Multivariate Analyseverfahren
Überblick

Dr. Alexander Preuß

2008

Überblick

Multivariate Analyseverfahren

- Multivariate Analyseverfahren sind statistische Verfahren, die die Beziehungsstruktur von **mehr als zwei Variablen** untersuchen.
- Als grobes Kriterium zur Unterscheidung der vielen unterschiedlichen multivariaten Verfahren dient die Einteilung in
  - **strukturenprüfende** und
  - **strukturenentdeckende (explorative)** Verfahren
- Strukturenprüfende Verfahren liefern in der Regel exakte Werte zur Lösung des Problems, während explorative Verfahren Strukturen und verdeckte Zusammenhänge aufzeigen und damit die Grundlage zur inhaltlichen **Interpretation** der Daten liefern.

Überblick

- Aufgrund der Einteilung in strukturenprüfende und strukturenentdeckende (explorative) Verfahren einerseits sowie in das zugrunde liegende Skalenniveau der Daten andererseits können die Verfahren wie folgt eingeteilt werden (keine vollständige Darstellung der Verfahren):

	nominal/ordinal	**metrisch**
strukturenprüfend		Varianzanalyse Multiple Regression Diskriminanzanalyse
strukturenentdeckend	Clusteranalyse Korrespondenzanalyse Optimal Scaling	PCA Faktorenanalyse MDS

Varianzanalyse

Varianzanalyse

Varianzanalyse (ANOVA)

- Gegeben sind metrische Daten, die sich einer Gruppierungs-variablen zuordnen lassen. