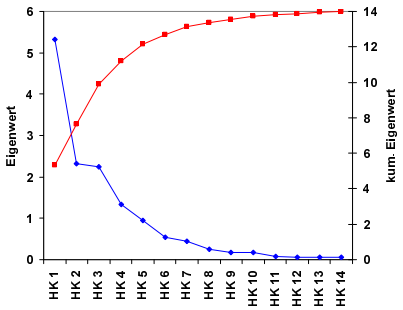
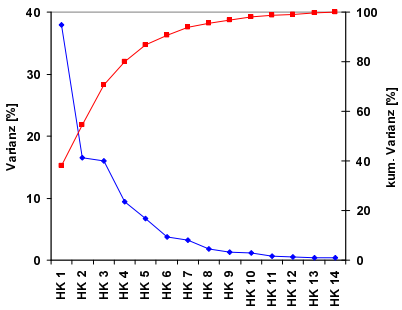
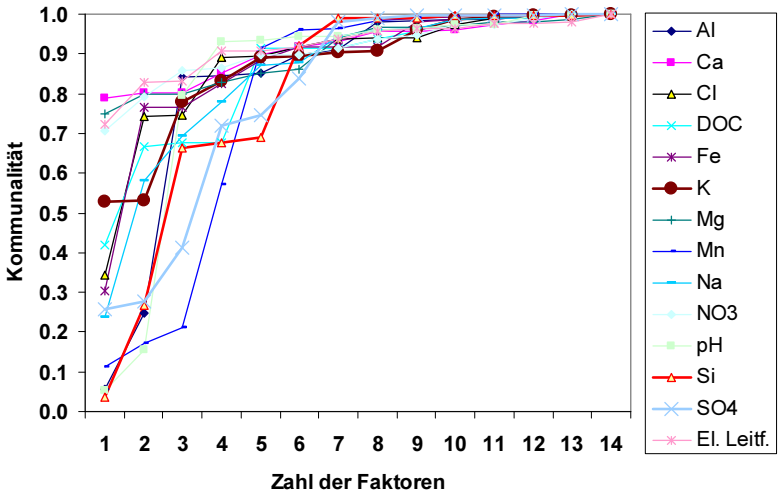
Eigenwerte

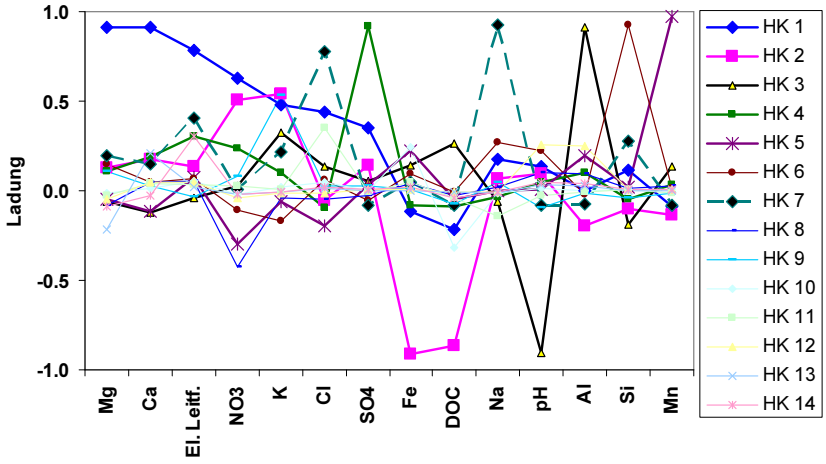




Kommunalitäten



Ladungen (Varimax-Rotation)



Messstellen

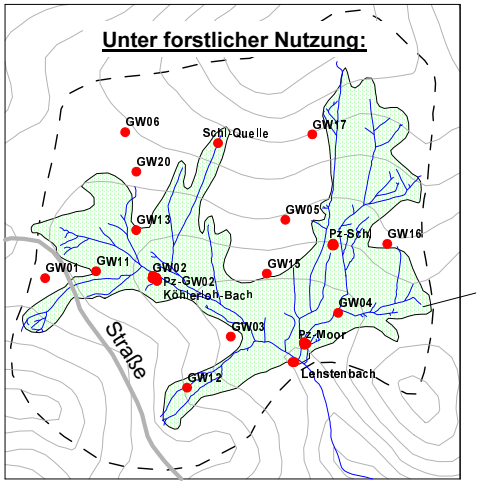
Unter landwirtschaftlicher Nutzung:

- B1-6
- B2-5 (reduziert)
- B3-7
- B7
- B7-5
- B7-9

Bezeichnungen

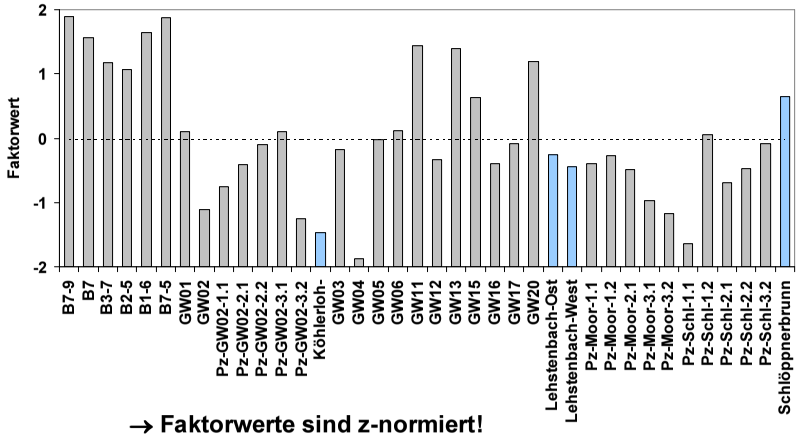
B, GW: Grundwasser-messstelle

Pz: Piezometer (max. 2 m Tiefe)



1. Hauptkomponente

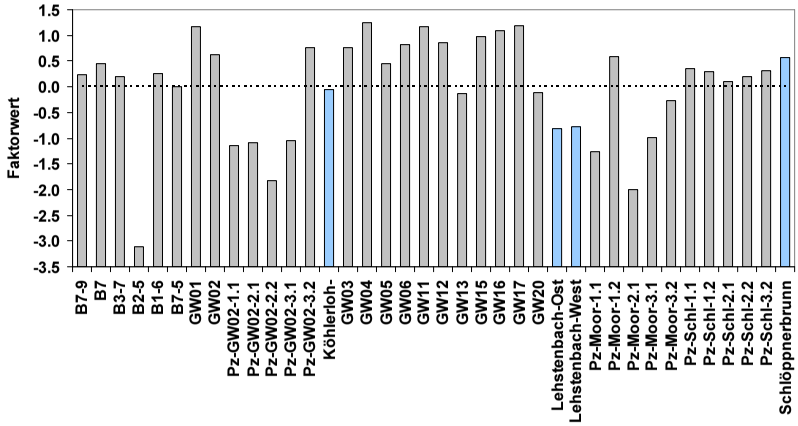
(+Ca, +Mg, +El. Lf., +NO_3)



→ **Faktorwerte sind z-normiert!**

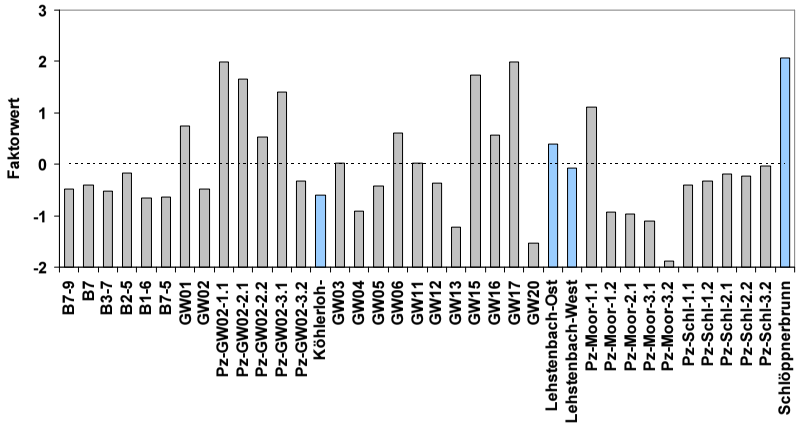
2. Hauptkomponente

(-Fe, -DOC, +K, +NO_3)



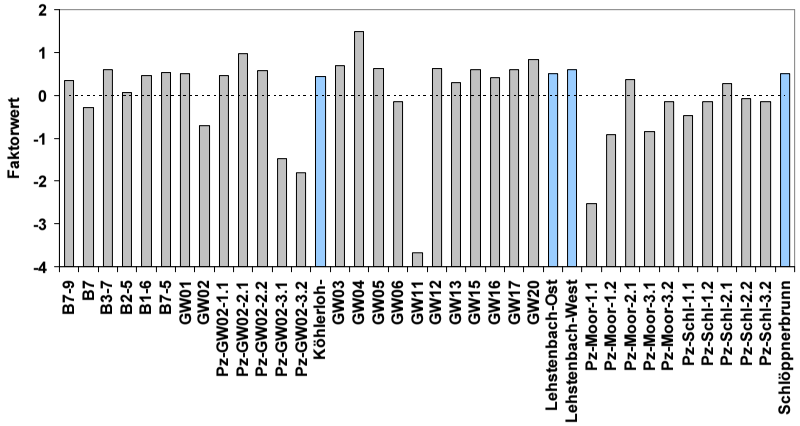
3. Hauptkomponente

(-pH, +Al)



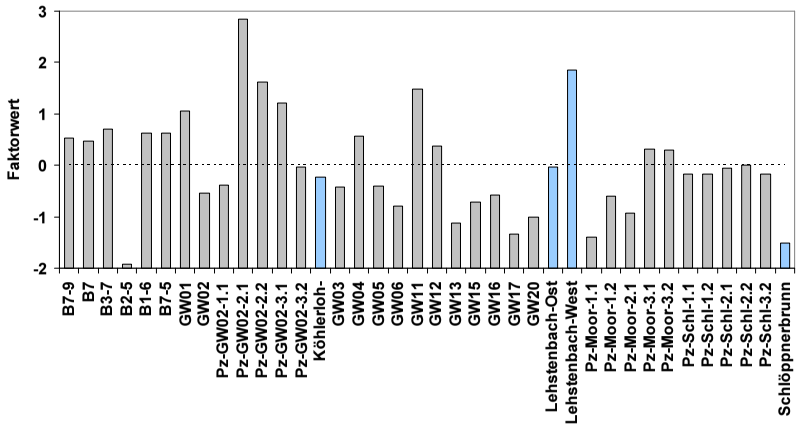
4. Hauptkomponente

(+SO_4, + El. Lf., +NO_3)



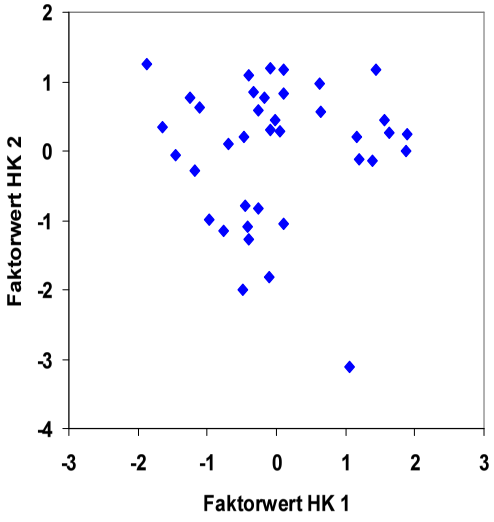
7. Hauptkomponente

(+Na, +Cl)



HK 1 vs. HK 2

(55% der Varianz)

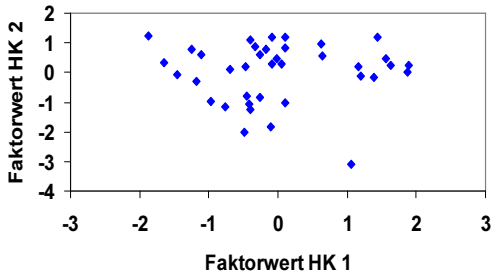


HK 1 vs. HK 2

(55% der Varianz)

Stauchung der HK 2-Achse

(16,5% / 38,0% der Varianz)



Eignung des Datensatzes für eine Hauptkomponentenanalyse

Inverse der Korrelationsmatrix A':

$$A\,A' = E$$

=> Eignung für Hauptkomponentenanalyse gegeben, wenn die Werte außerhalb der Hauptdiagonalen möglichst nahe 0

Bartlett-Test (Test of Sphericity):

- H_0-Hypothese: Variablen sind unkorreliert
- H_1-Hypothese: Variablen sind korreliert => Eignung für Hauptkomponenten-analyse
- Prüfgröße $\chi 2$-verteilt

Kaiser-Meyer-Olkin-Test

= **KMO**-Kriterium bzw. **MSA** = ***M***easure of ***S***ampling ***A***dequacy

= Summe der bivariaten quadrierten Korrelationen geteilt durch die Summen der quadrierten bivariaten *und partiellen* Korrelationen:

$$KMO\text{-}Kriterium = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j \neq k} r_{ij.k}^2} = [\ 0\ ;\ 1]$$

partielle Korrelation = Anteil der Korrelation r_{ij} zwischen den Variablen x_i und x_j, der nicht in den Korrelationen r_{ik}, r_{jk} enthalten ist („*Restkorrelation*“):

$$r_{ij.k} = \frac{r_{ij} - r_{ik} * r_{jk}}{\sqrt{1 - r_{ik}^2} * \sqrt{1 - r_{jk}^2}}$$

- angestrebt werden möglichst geringe „*Restkorrelationen*“
  => “wünschenswert“: => **KMO** > 0.8; "untragbar": **KMO** < 0.5

Kaiser-Meyer-Olkin-Test

KMO-Test für einzelne Variablen:

- ermittelt anhand der **Anti-Image-Korrelationsmatrix** (Guttmann 1953):
  Anteil der Varianz einer Variablen, der ***nicht*** durch Regression mit anderen Variablen erklärt werden kann
- wünschenswert:
  - hohe **KMO**-Werte („*geringe Restkorrelation*“) entlang der Hauptdiagonalen
  - niedrige **KMO**-Werte („*hohe Restkorrelation*“) außerhalb der Hauptdiagonalen

Hauptkomponentenanalyse: „Kochrezept“

hier: Verwendung der Hauptkomponentenanalyse zur explorativen Datenanalyse

1. Testen der Daten auf Normalverteilung, gegebenenfalls Transformation der Daten
2. z-Normierung der Daten
3. Überprüfung der Eignung des Datensatzes (Korrelationsmatrix, Kaiser-Meyer-Olkin-Kriterium, ...)
4. Bestimmung der Faktorenzahl
5. Durchführung der Hauptkomponentenanalyse
6. Varimax-Rotation
7. Ausgabe der Kommunalitäten der Variablen; gegeb. erneute Durchführung nach Ausschluss einzelner Variablen
8. Ausgabe der Eigenwerte bzw. der erklärten Varianz
9. Interpretation der Faktorladungen
10. Ausgabe der Faktorwerte

Multivariate Verfahren

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

Korrespondenzanalyse

→ 2-dimensionale grafische Darstellung (Visualisierung) eines multi-dimensionalen Datensatzes unter maximalem Erhalt der Varianz des Datensatzes

Untersuchung von ***Beobachtungen*** ("Zeilen") bzw. ***Parametern*** ("Spalten") hinsichtlich:

- Ähnlichkeit einzelner Beobachtungen/ Parameter
- Bildung von Gruppen
- Identifizierung von Ausreissern
- Zusammenhang zwischen Beobachtungen und Parametern

Korrespondenzanalyse
im Vergleich zur
Hauptkomponentenanalyse

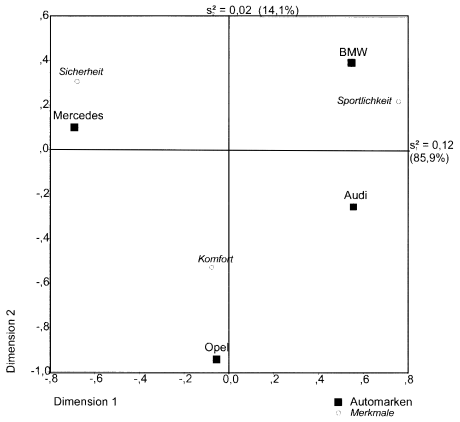
Korrespondenzanalyse

- für nominal skalierte Variablen (Häufigkeiten) entwickelt
- setzt keine Normalverteilung voraus
- dient i.d.R. der grafischen Darstellung (=> 2-dimensional)
- genutzt für die gleichzeitige Darstellung von Beobachtungen (Fällen) und Variablen

Beispiel: Umfrage

Häufigkeit der Assoziation:

	Sicherheit	Sportlichkeit	Komfort
Mercedes	9	3	
BMW	3	6	
Opel	1	1	
Audi	2	5	



(Backhaus et al. 2003)

Rechteckige Matrix: Beispiel

Mess-stelle	Ca	Mg	Na	K	F	Cl	NO_3
1	1.27	0.37	0.44	0.73	2.65	0.38	0.93
2	1.11	0.44	0.50	0.56	1.41	0.39	0.58
3	2.12	0.32	0.36	0.52	1.41	0.30	0.59
4	1.07	0.34	0.45	0.79	0.58	0.38	0.54
5	0.62	0.56	0.49	0.52	0.32	0.40	0.58
6	0.70	0.40	0.44	0.55	1.91	0.42	0.88
7	2.38	1.02	0.63	0.78	2.86	0.32	0.59
8	0.47	0.72	0.88	0.59	0.48	0.76	0.55
9	0.47	0.77	1.02	0.59	1.17	0.76	0.55
10	1.29	2.43	4.12	1.64	2.65	3.36	1.32

Rechteckige Matrix: Beispiel

Messstelle	Ca	Mg	Na	K	F	Cl	NO_3	Rand-Summe	Zeilen-Masse
1	1.27	0.37	0.44	0.73	2.65	0.38	0.93	6.77	0.10
2	1.11	0.44	0.50	0.56	1.41	0.39	0.58	4.98	0.08
3	2.12	0.32	0.36	0.52	1.41	0.30	0.59	5.62	0.09
4	1.07	0.34	0.45	0.79	0.58	0.38	0.54	4.15	0.06
5	0.62	0.56	0.49	0.52	0.32	0.40	0.58	3.47	0.05
6	0.70	0.40	0.44	0.55	1.91	0.42	0.88	5.30	0.08
7	2.38	1.02	0.63	0.78	2.86	0.32	0.59	8.60	0.13
8	0.47	0.72	0.88	0.59	0.48	0.76	0.55	4.46	0.07
9	0.47	0.77	1.02	0.59	1.17	0.76	0.55	5.32	0.08
10	1.29	2.43	4.12	1.64	2.65	3.36	1.32	16.81	0.26
Rand-Summe	11.50	7.36	9.32	7.28	15.45	7.47	7.11	65.49	
Spalten-Masse	0.18	0.11	0.14	0.11	0.24	0.11	0.11		1.00

Zeilen- und Spaltenmasse

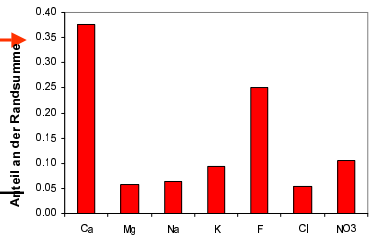
Messstelle	Ca	Mg	Na	K	F	Cl	NO_3	Rand-Summe	Zeilen-Masse
1	1.27	0.37	0.44	0.73	2.65	0.38	0.93	6.77	0.10
2	1.11	0.44	0.50	0.56	1.41	0.39	0.58	4.98	0.08
3	2.12	0.32	0.36	0.52	1.41	0.30	0.59	5.62	0.09
4	1.07	0.34	0.45	0.79	0.58	0.38	0.54	4.15	0.06
5	0.62	0.56	0.49	0.52	0.32	0.40	0.58	3.47	0.05
6	0.70	0.40	0.44	0.55	1.91	0.42	0.88	5.30	0.08
7	2.38	1.02	0.63	0.78	2.86	0.32	0.59	8.60	0.13
8	0.47	0.72	0.88	0.59	0.48	0.76	0.55	4.46	0.07
9	0.47	0.77	1.02	0.59	1.17	0.76	0.55	5.32	0.08
10	1.29	2.43	4.12	1.64	2.65	3.36	1.32	16.81	0.26
Rand-Summe	11.50	7.36	9.32	7.28	15.45	7.47	7.11	65.49	
Spalten-Masse	0.18	0.11	0.14	0.11	0.24	0.11	0.11		1.00

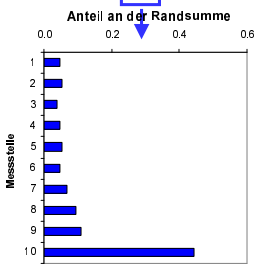
Zeilenmasse = Anteil der Zeile an der Gesamtsumme

Spaltenmasse = Anteil der Spalte an der Gesamtmasse

Zeilen- und Spalten-Profil

Mess-stelle	Ca	Mg	Na	K	F	Cl	NO_3	Rand-Summe	Zeilen-Masse
1	1.27	0.37	0.44	0.73	2.65	0.38	0.93		
2	1.11	0.44	0.50	0.56	1.41	0.39	0.58		
3	2.12	0.32	0.36	0.52	1.41	0.30	0.59		
4	1.07	0.34	0.45	0.79	0.58	0.38	0.54		
5	0.62	0.56	0.49	0.52	0.32	0.40	0.58		
6	0.70	0.40	0.44	0.55	1.91	0.42	0.88		
7	2.38	1.02	0.63	0.78	2.86	0.32	0.59		
8	0.47	0.72	0.88	0.59	0.48	0.76	0.55		
9	0.47	0.77	1.02	0.59	1.17	0.76	0.55		
10	1.29	2.43	4.12	1.64	2.65	3.36	1.32		
Rand-Summe					15	7.47	7.11		
Spalten-Masse					14	0.11	0.11	1.00	





χ^2-Statistik: Trägheit

$$\chi^2 = \sum \frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$$

für alle Zellen, wobei die

$$\text{erwartete Häufigkeit (in einer Zelle)} = \frac{\text{Summe der Zeile} \cdot \text{Summe der Spalte}}{\text{Gesamthäufigkeit}}$$

=> umso höherer χ^2-Wert, je stärker die tatsächlichen Häufigkeiten von den erwarteten abweichen

=> Division des χ^2-Wertes durch die Gesamthäufigkeit ergibt die ***Mittlere quadratische Kontingenz = totale Inertia = Gesamtträgheit***

(vergleiche „Trägheitsmoment" = Integral über dem Produkt aus Masse und der quadrierten Distanz zum Zentroid)

Zentrierung (Standardisierung)

$$\text{Zentrierter Wert} = \frac{\text{relative Häufigkeit (der Zelle)} - \text{erwartete relative Häufigkeit}}{\sqrt{\text{erwartete relative Häufigkeit}}}$$

$$\text{wobei relative Häufigkeit} = \frac{\text{Häufigkeit}}{\text{Gesamthäufigkeit}}$$

Singulärwert-Zerlegung

- **Eigenwert** (= **Trägheit = Principal Inertia)**
  = Anteil der durch eine Dimension (eine Zeile, eine Spalte) erklärten "Varianz"
  Summe der Eigenwerte = totale Inertia (Trägheit)
- **Singulärwert** = von einer Dimension repräsentierte "Streuung"
  = $\sqrt{\text{Eigenwert}}$
- **Singulärwertzerlegung**: Aufteilung der "Streuung" auf einzelne Zeilen, Spalten oder Dimensionen

→ Maximale Anzahl der Dimensionen = ***Min*** (Zeilenzahl, Spaltenzahl) - 1

Dimensionsreduktion

- Bestimmung eines 1-dimensionalen Unterraums so, dass die Summe der gewichteten Distanzen zwischen den Beobachtungen und ihrer Projektionen im Unterraum minimal wird,
  bzw. der Anteil erklärter Varianz maximal

  => Bestimmung der **Hauptachsen** (**Principal Axes**)

- Sukzessive Bestimmung weiterer, senkrecht dazu stehender Hauptachsen mit abnehmendem Beitrag zur Erklärung der Gesamtstreuung

SPSS: Auswertung

Auswertung

Dimension	Singulärwert	Auswertung für Trägheit	Anteil der Trägheit		Singulärwert für Konfidenz	
			Bedingen	Kumuliert	Standardabweichung	Korrelation 2
1	.039	.002	.183	.183	.000	.026
2	.034	.001	.137	.320	.000	
3	.030	.001	.106	.426		
4	.025	.001	.075	.501		
5	.020	.000	.051	.552		
6	.019	.000	.042	.594		
7	.018	.000	.041	.634		
8	.017	.000	.036	.671		
9	.015	.000	.029	.700		
10	.015	.000	.027	.727		
11	.014	.000	.024	.751		
12	.014	.000	.022	.773		
13	.013	.000	.020	.793		
14	.013	.000	.019	.812		
15	.012	.000	.016	.828		
16	.011	.000	.015	.844		
17	.011	.000	.014	.858		
18	.010	.000	.013	.871		
19	.010	.000	.013	.883		

Übersicht über Zeilenpunkte

Übersicht über Zeilenpunkte[a]

ZEILE	Masse	Wert in Dimension 1	Wert in Dimension 2	Übersicht über Trägheit	Beitrag des Punktes an der Trägheit der Dimension 1	Beitrag des Punktes an der Trägheit der Dimension 2	Beitrag der Dimension an der Trägheit des Punktes 1	Beitrag der Dimension an der Trägheit des Punktes 2	Gesamtübersicht
1	.002	.402	-.036	.000	.007	.000	.410	.003	.413
2	.002	-.093	-.244	.000	.000	.003	.043	.255	.297
3	.002	-.183	-.329	.000	.001	.005	.099	.275	.374
4	.002	-.069	-.249	.000	.000	.003	.018	.197	.215
5	.002	-.341	-.215	.000	.005	.002	.243	.083	.326
6	.002	-.232	-.160	.000	.002	.001	.200	.082	.283
7	.002	.005	-.257	.000	.000	.003	.000	.279	.279
8	.002	-.069	-.105	.000	.000	.001	.018	.035	.052
9	.002	.099	.038	.000	.000	.000	.076	.009	.085
10	.002	.064	-.347	.000	.000	.006	.016	.420	.436
11	.002	.042	-.259	.000	.000	.003	.008	.265	.273
12	.002	.105	-.084	.000	.000	.000	.056	.031	.087
13	.002	.153	-.165	.000	.001	.001	.111	.112	.223
14	.002	.126	-.267	.000	.001	.004	.047	.183	.229
15	.002	.016	-.170	.000	.000	.001	.001	.131	.132
16	.002	.037	-.120	.000	.000	.001	.009	.081	.090
17	.002	-.097	-.175	.000	.000	.002	.043	.123	.167
18	.002	-.127	-.125	.000	.001	.001	.083	.069	.152
19	.002	.247	-.342	.000	.003	.006	.170	.283	.453

