Übersicht

Normalverteilung

Korrelation

Standardverfahren der multivariaten Datenanalyse

- *Vorhersage*
- *Dimensionsreduktion*
- *Klassifizierung*

Versuchsauswertung

Neuere multivariate Verfahren

Normalverteilung

Bedeutung

Test auf Normalverteilung

- *χ^2-Test*
- *Kolmogorov-Smirnov mit Lilliefors-Korrektur*
- *Shapiro-Wilks*

Daten-Transformation

- *Box-Cox-Transformation*

$$T(x) = \frac{x^{\lambda} - 1}{\lambda} \quad \text{für} \quad \lambda > 0$$

$$T(x) = \ln x \quad \text{für} \quad \lambda = 0$$

Korrelation

- Varianz, Kovarianz, Korrelation
- Regression
- Produkt-Moment-Korrelation
- Spearman Rangkorrelation
- Kendalls τ
- Bestimmtheitsmaß

Standardverfahren der multivariaten Datenanalyse

	Lineare Regression	Hauptkomponentenanalyse	Korrespondenzanalyse	Clusteranalyse	Diskriminanzanalyse
Zweck:					
Vorhersage	x				
Dimensionsreduktion		x	x	**= Ordination**	
Klassifizierung				x	x
Eigenschaften:					
nicht-linear					
verteilungsfrei			x	x	
nominal skalierte Var.			x		x

Nomenklatur - Übersicht

	Multivariate Regression	**Hauptkomponenten-analyse**	**Korrespondenz-analyse**	**Clusteranalyse**	**Diskriminanz-analyse**
Name der synthetischen Variable	Regressand	Hauptkomponente	-	Clusterzugehörigkeit	Diskriminanzfunktion
Wert der synthetischen Variable	Schätzwert für reale Variable	Faktorwert	Wert in 1./ 2. Dimension	(Clusterzugehörigkeit)	Diskriminanzwert
durch synth. Variable **erklärte Gesamt-Streuung**	erklärte Varianz	Anteil der Eigenwerte	Anteil der Eigenwerte (der Trägheit, der Streuung, der Inertia)	(Anteil der Fehlerquadratsumme)	Anteil der Eigenwerte (Diskrimanzanteil)
durch synth. Variable **erkl. Streuung der einzelnen Variablen**	-	Kommunalität	Anteil der Eigenwerte (der Trägheit, der Streuung, der Inertia)	-	-
Korrelation zwischen realer und synthetischer Variable	partielle Korrelation	Ladung	-	-	Ladung

Zentrale Begriffe

Varianz = mittlere quadratische Abweichung

= durch Zahl der Freiheitsgrade geteilte Summe der quadratischen Abweichungen

Bestimmtheitsmaß = Anteil der erklärten Varianz

= erklärte Varianz / Gesamtvarianz

F-Wert = erklärte Varianz / nicht-erklärte Varianz

Ladung = Korrelation zwischen beobachteter Variabler und synthetischer Variabler

ANOVA

Prinzip: Vergleich der

Streuung ***zwischen*** den Zellen: ***erklärte Varianz*** (Treatmentvarianz, Wechselwirkungsvarianz)

Streuung ***innerhalb*** der Zellen: ***nicht-erklärte*** Varianz (Fehlervarianz)

= ***F-Wert***

Versuchsauswertung

	ANOVA	**Kruskal-Wallis**
Voraussetzungen	- Homoskedastizität - Normalverteilung - Unabhängigkeit der Fehlerkomponenten	- symmetrische Streuung um den Median
Erweiterung des	- t-Tests	- Wilcoxon-Tests
Vergleich der	- Mittelwerte	- Mediane
basierend auf	- metrisch skal. Werten	- Ränge
Posthoc-Test	- z.B. Scheffé	

Konfidenzintervall

= geschätzter Bereich, in dem mit einer gewissen Wahrscheinlichkeit ein bestimmter Wert liegt

Beispiele:

Schätzung des	wenn	
Erwartungswerts	- Normalverteilung - Varianz bekannt	$x_{KI} = \bar{x} \pm \frac{\sigma}{\sqrt{n}} \cdot z\left(1-\frac{\alpha}{2}\right)$
Erwartungswerts	- Normalverteilung - Varianz unbekannt	$x_{KI} = \bar{x} \pm \frac{s}{\sqrt{n}} \cdot t\left(1-\frac{\alpha}{2}; n-1\right)$
Erwartungswerts	- Verteilung unbekannt - Varianz unbekannt - n > 50	$x_{KI} = \bar{x} \pm \frac{s}{\sqrt{n}} \cdot z\left(1-\frac{\alpha}{2}\right)$

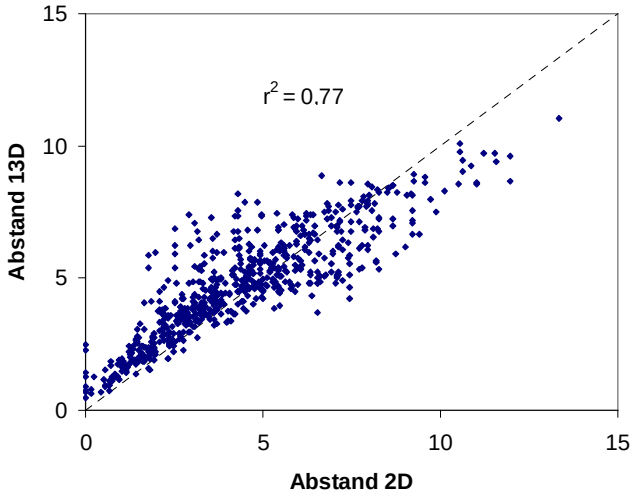
z(.): *Quantil der Normalverteilung*

s: *Standardabw. der Stichprobe*

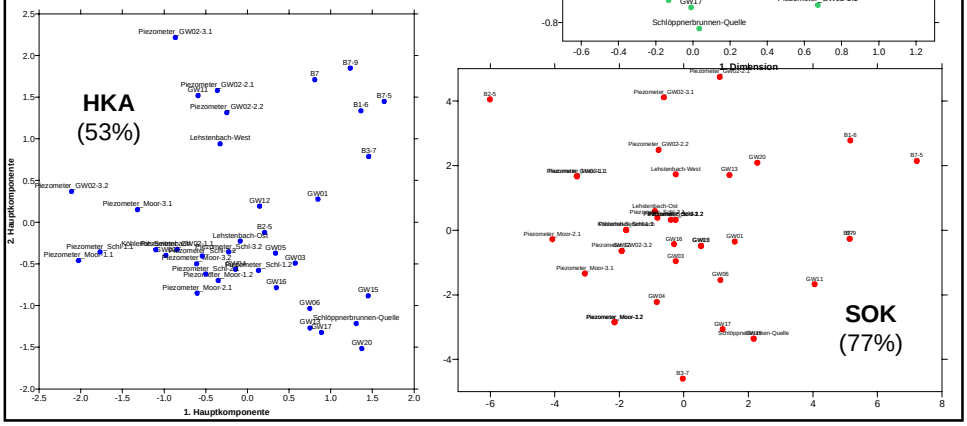
Neuere multivariate Verfahren

- Mehrschichtperzeptron
- Independent Component Analysis,
  Non-linear Principal Component Analysis
- Selbst-organisierende Karte, Lernende Vektor-Quantisierung

