems als interaktives Werkzeug zur Datenanalyse durch das verbesserte Klassifikationsergebnis, wenn man sie als „don't care"-Variablen spezifiziert (Gibt man diese dem System mit einer Partitionierung auf ebenfalls vier Fuzzy-Mengen als Eingaben vor und führt das Training mit derselben Anzahl von 2500 Epochen durch, so erhöht sich die Fehlerrate auf

40,98 % (25 Fehler).). Mit Hilfe der Diskriminanzanalyse wurde festgestellt, dass sich die Cl^-- und SO_4^{2-}-Ionen aufgrund der hohen Werte ihrer Diskriminanzkoeffizienten im ersten und zweiten NED offenbar hervorragend zur Separation der Sickerwassermessstellen eignen. Bei Anwendung des NEFCLASS-Systems lässt sich diesbezüglich keine eindeutige Aussage treffen. Bei keiner der (verbliebenen) vier Variablen sind die erlernten Fuzzy-Mengen optimal (im Sinne keiner oder nur geringer Überschneidungen) über dem Wertebereich verteilt, was ein Indiz für deren gute Separationsfähigkeit wäre. In den Regelprämissen spiegelt sich das darin wider, dass bei keiner der Eingabegrößen sämtliche der vier möglichen Fuzzy-Mengen verwendet werden. (Reduziert man die Anzahl der Fuzzy-Mengen von vornherein bei der Spezifizierung des Systems, so werden bestimmte Klassen nicht erkannt bzw. es kommt zu einem deutlichen Anstieg der Fehlerrate.). In Tab. 11-2 sind die erlernten Regeln eingetragen. Zur Vereinfachung erfolgte die Notation der Prämissen hier und nachfolgend in einer verkürzten Schreibweise, d. h., die Regel R_1 z. B. ist zu lesen als

if Na^+ is * and K^+ is * and Mg^{2+} is *small* and
Ca^{2+} is *v_small* and Cl^- is *small* and SO_4^{2-} is v_*small*
then *SWP 1*.

Enthält eine Regelprämisse für ein oder mehrere Merkmale statt eines linguistischen Terms einen „*“, so bedeutet dies, dass es bei der Auswertung dieser Regel unberücksichtigt bleibt (don't care). Die dreiecksförmigen Zugehörigkeitsfunktionen der mit den linguistischen Termen assoziierten Fuzzy-Mengen sind durch die entsprechenden Parameter eindeutig beschrieben. Somit ist es prinzipiell möglich, die Unschärfe dieser Aussagen zu präzisieren.

Die Regel R_6, welche als Konklusion den nur einmal fehlklassifizierten SWP 10 enthält, besitzt die höchste Bewertungszahl V_R. Die (umgangssprachliche und daher unscharfe) Aussage, nach der diese Messstelle sehr hoch (im Sinne von *v_large*) mit Mg^{2+}, Cl^- und SO_4^{2-} belastet ist, wird damit auch durch das System bestätigt.

Tab. 11-2. Klassifikationsregeln für den Lerndatensatz Sickerwassermessstellen.

Klassifikationsregel										V_R
R_1:	**if**	*(*,	*,	small,	v_small,	small,	v_small)*	**then**	*SWP_1*	-0,4856
R_2:	**if**	*(*,	*,	small,	small,	v_small,	v_small)*	**then**	*SWP_4*	4,7292
R_3:	**if**	*(*,	*,	v_small,	v_large,	small,	v_small)*	**then**	*SWP_5*	5,3712
R_4:	**if**	*(*,	*,	v_small,	v_large,	v_small,	v_small)*	**then**	*SWP_7*	-4,1175
R_5:	**if**	*(*,	*,	v_small,	small,	v_small,	v_small)*	**then**	*SWP_9*	-7,0424
R_6:	**if**	*(*,	*,	v_large,	small,	v_large,	v_large)*	**then**	*SWP_10*	5,6022
R_7:	**if**	*(*,	*,	v_small,	v_small,	v_small,	v_small)*	**then**	*SWP_11*	-10,0102

Die Anwendung des Modells auf die Daten der Grundwassermessstellen (110 Muster) führt zu einer Erhöhung der Klassifikationsfehlerrate auf 31,82 % (35 Fehler), s. Tab. 11-3.

Tab. 11-3. Fehlklassifikationen des Lerndatensatzes Grundwassermessstellen.

gesamt (Fehlerrate)	GWB 1 - 3 (A_Luppe)	GWB 4 - 6 (A_Nahle)	GWB 7 - 9 (Anstrom)
35 (31,82 %)	10	18	7

Damit werden die entsprechenden Untersuchungsergebnisse der Diskriminanzanalyse bestätigt - für den gleichen Datensatz wurde hier mit 29,09 % (32 Fehler) eine ebenfalls relativ hohe Fehlerrate ermittelt. Die Klassifizierung der Sickerwasserdaten bezüglich der Messstellen führt damit offensichtlich generell zu einem besseren Ergebnis gegenüber dem, welches man für die Grundwasserdaten bei einer gröberen Einteilung in einen Anstrombereich und zwei Abstrombereiche erzielt. Dies ist ein Indiz dafür, dass der Schadstoffaustrag des Deponiekörpers in den umgebenden Aquifer sich nicht in einer derart vereinfachten Form (ein Anstrombereich und zwei Abstrombereiche) einteilen lässt. Die hierfür möglichen Ursachen (örtlich inhomogene Verteilung der Müllbestandteile, unterschiedliches Alter der Deponiebereiche, differenzierte Wasserwegsamkeiten) wurden bereits mehrfach genannt.

Das Untersuchungsergebnis verdeutlicht einige interessante Aspekte. So führte die interaktive Datenanalyse auch hier zu der Erkenntnis, dass die Na^+- und K^+-Ionen nicht als Indikatoren für den Schadstofftransport geeignet sind, sie werden für die Trennung der Muster nicht benötigt. Die entsprechenden Untersuchungsergebnisse der Diskriminanzanalyse werden damit zumindest partiell bestätigt, denn auch hier kommt bezüglich des ersten NED diesen beiden Variablen eine nur geringe Bedeutung für eine erfolgreiche Trennung der Gruppen zu. Für die Wertebereiche der Eingabegrößen Mg^{2+} und Ca^{2+} wurde dem System eine Partitionierung auf vier (*v_small*, *small*, *large*, *v_large*) und für die der Eingabegrößen Cl^- und SO_4^{2-} eine auf zwei (*small*, *large*) Fuzzy-Mengen vorgegeben. Es erlernte die in Tab. 11-4 eingetragenen Regeln.

Auffallend an diesen ist, dass in den Prämissen für die Variablen Cl^- und SO_4^{2-} jeweils nur eine Fuzzy-Menge (small) in Gebrauch ist, bezüglich dieser beiden Merkmale also keine Aussage über die Unterschiedlichkeit der Gruppen getroffen werden kann. Dies führt zu der Vermutung, dass man beide Eingangsgrößen ebenfalls als „don't care"-Variablen spezifizieren kann, d. h. für diese unter ansonsten unveränderten Bedingungen dem System von vornherein eine Partitionierung auf nur eine Fuzzy-Menge vorgeben sollte. In den Untersuchungen wurde dies

ausprobiert, woraufhin die Fehlerrate auf 34,55 % (38 Fehler) anstieg. Daher wurden diese beiden Variablen mit jeweils zwei Fuzzy-Mengen im System belassen. Damit wird jedoch für diesen Datensatz ein Unterschied zum Ergebnis der Diskriminanzanalyse deutlich, wo sowohl die Cl^-- als auch die SO_4^{2-}-Ionen einen bedeutenden Beitrag zur Trennung der Gruppen liefern. Die Regeln R_6, R_7, und R_8, welche als Konklusion jeweils den Abstrombereich der Nahle enthalten, besitzen die höchsten Bewertungszahlen V_R. In den Prämissen von R_6 bzw. R_8 ist in der UND-Verknüpfung die Aussage enthalten, dass die Mg^{2+}-Werte sehr hoch (*v_large*) bzw. hoch (*large*) sein müssen und in der Prämisse von R_7 wurden für die Ca^{2+}-Ionen sehr hohe (*v_large*) Werte ermittelt. Dies lässt darauf schließen, dass diese beiden Elemente besonders signifikant für den Schadstoffeintrag im nordwestlichen Deponiebereich sind.

Tab. 11-4. Klassifikationsregeln für den Lerndatensatz Grundwassermessstellen.

	Klassifikationsregel									V_R
R_1:	**if**	*(*,*	**,*	*large,*	*large,*	*small,*	*small)*	**then**	*Anstrom*	1,0955
R_2:	**if**	*(*,*	**,*	*v_small,*	*small,*	*small,*	*small)*	**then**	*Anstrom*	-8,1073
R_3:	**if**	*(*,*	**,*	*small,*	*v_small,*	*small,*	*small)*	**then**	*A_Luppe*	2,0351
R_4:	**if**	*(*,*	**,*	*small,*	*small,*	*small,*	*small)*	**then**	*A_Luppe*	-0,2471
R_5:	**if**	*(*,*	**,*	*v_small,*	*v_small,*	*small,*	*small)*	**then**	*A_Luppe*	-5,2418
R_6:	**if**	*(*,*	**,*	*v_large,*	*v_small,*	*small,*	*small)*	**then**	*A_Nahle*	3,4122
R_7:	**if**	*(*,*	**,*	*small,*	*v_large,*	*small,*	*small)*	**then**	*A_Nahle*	2,7568
R_8:	**if**	*(*,*	**,*	*large,*	*v_small,*	*small,*	*small)*	**then**	*A_Nahle*	2,0677

2. Untersuchung

Die Anwendung des Systems auf die sechs Lerndatensätze zeigte auch hier, dass die Na^+- und K^+-Ionen generell nicht zur Trennung der Muster benötigt werden. Für die Mg^{2+}- bzw. SO_4^{2-}-Ionen trifft dies in einem bzw. zwei Fällen zu. Ansonsten wurde dem System für die Eingabegrößen eine Partitionierung auf vier (*v_small*, *small*, *large*, *v_large*), drei (*small*, *medium*, *large*) oder zwei (*small*, *large*) Fuzzy-Mengen vorgegeben, und es erlernte durch entsprechende Vorgabe in allen sechs Fällen jeweils vier Klassifikationsregeln, s. Tab. 11-5.

In Tab. 11-6 sind die Klassifikationsfehlerraten und Fehlzuordnungen der sechs Datensätze aufgeführt. Bis auf eine Ausnahme (Lerndatensatz vom Februar '95) liegt die mit dem System erzielte Fehlerrate über der, die bei Anwendung der KNN-Methode (hier bezüglich des optimalen k-Wertes, s. Abb. 7-2) und der Diskriminanzanalyse (s. Tab. 7-14) erreicht wurde. Die Überprüfung der Zuordnung des Testdatensatzes GWB 5 mit Hilfe des ermittelten Klassifikators hingegen ergab auch hier keine (Fehl-) Klassifikation in eine der Sickerwassergruppen.

Tab. 11-5. Klassifikationsregeln (sechs Lerndatensätze).

Lerndatensatz	Klassifikationsregel										V_R
März '94	R_1:	**if**	*(*,	*,	*small,*	*large,*	*small,*	*small)*	**then**	*SW*	2,2626
	R_2:	**if**	*(*,	*,	*small,*	*medium,*	*small,*	*small)*	**then**	*SW*	1,6796
	R_3:	**if**	*(*,	*,	*large,*	*medium,*	*large,*	*large)*	**then**	*SW*	0,7754
	R_4:	**if**	*(*,	*,	*small,*	*small,*	*small,*	*small)*	**then**	*GW*	8,6950
August '94	R_1:	**if**	*(*,	*,	*small,*	*large,*	*small,*	**)*	**then**	*SW*	0,9471
	R_2:	**if**	*(*,	*,	*large,*	*small,*	*large,*	**)*	**then**	*SW*	0,7797
	R_3:	**if**	*(*,	*,	*small,*	*medium,*	*small,*	**)*	**then**	*SW*	-0,2165
	R_4:	**if**	*(*,	*,	*small,*	*small,*	*small,*	**)*	**then**	*GW*	9,4122
Februar '95	R_1:	**if**	*(*,	*,	**,*	*large,*	*large,*	*large)*	**then**	*SW*	3,4543
	R_2:	**if**	*(*,	*,	**,*	*medium,*	*large,*	*large)*	**then**	*SW*	1,8951
	R_3:	**if**	*(*,	*,	**,*	*small,*	*small,*	*small)*	**then**	*GW*	11,9049
	R_4:	**if**	*(*,	*,	**,*	*medium,*	*small,*	*small)*	**then**	*GW*	3,9907
April '95	R_1:	**if**	*(*,	*,	*small,*	*large,*	*small,*	*small)*	**then**	*SW*	2,1735
	R_2:	**if**	*(*,	*,	*small,*	*medium,*	*medium,*	*small)*	**then**	*SW*	0,9670
	R_3:	**if**	*(*,	*,	*large,*	*medium,*	*large,*	*large)*	**then**	*SW*	0,9025
	R_4:	**if**	*(*,	*,	*small,*	*small,*	*small,*	*small)*	**then**	*GW*	10,5279
September '95	R_1:	**if**	*(*,	*,	*small,*	*large,*	**,*	**)*	**then**	*SW*	3,2872
	R_2:	**if**	*(*,	*,	*small,*	*v_large,*	**,*	**)*	**then**	*SW*	1,7244
	R_3:	**if**	*(*,	*,	*large,*	*small,*	**,*	**)*	**then**	*SW*	1,5325
	R_4:	**if**	*(*,	*,	*small,*	*v_small,*	**,*	**)*	**then**	*GW*	10,7000
November '97	R_1:	**if**	*(*,	*,	*small,*	*large,*	*small,*	*small)*	**then**	*SW*	1,2315
	R_2:	**if**	*(*,	*,	*large,*	*medium,*	*small,*	*small)*	**then**	*SW*	0,8172
	R_3:	**if**	*(*,	*,	*large,*	*medium,*	*large,*	*large)*	**then**	*SW*	0,8151
	R_4:	**if**	*(*,	*,	*small,*	*small,*	*small,*	*small)*	**then**	*GW*	9,3896

Betrachtet man die erlernten Klassifikationsregeln, so war diese jeweils korrekte Zuordnung nicht zu erwarten, da zur Klassifizierung der Grundwassermessstellen mit einer Ausnahme jeweils nur eine Regel erlernt wurde, in deren Prämisse zudem ausschließlich die durch Fuzzy-Mengen repräsentierten linguistischen Terme *small* bzw. *v_small* enthalten sind. Unter dem Aspekt, dass bei der Fuzzy-Partitionierung der Eingabegrößen die Intervallgrenzen der Wertebereiche auf den Maximal- bzw. Minimalwerten beruhen, welche sich bei Zusammenfassung der Sicker- und Grundwassermessstellen ergeben, ist die hohe Belastung des GWB 5 letztlich jedoch nicht ausreichend für eine Fehlkassifikation in eine der Sickerwassergruppen.

Aus Tab. 11-6 ist allerdings zu ersehen, dass eine Fehlzuordnung des im Lerndatensatz des Monats März '94 enthaltenen Objekts GWB 4 zur Gruppe der Sickerwassermessstellen ermittelt wurde. Diese ist insofern hervorhebenswert, da diese Probenahmestelle bezüglich der Salzfrachten gemeinhin zwar relativ hohe Werte aufweist (s. Abb. 6-4 und 6-10), hier jedoch zudem offensichtlich jahreszeitliche Effekte (Einflussfaktor DATUM) zu dieser örtlich signifikanten Belastung und damit Fehlklassifikation geführt haben - eine Aussage, die aus den Ergebnissen der einfaktoriellen Varianzanalysen zum Haupteffekt DATUM nicht abzuleiten ist und auf die Notwendigkeit differenzierter Betrachtungsweisen hindeutet.

Tab. 11-6. Klassifikationsergebnis (sechs Lerndatensätze, je ein Testdatensatz GWB 5).

Lern-, Testdatensatz	Fehlerrate des Lerndatensatzes	Fehlzuordnungen des Lern- bzw. Testdatensatzes
März '94	15,79 %	SWP 01, SWA, GWB 04
August '94	9,52 %	SWP 11, SWA
Februar '95	0 %	---
April '95	13,04 %	SWP 02, SWA, SWA-Pegel
September '95	4,76 %	SWA
November '97	10,53 %	SWP 01, SWA

Das System konnte in allen sechs Fällen so spezifiziert werden, dass es bei allen Variablen jede der vorgegebenen Fuzzy-Mengen verwendet. Die erlernten Regeln verdeutlichen, dass insbesondere Ca^{2+} als Indikator zur Unterscheidung zwischen Sicker- und Grundwasser angesehen werden kann. Es wird in allen Datensätzen zur Klassifizierung benötigt und weist hier in den Regeln mit der jeweils höchsten Bewertung auf eine hohe Belastung (im Sinne von *large*) für das Sickerwasser und eine nur geringe (im Sinne von *small*) für das Grundwasser hin.
Die bei Anwendung der Diskriminanzanalyse ermittelte Signifikanz der Cl^-- und SO_4^{2-}-Ionen zur Trennung der Gruppen wird hier zumindest teilweise bestätigt. In den Fällen, wo das System diese beiden Größen zur Klassifizierung benötigt, ist in der entsprechenden zweidimensionalen Darstellung eine relativ gute Separation der Gruppen zu erkennen und zudem wird der Wertebereich beider Variablen durch die im Ergebnis des Trainings entstandenen Fuzzy-Mengen optimal aufgeteilt. Abb. 11-2 zeigt dies für den Datensatz des Monats November '97.

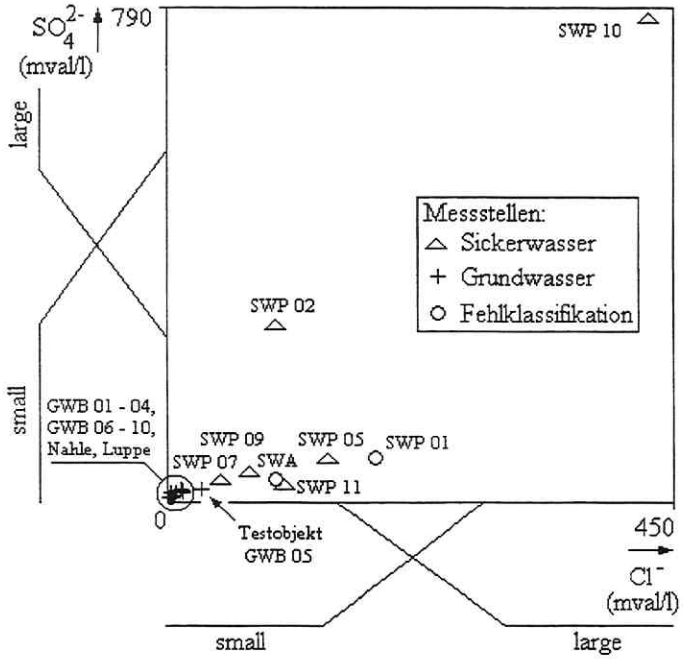

Abb. 11-2. Zweidimensionale Objektdarstellung.

Des Weiteren spiegelt sich in dieser Darstellung die hohe Belastung des SWP 10 in dessen deutlichem Abstand von allen anderen Objekten wider. Die für diesen Datensatz vom System erlernte Regel R_3 (s. Tab. 11-5 bzw. Abb. 11-3) dient zur Erkennung nur dieses einen Musters. In Abb. 11-3 ist beispielhaft das für den Lerndatensatz des Monats November '97 zur Klassifizierung verwendete Modell dargestellt. Aus dieser ist ersichtlich, dass das System bei drei der vier Eingabegrößen (die als „don't care"-Variablen spezifizierten Na^+ und K^+ wurden zur Vereinfachung weggelassen) lediglich zwei Fuzzy-Mengen zur Klassifizierung benötigt und zur Identifizierung der Muster des Grundwassers nur eine Regel erlernt.

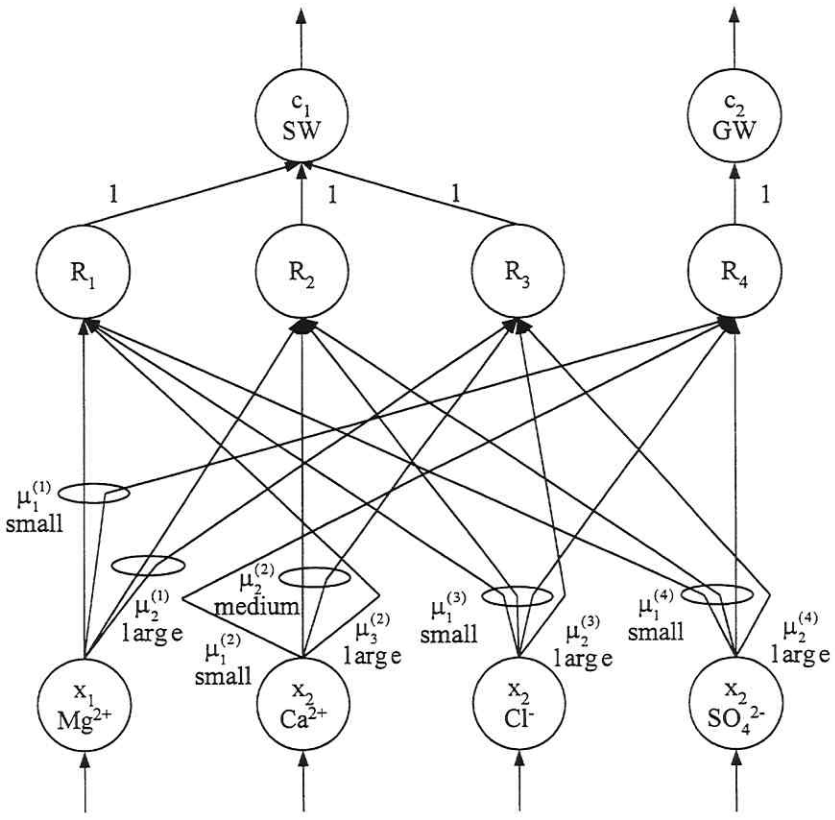

Abb. 11-3. NEFCLASS-System für den Lerndatensatz des Monats November '97.

Abschließend sei darauf hingewiesen, dass ein expliziter Vergleich der Klassifikationsleistung der angewandten Methoden der überwachten Klassifikation (NEFCLASS-Modell als Vertreter eines wissensbasierten Neuro-Fuzzy-Ansatzes und KNN-Methode sowie Diskriminanzanalyse als Vertreter der konventionellen statistischen Verfahren) in Abschnitt 12.2 erfolgt.

12 Zusammenfassung

12.1 Komplexe Bewertung der Untersuchungsergebnisse

Die mit der relativ großen Anzahl von Methoden erzielten Untersuchungsergebnisse stellen sich als sehr umfangreich, vielfältig und teilweise auch widersprüchlich dar. Es werden daher nachfolgend die wesentlichsten Aspekte der in den einzelnen Abschnitten diskutierten Ergebnisse zusammengefasst:

- Durch den Deponiekörper erfolgt prinzipiell ein Salzfrachteneintrag in das Grundwasser, dieser ist im Abstrombereich der Nahle höher als im Abstrombereich der Luppe.
- Als Probenahmestelle mit den höchsten Salzfrachten des Sickerwassers bzw. des Grundwassers wurde der SWP 10 bzw. der GWB 5 ermittelt.
- Die im Abstrombereich der Nahle gelegene Grundwassermessstelle GWB 5 befindet sich im nordwestlichen Teil des Deponiegeländes, durch diesen besteht offensichtlich eine erhöhte Gefährdung für den umgebenden Aquifer (Diskussion über mögliche Transportpfade des Sickerwassers s. u.).
- Signifikante saisonale Tendenzen liegen sowohl für die schadstoffrelevanten Parameter des Sickerwassers als auch für die des Grundwassers im Wesentlichen nicht vor.
- Der Wasserspiegelstand des Sickerwassers ist jahreszeitlich unabhängig (kein signifikanter Einfluss des Faktors DATUM), der des Grundwassers hingegen nicht, wobei jedoch sowohl beim Sickerwasser als auch beim Grundwasser praktisch keine signifikante Korrelation zwischen dem Wasserspiegelstand und den analytischen Parametern besteht.
- Eine Erhöhung der Salzfrachten hat eine Erhöhung des Anteils transportierter organischer Stoffe zur Folge.
- Es besteht die Möglichkeit einer schnellen Ionenbestimmung über die Feldmessung der Summenparameter Leitfähigkeit und Wasserhärte, die Aussage über die Zulässigkeit dieser Summenparameter zur Langzeitüberwachung kann als gesichert angesehen werden.
- Eine Normalverteilung der die Schadstoffbelastung charakterisierenden Einzelionen Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^- und SO_4^{2-} ist für die jeweils von allen Sickerwassermessstellen zusammengefassten Werte des Untersuchungszeitraums nicht vorhanden. Darüber hinaus ist deren relativ breite Streuung ein Indiz für lokal unterschiedliche Belastungen. Hauptursache hierfür ist die örtlich inhomogene Verteilung der Müllbestandteile im Deponiekörper

(stoffliche Zusammensetzung, Ablagerungsalter, verschiedene Korngrößenverteilungen). Diese führt zu bevorzugten Austrittspfaden des Sickerwassers in den umgebenden Aquifer. Infolgedessen sowie aufgrund wechselnder Grundwasserfließrichtungen liegen für die zusammengefassten Werte der Grundwassermessstellen ebenfalls keine Normalverteilung sowie eine relativ breite Streuung vor. Darüber hinaus weisen die Ergebnisse der einfaktoriellen Varianzanalyse zum Haupteffekt MESSSTELLE bei sämtlichen der untersuchten Parameter auf signifikante lokale Unterschiede hin.

- Bezüglich der Annahmen über eine Substitution zum einen des CSB durch den DOC und zum anderen der aufwendigen Titrationsbestimmung des CSB durch Küvettentests kann keine eindeutige Aussage getroffen werden: Die Ergebnisse zur linearen Korrelation und Regression bestärken diese Annahmen, die Resultate der einfaktoriellen Varianzanalyse zum Haupteffekt BESTIMMUNGSMETHODE hingegen stellen sie in Frage.
- Die Untersuchungsergebnisse zur überwachten Klassifikation (KNN-Methode, Diskriminanzanalyse, NEFCLASS-System) verdeutlichen, dass trotz der hohen Belastung im Bereich von GWB 5 dessen aus den Salzfrachten induzierte Merkmalsmuster denen der Sickerwassermessstellen nicht ähnlich sind.
- Durch die Methoden der überwachten Klassifikation zeigt sich des Weiteren für die Sickerwasserdaten eine hervorragende Trennung der a priori vorgegebenen Gruppen der Messstellen bezüglich der Einzelionen. Damit wird nochmals die lokal unterschiedliche Schadstoffbelastung im Deponiekörper hervorgehoben. Demhingegen führt die Anwendung dieser Verfahren auf den Datensatz der Grundwassermessstellen mit einer Einteilung in drei Gruppen (Anstrombereich des Deponiekörpers, Abstrombereich der Luppe und der Nahle) zu relativ hohen Fehlerraten. Dies ist ein Indiz dafür, dass der Schadstoffaustrag aus dem Deponiekörper in den umgebenden Aquifer sich nicht in einer derart vereinfachten Form (ein Anstrombereich und zwei Abstrombereiche) einteilen lässt.
- Der Informationsgehalt der betrachteten analytischen Parameter (Einzelionen) zur Beschreibung der Schadstoffverteilung ist verschieden: Die Na^+- und K^+-Ionen liefern im Vergleich zu den Mg^{2+}, Ca^{2+}, Cl^- und SO_4^{2-}-Ionen die geringste Information. Für diese wiederum lassen sich aufgrund der Vielfalt der Untersuchungsergebnisse gewisse Differenzierungen vornehmen: Die Diskriminanzanalyse führt zu dem Ergebnis, dass im Deponiekörper selber offensichtlich insbesondere die Mg^{2+}-Ionen als Schadstofftracer relevant sind, für den Grundwasserbereich trifft dies auf die Ca^{2+}-Ionen zu. Zur unmittelbaren Beschreibung

der Transportrichtung des Sickerwassers in den Grundwasserleiter jedoch können alle vier Parameter, sowohl die Cl^-- und SO_4^{2-}-Ionen (nachgewiesen durch die Diskriminanzanalyse) als auch die Mg^{2+}- und Ca^{2+}-Ionen (nachgewiesen durch Anwendung des NEFCLASS-Systems), als gleichermaßen relevant angesehen werden.

- Aus den Ergebnissen der zur Anwendung gekommenen Methoden der automatischen Klassifikation sowie der Hauptkomponentenanalyse lassen sich die folgenden Aussagen über mögliche Verteilungen der Schadstoffe (Salzfrachten) im Deponiebereich ableiten:
  - Im südöstlichen Deponiebereich (SWP 9) beginnend erfolgt der Transport der Salzfrachten entsprechend der Grundwasserfließrichtung in Richtung Nordwesten (SWP 7), und es kommt hier im Bereich des von allen Grundwassermessstellen am höchstbelasteten GWB 5 zu einem bevorzugten Austrag in den umgebenden Aquifer (insgesamt sechs Hinweise).
  - Der hochbelastete Deponiebereich von SWP 10 muss als die wesentlichste Kontaminationsquelle angesehen werden, von hier ausgehend erfolgt der Salzfrachtenaustrag in Richtung Nahle zu GWB 5 (insgesamt fünf Hinweise).
  - Der hochbelastete SWP 10 befindet sich geographisch in unmittelbarer Nähe zum Abstrombereich der Luppe, ein bevorzugter Schadstoffaustritt in das Grundwasser erfolgt hier im Bereich von GWB 1 und 2 (insgesamt zwei Hinweise).
- Die als Versuchsmodell konzipierten Untersuchungen zur Zeitreihenanalyse zeigten zum einen, dass sich insbesondere für den pH-Wert gute Prognosen treffen lassen, und zum anderen, dass die exponentielle Glättung und die Prozessanpassung annähernd die gleiche Prognosegüte liefern, wobei für ähnliche praktische Anwendungen aufgrund des geringeren Aufwands in der Phase der Modellidentifikation die exponentielle Glättung vorgezogen werden sollte. Unter der Voraussetzung einer zeitkontinuierlichen Messung der Parameter am Deponiekörper lassen sowohl die in den Untersuchungen angewandten klassischen Prognoseverfahren (exponentielle Glättung, Prozessanpassung) als auch solche, die auf dem Konzept der Künstlichen Neuronalen Netze beruhen, einen erfolgreichen Einsatz erwarten.

Es kann eingeschätzt werden, dass der konzipierte Algorithmus zur Auswertung der Daten sich als ein geeigneter Lösungsansatz der gegebenen Aufgabenstellung erwiesen hat. Insbesondere sei dabei hervorgehoben, dass die zur Anwendung gekommenen Methoden der Mustererkennung die Möglichkeit bieten, über die Verteilung der Sicker- und Grundwassermessstellen Rückschlüsse auf Transportpfade der Schadstoffe zu ziehen.

12.2 Vergleichende Bewertung der angewandten Methoden

In diesem Abschnitt wird eine vergleichende Bewertung der angewandten Methoden sowohl der überwachten (KNN-Methode, Diskriminanzanalyse, NEFCLASS-Modell) als auch der automatischen (hierarchisch agglomerative und nichthierarchisch optimierende Clusteranalyse) Klassifikation vorgenommen, um damit zusammenfassend Aussagen bezüglich des mit ihnen erzielten Informationsgehaltes und ihrer Adaptionsfähigkeit an die gegebene Aufgabenstellung treffen zu können.

Die bei einer Anwendung der drei Methoden der **überwachten Klassifikation** auf identische Datensätze erzielten Klassifikationsergebnisse (Fehleranzahl und Fehlerrate) sind in Tab. 12-1 zusammengestellt, wobei einzelne Resultate bereits in den entsprechenden vorangegangenen Abschnitten vorgestellt und diskutiert wurden. Es wurden insgesamt neun Datensätze untersucht, als Variablen enthalten diese jeweils die Einzelionen Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^- und SO_4^{2-}:

- Datensatz der Sickerwassermessstellen (Zusammenfassung der Werte aller Messkampagnen, Unterteilung in g = 7 Gruppen, s. Tab. 7-6 des Abschnitts 7.3.5.1)
- Datensatz der Grundwassermessstellen (Zusammenfassung der Werte aller Messkampagnen, Unterteilung in g = 3 Gruppen, s. Tab. 7-7 des Abschnitts 7.3.5.1)
- sechs Datensätze der zusammengefassten Sicker- und Grundwassermessstellen jeweils eines Monats (Unterteilung in g = 2 Gruppen, s. Tab. 7-1 des Abschnitts 7.2.3.1)
- ein Datensatz der zusammengefassten Sicker- und Grundwassermessstellen von insgesamt sechs Monaten (Unterteilung in g = 12 Gruppen, s. Tab. 7-1 des Abschnitts 7.2.3.1)

Tab. 12-1. Klassifikationsergebnisse der Methoden der überwachten Klassifikation.

Datensatz der Messstellen			KNN-Methode		Diskriminanzanalyse		NEFCLASS-System	
Bezeichnung	Objekt-anzahl n	Gruppen-anzahl g	Fehler-anzahl	Fehler-rate	Fehler-anzahl	Fehler-rate	Fehler-anzahl	Fehler-rate
Sickerwasser	61	7	6	9,84 %	3	4,29 %	13*)	21,31 %
Grundwasser	110	3	32	29,09 %	32	29,09 %	35*)	31,82 %
März '94	20	2	1	5,00 %	1	5,00 %	3*)	15,00 %
August '94	22	2	2	9,09 %	1	4,55 %	2*)	9,09 %
Februar '95	20	2	0	0,00 %	0	0,00 %	0*)	0,00 %
April '95	24	2	0	0,00 %	0	0,00 %	3*)	12,50 %
September '95	22	2	0	0,00 %	0	0,00 %	1*)	4,55 %
November '97	20	2	1	5,00 %	0	0,00 %	2*)	10,00 %
sechs Monate	128	12	100	78,13 %	91	71,09 %	78	60,94 %

Bezüglich der mit den drei Methoden erzielten Ergebnisse ist folgendes anzumerken:

- Der KNN-Methode liegt der mittels Leave-one-out-Verfahren (LACHENBRUCH, 1967) ermittelte optimale k-Wert zugrunde.
- Bei der Diskriminanzanalyse wurden alle NED unabhängig von ihrer Signifikanz und alle Ausgangsmerkmale unabhängig von ihrer Trennfähigkeit zur Klassifizierung herangezogen.
- Bei dem zur Klassifizierung einiger Datensätze als letztendlich optimal ermittelten NEFCLASS-System wurden einzelne Merkmale als „don't care"-Variablen spezifiziert, d. h., sie wurden zur Klassifizierung nicht benötigt; in diesen Fällen ist die Fehleranzahl in Tab. 12-1 mit einem „*" markiert.

Bei allen drei Methoden sind deutliche Unterschiede zwischen den Fehlerraten der Datensätze zu erkennen: Die a priori gegebenen Gruppen des Datensatzes der Sickerwassermessstellen unterscheiden sich bezüglich der gemessenen Merkmale gut voneinander, demhingegen wurde für den nur grob partitionierten Datensatz der Messstellen des Grundwassers jeweils eine Fehlerrate von ca. 30 % ermittelt. Bei den sechs kleineren (Monats-) Datensätzen mit lediglich zwei Gruppen (Sicker- und Grundwasser) zeigen alle drei Verfahren die jeweils beste Separationsfähigkeit. Bei dem größeren Datensatz der von sechs Monaten zusammengefassten Messstellen hingegen ist die Fehlerrate wiederum jeweils sehr hoch, wobei hier jedoch zu berücksichtigen ist, dass der Gruppeneinteilung auch die verschiedenen Messzeitpunkte zugrunde liegen. Mittels Varianzanalyse wurde jedoch festgestellt, dass der Einfluss des Faktors DATUM auf die Salzfrachten nicht signifikant ist, so dass diesbezüglich wohl nur geringe Kontraste zwischen den Gruppen bestehen. Diese für alle drei Verfahren einheitlichen Unterschiede sind ein Indiz dafür, dass die in den entsprechenden Abschnitten zu den Datensätzen getroffenen Aussagen bezüglich der Untersuchungsergebnisse (Fehlklassifizierungen von Objekten, Separationsfähigkeit der Ausgangsmerkmale usw.) als gesichert angesehen werden können.