Grafische Darstellung

- **Asymmetrische Darstellung**: unterschiedliche Skalierungen der Zeilen- (und Spalten)punkte
  - **Zeilen-Prinzipal-Normalisierung**:
    Zeilen als Punkte in einem Raum, der durch die Spalten aufgespannt wird (Zeilenprofile)

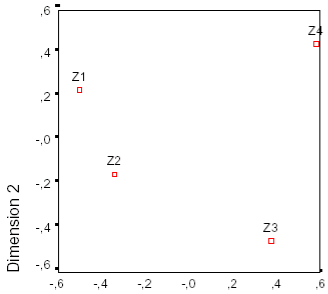
    Prinzipal-Koordinaten: Profile der Zeilen

    standardisierte Koordinaten: auf eins normierte Spalten
- **Symmetrische Darstellung**

Asymmetrische Darstellung: Zeilen-/ Spalten-Prinzipal

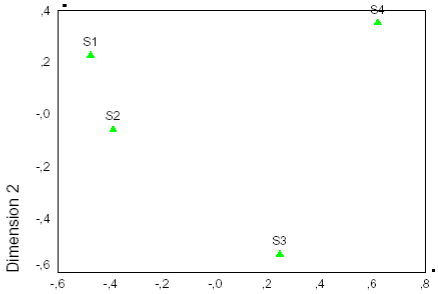
Zeilenpunkte für Zeilen der Tabelle

Prinzipal-Normalisierung



Spaltenpunkte für Spalten der Tabelle

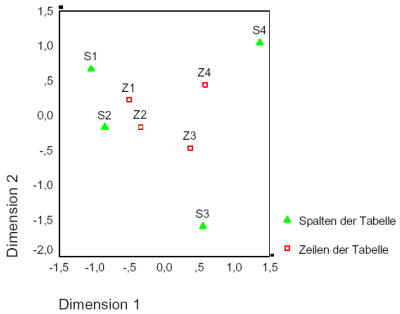
Prinzipal-Normalisierung



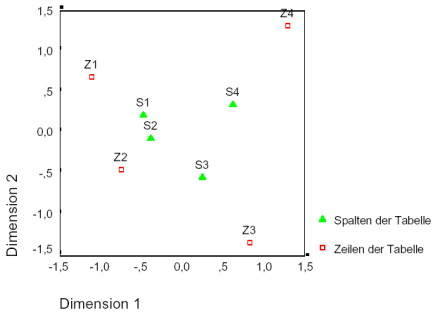
(http://www.wiwi.uni-wuppertal.de/kappelhoff/papers/korrespondenzanalyse2.pdf)

Asymmetrische Darstellung: Prinzipal-Normalisierung

Zeilen-Prinzipal-Normalisierung



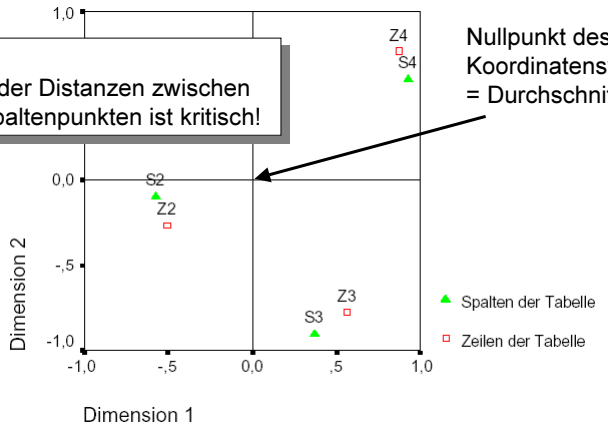
Spalten-Prinzipal-Normalisierung



(http://www.wiwi.uni-wuppertal.de/kappelhoff/papers/korrespondenzanalyse2.pdf)

Symmetrische Darstellung

Warnung:
Interpretation der Distanzen zwischen Zeilen- und Spaltenpunkten ist kritisch!