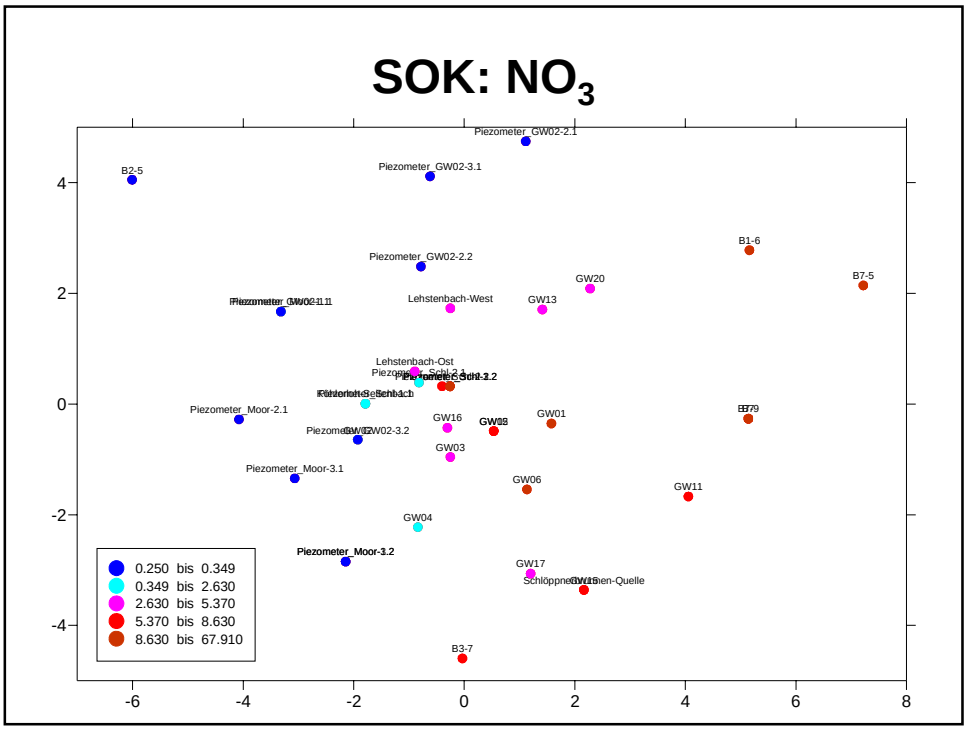
SOK: Abstände



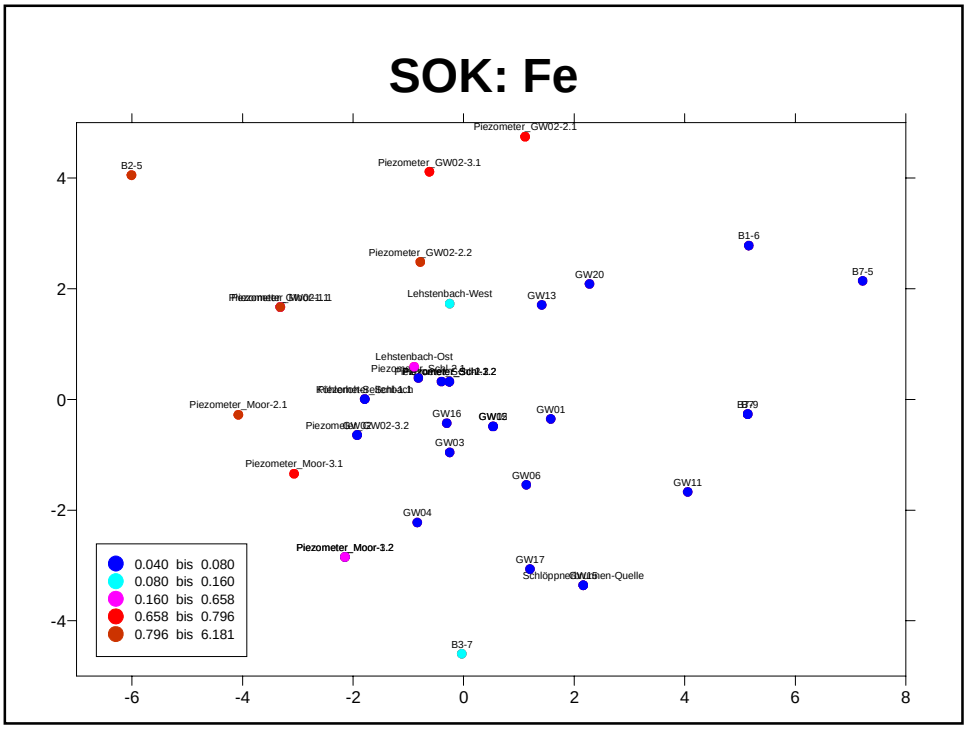
Vergleich



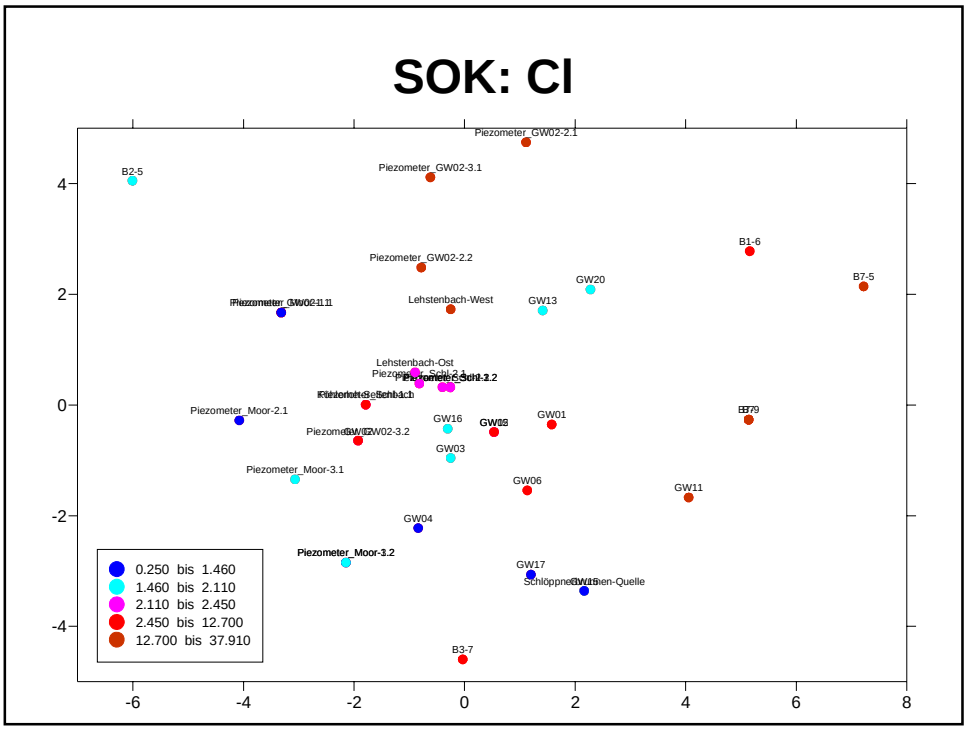
SOK: NO_3



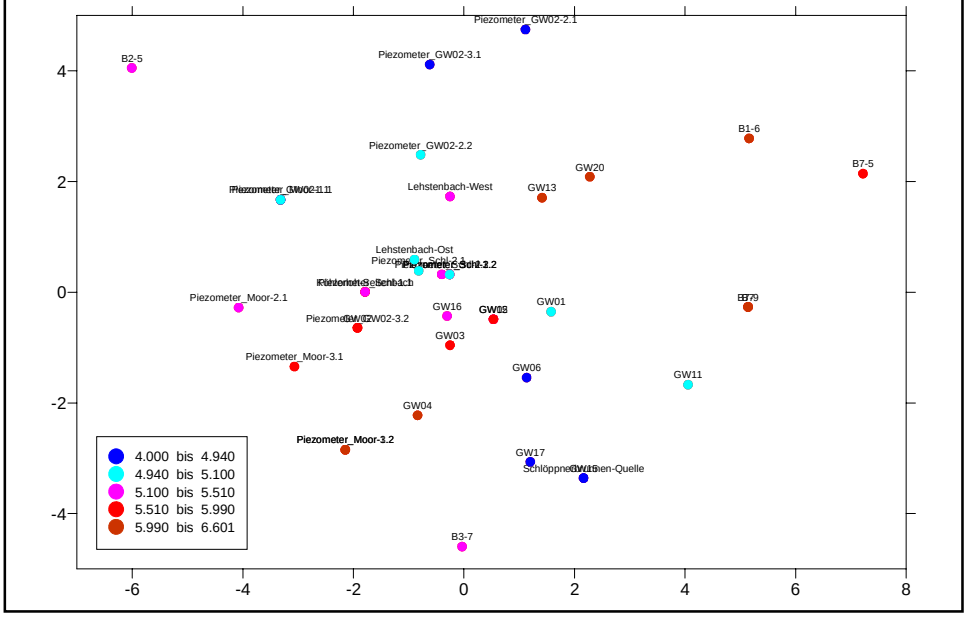
SOK: Fe

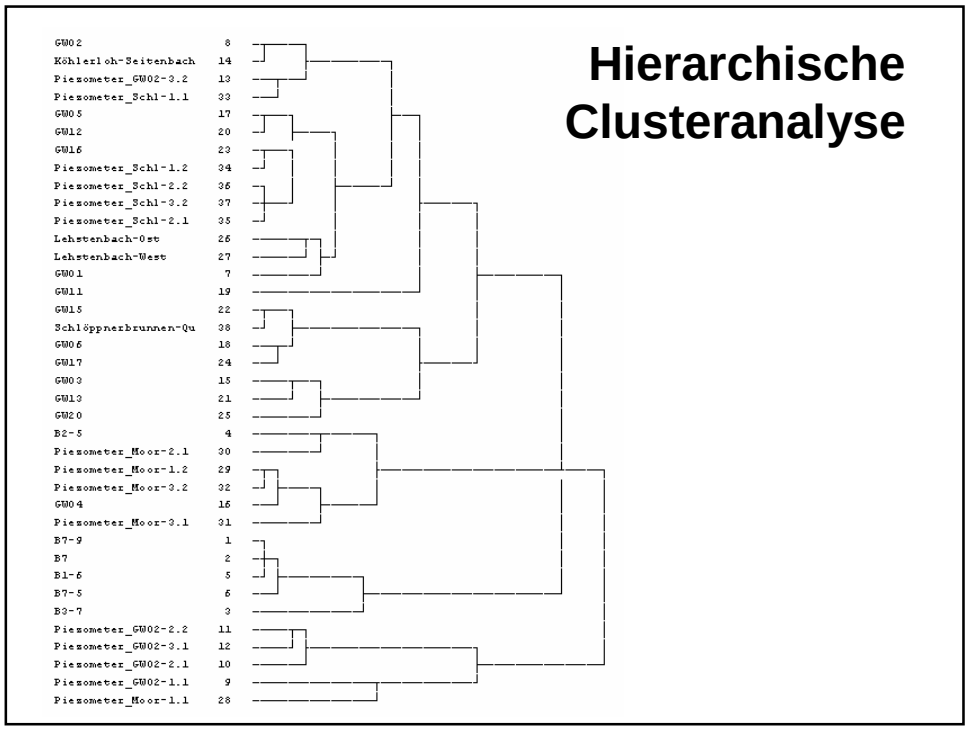


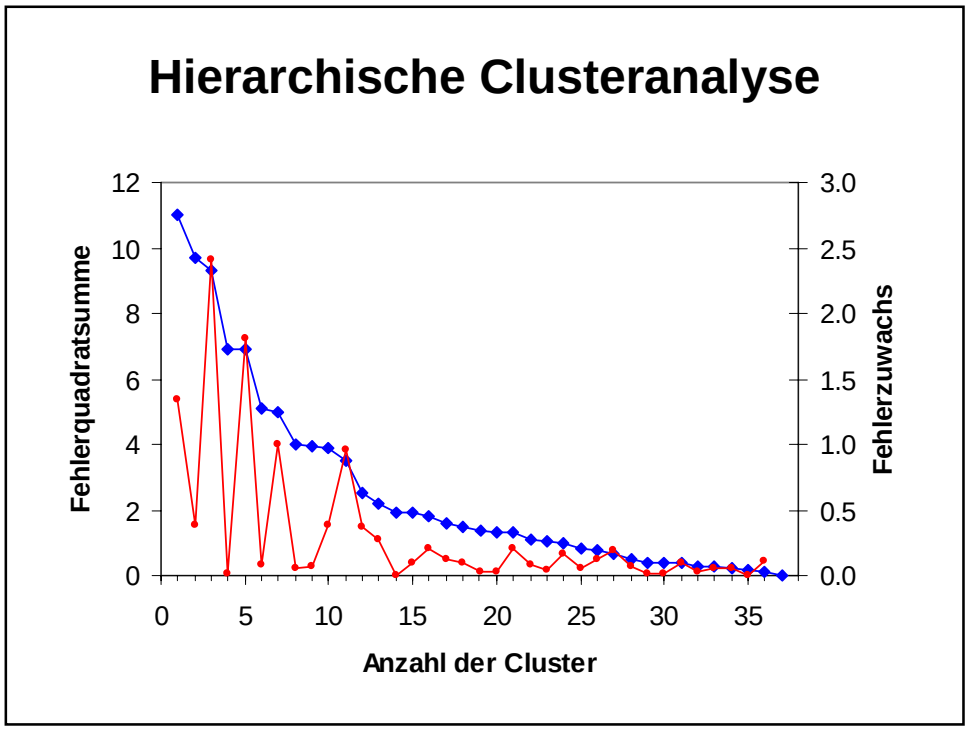
SOK: Cl


Piezometer_GW02-2.1
B2-5
Piezometer_GW02-3.1
B1-6
Piezometer_GW02-2.2
B7-5
GW20
Lehstenbach-West
GW13
Lehstenbach-Ost
Piezometer_Moor-2.1
GW16
GW01
GW03
Piezometer_Moor-3.1
GW06
GW11
GW04
Piezometer_Moor-3.2
GW17
B3-7
0.250 bis 1.460
1.460 bis 2.110
2.110 bis 2.450
2.450 bis 12.700
12.700 bis 37.910

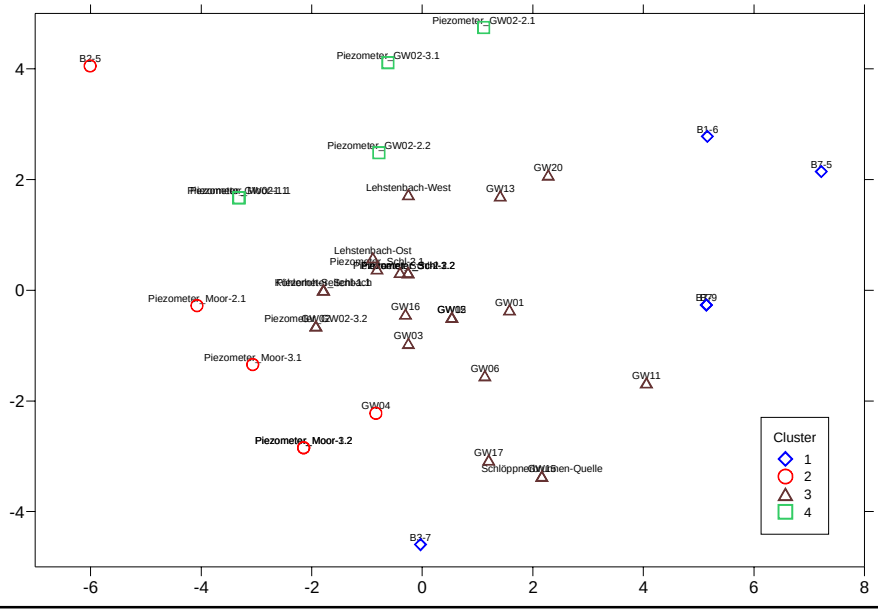
SOK: pH


Piezometer_GW02-2.1
B2-5
Piezometer_GW02-3.1
B1-6
Piezometer_GW02-2.2
B7-5
GW20
Lehstenbach-West
GW13
Lehstenbach-Ost
Piezometer_Moor-2.1
GW16
GW01
GW03
Piezometer_Moor-3.1
GW06
GW11
GW04
Piezometer_Moor-3.2
GW17
B3-7
4.000 bis 4.940
4.940 bis 5.100
5.100 bis 5.510
5.510 bis 5.990
5.990 bis 6.601

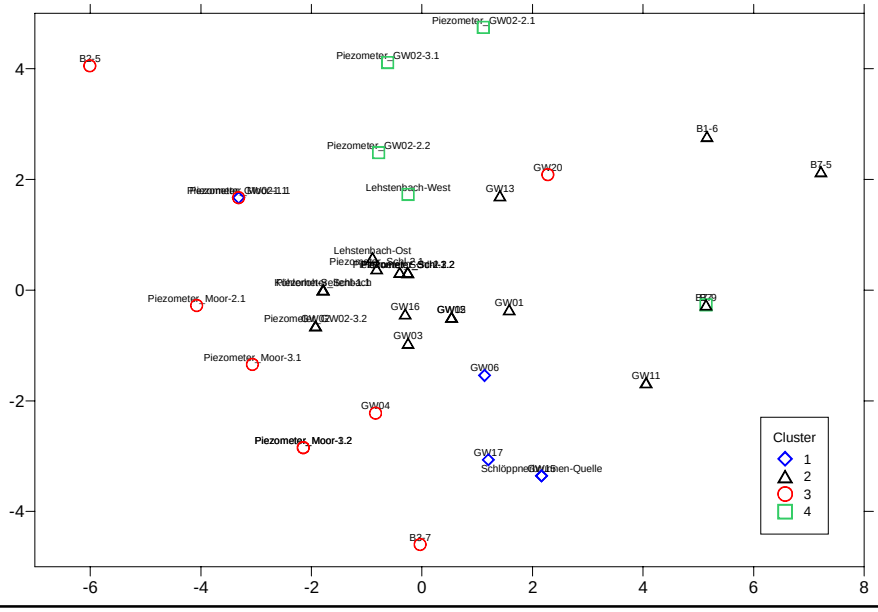

Hierarchische Clusteranalyse
GW02 8
Köhlerloh-Seitenbach 14
Piezometer_GW02-3.2 13
Piezometer_Schl-1.1 33
GW05 17
GW12 20
GW16 23
Piezometer_Schl-1.2 34
Piezometer_Schl-2.2 36
Piezometer_Schl-3.2 37
Piezometer_Schl-2.1 35
Lehstenbach-Ost 26
Lehstenbach-West 27
GW01 7
GW11 19
GW15 22
Schlöppnerbrunnen-Qu 38
GW06 18
GW17 24
GW03 15
GW13 21
GW20 25
B2-5 4
Piezometer_Moor-2.1 30
Piezometer_Moor-1.2 29
Piezometer_Moor-3.2 32
GW04 16
Piezometer_Moor-3.1 31
B7-9 1
B7 2
B1-6 5
B7-5 6
B3-7 3
Piezometer_GW02-2.2 11
Piezometer_GW02-3.1 12
Piezometer_GW02-2.1 10
Piezometer_GW02-1.1 9
Piezometer_Moor-1.1 28


Hierarchische Clusteranalyse
Fehlerquadratsumme
Fehlerzuwachs
Anzahl der Cluster
0
2
4
6
8
10
12
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
25
30
35