Ein individueller Vergleich von KNN-Methode und Diskriminanzanalyse zeigt, dass letztere eine höhere Klassifikationsleistung liefert (bei fünf von neun Datensätzen liegt die ermittelte Fehlerrate unter der der KNN-Methode und in nicht einem Fall darüber), besonders unter dem Aspekt, dass hier keine Schätzung des Diskriminationsfehlers in Abhängigkeit der Signifikanz der NED sowie der Trennfähigkeit der Ausgangsmerkmale erfolgte (bei der KNN-Methode wurde der optimale k-Wert berücksichtigt), die eine weitere Verbesserung erwarten ließe. Bezüglich des Informationsgehaltes und der Adaptionsfähigkeit schneidet die Diskriminanzanalyse ebenfalls besser ab. Zum einen kann mit dieser nicht nur die Klassifikationsaufgabe

(Klassifizierung des Lerndatensatzes, Überprüfung der Zuordnung von Pseudo- bzw. echten Testobjekten) an sich gelöst werden, sondern es können auch weitere Aspekte wie die Signifikanz der Ausgangsmerkmale zur Trennung der Gruppen und die Verschiedenheit der Gruppen (durch Dimensionserniedrigung und Darstellung im LDA-Display) untersucht werden. Zum anderen lässt sich diese besser an eine gegebene Aufgabenstellung anpassen: Bei der KNN-Methode kann zur Adaption lediglich der optimale k-Wert ermittelt werden, hier jedoch kann eine (in den Untersuchungen allerdings nicht durchgeführte) vorherige Schätzung des Diskriminationsfehlers in Abhängigkeit von der Signifikanz der NED sowie der Trennfähigkeit der Ausgangsmerkmale erfolgen.

Das NEFCLASS-System führt mit nur einer Ausnahme in allen Fällen gegenüber den beiden konventionellen statistischen Verfahren zu einem schlechteren Klassifikationsergebnis. Die wesentlichen Ursachen hierfür können sein:

- Der durch die interaktive Datenanalyse ermittelte Klassifikator ist nicht der optimale, d. h., er könnte weiter verbessert werden (Veränderung der Anzahl von Fuzzy-Mengen je Variable, Elimination von Variablen, Veränderung der Anzahl von Regeln usw.).
- Die Regelbasis wurde vom System vollständig erlernt. Sie könnte durch eine partielle oder auch vollständige Initialisierung mit tatsächlichem a priori-Wissen, welches man z. B. durch Befragung von Experten erhält, dahingehend verändert werden, dass das anschließende Training der in den Regeln enthaltenen Fuzzy-Mengen zu einer Verbesserung der Klassifikationsleistung führt. Bei den vom System erlernten Regeln ist dabei insbesondere von Nachteil, dass grundsätzlich zunächst alle Eingangsgrößen in den Prämissen verwendet werden. Die Spezifizierung von „Don't Care"-Variablen zur Verbesserung der Klassifikationsleistung ist vom Anwender vorzugeben.

Bezüglich des einen Falls (Datensatz der von sechs Monaten zusammengefassten Messstellen), in dem das NEFCLASS-System ein besseres Klassifikationsergebnis als die KNN-Methode und die Diskriminanzanalyse liefert, ist anzumerken, dass zum einen zum Erreichen dieser vergleichsweise niedrigen Fehlerrate dem System alle sechs Merkmale als Eingabegrößen mit einer Fuzzy-Partitionierung auf sieben Fuzzy-Mengen vorgegeben werden mussten, womit die Interpretation des erlernten Klassifikators problematisch ist, und zum anderen einige der Klassen durch die erlernten Regeln nicht abgedeckt werden.

Die Tatsache, dass das NEFCLASS-System für fast alle gegebenen Lernaufgaben ein schlechteres Klassifikationsergebnis als die beiden anderen Methoden geliefert hat, ist jedoch nicht

überzubewerten. NAUCK et al. (1996a) weisen ausdrücklich darauf hin, dass der Einsatz dieses Modells nicht notwendigerweise zu einer verbesserten Klassifikationsleistung gegenüber der eines statistischen Regressionsverfahrens führen muss. Des Weiteren haben NAUCK und KLAWONN (1996) die bei Anwendung verschiedener Klassifikationsansätze auf den sog. „Wisconsin-Breast-Cancer"- Datensatz erzielten Resultate veröffentlicht. Auch hier schneidet ein vollständig durch den Lernvorgang erzeugtes NEFCLASS-System am schlechtesten ab (Fehlerrate 19,6 %), gefolgt vom Fuzzy-Clusteranalyseverfahren nach GUSTAFSON und KESSEL (1979) mit 13,8 %. Ein hervorragendes Klassifikationsergebnis (Fehlerrate 7,3 %) wird hingegen für ein mit den aus der Fuzzy-Clusteranalyse gewonnenen Regeln initialisiertes NEFCLASS-System angegeben.

Der entscheidende Nachteil des Systems ist darin zu sehen, dass es vom Nutzer experimentell (d. h. von Hand) parametrisiert werden muss. Insbesondere bei größeren Datensätzen mit vielen (Eingabe-) Variablen und (Ausgabe-) Klassen kann dies sehr aufwendig sein. Erinnert sei allein an die Anzahl möglicher Regeln, mit denen man das System initialisieren kann: Liegen beispielsweise vier Eingabegrößen vor, für deren Wertebereiche man eine Partitionierung auf drei Fuzzy-Mengen vornimmt, so ergeben sich bereits 3^4 Kombinationsmöglichkeiten für eine Prämisse, also 81 mögliche Regeln. Es sei nochmals darauf hingewiesen, dass unter einer „optimalen Parametrisierung des Systems" nicht nur das Erreichen einer geringen Fehlerrate zu verstehen ist, sondern auch eine möglichst einfache Interpretierbarkeit des Klassifikators, d. h., die Anzahl der in den Regelprämissen enthaltenen Größen sowie die Anzahl der Regeln selber sollte gering gehalten werden; dennoch sollten alle Klassen abgedeckt werden.

Der entscheidende Vorteil eines Neuro-Fuzzy-Klassifikationsverfahrens wie NEFCLASS gegenüber den konventionellen statistischen Verfahren besteht darin, dass man einen Klassifikator auf der Grundlage linguistischer Regeln erhält, der zum einen interpretierbar und zum anderen initialisierbar mit a priori-Wissen ist. Ein weiterer Aspekt soll ebenfalls hervorgehoben werden: Wird das System vollständig durch das Lernverfahren erzeugt (d. h. Erlernen der Wissensbasis in Form linguistischer Regeln und nachfolgende Optimierung der in den Regeln enthaltenen Fuzzy-Mengen), so erhält man nicht nur eine Lösung des (Klassifikations-) Problems, sondern zusätzlich noch Wissen über diese Lösung. Das Neuro-Fuzzy-System dient somit auch dem Wissenserwerb, einer wichtigen Eigenschaft der sog. intelligenten Systeme. Es entwickelt, ähnlich wie ein Mensch, ausgehend von Beobachtungswerten und mit einer Zielvorgabe, eine geeignete Vorgehensweise.

Betrachtet man die im Rahmen der **automatischen Klassifikation** mit Hilfe der hierarchisch agglomerativen und der nichthierarchisch optimierenden Clusteranalyse untersuchten Datensätze (s. Tab. 8-1 des Abschnitts 8.2.2.1), so lässt sich auch hier eine Gegenüberstellung der Klassifikationsergebnisse vornehmen. Hierzu wird die in den sechs Datensätzen (jeweils mit n = 19 Objekten) a priori vorhandene Gruppierung in Sicker- und Grundwassermessstellen mit der verglichen, die die Clusteranalyseverfahren für g = 2 liefern. Die sich hierbei ergebenden Fehleranzahlen und Fehlerraten sind in Tab. 12-2 zusammengefasst.

Tab. 12-2. Klassifikationsergebnisse der Methoden der automatischen Klassifikation.

Datensatz der Messstellen		Hierarchisch agglomerative CA				Nichthierarchisch optimierende CA	
		WARDs Methode		Average Linkage			
Variablen	Anzahl Variablenwerte	Fehler-anzahl	Fehler-rate	Fehler-anzahl	Fehler-rate	Fehler-anzahl	Fehler-rate
Na^+, K^+	16	2	10,53 %	6	31,58 %	6	31,58 %
Mg^{2+}, Ca^{2+}	16	2	10,53 %	6	31,58 %	4	21,05 %
Cl^-, SO_4^{2-}	7	6	31,58 %	6	31,58 %	6	31,58 %
Na^+/K^+	8	9	47,37 %	9	47,37 %	9	47,37 %
Mg^{2+}/Ca^{2+}	8	6	31,58 %	6	31,58 %	6	31,58 %
Cl^-/SO_4^{2-}	3	3	15,79 %	5	26,32 %	5	26,32 %

Vergleicht man die mit den beiden konservativen hierarchischen Verfahren ermittelten Fehlerraten, so schneidet WARDs Methode in drei von sechs Fällen deutlich besser ab. Ursache hierfür ist, dass in den betreffenden Datensätzen jeweils Ausreißer auftreten, die bei Anwendung von Average Linkage zu einer separaten Clusterbildung und damit aufgrund der Einteilung in lediglich zwei a priori-Gruppen zu einer relativ hohen Anzahl von Fehlklassifikationen führen. WARDs Methode hingegen nivelliert die Abstände zwischen den Objekten besser, es tritt ein abgestufteres Verhalten auf, wodurch die Ausreißer gewissermaßen in die Objektmenge integriert werden und damit Ähnlichkeiten zwischen diesen und anderen Objekten besser erkannt werden können. Abb. 12-1 verdeutlicht diesen Unterschied für die Clusterung in Abhängigkeit von den Na^+- und K^+-Ionen. Während bei Average Linkage für g = 2 der SWP 10 ein separates Cluster bildet, was zu einer relativ hohen Fehlerrate führt, wird bei WARDs Methode diese Messstelle in die vorhandene Objektmenge eingebunden und bleibt dennoch gut als (hochbelasteter) Ausreißer zu erkennen. Für g = 2 liegt hier das weitaus „realistischere" Cluster {SWP 1; SWP; 4; SWP 5; SWP 10; SWP 11} vor, es werden lediglich zwei Objekte (SWP 7 und 9) fehlklassifiziert, und die Fehlerrate sinkt damit auf 10,53 %.

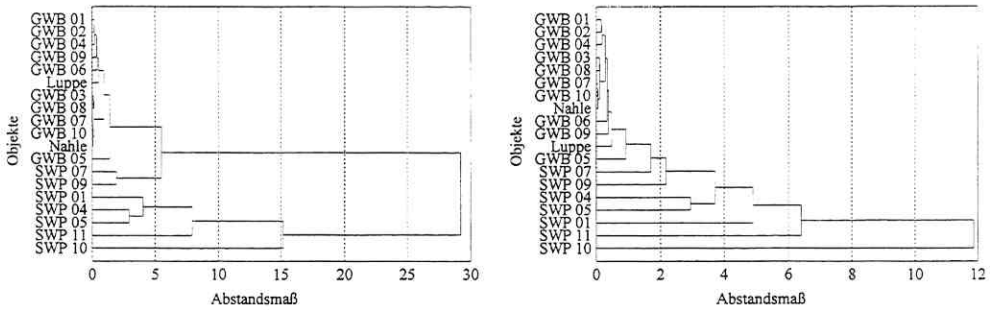

Abb. 12-1. Vergleich der Dendrogramme von WARDs Methode (l.) und Average Linkage (r.).

Unabhängig vom erläuterten Vorteil von WARDs Methode gegenüber Average Linkage sei jedoch hervorgehoben, dass die in den entsprechenden Untersuchungen (s. Abschnitt 8.2.2) ermittelten und diskutierten relevanten Ähnlichkeiten zwischen einzelnen Objekten mit beiden Verfahren festgestellt wurden und die Unterschiede auch insgesamt gesehen relativ gering waren. Die hier bereits getroffene Empfehlung, in ähnlichen praktischen Anwendungsfällen ebenfalls konservative Verfahren einzusetzen, kann nach den obigen Ausführungen nun dahingehend ergänzt werden, dass von diesen im Entscheidungsfall WARDs Methode vorzuziehen ist. Über die allgemeinen Vor- und Nachteile hierarchischer und nichthierarchischer Verfahren wurde in den entsprechenden Abschnitten 8.2 und 8.3 bereits diskutiert. Bezüglich der untersuchten Datensätze kann die Klassifikationsleistung der nichthierarchisch optimierenden Clusteranalyse der der beiden hierarchischen Verfahren als annähernd gleichwertig angesehen werden: Sie liegt etwas unter der von WARDs Methode, ist aber wiederum besser als die von Average Linkage, s. Tab. 12-2. Unter diesem Aspekt könnte sich - um den Aufwand der Untersuchungen möglichst gering zu halten - in ähnlichen Anwendungsfällen auf die Anwendung hierarchischer Verfahren (und hier wiederum auf die von WARDs Methode) beschränkt werden. Bei annähernd gleicher Klassifikationsleistung erhält man hier für die Untersuchungsergebnisse ein eindeutiges Gesamtbild in Form eines Dendrogramms, wodurch die Interpretation wesentlich vereinfacht wird.

Es muss abschließend betont werden, dass zur vergleichenden Bewertung der Verfahren bezüglich ihrer Klassifikationsleistung ausschließlich die in den Tab. 12-1 bzw. 12-2 genannten Datensätze untersucht wurden. Die daraus abgeleiteten Empfehlungen, welche Verfahren vorzuziehen sind, können verallgemeinert werden und sind damit für ähnliche Anwendungsfälle relevant. Sie sollten nach Möglichkeit jedoch auch hier zunächst durch individuelle Untersuchungen überprüft werden.

13 Literaturverzeichnis

AHRENS, H. (1967): *Varianzanalyse*.- Akademie-Verlag, Berlin

AHRENS, H. und J. LÄUTER (1981): *Mehrdimensionale Varianzanalyse*.- 2. Aufl., Akademie-Verlag, Berlin

ANDREAS, L. (1993): *Deponieforschung im Land Sachsen - Untersuchung der Ausbreitung von Schadstoffen in Hausmülldeponien*.- Diplomarbeit, TU Dresden, Fachbereich Wasserwirtschaft, Dresden

BARKOWSKI, D., P. GÜNTHER, E. HINZ und R. RÖCHERT (1993): *Altlasten. Handbuch zur Ermittlung und Abwehr von Gefahren durch kontaminierte Standorte*.- 4. Aufl., Verlag C. F. Müller, Karlsruhe

BAUMANN, Th., M. BAUMANN und R. NIESSNER (1993): *Hydrogeologische und hydrochemische Untersuchungen im Einflußbereich einer Hausmülldeponie. Teil 1: Räumliche Erfassung der Kontamination*.- In: Vom Wasser **81**, S. 105 - 122

BERGS, C.-A. (1997): *TA Siedlungsabfall. Technische Anleitung zur Verwertung, Behandlung und sonstigen Entsorgung von Siedlungsabfällen*.- 2. Aufl., Erich Schmidt Verlag GmbH & Co., Berlin

BEZDEK, J. C. (1981): *Pattern Recognition with fuzzy objective function algorithms*.- Plenum-Verlag, New York

BEZDEK, J. C., R. EHRLICH und W. FULL (1984): *FCM: The Fuzzy c-Means Clustering Algorithm*.- In: Computers & Geosciences **10**, S. 191 - 203

BOCKLISCH, St. F. (1987): *Prozeßanalyse mit unscharfen Verfahren*.- VEB Verlag Technik, Berlin

BORTZ, J. (1993): *Statistik für Sozialwissenschaftler*.- 4. Aufl., Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona und Budapest

BOX, G. E. P. und G. M. JENKINS (1976): *Time Series Analysis, Forecasting and Control*.- Verlag Holden-Day, San Francisco

BOX, G. E. P. und D. A. PIERCE (1970): *Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models*.- In: J. Amer. Statist. Assoc. **65**, S. 1509 - 1526

BROEKAERT, J. A. C. (1992): *Einsatz der ICP-Atomspektrometrie in der Wasseranalytik*.- In: Technisches Messen **4**, S. 147 - 153

BRONSTEIN, I. N. und K. A. SEMENDJAJEW (1983): *Taschenbuch der Mathematik.*- 21. Aufl., Verlag Nauka und BSB B. G. Teubner Verlagsgesellschaft, Moskau und Leipzig

BUNDESANSTALT FÜR GEOWISSENSCHAFTEN UND ROHSTOFFE (Hrsg.) (1995): *Forschungsverbundvorhaben „Methoden zur Erkundung und Beschreibung des Untergrundes von Deponien und Altlasten" - kurz „Deponieuntergrund" - Schlußbericht für den Bewilligungszeitraum vom 01.11.1989 bis zum 31.01.1995.*- Hannover

CZURDA, K. A. (1992): *Relevante Schadstoff-Ausbreitungspfade in den Systemen Boden und Grundwasser.*- In: Czurda, K. A. (Hrsg.), Deponie und Altlasten. Sickerwasser- und Grundwassersanierung, S. 1 - 16, EF-Verlag für Energie- und Umwelttechnik GmbH, Berlin

DASSOW, W., J. FISCHER und H.-J. KRAMER (1992): *Gutachten zur Gefährdungsabschätzung der Deponie Möckern.*- Gutachten im Auftrag des Amtes für Umwelt der Stadt Leipzig, G.E.O.S. Freiberg Ingenieurgesellschaft mbH, Niederlassung Leipzig, Leipzig

DAUS, B. (1996): *Charakterisierung der Bindungsform von Schwermetallen in regionalen Flußsedimenten und deren chemometrische Interpretation.*- Dissertation, Martin-Luther-Universität Halle-Wittenberg, Mathematisch-Naturwissenschaftlich-Technische Fakultät, Halle (Saale)

DENGEL, A. (1994): *Künstliche Intelligenz. Allgemeine Prinzipien und Modelle.*- B.I.-Taschenbuchverlag, Mannheim, Leipzig, Wien und Zürich

DER RAT VON SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN (Hrsg.) (1990): *Altlasten. Sondergutachten Dezember 1989.*- Verlag Metzler-Poeschel, Stuttgart

DER RAT VON SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN (Hrsg.) (1995): *Altlasten II. Sondergutachten Februar 1995.*- Verlag Metzler-Poeschel, Stuttgart

DILGER, W. 1997: *Einführung in die Künstliche Intelligenz.*- Skript zur Vorlesung im Sommersemester 1997, TU Chemnitz-Zwickau, Fakultät für Informatik, Chemnitz

DIN 38402 - A 13 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Allgemeine Angaben (Gruppe A). Probenahme aus Grundwasserleitern (A 13).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38402 - A 15 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Allgemeine Angaben (Gruppe A). Probenahme aus Fließgewässern (A 15).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38404 - C 4 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Physikalische und physikalisch-chemische Kenngrößen (Gruppe C). Bestimmung der Temperatur (C 4)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38404 - C 5 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Physikalische und physikalisch-chemische Kenngrößen (Gruppe C). Bestimmung des pH-Wertes (C 5)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38404 - C 8 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Physikalische und physikalisch-chemische Kenngrößen (Gruppe C). Bestimmung der elektrischen Leitfähigkeit (C 8)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38405 - D 1 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Anionen (Gruppe D). Bestimmung des Chlorid-Ions (D 1)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38405 - D 5 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Anionen (Gruppe D). Bestimmung der Sulfat-Ionen (D 5)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38405 - D 19 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Anionen (Gruppe D). Bestimmung der Anionen Fluorid, Chlorid, Nitrit, Phosphat (ortho-), Bromid, Nitrat und Sulfat in wenig belasteten Wässern mit der Ionenchromatographie (D 19)*.- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38406 - E 3 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Kationen (Gruppe E). Bestimmung von Calcium und Magnesium (E 3).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38406 - E 22 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Kationen (Gruppe E). Bestimmung der 33 Elemente Ag, Al, As, B, Ba, Be, Bi, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Se, Si, Sn, Sr, Ti, V, W, Zn und Zr durch Atomemissionsspektrometrie mit induktiv gekoppeltem Plasma (ICP-OES) (E 22).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38409 - H 3 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Summarische Wirkungs- und Stoffkenngrößen (Gruppe H). Bestimmung des gesamten organisch gebundenen Kohlenstoffs (TOC) (H 3).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38409 - H 6 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Summarische Wirkungs- und Stoffkenngrößen (Gruppe H). Härte eines Wassers (H 6).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN 38409 - H 41 (1989): *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Summarische Wirkungs- und Stoffkenngrößen (Gruppe H). Bestimmung des Chemischen Sauerstoffbedarfs (CSB) im Bereich über 15 mg/l (H 41).*- In: Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker in Gemeinschaft mit dem Normenausschuß Wasserwesen (NAW) im DIN Deutsches Institut für Normung e. V. (Hrsg.), Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, 22. Lieferung, VCH Verlagsgesellschaft mbH, Weinheim

DIN ISO 5725 (1988): *Präzision von Meßverfahren. Ermittlung der Wiederhol- und Vergleichspräzision von festgelegten Meßverfahren durch Ringversuche.*- Beuth Verlag GmbH, Berlin

EHRIG, H. J. (1989): *Sickerwasser aus Hausmülldeponien - Menge und Zusammensetzung*.- In: Hösel, G. (Hrsg.), Müll-Handbuch. Sammlung und Transport, Behandlung und Ablagerung sowie Vermeidung und Verwertung von Abfällen. Ergänzbares Handbuch für die kommunale und industrielle Abfallwirtschaft, Erich Schmidt Verlag GmbH & Co., Berlin

EINAX, J. W., K. OSWALD und K. DANZER (1990): *Analytical investigations and chemometrical characterization of polluted soils*.- In: Fresenius J. Anal. Chem. **336**, S. 394 - 399

EINAX, J. W., H. W. ZWANZIGER und S. GEIß (1997): *Chemometrics in Environmental Analysis*.- VCH Verlagsgesellschaft mbH, Weinheim

EISENHART, Ch. (1947): *The assumptions underlying the analysis of variance*.- In: Biometrics **3**, S. 1 - 21

ERTEL, Th., B. BESSEY, F. KERN und A. MAURER (1997*): Vor-Ort-Analytik bei der Altlastenbearbeitung - Ergebnisse eines Feldtests*.- In: TerraTech **3**, S. 19 - 22

EVERITT, B. S. (1979): *Unresolved Problems in Cluster Analysis*.- In: Biometrics **35**, S. 169 - 181

FISCHER, J. (1993): *Bericht über orientierende Untersuchungen zur Sickerwassersituation an der Deponie Leipzig-Möckern*.- Bericht im Auftrag des Amtes für Umwelt der Stadt Leipzig, G.E.O.S. Freiberg Ingenieurgesellschaft mbH, Niederlassung Leipzig, Leipzig

FISHER, R. A. (1936): *The Use of Multiple Measurements in Taxonomic Problems*.- In: Annals of Eugenics **7**, S. 179 - 188

FIX, E. und J. HODGES (1951): *Discriminatory analysis, nonparametric discrimination: consistency properties*.- In: USAF School of Aviation Medicine **4**, Randolph Field

FLACHOWSKY, J. (1996): *Modellhafte Bestimmung des Gefährdungspotentials der Altdeponie Möckern bezüglich der Gewässer des nördlichen Leipziger Auwaldes*.- Abschlussbericht zum Teilprojekt REGNAL III.3, Umweltforschungszentrum Leipzig-Halle GmbH, Sektion Analytik, Leipzig

FLACHOWSKY, J. (1998): *Mobile Umweltanalytik*.- In: Günzler, H. (Hrsg.): Analytiker Taschenbuch 18, S. 143 - 179, Springer-Verlag, Berlin, Heidelberg und New York

FLEISCHER, W. und M. NAGEL (1989): *Datenanalyse mit dem Personalcomputer*.- VEB Verlag Technik, Berlin

FLETCHER, R. (1969): *Optimization*.- Verlag Academic Press, New York

FLETCHER, R. und M. J. D. POWELL (1963): *A rapidly convergent descent method for minimization*.- In: Computer Journal **6**, S. 163 - 168

FORGY, E. W. (1965): *Cluster analysis of multivariate data. Efficiency versus interpretability of classifications*.- In: Biometrics **21**, S. 768 - 769

FRIEDERICHS, M., O. FRÄNZLE und A. SALSKI (1996): *Classifying Existing Chemicals According to Their Ecotoxicological Properties - A Fuzzy Clustering Approach*.- In: ECOSYS 4, S. 107 - 120, Kiel

FUNK, P. und H. KLEIN (1994): *Steuerungsparameter CSB contra Überwachungsparameter TOC?*.- Anwendungsbericht Ch. No. 35, Fa. Dr. Bruno Lange GmbH, Berlin

GADE, B. (1994): *Bilanzierung einer modernen Sonderabfalldeponie am Beispiel der SAD Raindorf*.- In: Wasser, Luft und Boden **5**, S. 84 - 89

GARDNER, E. S., Jr. (1985): *Exponential smoothing: The state of the art*.- In: Journal of Forecasting **4**, S. 1 - 28

GÖHLER, W. (1989*): Höhere Mathematik. Formeln und Hinweise. Kleiner Wissensspeicher*.- 10. Aufl., VEB Deutscher Verlag für Grundstoffindustrie, Leipzig

GROTELÜSCHEN, M. (1997): *Data Mining findet Informationen in riesigen Datenlagern*.- In: VDI nachrichten **13**, S. 10

GUHL, W. und U. WERNER (1997): *Wasseranalytik-Summenparameter*.- In: Nachrichten aus Chemie, Technik und Laboratorium **4**, S. 15 - 19

GUSTAFSON, E. E. und W. C. KESSEL (1979*): Fuzzy Clustering with a Fuzzy Covariance Matrix*.- In: Proc. IEEE CDC, S. 761 - 766, San Diego

HEIN, H. und W. KUNZE (1995): *Umweltanalytik mit Spektrometrie und Chromatographie. Von der Laborgestaltung bis zur Dateninterpretation*.- 2. Aufl., VCH Verlagsgesellschaft mbH, Weinheim, New York, Basel, Cambridge und Tokyo

HEINRICH, I. (1995): *Anforderungen an den Einsatz künstlicher neuronaler Netze zur Lastprognose in Stromversorgungsunternehmen*.- In: Elektrizitätswirtschaft **6**, S. 303 - 306

HENRION, G., A. HENRION und R. HENRION (1988): *Beispiele zur Datenanalyse mit BASIC-Programmen*.- VEB Deutscher Verlag der Wissenschaften, Berlin

HENRION, R. (1998): *Simultaneous simplification of loading and core matrices in N-way PCA: application to chemometric data arrays*.- In: Fresenius J. Anal. Chem. **361**, S. 15 - 22

HENRION, R. und G. HENRION (1994): *Multivariate Datenanalyse. Methodik und Anwendung in der Chemie und verwandten Gebieten*.- Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong; Barcelona und Budapest

HERING, E., J. GUTEKUNST und U. DYLLONG (1995): *Informatik für Ingenieure*.- VDI-Verlag, Düsseldorf

HÖLTING, B. (1995): *Hydrogeologie. Einführung in die allgemeine und angewandte Hydrogeologie*.- 5. Aufl., Enke-Verlag, Stuttgart

HOFERICHTER, T. (1996): *Entwurf und Implementierung eines UNIX-basierten Softwaretools zur Realisierung des Neuro-Fuzzy-Klassifikationsansatzes NEFCLASS*.- Diplomarbeit, TU Braunschweig, Institut für Betriebssysteme und Rechnerverbund, Braunschweig

JANDEL, A. S. (1998): *Immer mehr Schadstoffe verschmutzen Grundwasser*.- In: VDI nachrichten **1/2**, S. 33

JOHNSON, R. A. und D. W. WICHERN (1982): *Applied Multivariate Statistical Analysis*.- Verlag Prentice Hall, New Jersey

KAISER, U. (1995): *Energieüberwachungs- und Lastprognosesystem*.- Diplomarbeit, TU Chemnitz-Zwickau, Fakultät für Elektrotechnik und Informationstechnik, Chemnitz

KANNE, M. (1996): *Neuro-Fuzzy-Methoden in der Datenanalyse*.- Diplomarbeit, TU Braunschweig, Institut für Betriebssysteme und Rechnerverbund, Braunschweig

KNOBLOCH, S. und H. W. ZWANZIGER (1995*): Probleme und Artefakte bei der chemometrischen Aufbereitung von Umweltdaten*.- In: GIT Fachzeitschrift für das Laboratorium **6**, S. 535 - 541

KOBER, A. J. (1997): *Untersuchung und Bewertung von Einflußfaktoren auf die Lastprognose mittels Fuzzy Pattern Klassifikation*.- Diplomarbeit, TU Chemnitz-Zwickau, Fakultät für Elektrotechnik und Informationstechnik, Chemnitz

KOEHNE, W. (1948): *Grundwasserkunde*.- 2. Aufl., E. Schweizerbart' sche Verlagsbuchhandlung, Stuttgart

KRUSE, R., D. NAUCK und F. KLAWONN (1997): *Neuronale Fuzzy-Systeme*.- In: Spektrum der Wissenschaft, Dossier **4**, S. 92 - 99

LACHENBRUCH, P. A. (1967): *An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis*.- In: Biometrics **23**, S. 639 - 649

LANCE, G. N. und W. T. WILLIAMS (1966): *A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems*.- In: Computer Journal **9**, S. 373 - 380

LEIDRAAD BODEMSANERING (1995): *Niederländischer Leitfaden zur Bodenbewertung und Bodensanierung*.- In: Rosenkranz, D., G. Bachmann, G. Einsele und H.-M. Harres (Hrsg.), Bodenschutz. Ergänzbares Handbuch der Maßnahmen und Empfehlungen für Schutz, Pflege und Sanierung von Böden, Landschaft und Grundwasser **8936**, 18. Lieferung, Erich Schmidt Verlag GmbH & Co., Berlin

LESCHBER, R., K.-D. PERNAK und U. ZIMMERMANN (1993): *Anorganische und organische Inhaltsstoffe im Regenwasserabfluß und ihr Verhalten in der Untergrundpassage*.- In: Leschber, R., U. Müller-Wegener und R. Schmidt (Hrsg.), Boden- und Grundwasserverunreinigungen aus Punkt- und Flächenquellen, S. 71 - 91, Gustav Fischer Verlag, Stuttgart und New York

LINDMAN, H. R. (1974): *Analysis of variance in complex experimental designs*.- Verlag W. H. Freemann & Co., San Francisco

LÖHN, A. und W. SCHAEFER (1998): *Wissensbasierte Überwachung von Analysensystemen*.- In: GIT Labor-Fachzeitschrift **2**, S. 126 - 127

LUTHER, Th. (1997): *Kurse aus dem Chaos*.- In: FINANZtest **12**, S. 50 - 51

MACQUEEN, J. (1967): *Some Methods for Classification and Analysis of Multivariate Observations*.- In: Lecam, L. M. und J. Neymann (Hrsg.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I: Theory of Statistics, S. 281 - 297, Verlag University of California Press, Berkeley und Los Angeles

MARTIN, G. (1998): *Electronic Commerce: Aufschwung mit künstlicher Intelligenz. Im Internet auf Schnäppchenjagd*.- In: VDI nachrichten **24**, S. 14

MATTHES, W. (1995): *Bestimmung der Wechselwirkung zwischen Sickerwasseraustritt einer Deponie und der Untergrundbelastung als Funktion von Elementverteilung und Korngröße unter Einsatz mobiler Analysentechniken*.- Diplomarbeit, TU Bergakademie Freiberg, Fachbereich Geowissenschaften, Freiberg (Sachsen)

MELARD, G. (1984): *A fast algorithm for the exact likelihood of autoregressive-moving average models*.- In: Applied Statistics **33**, S. 104 - 119

NAGEL, M., W. FLEISCHER und K. HENSCHKE (1988): *Explorative Datenanalyse*. In: Wissenschaft und Fortschritt **11**, S. 299 - 302

NAUCK, D. und F. KLAWONN (1996): *Neuro-Fuzzy Classification Initialized by Fuzzy Clustering*.- In: Proc. Fourth European Congress on Intelligent Techniques and Soft Computing (EUFIT '96), Aachen

NAUCK, D., F. KLAWONN und R. KRUSE (1996a): *Neuronale Netze und Fuzzy-Systeme. Grundlagen des Konnektionismus, Neuronaler Fuzzy-Systeme und der Kopplung mit wissensbasierten Methoden*.- 2. Aufl., Verlag Vieweg, Braunschweig und Wiesbaden

NAUCK, D., U. NAUCK und R. KRUSE (1996b): *Generating Classification Rules with the Neuro-Fuzzy-System NEFCLASS*.- In: Proc. Biennial Conf. of the North American Fuzzy Information Processing Society (NAFIPS '96), Berkeley

NEUMANN, I. und B. ZIELONKA (1992): *Neuronale Netze zur Lastprognose*.- In: Elektrotechnische Zeitschrift **6-7**, S. 374 - 378

OSE, G., G. SCHIEMANN, H. BAUMANN, F. STOPP, W. KÖRNER und G. LOCHMANN (1974): *Ausgewählte Kapitel der Mathematik*.- 8. Aufl., VEB Fachbuchverlag, Leipzig

PAASCH, J. (1994): *Weiterentwicklung und Portierung eines Unterstützungssystems zur Fuzzy-Cluster-Analyse*.- Diplomarbeit, Christian-Albrechts-Universität Kiel, Institut für Informatik und Praktische Mathematik, Kiel

PASS, U. und K. D. SCHMIDT (1998): *Zeit- und kostenbewußt analysieren. Die Rolle der Summenparameter in der Umweltanalytik*.- In: Umwelt **5/6**, S. 33 - 36

RIESE, B. (1998): *Deponie mit laschen Standards*.- In: VDI nachrichten **27**, S. 5

ROUBENS, M. (1982): *Fuzzy clustering algorithms and their cluster validity*.- In: European Journal of Operational Research **10**, S. 294 - 301

RUBIN, J. (1967*): Optimal Classification into Groups. An Approach for Solving the Taxonomy Problem*.- In: Journal of Theoretical Biology **15**, S. 103 - 144

RUDOLPH, M. (1998a): *Anwendung von Methoden der Mustererkennung zur Bewertung der Schadstoffverteilung im Einzugsbereich von Deponien*.- Zwischenbericht zum Stand der Dissertationsarbeit, Umweltforschungszentrum Leipzig-Halle GmbH, Sektion Analytik, Leipzig

RUDOLPH, M. (1998b): *Expertise zur Anwendung eines Verfahrens zur Lastvorhersage von Elektroenergie im Kurzfristbereich*.- Studie zur vergleichenden Bewertung von Lastprognoseverfahren im Auftrag der Energieversorgung Südsachsen AG, Gesellschaft für Wissens- und Technologietransfer der TU Dresden mbH, Chemnitz

RUMP, H. H. (1998): *Laborhandbuch für die Untersuchung von Wasser, Abwasser und Boden*.- 3. Aufl., WILEY-VCH Verlag GmbH, Weinheim, Chichester, New York, Toronto, Brisbane und Singapore

SÄCHSISCHES LANDESAMT FÜR UMWELT UND GEOLOGIE (Hrsg.) (1998): *Grundwasser*.- In: Sächsisches Landesamt für Umwelt und Geologie (Hrsg.), Materialien zur Altlastenbehandlung 3/1998. Probenahme bei der Technischen Erkundung von Altlasten, S. 24 - 43, Lößnitz-Druck GmbH, Radebeul

SCHÄFER, W. (1978): *Grundlagen der Ingenieurmathematik. Band 1*.- Vorlesungsauszug, TH Leipzig, Sektion Mathematik und Rechentechnik, Leipzig

SCHÄFER, W. und H.-G. BUSSE (1978): *Grundlagen der Ingenieurmathematik. Band 3.*- Vorlesungsauszug, TH Leipzig, Sektion Mathematik und Rechentechnik, Leipzig

SCHIELE-TRAUTH, U. (1997): *Neuronale Netze als Wetterfrosch.*- In: VDI nachrichten **44**, S. 28

SCHMIDT, E. (1997): *Autonome Agenten sollen Menschen entlasten.*- In: Computer und Kommunikation, Sonderteil der VDI nachrichten **43** zur Messe Systems, S. S5

SCHULZE, P. M. (1998): *Beschreibende Statistik.*- 3. Aufl., R. Oldenbourg-Verlag, München und Wien

STATSOFT, Inc. (Hrsg.) (1996): *STATISTICA für Windows.*- Computer-Handbuch, Tulsa

STEINHAUSEN, D. und K. LANGER (1977): *Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation.*- de Gruyter-Verlag, Berlin und New York

STORM, R. (1995): *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle.*- 10. Aufl., Fachbuchverlag, Leipzig und Köln

TUKEY, J. W. (1962): *The future of data analysis.*- In: Ann. Math. Statist. **33**, S. 1 - 67

UMWELTBUNDESAMT (Hrsg.) (1995): *Verbundvorhaben Deponiekörper. Tagungsband zum 1. Statusseminar am 25. und 26. April 1995 in Wuppertal.*- Berlin

UMWELTBUNDESAMT (Hrsg.) (1997): *Verbundvorhaben Deponiekörper. Tagungsband zum 2. Statusseminar am 4. und 5. Februar 1997 in Wuppertal.*- Berlin

WAGNER, K. (1995): *TA Abfall. Die Verwaltungsvorschriften zum Abfallgesetz.*- In: Müller, R. und G. Schmitt-Gleser (Hrsg.), Handbuch der Abfallentsorgung. Abfallrecht, TA Abfall, Entsorgungstechnologie, Altlasten, angrenzende Rechtsbereiche, Management, Loseblatt-Ausgabe, Ecomed-Verlagsgesellschaft, Landsberg

WINDHAM, M. P. (1981): *Cluster validity for fuzzy clustering algorithms.*- In: Fuzzy Sets and Systems **5**, S. 177 - 185

WÜBBOLD, S., W. SPICKERMANN und G. STORK (1987): *Untersuchungen an Lahnsedimenten. 3. Sättigungskapazität und Sorptionskonstante.*- In: Fresenius Z. Anal. Chem. **328**, S. 648 - 652

ZADEH, L. A. (1965): *Fuzzy-Sets.*- In: Information and Control **8**, S. 338 - 353

ZWANZIGER, H. W. (1988): *Mustererkennung und Mehrkomponentenanalytik mit multivariaten chemometrischen Methoden.*- Dissertation (B), Karl-Marx-Universität Leipzig, Sektion Chemie, Leipzig

14 Anhang

Tab. 14-1 verschafft eine Gesamtübersicht über sämtliche der im Anhang enthaltenen Tabellen.

Tab. 14-1. Gesamtübersicht über die im Anhang enthaltenen Tabellen.

Tab. A-1. Statistische Maßzahlen (Sickerwasser).

Parameter			Anzahl	Mittelw.	Median	Minimum	Maximum	unteres	oberes	range	Quartils-	Std.-abw.	Varianz
Nr.	Name	Methode	n	$\bar{x}$	$\tilde{x}$	x_{min}	x_{max}	Quartil	Quartil	R	abstand	s	s^2
1	Na^+ (mval/l)	AAS	119	140,1304	83,8198	2,2880	866,2972	51,0226	129,7532	864,0093	78,7306	180,461	32565,984
2	K^+ (mval/l)	AAS	118	17,5662	10,1056	2,2250	181,9600	6,8794	19,7688	179,7351	12,8894	21,051	443,159
3	Mg^{2+} (mval/l)	ICP-AES	119	51,2731	38,4172	0,0541	233,7940	6,4084	55,4788	233,7399	49,0704	58,543	3427,227
4	Ca^{2+} (mval/l)	ICP-AES	119	28,0570	26,2974	1,3398	67,2655	21,9760	34,6307	65,9256	12,6547	11,368	129,242
5	Cl^- (mval/l)	IC	82	127,4533	95,7391	0,9809	484,4978	79,0745	152,5002	483,5169	73,4257	100,111	10022,233
		Titration	71	121,7330	91,4118	22,9525	409,4216	77,7320	150,5317	386,4691	72,7997	87,410	7640,430
		Elektroden	26	127,7206	96,3247	67,6953	518,9970	84,9011	132,8519	451,3017	47,9508	89,247	7965,050
6	SO_4^{2-}	IC	85	160,0233	72,1148	0,5208	955,5702	44,5945	152,6414	955,0494	108,0469	226,982	51520,807
	(mval/l)	Küvetten	17	197,3776	137,4118	28,1070	709,4406	49,9679	152,7874	681,3336	102,8194	246,750	60885,342
		Säule	46	139,0026	57,4133	26,4310	777,3554	44,3946	124,2188	750,9244	79,8242	186,345	34724,534
		gravimetrisch	3	94,0269	109,6078	39,9558	132,5171	---	---	92,5612	---	48,208	2323,9690
7	pH-Wert	HORIBA	130	7,9193	7,2500	6,5050	10,1000	7,0000	9,4000	3,5950	2,4000	1,203	1,448
		Elektrode	15	7,5890	7,4600	6,8200	9,0100	7,2100	7,6750	2,1900	0,4650	0,617	0,381
8	Leitfähigkeit	HORIBA	130	25,4512	19,0000	4,3000	97,1000	14,7000	31,9000	92,8000	17,2000	18,727	350,702
	(mS/cm)	Elektrode	15	18,4853	16,9700	12,4900	27,5000	14,7200	25,5000	15,0100	10,7800	5,159	26,614
9	Temperatur	HORIBA	126	20,3456	21,0000	7,2000	29,6000	17,9000	24,0000	22,4000	6,1000	4,826	23,294
	(°C)	Temp.-fühler	5	16,1500	18,8500	8,3000	21,0000	13,1000	19,5000	12,7000	6,4000	5,314	28,240
10	Härte (°dH)	Titration	124	241,8548	185,3320	59,7551	795,4687	118,3000	241,4400	735,7136	123,1400	174,737	30533,077
11	CSB (mg/l)	Titration	90	487,2204	257,8998	81,6857	3999,236	214,9628	587,7745	3917,5502	372,8117	545,211	297255,369
		Küvetten	39	482,0511	320,0000	66,0000	2006,000	270,0000	656,0000	1940,0000	386,0000	411,389	169241,149
		Küvettensatz	27	832,1599	564,0000	71,9000	3514,000	235,0000	1306,000	3442,1000	1071,0000	805,381	648638,625
12	DOC (mg/l)	liquiTOC	120	99,4160	50,7325	17,6850	398,6000	36,2250	147,9750	380,9150	111,7500	99,252	9850,963
13	W.-stand (m)	Lichtlot	121	105,0293	105,1600	102,9700	114,5000	104,4700	105,5000	11,5300	1,0300	1,3339	1,7793

Tab. A-2. Statistische Maßzahlen (Grundwasser).

Parameter			Anzahl	Mittelw.	Median	Minimum	Maximum	unteres	oberes	range	Quartils-	Std.-abw.	Varianz
Nr.	Name	Methode	n	$\bar{x}$	$\tilde{x}$	x_{min}	x_{max}	Quartil	Quartil	R	abstand	s	s^2
1	Na^+ (mval/l)	AAS	170	8,2908	4,9870	0,7395	45,2375	2,6099	10,9614	44,4980	8,3515	8,275	68,478
2	K^+ (mval/l)	AAS	170	0,7540	0,3325	0,0153	9,1300	0,1867	0,8900	9,1146	0,7033	1,034	1,068
3	Mg^{2+} (mval/l)	ICP-AES	170	8,1699	5,4541	1,2134	31,1780	4,0556	10,7601	29,9646	6,7045	6,494	42,174
4	Ca^{2+} (mval/l)	ICP-AES	170	10,1487	9,4548	3,0888	31,0878	7,1507	12,0933	27,9990	4,9426	4,879	23,799
5	Cl^- (mval/l)	IC	157	8,1710	3,7531	0,2970	154,0772	1,6949	10,2843	153,7802	8,5894	14,750	217,572
		Titration	105	8,9544	5,3169	0,7032	47,2575	2,9971	11,5844	46,5542	8,5873	8,799	77,428
		Elektroden	69	9,8044	6,7977	0,0313	38,3606	3,8361	14,3006	38,3293	10,4646	8,388	70,357
6	SO_4^{2-}	IC	155	15,6647	13,1370	1,0604	86,6486	7,6724	20,9434	85,5882	13,2711	11,365	129,171
	(mval/l)	Küvetten	11	16,9220	16,1355	6,2044	31,6464	7,6305	23,4225	25,4420	15,7920	8,680	75,346
		Säule	76	15,4482	14,5262	0,3198	43,0765	9,5212	20,1354	42,7567	10,6142	8,040	64,645
		gravimetrisch	11	15,9576	16,5116	7,3747	23,4319	10,8430	21,2249	16,0572	10,3819	5,720	32,714
7	pH-Wert	HORIBA	240	6,6874	6,6000	5,7000	8,0700	6,5000	6,8000	2,3700	0,3000	0,331	0,110
		Elektrode	50	6,9491	6,9200	6,4650	7,8500	6,7700	7,0800	1,3850	0,3100	0,266	0,071
8	Leitfähigkeit	HORIBA	240	2,7228	2,1050	0,0200	12,0000	1,5525	3,4150	11,9800	1,8625	1,832	3,355
	(mS/cm)	Elektrode	50	2,7271	2,1850	0,6800	6,8100	1,6655	3,2600	6,1300	1,5945	1,392	1,937
9	Temperatur	HORIBA	240	12,3577	12,0000	6,0000	22,8500	10,3000	14,0000	16,8500	3,7000	2,912	8,482
	(°C)	Temp.-fühler	22	11,5261	10,7000	6,4750	16,8000	10,2000	12,3000	10,3250	2,1000	2,399	5,757
10	Härte (°dH)	Titration	185	53,9758	47,3200	13,1797	150,1187	37,5648	66,9200	136,9391	29,3552	28,009	784,483
11	CSB (mg/l)	Titration	119	24,8411	18,3479	4,6557	124,3716	13,9509	28,1438	119,7159	14,1930	19,188	368,187
		Küvetten	99	50,2545	29,8750	6,5000	763,5000	19,0000	49,0000	757,0000	30,0000	88,784	7882,607
		Küvettensatz	37	14,8432	12,5250	0,0000	62,0000	2,8800	21,9667	62,0000	19,0867	14,135	199,801
12	DOC (mg/l)	liquiTOC	164	7,4186	6,3119	2,4015	23,4200	4,5097	9,1285	21,0185	4,6187	4,050	16,401
13	W.-stand (m)	Lichtlot	242	100,8322	100,7450	99,4200	102,4500	100,5000	101,1400	3,0300	0,6400	0,506	0,256

Tab. A-3. Korrelationsmatrix ausgewählter Messgrößen (Sickerwasser).

Variable	Na^+	K^+	Mg^{2+}	Ca^{2+}	Cl^-	SO_4^{2-}	pH-Wert	Leitf.	W.-härte	CSB	DOC	W.-stand
Na^+ (mval/l)	1,0000 (n = 119)	**0,7913** (n = 118)	**0,4261** (n = 119)	-0,00357 (n = 119)	**0,5783** (n = 82)	**0,6325** (n = 85)	**-0,2506** (n = 119)	**0,7563** (n = 119)	**0,5042** (n = 119)	**0,4072** (n = 89)	**0,6636** (n = 119)	0,1939 (n = 101)
K^+ (mval/l)		1,0000 (n = 118)	**0,3737** (n = 118)	-0,12703 (n = 118)	**0,4166** (n = 81)	**0,4505** (n = 84)	**-0,3624** (n = 118)	**0,5296** (n = 118)	**0,3874** (n = 118)	**0,2554** (n = 89)	**0,5309** (n = 118)	0,0783 (n = 100)
Mg^{2+} (mval/l)			1,0000 (n = 119)	**-0,25924** (n = 119)	**0,5049** (n = 82)	**0,7282** (n = 85)	**-0,5427** (n = 119)	**0,6162** (n = 119)	**0,9002** (n = 119)	**0,4337** (n = 89)	**0,6195** (n = 119)	0,0462 (n = 101)
Ca^{2+} (mval/l)				1,00000 (n = 119)	-0,0225 (n = 82)	-0,0766 (n = 85)	**0,5453** (n = 119)	-0,0171 (n = 119)	-0,0873 (n = 119)	-0,1615 (n = 89)	-0,0999 (n = 119)	0,1033 (n = 101)
Cl^- (mval/l)					1,0000 (n = 82)	**0,8569** (n = 82)	**-0,3616** (n = 82)	**0,9321** (n = 82)	**0,5398** (n = 82)	**0,7196** (n = 62)	**0,7087** (n = 82)	**0,2564** (n = 69)
SO_4^{2-} (mval/l)						1,0000 (n = 85)	**-0,3412** (n = 85)	**0,9020** (n = 85)	**0,7442** (n = 85)	**0,6689** (n = 65)	**0,7178** (n = 85)	0,2160 (n = 72)
pH-Wert							1,0000 (n = 130)	**-0,3210** (n = 130)	**-0,4671** (n = 120)	**-0,3239** (n = 90)	**-0,3506** (n = 120)	**0,2865** (n = 111)
Leitfähigkeit (mS/cm)								1,0000 (n = 130)	**0,7204** (n = 120)	**0,7468** (n = 90)	**0,8610** (n = 120)	**0,2657** (n = 111)
Wasserhärte (°dH)									1,0000 (n = 124)	**0,5646** (n = 90)	**0,6966** (n = 120)	0,1178 (n = 102)
CSB (mg/l)										1,0000 (n = 90)	**0,7260** (n = 90)	**0,2622** (n = 76)
DOC (mg/l)											1,0000 (n = 120)	**0,3563** (n = 102)
Wasserspiegelstand (m)												1,0000 (n = 121)

Tab. A-4. Korrelationsmatrix ausgewählter Messgrößen (Grundwasser).

Variable	Na^+	K^+	Mg^{2+}	Ca^{2+}	Cl^-	SO_4^{2-}	pH-Wert	Leitf.	W.-härte	CSB	DOC	W.-stand
Na^+ (mval/l	1,0000 (n = 170)	**0,8067** (n = 170)	**0,8143** (n = 170)	**0,2980** (n = 170)	**0,5716** (n = 157)	**0,5989** (n = 155)	-0,0048 (n = 170)	**0,8183** (n = 170)	**0,7171** (n = 156)	**0,7063** (n = 119)	**0,5533** (n = 164)	**-0,1833** (n = 144)
K^+ (mval/l)		1,0000 (n = 170)	**0,7318** (n = 170)	0,0969 (n = 170)	**0,5773** (n = 157)	**0,5436** (n = 155)	0,0738 (n = 170)	**0,7573** (n = 170)	**0,5803** (n = 156)	**0,7445** (n = 199)	**0,5625** (n = 164)	-0,1013 (n = 144)
Mg^{2+} (mval/l			1,0000 (n = 170)	**0,5383** (n = 170)	**0,6355** (n = 157)	**0,7653** (n = 155)	-0,1373 (n = 170)	**0,9606** (n = 170)	**0,9362** (n = 156)	**0,6876** (n = 119)	**0,5735** (n = 164)	-0,0677 (n = 144)
Ca^{2+} (mval/l)				1,0000 (n = 170)	0,1453 (n = 157)	**0,4806** (n = 155)	**-0,4466** (n = 170)	**0,4614** (n = 170)	**0,7535** (n = 156)	0,0882 (n = 119)	0,0704 (n = 164)	-0,0066 (n = 144)
Cl^- (mval/l)					1,0000 (n = 157)	**0,7837** (n = 155)	0,1117 (n = 157)	**0,6541** (n = 157)	**0,5332** (n = 143)	**0,7733** (n = 107)	**0,4059** (n = 151)	-0,1586 (n = 133)
SO_4^{2-} (mval/l)						1,0000 (n = 155)	-0,1143 (n = 155)	**0,7368** (n = 155)	**0,7359** (n = 141)	**0,4997** (n = 107)	**0,3990** (n = 149)	-0,0358 (n = 131)
pH-Wert							1,0000 (n = 240)	-0,0812 (n = 240)	**-0,2756** (n = 156)	0,0473 (n = 119)	0,0544 (n = 164)	-0,0973 (n = 206)
Leitfähigkeit (mS/cm)								1,0000 (n = 240)	**0,8955** (n = 156)	**0,7480** (n = 119)	**0,6112** (n =164)	-0,0956 (n = 206)
Wasserhärte (°dH)									1,0000 (n = 185)	**0,5633** (n = 119)	**0,5055** (n = 154)	-0,0962 (n = 157)
CSB (mg/l)										1,0000 (n = 119)	**0,8317** (n = 117)	-0,1051 (n = 101)
DOC (mg/l)											1,0000 (n = 164)	-0,1245 (n = 138)
Wasserspiegelstand (m)												1,0000 (n = 242)

Tab. A-5. Einfaktorielle Varianzanalyse zum Haupteffekt DATUM (Sickerwasser).

abhängige Variable	MQ_Z	MQ_I	F_{exp}	α_{calc}
Na^+ (mval/l)	21125,1058	34232,1310	0,6171	0,85499630
K^+ (mval/l)	439,6182	443,6405	0,9909	0,46804912
Na^+/K^+	34,2648	32,3412	1,0595	0,40280282
Mg^{2+} (mval/l)	1659,5638	3684,6541	0,4504	0,95904780
Ca^{2+} (mval/l)	85,8448	135,5616	0,6333	0,84129390
Mg^{2+}/Ca^{2+}	21,7301	22,0450	0,9857	0,47565143
Cl^- (mval/l)	4564,3954	10971,4225	0,4160	0,95235875
SO_4^{2-} (mval/l)	11127,3747	58916,7881	0,1889	0,99904068
Cl^-/SO_4^{2-}	0,5833	1,1814	0,4937	0,91164433
pH-Wert	0,3104	1,6321	0,1902	0,99988772
Leitfähigkeit (mS/cm)	67,9425	396,5544	0,1713	0,99994807
Temperatur (°C)	68,6342	15,6665	4,3809	**0,00000061**
Wasserhärte (°dH)	12423,2094	33637,6256	0,3693	0,99072577
CSB (mg/l)	372142,5844	287775,9751	1,2932	0,24885915
DOC (mg/l)	2289,5450	10941,5520	0,2093	0,99930055
Wasserspiegelstand (m)	1,6816	1,7977	0,9354	0,54198435

Tab. A-6. Einfaktorielle Varianzanalyse zum Haupteffekt DATUM (Grundwasser).

abhängige Variable	MQ_Z	MQ_I	F_{exp}	α_{calc}
Na^+ (mval/l)	108,3559	64,3081	1,6849	0,05486777
K^+ (mval/l)	1,3753	1,0361	1,3274	0,18715227
Na^+/K^+	923,4690	654,8482	1,4102	0,14352235
Mg^{2+} (mval/l)	52,6493	41,0780	1,2817	0,21542218
Ca^{2+} (mval/l)	31,7511	22,9678	1,3824	0,15711578
Mg^{2+}/Ca^{2+}	0,8481	0,6140	1,3813	0,15767981
Cl^- (mval/l)	1533,1308	77,6194	19,7519	**0,00000000**
SO_4^{2-} (mval/l)	488,4266	90,4027	5,4028	**0,00000002**
Cl^-/SO_4^{2-}	0,2703	0,0643	4,2026	**0,00000233**
pH-Wert	0,2326	0,0983	2,3647	**0,00127415**
Leitfähigkeit (mS/cm)	2,4037	3,4422	0,6983	0,82600157
Temperatur (°C)	53,9068	4,3335	12,4397	**0,00000000**
Wasserhärte (°dH)	880,6886	774,6899	1,1368	0,32318838
CSB (mg/l)	575,8284	351,0417	1,6403	0,11265641
DOC (mg/l)	17,2318	16,3284	1,0553	0,40248705
Wasserspiegelstand (m)	2,0015	0,0541	36,9692	**0,00000000**

Tab. A-7. Einfaktorielle Varianzanalyse zum Haupteffekt MESSSTELLE (Sickerwasser).

abhängige Variable	MQ_Z	MQ_I	F_{exp}	α_{calc}
Na^+ (mval/l)	239121,4274	13440,4797	17,7911	**0,00000000**
K^+ (mval/l)	1966,9147	300,7521	6,5400	**0,00000008**
Na^+/K^+	124,3030	23,9984	5,1796	**0,00000364**
Mg^{2+} (mval/l)	34126,6939	584,6841	58,3677	**0,00000000**
Ca^{2+} (mval/l)	1072,6292	41,8910	25,6053	**0,00000000**
Mg^{2+}/Ca^{2+}	73,6985	17,2185	4,2802	**0,00004941**
Cl^- (mval/l)	74070,4441	1001,3586	73,9700	**0,00000000**
SO_4^{2-} (mval/l)	408441,9127	3288,2257	124,2135	**0,00000000**
Cl^-/SO_4^{2-}	5,0377	0,5371	9,3789	**0,00000000**
pH-Wert	17,1703	0,1265	135,7768	**0,00000000**
Leitfähigkeit (mS/cm)	4324,0061	16,8104	257,2219	**0,00000000**
Temperatur (°C)	153,2210	11,9959	12,7728	**0,00000000**
Wasserhärte (°dH)	365361,6819	902,2269	404,9554	**0,00000000**
CSB (mg/l)	1808471,9518	127243,5039	14,2127	**0,00000000**
DOC (mg/l)	102119,0723	1385,9987	73,6791	**0,00000000**
Wasserspiegelstand (m)	13,7547	0,9239	14,8870	**0,00000000**

Tab. A-8. Einfaktorielle Varianzanalyse zum Haupteffekt MESSSTELLE (Grundwasser).

abhängige Variable	MQ_Z	MQ_I	F_{exp}	α_{calc}
Na^+ (mval/l)	566,9416	26,9397	21,0449	**0,00000000**
K^+ (mval/l)	9,3721	0,3762	24,9112	**0,00000000**
Na^+/K^+	1364,4884	623,2624	2,1893	**0,01226943**
Mg^{2+} (mval/l)	383,3274	13,7440	27,8906	**0,00000000**
Ca^{2+} (mval/l)	143,7096	13,8068	10,4086	**0,00000000**
Mg^{2+}/Ca^{2+}	4,4594	0,3176	14,0426	**0,00000000**
Cl^- (mval/l)	1313,4672	117,9456	11,1362	**0,00000000**
SO_4^{2-} (mval/l)	805,1357	66,8483	12,0442	**0,00000000**
Cl^-/SO_4^{2-}	0,7320	0,0247	29,6709	**0,00000000**
pH-Wert	1,2171	0,0459	26,5346	**0,00000000**
Leitfähigkeit (mS/cm)	44,9038	0,9653	46,5168	**0,00000000**
Temperatur (°C)	31,9949	7,1293	4,4878	**0,00000086**
Wasserhärte (°dH)	6700,0780	334,7597	20,0146	**0,00000000**
CSB (mg/l)	1966,0108	170,3606	11,5403	**0,00000000**
DOC (mg/l)	112,4523	8,0760	13,9243	**0,00000000**
Wasserspiegelstand (m)	0,5065	0,2442	2,0742	**0,02305382**

Tab. A-9. Zweifaktorielle Varianzanalyse mit einfacher Besetzung zu den Haupteffekten DATUM (A) und MESSSTELLE (B) (Sickerwasser).

abhängige Variable	Effekt	MQ_A, MQ_B	MQ_R	F_{exp}	α_{calc}
Na^+ (mval/l)	A	26714,7500	17856,8600	1,4960	0,16332030
	B	300176,1000	17856,8600	16,8101	**0,00000000**
K^+ (mval/l)	A	667,7207	417,7450	1,5984	0,12910110
	B	2724,7960	417,7450	6,5226	**0,00002447**
Na^+/K^+	A	39,0844	31,2955	1,2489	0,27987920
	B	174,5843	31,2955	5,5786	**0,00011959**
Mg^{2+} (mval/l)	A	162,4387	896,5360	0,1812	0,99703840
	B	26292,9000	896,5360	29,3272	**0,00000000**
Ca^{2+} (mval/l)	A	109,7577	42,2035	2,6007	**0,01082235**
	B	752,7438	42,2035	17,8361	**0,00000000**
Mg^{2+}/Ca^{2+}	A	21,7934	24,1297	0,9032	0,53600910
	B	66,6312	24,1297	2,7614	**0,01947191**
Cl^- (mval/l)	A	1734,6670	965,0894	1,7974	0,14358830
	B	94522,0200	965,0894	97,9412	**0,00000000**
SO_4^{2-} (mval/l)	A	3929,1850	3577,3330	1,0984	0,38185980
	B	474878,1000	3577,3330	132,7464	**0,00000000**
Cl^-/SO_4^{2-}	A	0,0283	0,1302	0,2171	0,95243440
	B	5,1621	0,1302	39,6384	**0,00000000**
pH-Wert	A	0,1125	0,1170	0,96153	0,48949550
	B	22,0531	0,1170	188,47140	**0,00000000**
Leitfähigkeit (mS/cm)	A	43,5367	15,2096	2,8624	**0,00397524**
	B	6647,9850	15,2096	437,0906	**0,00000000**
Temperatur (°C)	A	50,8318	2,9535	17,2107	**0,00000000**
	B	86,8077	2,9535	29,3915	**0,00000000**
Wasserhärte (°dH)	A	315,2827	882,7819	,3571	0,96779180
	B	332993,3000	882,7819	377,2090	**0,00000000**
CSB (mg/l)	A	440767,5000	113944,7000	3,8683	**0,00078351**
	B	2510113,0000	113944,7000	22,0292	**0,00000000**
DOC (mg/l)	A	1516,7670	1413,5500	1,0730	0,39620880
	B	138484,5000	1413,5500	97,9693	**0,00000000**
Wasserspiegelstand (m)	A	0,0806	0,0461	1,745550	0,07470490
	B	10,3046	0,0461	223,303000	**0,00000000**

Tab. A-10. Zweifaktorielle Varianzanalyse mit einfacher Besetzung zu den Haupteffekten DATUM (A) und MESSSTELLE (B) (Grundwasser).

abhängige Variable	Effekt	MQ_A, MQ_B	MQ_I	F_{exp}	α_{calc}
Na^+ (mval/l)	A	31,1609	29,2345	1,0659	0,39431990
	B	401,1032	29,2345	13,7202	**0,00000000**
K^- (mval/l)	A	1,5063	0,4528	3,3268	**0,00227364**
	B	7,3008	0,4528	16,1241	**0,00000000**
Na^+/K^+	A	1450,987000	868,6094	1,6705	0,11687970
	B	1805,360000	868,6094	2,0784	**0,03009557**
Mg^{2+} (mval/l)	A	15,0250	15,7840	,9519	0,47875330
	B	296,7478	15,7840	18,8006	**0,00000000**
Ca^{2+} (mval/l)	A	29,0946	13,2192	2,2009	**0,03474038**
	B	146,1629	13,2192	11,0569	**0,00000000**
Mg^{2+}/Ca^{2+}	A	0,3162	0,3657	0,8645	0,54952230
	B	3,0367	0,3657	8,3034	**0,00000000**
Cl^- (mval/l)	A	28,4640	19,7814	1,4389	0,21325230
	B	555,0506	19,7814	28,0592	**0,00000000**
SO_4^{2-} (mval/l)	A	181,3829	53,4149	3,3957	**0,00557558**
	B	497,2336	53,4149	9,3089	**0,00000000**
Cl^-/SO_4^{2-}	A	0,1357	0,0176	7,7258	**0,00000261**
	B	0,4176	0,0176	23,7761	**0,00000000**
pH-Wert	A	0,1081	0,0316	3,4227	**0,00036700**
	B	1,0227	0,0316	32,3923	**0,00000000**
Leitfähigkeit (mS/cm)	A	1,1580	1,0105	1,1460	0,33220440
	B	36,6733	1,0105	36,2934	**0,00000000**
Temperatur (°C)	A	46,0809	2,9332	15,7100	**0,00000000**
	B	28,8683	2,9332	9,8419	**0,00000000**
Wasserhärte (°dH)	A	460,3734	399,9840	1,1510	0,33804460
	B	5411,8780	399,9840	13,5302	**0,00000000**
CSB (mg/l)	A	420,4184	165,2517	2,5441	**0,02071769**
	B	1818,3520	165,2517	11,0035	**0,00000000**
DOC (mg/l)	A	12,4636	9,5981	1,2985	0,25478240
	B	92,4853	9,5981	9,6358	**0,00000000**
Wasserspiegelstand (m)	A	1,6905	0,0326	51,7995	**0,00000000**
	B	0,3100	0,0326	9,4996	**0,00000000**

Tab. A-11. Zweifaktorielle Varianzanalyse mit mehrfacher Besetzung zu den Haupteffekten DATUM (A) und MESSSTELLE (B) (Sickerwasser).

abhängige Variable	Effekt	MQ_A, MQ_B, MQ_{AB}	MQ_R	F_{exp}	α_{calc}
Na^+ (mval/l)	A	21,8872	282,9222	0,0774	0,78657210
	B	2332,3870	282,9222	8,2439	**0,00329931**
	A-B	133,5550	282,9222	0,4721	0,75557200
K^+ (mval/l)	A	7,6543	3,9961	1,9154	0,19647040
	B	204,3613	3,9961	51,1397	**0,00000127**
	A-B	4,6436	3,9961	1,1620	0,38355090
Na^+/K^+	A	0,4496	4,2495	0,1058	0,75167500
	B	88,1055	4,2495	20,7331	**0,00007880**
	A-B	0,2283	4,2495	0,0537	0,99372450
Mg^{2+} (mval/l)	A	28,6336	258,0021	0,1110	0,74590810
	B	19180,1500	258,0021	74,3411	**0,00000021**
	A-B	285,1589	258,0021	1,1053	0,40609390
Ca^{2+} (mval/l)	A	1740,4530	331,5956	5,2487	**0,04493652**
	B	1091,1900	331,5956	3,2907	0,05761323
	A-B	89,6699	331,5956	0,2704	0,89040850
Mg^{2+}/Ca^{2+}	A	19,1945	0,4767	40,2685	**0,00008398**
	B	30,0406	0,4767	63,0226	**0,00000047**
	A-B	5,3012	0,4767	11,1214	**0,00105842**
Cl^- (mval/l)	A	6101,2660	20,8119	293,1629	**0,00000014**
	B	5314,2210	20,8119	255,3458	**0,00000003**
	A-B	5291,1040	20,8119	254,2350	**0,00000003**
SO_4^{2-} (mval/l)	A	519,8495	4113,7380	0,1264	0,73141770
	B	22118,1300	4113,7380	5,3766	**0,02546979**
	A-B	10478,0000	4113,7380	2,5471	0,12914600
Cl^-/SO_4^{2-}	A	1,9274	2,1022	0,9168	0,36634230
	B	1,5386	2,1022	0,7319	0,56145260
	A-B	2,1210	2,1022	1,0090	0,43759760
pH-Wert	A	0,0070	0,0009	7,3709	**0,01087485**
	B	25,2758	0,0009	26782,2600	**0,00000000**
	A-B	0,3758	0,0009	398,1523	**0,00000000**
Leitfähigkeit (mS/cm)	A	50,5517	0,0712	709,4971	**0,00000000**
	B	357,2100	0,0712	5013,4740	**0,00000000**
	A-B	55,1617	0,0712	774,1989	**0,00000000**
Temperatur (°C)	A	27,1563	0,0785	345,9395	**0,00000000**
	B	12,6405	0,0785	161,0255	**0,00000017**
	A-B	0,7617	0,0785	9,7038	**0,00179505**
Wasserhärte (°dH)	A	380,1435	27,9070	13,6218	**0,00000000**
	B	503046,3000	27,9070	18025,8200	**0,00000000**
	A-B	1789,9390	27,9070	64,1395	**0,00000000**
CSB (mg/l)	A	17634,2600	1165,2180	15,1339	**0,00000001**
	B	1789756,0000	1165,2180	1535,9840	**0,00000000**
	A-B	16805,6400	1165,2180	14,4227	**0,00000000**
DOC (mg/l)	A	2837,1860	10,0886	281,2282	**0,00000000**
	B	227672,6000	10,0886	22567,4200	**0,00000000**
	A-B	2668,2570	10,0886	264,4836	**0,00000000**

Tab. A-12. Zweifaktorielle Varianzanalyse mit mehrfacher Besetzung zu den Haupteffekten DATUM (A) und MESSSTELLE (B) (Grundwasser).

abhängige Variable	Effekt	MQ_A, MQ_B, MQ_{AB}	MQ_R	F_{exp}	α_{calc}
Na^+ (mval/l)	A	16,9416	12,7638	1,3273	0,33320160
	B	238,2284	12,7638	18,6644	**0,00498144**
	A-B	12,4123	12,7638	0,9725	0,43070980
K^+ (mval/l)	A	0,0247	0,0417	0,5918	0,58268730
	B	15,5361	0,0417	372,6377	**0,00000125**
	A-B	0,2677	0,0417	6,4218	**0,03228207**
Na^+/K^+	A	93,8494	3,7725	24,8771	**0,00124629**
	B	279,3736	3,7725	74,0550	**0,00013544**
	A-B	69,1588	3,7725	18,3323	**0,00278134**
Mg^{2+} (mval/l)	A	3,1972	0,5370	5,9541	**0,03760921**
	B	242,9822	0,5370	452,5090	**0,00000070**
	A-B	5,2381	0,5370	9,7550	**0,01301122**
Ca^{2+} (mval/l)	A	13,7270	1,1384	12,0586	**0,00790694**
	B	13,2191	1,1384	11,6125	**0,01435908**
	A-B	25,7447	1,1384	22,6157	**0,00160637**
Mg^{2+}/Ca^{2+}	A	0,0056	0,0205	0,2713	0,77125550
	B	2,1039	0,0205	102,7887	**0,00005355**
	A-B	0,0349	0,0205	1,7065	0,25897570
Cl^- (mval/l)	A	26,7790	1,2447	21,5149	**0,00034750**
	B	524,9990	1,2447	421,7971	**0,00000003**
	A-B	14,3987	1,2447	11,5682	**0,00279404**
SO_4^{2-} (mval/l)	A	65,2369	6,1593	10,5916	**0,00368855**
	B	187,6087	6,1593	30,4593	**0,00056104**
	A-B	35,7138	6,1593	5,7983	**0,02094900**
Cl^-/SO_4^{2-}	A	0,0796	0,0170	4,6706	**0,03612946**
	B	0,3549	0,0170	20,8209	**0,00184267**
	A-B	0,0346	0,0170	2,0297	0,18834420
pH-Wert	A	0,8850	0,0018	492,5810	**0,00000000**
	B	0,2064	0,0018	114,9041	**0,00000000**
	A-B	0,0400	0,0018	22,2868	**0,00000000**
Leitfähigkeit (mS/cm)	A	3,8120	0,0022	1749,5040	**0,00000000**
	B	29,3951	0,0022	13490,8600	**0,00000000**
	A-B	1,7417	0,0022	799,3412	**0,00000000**
Temperatur (°C)	A	38,4821	0,0031	12369,2300	**0,00000000**
	B	19,5709	0,0031	6290,6340	**0,00000000**
	A-B	3,0778	0,0031	989,2946	**0,00000000**
Wasserhärte (°dH)	A	1063,3160	0,1859	5720,5610	**0,00000000**
	B	9006,9330	0,1859	48456,6100	**0,00000000**
	A-B	868,6570	0,1859	4673,3090	**0,00000000**
CSB (mg/l)	A	260,7062	2,4711	105,5020	**0,00000000**
	B	2095,7720	2,4711	848,1129	**0,00000000**
	A-B	144,7541	2,4711	58,5788	**0,00000000**
DOC (mg/l)	A	22,3200	0,0493	452,2914	**0,00000000**
	B	167,6014	0,0493	3396,2660	**0,00000000**
	A-B	13,3836	0,0493	271,2050	**0,00000000**

Tab. A-13. Hierarchisch agglomerative Clusteranalysen der Messstellen in Abhängigkeit von den Ionenpaaren. - Clusterbildungen bei variierenden Indexwerten.

Ionenpaare (Agglomeration)	Clusterbildungen bei variierenden Indexwerten (Aufteilung in zehn, fünf und drei Gruppen)		
	g = 10 (Index 1)	g = 5 (Index 2)	g = 3 (Index 3)
Na^+ und K^+ (WARDs Methode)	{SWP 01}; {SWP 04}; {SWP 05}; {SWP 07}; {SWP 09}; **{SWP 10}**; {SWP 11}; {GWB 05}; {GWB 03; GWB 07 - 08; GWB 10; Nahle}; {GWB 01 - 02; GWB 04; GWB 06; GWB 09; Luppe}	**{SWP 10}**; {SWP 11}; {SWP 07; SWP 09}; {SWP 01; SWP 04 - 05}; {GWB 01 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04; SWP 05; SWP 11}; {SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}
Na^+ und K^+ (Average Linkage)	{SWP 01}; {SWP 04}; {SWP 05}; {SWP 07}; {SWP 09}; **{SWP 10}**; **{SWP 11}**; {GWB 05}; {Luppe}; {GWB 01 - 04; GWB 06 - 10; Nahle}	{SWP 01}; **{SWP 10}**; **{SWP 11}**; {SWP 04 - 05}; {SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}	**{SWP 10}**; **{SWP 11}**; {SWP 01; SWP 04 - 05; SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}
Mg^{2+} und Ca^{2+} (WARDs Methode)	{SWP 01}; {SWP 04}; {SWP 05}; **{SWP 10}**; {SWP 11}; {SWP 07; SWP 09}; {GWB 04}; {GWB 06}; {GWB 01; GWB 05; GWB 07; GWB 09}; {GWB 02 - 03; GWB 08; GWB 10; Nahle; Luppe}	{SWP 05}; **{SWP 10}**; {SWP 01; SWP 04; SWP 11}; {SWP 07; SWP 09; GWB 04; GWB 06}; {GWB 01 - 03; GWB 05; GWB 07 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04; SWP 05; SWP 07; SWP 09; SWP 11; GWB 04; GWB 06}; {GWB 01 - 03; GWB 05; GWB 07 - 10; Nahle; Luppe}
Mg^{2+} und Ca^{2+} (Average Linkage)	{SWP 01}; {SWP 04}; **{SWP 05}**; **{SWP 10}**; {SWP 11}; {GWB 04}; {GWB 06}; {SWP 07; SWP 09}; {Nahle; Luppe}; {GWB 01 - 03; GWB 05; GWB 07 - 10}	**{SWP 05}**; **{SWP 10}**; {SWP 11}; {SWP 01; SWP 04}; {SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}	**{SWP 05}**; **{SWP 10}**; {SWP 01; SWP 04; SWP 07; SWP 09; SWP 11; GWB 01 - 10; Nahle; Luppe}
Cl^- und SO_4^{2-} (WARDs Methode)	{SWP 01}; {SWP 04}; {SWP 05}; **{SWP 10}**; {SWP 11}; {GWB 05}; {SWP 07; SWP 09}; {GWB 01 - 02; GWB 04}; {GWB 06 - 07; GWB 09}; {GWB 03; GWB 08; GWB 10; Nahle; Luppe}	{SWP 01}; **{SWP 10}**; {SWP 04- 05}; {SWP 07; SWP 09; SWP 11; GWB 05}; {GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04 - 05}; {SWP 07; SWP 09; SWP 11; GWB 01 - 10; Nahle; Luppe}
Cl^- und SO_4^{2-} (Average Linkage)	{SWP 01}; {SWP 04}; {SWP 05}; {SWP 07}; {SWP 09}; **{SWP 10}**; {SWP 11}; {GWB 04}; {GWB 05}; {GWB 01 - 03; GWB 06 - 10; Nahle; Luppe}	{SWP 01}; {SWP 04}; {SWP 05}; **{SWP 10}**; {SWP 07; SWP 09; SWP 11; GWB 01 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04 - 05}; {SWP 07; SWP 09; SWP 11; GWB 01 - 10; Nahle; Luppe}

Tab. A-14. Hierarchisch agglomerative Clusteranalysen der Messstellen in Abhängigkeit von den Ionenverhältnissen. - Clusterbildungen bei variierenden Indexwerten.

Ionenverhältnisse (Agglomeration)	Clusterbildungen bei variierenden Indexwerten (Aufteilung in zehn, fünf und drei Gruppen)		
	g = 10 (Index 1)	g = 5 (Index 2)	g = 3 (Index 3)
Na^+/K^+ (WARDs Methode)	{SWP 07}; {SWP 10}; {GWB 04}; {GWB 08}; {GWB 10}; {Luppe}; {GWB 01; Nahle}; {GWB 02 - 03}; {SWP 05; GWB 06 - 07; GWB 09}; {SWP 01; SWP 04; SWP 09; SWP 11; GWB 05}	{GWB 04}; {SWP 07; Luppe}; {GWB 08; GWB 10}; {SWP 01; SWP 04; SWP 09 - 11; GWB 05}; {SWP 05; GWB 01 - 03; GWB 06 - 07; GWB 09; Nahle}	{SWP 07; Luppe}; {GWB 04; GWB 08; GWB 10}; {SWP 01; SWP 04 - 05; SWP 09 - 11; GWB 01 - 03; GWB 05 - 07; GWB 09; Nahle}
Na^+/K^+ (Average Linkage)	{SWP 05}; {SWP 07}; {SWP 10}; **{GWB 04}**; {GWB 08}; {GWB 10}; {Luppe}; {GWB 02 - 03}; {SWP 01; SWP 04; SWP 09; SWP 11; GWB 05}; {GWB 01; GWB 06 - 07; GWB 09; Nahle}	{SWP 07}; **{GWB 04}**; {Luppe}; {GWB 08; GWB 10}; {SWP 01; SWP 04 - 05; SWP 09 - 11; GWB 01 - 03; GWB 05 - 07; GWB 09; Nahle}	**{GWB 04}**; {GWB 08; GWB 10}; {SWP 01; SWP 04 - 05; SWP 07; SWP 09 - 11; GWB 01 - 03; GWB 05 - 07; GWB 09; Nahle; Luppe}
Mg^{2+}/Ca^{2+} (WARDs Methode)	{SWP 01}; {SWP 04}; **{SWP 10}**; {SWP 11}; {GWB 05}; {GWB 09}; {SWP 07; SWP 09}; ; {GWB 01 - 02}; {GWB 04; GWB 06}; {SWP 05; GWB 03; GWB 07 - 08; GWB 10; Nahle; Luppe}	{SWP 01}; **{SWP 10}**; {SWP 11}; {SWP 04; GWB 05}; {SWP 05; SWP 07; SWP 09}; {GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04; SWP 11; GWB 05}; {SWP 05; SWP 07; SWP 09; GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}
Mg^{2+}/Ca^{2+} (Average Linkage)	**{SWP 01}**; {SWP 04}; {SWP 05}; **{SWP 10}**; {SWP 11}; {GWB 05}; {GWB 09}; {SWP 07; SWP 09}; {GWB 01 - 02}; {GWB 03 - 04; GWB 06; GWB 07 - 08; GWB 10; Nahle; Luppe}	**{SWP 01}**; **{SWP 10}**; {SWP 11}; {SWP 04; GWB 05}; {SWP 05; SWP 07; SWP 09; GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}	**{SWP 01}**; **{SWP 10}**; {SWP 04; SWP 05; SWP 07; SWP 09; SWP 11; GWB 01 - 10; Nahle; Luppe}
Cl^-/SO_4^{2-} (WARDs Methode)	{SWP 01}; {SWP 04}; {SWP 05}; {SWP 11}; {GWB 03}; {GWB 05}; {SWP 07; SWP 09}; {GWB 07; GWB 10}; {SWP 10; GWB 01 - 02; GWB 04}; {GWB 06; GWB 08 - 09; Nahle; Luppe}	{SWP 05}; {SWP 11}; {SWP 01; SWP 07; SWP 09; GWB 05}; {SWP 04; SWP 10; GWB 01 - 02; GWB 04}; {GWB 03; GWB 06 - 10; Nahle; Luppe}	{SWP 05; SWP 11}; {SWP 01; SWP 07; SWP 09; GWB 05}; {SWP 04; SWP 10; GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}
Cl^-/SO_4^{2-} (Average Linkage)	{SWP 01}; {SWP 04}; **{SWP 05}**; {SWP 07}; **{SWP 11}**; {GWB 03}; {GWB 05}; {GWB 07; GWB 10}; {SWP 09 - 10; GWB 01 - 02; GWB 04}; {GWB 06; GWB 08 - 09; Nahle; Luppe}	**{SWP 05}**; **{SWP 11}**; {SWP 01; GWB 05}; {SWP 07; SWP 09}; {SWP 04; SWP 10; GWB 01 - 04; GWB 06- 10; Nahle; Luppe}	**{SWP 05}**; **{SWP 11}**; {SWP 01; SWP 04; SWP 07; SWP 09 - 10; GWB 01 - 10; Nahle; Luppe}

Tab. A-15. Nichthierarchisch optimierende Clusteranalysen der Messstellen in Abhängigkeit von den Ionenpaaren und den Ionenverhältnissen. - Objektpartitionierungen bei variierenden Gruppenanzahlen g.

Ionenpaare bzw. Ionenverhältnisse	Objektpartitionierungen bei variierenden Gruppenanzahlen g (Aufteilung in zehn, fünf und drei Gruppen)		
	g = 10	g = 5	g = 3
Na^+ und K^+	{ }; {SWP 05}; {SWP 09}; **{SWP 10}**; {SWP 11}; {SWP 01; SWP 04}; {SWP 07; GWB 05}; {GWB 02; Luppe}; {GWB 01; GWB 04; GWB 06; GWB 09}; {GWB 03; GWB 07; GWB 08; GWB 10; Nahle}	{ }; **{SWP 10}**; {SWP 11}; {SWP 01; SWP 04; SWP 05; SWP09}; {SWP 07; GWB 01 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04; SWP 05; SWP 09; SWP 11}; {SWP07; GWB 01 - 10; Nahle; Luppe}
Mg^{2+} und Ca^{2+}	{SWP 01}; {SWP 04}; {SWP 05}; {SWP 10}; {SWP 11}; {GWB 04}; {GWB 06} {SWP 07; SWP 09}; {GWB 02; GWB 03; GWB 10; Nahle; Luppe}; {GWB 01; GWB 05; GWB 07 - 09}	{SWP 10}; {GWB 06}; {SWP 01; SWP 04; SWP 11}; {SWP 05; SWP 07; SWP 09}; {GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}	{SWP 01; SWP 04; SWP 10}; {SWP 05; SWP 07; SWP 09; SWP 11}; {GWB 01 - 10; Nahle; Luppe}
Cl^- und SO_4^{2-}	{ }; {SWP 07}; **{SWP 10}**; {SWP 11}; {GWB 04}; {SWP 09; GWB 05}; {SWP 01; SWP 04; SWP 05}; {GWB 03; GWB 08; GWB 09}; {GWB 10; Nahle; Luppe}; {GWB 01; GWB 02; GWB 06; GWB 07}	**{SWP 10}**; {SWP 01; SWP 05}; {GWB 01; GWB 02; GWB 04; GWB 09}; {SWP 04; SWP 07; SWP 09; SWP 11; GWB 05}; {GWB 03; GWB 06; GWB 07; GWB 08; GWB 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 04; SWP 05; SWP 11}; {SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}
Na^+/K^+	{ }; {SWP 07}; {GWB 04}; {GWB 08}; {GWB 10}; {Luppe}; { SWP 10; GWB 05}; {SWP 01; SWP 04; SWP 09; SWP 11}; {SWP 05; GWB 06; GWB 07; GWB 09}; {GWB 01 - 03; Nahle}	{SWP 07}; {GWB 04}; {GWB 08; GWB 10}; {SWP 01; SWP 04; SWP 09; SWP 11; GWB 02; GWB 03; Luppe}; {SWP 05; SWP 10; GWB 01; GWB 05 - 07; GWB 09; Nahle}	{Luppe}; {GWB 04; GWB 08; GWB 10}; {SWP 01; SWP 04; SWP 05; SWP 07 SWP 09 - 11; GWB 01 - 03; GWB 05 - 07; GWB 09; Nahle}
Mg^{2+}/Ca^{2+}	{ }; { }; {SWP 01}; **{SWP 10}**; {SWP11}; {GWB 09}; {SWP 04; GWB 05}; {SWP 05; SWP 07; SWP 09}; {GWB 01; GWB 02; GWB 04; GWB 06}; {GWB 03; GWB 07; GWB 08; GWB 10; Nahle; Luppe}	**{SWP 10}**; {SWP 11}; {SWP 01; SWP 04; GWB 05}; {SWP 05; SWP 07; SWP 09}; {GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}	**{SWP 10}**; {SWP 01; SWP 11}; {SWP 04; SWP 05; SWP 07; SWP 09; GWB 01 - 10; Nahle; Luppe}
Cl^-/SO_4^{2-}	{ }; { }; {SWP 05}; {SWP 11}; {GWB 03}; {GWB 01; GWB 02}; {GWB 07; GWB 10}; {SWP 04; SWP 10; GWB 04}; {SWP 01; SWP 07; SWP 09; GWB 05}; {GWB 06; GWB 08; GWB 09; Nahle; Luppe}	{SWP 01; GWB 05}; {SWP 05; SWP 11}; {SWP 07; SWP 09}; {SWP 04; SWP 10; GWB 01; GWB 02; GWB 04}; {GWB 03; GWB 06 - 10; Nahle; Luppe}	{SWP 05; SWP 11}; {SWP 01; SWP 07; SWP 09; GWB 05}; {SWP 04; SWP 10; GWB 01 - 04; GWB 06 - 10; Nahle; Luppe}

Tab. A-16. Fuzzy-Clusteranalyse in Abhängigkeit von den Na^+- und K^+-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 01	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
SWP 10	0,00	**1,00**	0,00	0,00	0,00
Cluster 3					
SWP 11	0,00	0,00	**1,00**	0,00	0,00
Cluster 4					
SWP 04	0,12	0,00	0,01	**0,85**	0,02
SWP 05	0,04	0,00	0,01	**0,94**	0,02
SWP 09	0,02	0,00	0,02	**0,71**	0,25
Cluster 5					
SWP 07	0,01	0,00	0,03	0,10	**0,85**
GWB 01	0,00	0,00	0,00	0,00	**1,00**
GWB 02	0,00	0,00	0,00	0,00	**1,00**
GWB 03	0,00	0,00	0,00	0,00	**1,00**
GWB 04	0,00	0,00	0,00	0,00	**1,00**
GWB 05	0,00	0,00	0,00	0,00	**0,99**
GWB 06	0,00	0,00	0,00	0,00	**1,00**
GWB 07	0,00	0,00	0,00	0,00	**1,00**
GWB 08	0,00	0,00	0,00	0,00	**1,00**
GWB 09	0,00	0,00	0,00	0,00	**1,00**
GWB 10	0,00	0,00	0,00	0,00	**1,00**
Nahle	0,00	0,00	0,00	0,00	**1,00**
Luppe	0,00	0,00	0,00	0,00	**1,00**

Tab. A-17. Fuzzy-Clusteranalyse in Abhängigkeit von den Mg^{2+}- und Ca^{2+}-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 05	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
SWP 10	0,00	**1,00**	0,00	0,00	0,00
Cluster 3					
SWP 01	0,01	0,00	**0,91**	0,05	0,03
SWP 04	0,01	0,01	**0,93**	0,04	0,01
SWP 11	0,09	0,01	**0,39**	0,38	0,12
Cluster 4					
SWP 07	0,04	0,00	0,03	**0,92**	0,01
SWP 09	0,00	0,00	0,00	**0,99**	0,00
GWB 04	0,03	0,00	0,15	**0,44**	0,37
Cluster 5					
GWB 01	0,00	0,00	0,00	0,00	**1,00**
GWB 02	0,00	0,00	0,00	0,00	**1,00**
GWB 03	0,00	0,00	0,00	0,00	**1,00**
GWB 05	0,00	0,00	0,03	0,02	**0,95**
GWB 06	0,02	0,01	0,12	0,21	**0,65**
GWB 07	0,00	0,00	0,00	0,01	**0,99**
GWB 08	0,00	0,00	0,00	0,00	**1,00**
GWB 09	0,00	0,00	0,01	0,02	**0,97**
GWB 10	0,00	0,00	0,00	0,00	**1,00**
Nahle	0,00	0,00	0,00	0,00	**0,99**
Luppe	0,00	0,00	0,01	0,01	**0,98**