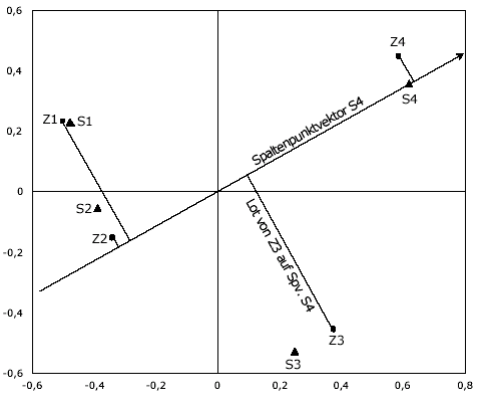


Nullpunkt des Koordinatensystems = Durchschnittsprofil

(http://www.wiwi.uni-wuppertal.de/kappelhoff/papers/korrespondenzanalyse2.pdf)

Verknüpfung von Zeilen- und Spaltenpunkten

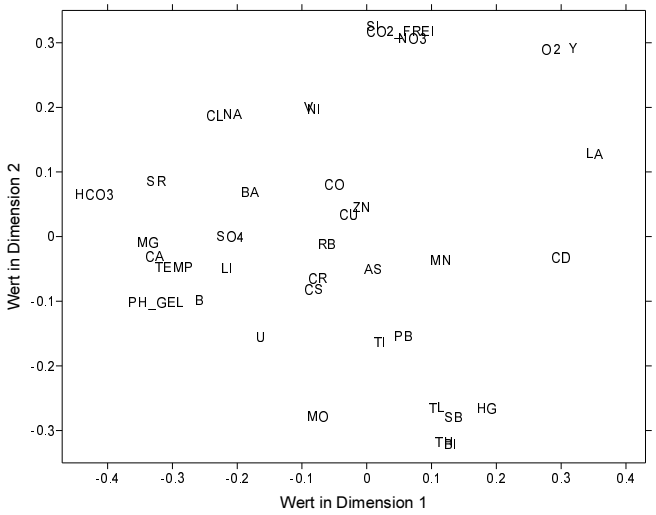
- Abstände zwischen Zeilen- und Spaltenpunkte sind nicht zu interpretieren
- stattdessen: Fällung des Lots der Zeilenpunkte auf den Vektor, der Spaltenpunkt und Ursprung verbindet
- je weiter *vom Ursprung entfernt* das Lot auf den Vektor trifft, desto enger korrespondieren Zeilen- und Spaltenpunkt



(http://www.wiwi.uni-wuppertal.de/kappelhoff/papers/tabellenanalyse9.pdf)

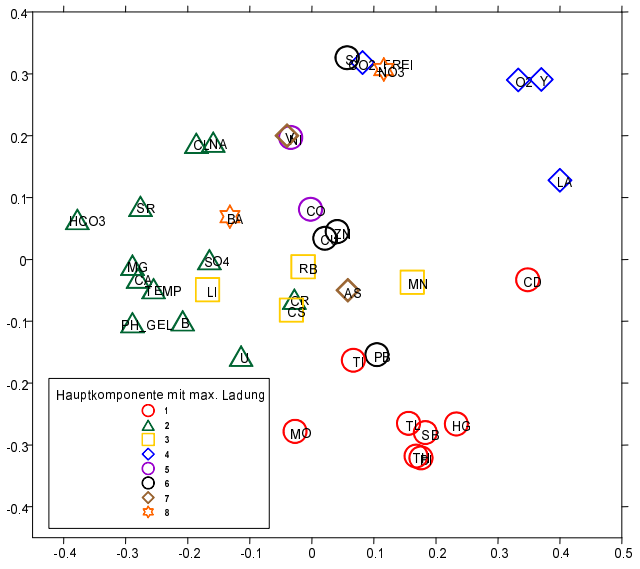
Nordbayerische Grundwässer

(Klaas 2003)



Korrespondenz- vs. HK-Analyse

(Klaas 2003)



Aufgabe

1. Transformieren Sie die Rohdaten durch Division durch die jeweilige Standardabweichung.
2. Führen Sie eine Korrespondenzanalyse durch. Wie hoch ist der Anteil der ersten zwei Dimensionen an der Gesamtträgheit? Was bedeutet das für die Interpretation der Ergebnisse?
3. Stellen Sie die Zeilenpunkte grafisch dar (Prinzipal-Normalisierung). Welche Messstellen weisen ähnliche Profile auf? Weisen diese auch ähnliche Faktorenwerte (der Hauptkomponentenanalyse) auf?
4. Stellen Sie die Spaltenpunkte grafisch dar (Prinzipal-Normalisierung). Weisen "benachbarte" Parameter auch ähnliche Ladungen der Hauptkomponenten auf?
5. Interpretieren Sie die Ergebnisse anhand der symmetrischen Darstellung.