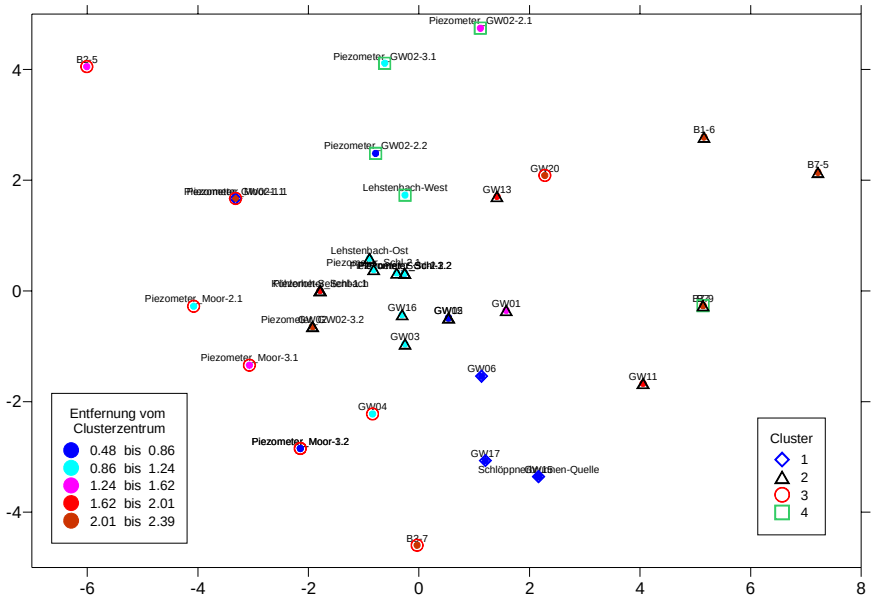
Hierarchische Clusteranalyse



Clusterzentrenanalyse



Clusterzentrenanalyse

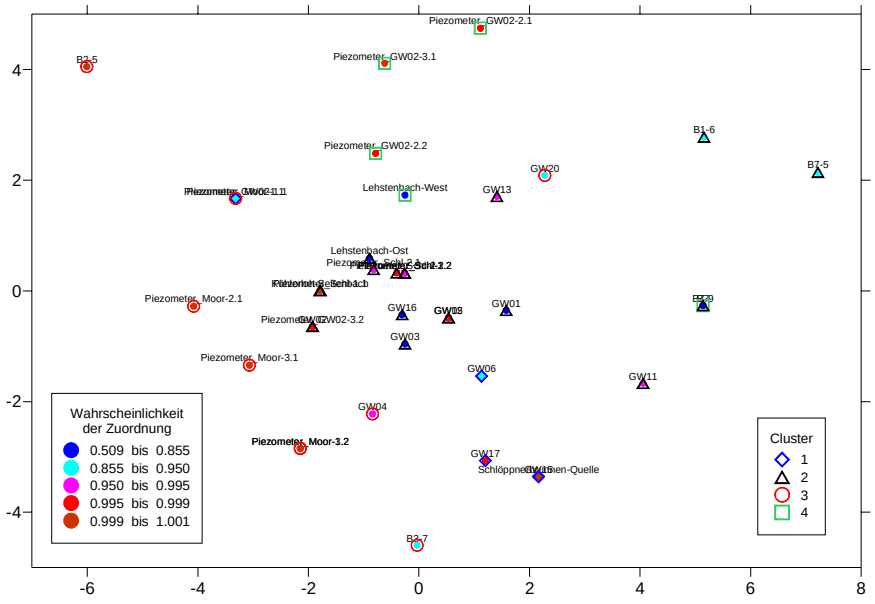


Diskriminanzanalyse

Klassifizierungsergebnisse[b,c]

		Cluster-Nr. des Falls	Vorhergesagte Gruppenzugehörigkeit				Gesamt
			1	2	3	4	
Original	Anzahl	1	5	0	0	0	5
		2	0	19	0	0	19
		3	0	0	9	0	9
		4	0	0	0	5	5
	%	1	100.0	.0	.0	.0	100.0
		2	.0	100.0	.0	.0	100.0
		3	.0	.0	100.0	.0	100.0
		4	.0	.0	.0	100.0	100.0
Kreuzvalidiert[a]	Anzahl	1	5	0	0	0	5
		2	1	17	0	1	19
		3	0	1	8	0	9
		4	0	1	0	4	5
	%	1	100.0	.0	.0	.0	100.0
		2	5.3	89.5	.0	5.3	100.0
		3	.0	11.1	88.9	.0	100.0
		4	.0	20.0	.0	80.0	100.0

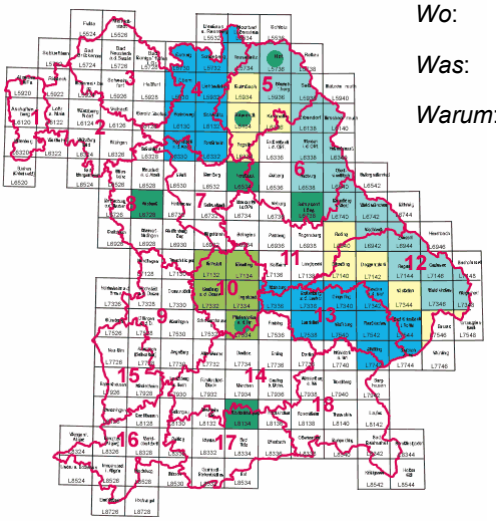
Diskriminanzanalyse



Multivariate Analyse: Take-Home Message

1. Die meisten Standardmethoden basieren auf ***linearen Zusammenhängen*** im Datensatz.
2. In der Regel gibt es verschiedene ***Maßzahlen für die Güte des Verfahrens***, die unbedingt beachtet werden sollten (ein großes r^2 alleine sagt noch nicht viel aus!).
3. Multivariate Verfahren werden überwiegend eingesetzt, um
   - einzelne Werte ***vorherzusagen*** (Regression)
   - die ***Dimension*** des Datensatzes zu ***reduzieren*** (Prozessanalyse, Visualisierung)
   - ***Gruppen*** zu identifizieren (klassifizieren).
4. Die Verfahren verlangen i.d.R. ***±subjektive Entscheidungen*** des Anwenders, die zu begründen sind.
5. Die Ergebnisse der hier vorgestellten ***Verfahren sind nicht unabhängig*** voneinander.

Diplomarbeit



Wo: Bayer. Geologisches Landesamt

Was: Multivariate Statistik, Geostatistik

Warum: EU-Wasserrahmenrichtlinie und Hydrogeologische Landesaufnahme