Tab. A-18. Fuzzy-Clusteranalyse in Abhängigkeit von den Cl^-- und SO_4^{2-}-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 01	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
SWP 10	0,00	**1,00**	0,00	0,00	0,00
Cluster 3					
SWP 04	0,01	0,00	**0,92**	0,05	0,01
SWP 05	0,10	0,00	**0,85**	0,03	0,01
Cluster 4		0,00			
SWP 07	0,00	0,00	0,00	**0,98**	0,02
SWP 09	0,00	0,00	0,00	**0,99**	0,01
SWP 11	0,02	0,00	0,20	**0,70**	0,08
GWB 05	0,00	0,00	0,00	**0,96**	0,03
Cluster 5					
GWB 01	0,00	0,00	0,00	0,00	**1,00**
GWB 02	0,00	0,00	0,00	0,00	**1,00**
GWB 03	0,00	0,00	0,00	0,00	**1,00**
GWB 04	0,00	0,00	0,00	0,03	**0,97**
GWB 06	0,00	0,00	0,00	0,00	**1,00**
GWB 07	0,00	0,00	0,00	0,00	**1,00**
GWB 08	0,00	0,00	0,00	0,00	**1,00**
GWB 09	0,00	0,00	0,00	0,00	**1,00**
GWB 10	0,00	0,00	0,00	0,00	**1,00**
Nahle	0,00	0,00	0,00	0,00	**1,00**
Luppe	0,00	0,00	0,00	0,00	**1,00**

Tab. A-19. Fuzzy-Clusteranalyse in Abhängigkeit der Verhältnisse von Na^+- zu K^+-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 07	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
GWB 04	0,00	**1,00**		0,00	0,00
Cluster 3					
GWB 08	0,01	0,01	**0,92**	0,00	0,04
GWB 10	0,01	0,02	**0,94**	0,01	0,02
Cluster 4					
SWP 01	0,00	0,00	0,00	**0,98**	0,02
SWP 04	0,00	0,00	0,00	**0,99**	0,00
SWP 09	0,00	0,00	0,00	**1,00**	0,00
SWP 10	0,02	0,00	0,01	**0,71**	0,26
SWP 11	0,00	0,00	0,00	**0,99**	0,01
GWB 05	0,00	0,00	0,00	**0,90**	0,10
Cluster 5					
SWP 05	0,01	0,00	0,01	0,07	**0,91**
GWB 01	0,00	0,00	0,00	0,05	**0,95**
GWB 02	0,01	0,00	0,01	0,21	**0,77**
GWB 03	0,01	0,00	0,00	0,21	**0,77**
GWB 06	0,00	0,00	0,00	0,01	**0,99**
GWB 07	0,00	0,00	0,00	0,03	**0,97**
GWB 09	0,01	0,00	0,00	0,17	**0,82**
Nahle	0,00	0,00	0,00	0,03	**0,97**
Luppe	0,12	0,06	0,08	0,30	**0,43**

Tab. A-20. Fuzzy-Clusteranalyse in Abhängigkeit der Verhältnisse von Mg^{2+}- zu Ca^{2+}-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 01	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
SWP 10	0,00	**1,00**	0,00	0,00	0,00
Cluster 3					
SWP 11	0,00	0,00	**1,00**	0,00	0,00
Cluster 4					
SWP 04	0,00	0,00	0,00	**0,99**	0,01
GWB 05	0,00	0,00	0,00	**0,99**	0,01
Cluster 5					
SWP 05	0,00	0,00	0,00	0,00	**1,00**
SWP 07	0,00	0,00	0,01	0,01	**0,98**
SWP 09	0,00	0,00	0,00	0,01	**0,99**
GWB 01	0,00	0,00	0,00	0,04	**0,95**
GWB 02	0,00	0,00	0,00	0,01	**0,99**
GWB 03	0,00	0,00	0,00	0,00	**1,00**
GWB 04	0,00	0,00	0,00	0,00	**1,00**
GWB 06	0,00	0,00	0,00	0,00	**1,00**
GWB 07	0,00	0,00	0,00	0,00	**1,00**
GWB 08	0,00	0,00	0,00	0,00	**1,00**
GWB 09	0,01	0,00	0,02	0,15	**0,83**
GWB 10	0,00	0,00	0,00	0,00	**1,00**
Nahle	0,00	0,00	0,00	0,00	**1,00**
Luppe	0,00	0,00	0,00	0,00	**1,00**

Tab. A-21. Fuzzy-Clusteranalyse in Abhängigkeit der Verhältnisse von Cl^-- zu SO_4^{2-}-Ionen.

	1	2	3	4	5
Cluster 1					
SWP 05	**1,00**	0,00	0,00	0,00	0,00
Cluster 2					
SWP 11	0,00	**1,00**	0,00	0,00	0,00
Cluster 3					
SWP 01	0,01	0,00	**0,97**	0,01	0,01
SWP 07	0,00	0,00	**0,48**	0,44	0,07
SWP 09	0,00	0,00	**0,60**	0,33	0,07
GWB 05	0,00	0,00	**0,99**	0,00	0,00
Cluster 4					
SWP 04	0,00	0,00	0,00	**1,00**	0,00
SWP 10	0,00	0,00	0,00	**0,99**	0,01
GWB 01	0,00	0,00	0,00	**0,98**	0,02
GWB 02	0,00	0,00	0,00	**0,94**	0,06
GWB 04	0,00	0,00	0,00	**1,00**	0,00
Cluster 5					
GWB 03	0,00	0,00	0,00	0,16	**0,83**
GWB 06	0,00	0,00	0,00	0,01	**0,99**
GWB 07	0,00	0,00	0,00	0,03	**0,97**
GWB 08	0,00	0,00	0,00	0,00	**1,00**
GWB 09	0,00	0,00	0,00	0,10	**0,90**
GWB 10	0,00	0,00	0,00	0,03	**0,97**
Nahle	0,00	0,00	0,00	0,00	**1,00**
Luppe	0,00	0,00	0,00	0,03	**0,97**

Versicherung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützung von folgenden Personen erhalten:

- Herrn Prof. Dr. sc. techn. St. F. Bocklisch
- Herrn Prof. Dr. rer. nat. habil. W. Dilger
- Herrn Prof. Dr. rer. nat. habil. H. W. Zwanziger
- Herrn Dr. rer. nat. J. Flachowsky
- Herrn Dr. rer. nat. H. Borsdorf
- Herrn Dr. rer. nat. G. Schulte
- Herrn Dipl.-Math. L. Brüggemann

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Leipzig, 1. November 1998 ..... Matthias Rudolph .....

Ort, Datum Unterschrift

Thesen

Eine Möglichkeit, die Zusammenhänge zwischen abgelagerten Stoffen einer Mülldeponie und den zu erwartenden Emissionen aufzuklären, besteht darin, im An- und Abstrom des unter dem Deponiekörper befindlichen Grundwasserleiters anthropogene Schadstoffe nachzuweisen. Die Bestimmung des Ausbreitungsverhaltens von Schadstoffen aus einem Deponiekörper als punktueller Schadstoffquelle in den umgebenden Aquifer ist jedoch generell problematisch. Zum einen liefern die möglichen zu messenden analytischen Parameter hinsichtlich ökologischer Auswirkungen sehr vielfältige Aussagen; zum anderen steht gemeinhin eine relativ große Anzahl von interessierenden Messpunkten (hier: Sickerwasserpegel SWP und Grundwasserbeobachtungsbrunnen GWB) zur Verfügung, an denen aufgrund der zunehmenden Automatisierung der Bestimmungsmethoden in immer kürzerer Zeit immer mehr Parameter simultan bestimmt werden können. Das auszuwertende Datenmaterial stellt sich daher gemeinhin als sehr umfangreich und damit unübersichtlich dar.

Eine spezifische Problematik besteht zudem bei Mülldeponien der ehemaligen DDR, den Untersuchungsstandort, eine Altdeponie bei Leipzig, eingeschlossen. Hier ist die Ausbreitung der Schadstoffe über den „Wasserpfad" durch die in der Regel fehlenden Abdeckungen, Abdichtungen und Sickerwassererfassungen begünstigt, und aufgrund der örtlich inhomogenen Verteilungen der Müllbestandteile, des unterschiedlichen Alters der Deponiebereiche und der differenzierten Wasserwegsamkeiten im Deponiekörper ist eine Vorbestimmung des Sickerwassertransfers nicht möglich.

Ein Ansatz zur Lösung dieses Problems ist durch die Anwendung von Methoden der Mustererkennung gegeben, indem anhand der Verteilung der Objekte - unter einem solchen ist die geordnete Gesamtheit der an einer Probenahmestelle gemessenen schadstoffrelevanten Parameter zu verstehen - Rückschlüsse auf Kontaminationsquellen und Transportpfade des Sickerwassers gezogen werden. Darüber hinaus ist es durch die Anwendung dieser und anderer (entsprechend ausgewählter) datenanalytischer Auswertemethoden möglich, weitere Problemstellungen zu diskutieren, z. B., welche analytischen Parameter den Sickerwassertransport am besten beschreiben, um hieraus Aussagen über die Zulässigkeit der Substitution besonders zeit- und kostenaufwendig zu bestimmender Parameter abzuleiten und ob die für bestimmte Parameter aufwendigen Analyseverfahren durch einfachere Methoden bei vergleichbarem Informationsgehalt ersetzt werden können.

Es existiert eine Vielzahl von Methoden der Mustererkennung. Die mit ihnen erzielbaren Untersuchungsergebnisse sind in der Regel sehr umfangreich, vielfältig und (bei Anwendung auf identische Datensätze) teilweise auch widersprüchlich. Aus dieser Problemstellung resultierend ist die in der Arbeit vorgenommene vergleichende Bewertung der Methoden der überwachten bzw. der automatischen Klassifikation hinsichtlich ihres Informationsgehaltes und ihrer Adaptionsfähigkeit an die gegebene Aufgabenstellung von besonderer Bedeutung, da die diesbezüglichen Erkenntnisse verallgemeinert werden können und damit auch für andere praktische Anwendungsfälle relevant sind.

1. Bei Vorliegen umfangreicher und damit unübersichtlicher Datenmengen ist unabhängig von der konkreten Aufgaben- bzw. Problemstellung eine Erstauswertung mit den Methoden der beschreibenden mathematischen Statistik notwendig. Für den Modellstandort kann damit u. a. nachgewiesen werden, dass durch den Deponiekörper prinzipiell ein Salzfrachteneintrag in das Grundwasser erfolgt und dass eine hohe Schadstoffbelastung im Bereich von SWP 10 und GWB 5 vorliegt.

2. Die univariate Varianzanalyse zeigt, dass für die schadstoffrelevanten Parameter sowohl des Sickerwassers als auch des Grundwassers keine saisonalen Tendenzen vorliegen (d. h. kein signifikanter Einfluss des Faktors DATUM), demhingegen für diese jedoch deutliche lokale Unterschiede bestehen (d. h. signifikanter Einfluss des Faktors MESSSTELLE). Letzteres ist ein Beleg für die differenzierten Wasserwegsamkeiten im Deponiekörper und weist auf die angesprochene, insbesondere bei ostdeutschen Mülldeponien, problematische Vorbestimmung des Sickerwassertransfers hin.

3. Es besteht die Möglichkeit einer schnellen Ionenbestimmung über die Feldmessung der Summenparameter Leitfähigkeit und Wasserhärte, die Untersuchungsergebnisse der linearen Korrelation und Regression lassen die Aussage über die Zulässigkeit dieser Summenparameter zur Langzeitüberwachung als gesichert ansehen.
   Bezüglich der Annahmen über eine Substitution zum einen des CSB (= Chemischer Sauerstoffbedarf) durch den DOC (= gelöster organischer Kohlenstoff) und zum anderen der aufwendigen Titrationsbestimmung des CSB durch Küvettentests kann keine eindeutige Aussage getroffen werden. Die Resultate der linearen Korrelation und Regression bestätigen diese Annahmen, die der einfaktoriellen Varianzanalyse zum Haupteffekt BESTIMMUNGSMETHODE hingegen stellen sie in Frage.

4. Der Informationsgehalt der betrachteten analytischen Parameter (Einzelionen) zur Beschreibung der Schadstoffverteilung kann durch die Anwendung von Methoden der überwachten Klassifikation (Diskriminanzanalyse, Neuro-Fuzzy-Klassifikationssystem NEFCLASS) ermittelt werden. Die Na^+- und K^+-Ionen liefern im Vergleich zu den Mg^{2+}-, Ca^{2+}-, Cl^-- und SO_4^{2-}-Ionen die geringste Information. Für diese wiederum führen beide Verfahren bzw. deren Anwendung auf unterschiedlich objekt-strukturierte Datensätze zu mehrdeutigen Aussagen.

5. Aus der sich durch Anwendung von Methoden der automatischen Klassifikation (hierarchisch agglomerative und nichthierarchisch optimierende Clusteranalyse, Fuzzy-Clusteranalyse) und der Hauptkomponentenanalyse ergebenden Verteilung (Clusterbildung) der Objekte (Messstellen) lassen sich drei Aussagen über mögliche Kontaminationsquellen und Transportpfade der Schadstoffe (Salzfrachten) ableiten:
   - Im südöstlichen Deponiebereich beginnend erfolgt der Transport der Salzfrachten entsprechend der Grundwasserfließrichtung in Richtung Nordwesten, und es kommt hier im Bereich des von allen Grundwassermessstellen am höchstbelasteten GWB 5 zu einem bevorzugten Austrag in den umgebenden Aquifer.
   - Der hochbelastete Deponiebereich von SWP 10 ist als die wesentlichste Kontaminationsquelle anzusehen, von hier ausgehend erfolgt der Salzfrachtenaustrag in Richtung Nahle (westlich des Deponiekörpers gelegener Vorfluter) zu GWB 5.
   - Der hochbelastete SWP 10 befindet sich geographisch in unmittelbarer Nähe zum Abstrombereich der Luppe (östlich des Deponiekörpers gelegener Vorfluter), ein bevorzugter Schadstoffaustritt in das Grundwasser erfolgt in diesem Bereich.

6. Unter der Voraussetzung einer zeitkontinuierlichen Messung der Parameter an den Probenahmestellen lassen sowohl die im Rahmen der Untersuchungen zur Zeitreihenanalyse angewandten klassischen Prognoseverfahren (exponentielle Glättung, Anpassung eines multiplikativen saisonalen ARIMA-Prozesses) als auch solche, die auf dem Konzept der Künstlichen Neuronalen Netze beruhen, aussagerelevante Ergebnisse erwarten.

7. Von den beiden konventionellen Verfahren der überwachten Klassifikation liefert die Diskriminanzanalyse gegenüber der KNN-Methode eine mit Abstand höhere Klassifikationsleistung und ist zudem aufgrund des höheren Informationsgehaltes der Ergebnisse und der

besseren Adaptionsfähigkeit an eine gegebene (Klassifikations-) Aufgabenstellung der KNN-Methode vorzuziehen.

Das Neuro-Fuzzy-Klassifikationssystem zeigt im Vergleich zu den beiden konventionellen Verfahren der überwachten Klassifikation eine relativ schlechte Klassifikationsleistung, besitzt jedoch zwei entscheidende Vorteile: Man erhält hier einen Klassifikator auf der Grundlage linguistischer Regeln, der zum einen interpretierbar und zum anderen initialisierbar mit a priori-Wissen ist.

8. Bei der automatischen Klassifikation liefert das nichthierarchisch optimierende (Minimaldistanz-) Verfahren eine annähernd gleiche Klassifikationsleistung wie die beiden konservativen hierarchischen Verfahren Wards Methode und Average Linkage. Deren (allgemeiner) Vorteil besteht darin, dass die mit ihnen erzielten Ergebnisse relativ einfach interpretierbar sind (Gesamtbild in Form eines Dendrogramms).

   Wards Methode zeigt gegenüber Average Linkage eine besser interpretierbare („realistischere") Clusterbildung.

9. Der konzipierte Algorithmus zur Auswertung der Daten (in der Reihenfolge der Anwendung: beschreibende mathematische Statistik, lineare Korrelation und Regression, Varianzanalyse (univariat), Methoden der überwachten Klassifikation, Methoden der automatischen Klassifikation, Hauptkomponentenanalyse, Zeitreihenanalyse, (überwachte) Neuro-Fuzzy-Klassifikation) ist ein geeigneter Lösungsansatz und übertragbar auf andere Anwendungsgebiete mit ähnlichen Aufgaben- bzw. Problemstellungen. Zu diskutieren ist dabei zum einen eine Anwendung der Methoden in erweiterter und/oder verbesserter Form (z. B. univariate Varianzanalyse erweitern auf multivariate, bei Diskriminanzanalyse Schätzung des Diskriminationsfehlers) und zum anderen die Einbeziehung weiterer Methoden (z. B. Neuronale Netze zur Prognose).

Lebenslauf

Name:	Mathias Rudolph
Anschrift:	Johannes-R.-Becher-Str. 8 04279 Leipzig
Geburtsdatum:	23. Januar 1968
Geburtsort:	Leipzig
Staatsangehörigkeit:	deutsch
Familienstand:	ledig

1974 - 1984	Allgemeinbildende Polytechnische Oberschule in Leipzig Abschluss: 10. Klasse
1984 - 1987	Berufsausbildung mit Abitur im Werkzeugmaschinenbetrieb Mikrosa Leipzig Abschluss: Maschinenbauzeichner mit Abitur
1987	Maschinenbauzeichner im Werkzeugmaschinenbetrieb Mikrosa Leipzig
1987 - 1989	Grundwehrdienst bei der Nationalen Volksarmee in Oranienburg
1989	Technischer Zeichner an der Akademie der Wissenschaften Leipzig im Fachbereich Wissenschaftlicher Gerätebau
1989 - 1990	Grundlagenstudium der Informationstechnik, Fachrichtung Gerätetechnik, an der Technischen Universität Karl-Marx-Stadt
1990 - 1995	Studium der Elektrotechnik, Fachrichtung Mess-, Steuerungs- und Regelungstechnik, an der Technischen Hochschule Leipzig Abschluss: Diplomingenieur
1993 - 1995	Wissenschaftliche Hilfskraft an der Technischen Hochschule Leipzig im Fachbereich Elektrotechnik, Fachgruppe Messtechnik
seit 1993	Lehrtätigkeit in den Fächern Mathematik und Physik bei der Schülerhilfe Leipzig GmbH
seit 1995	Wissenschaftliche Hilfskraft am Umweltforschungszentrum Leipzig-Halle in der Sektion Analytik

Leipzig, 1. November 1998

UFZ-Umweltforschungszentrum Leipzig-Halle GmbH
Sektion Analytik
Permoserstraße 15
D-04318 Leipzig
Telefon 0341/235-2370
Telefax 0341/235-2625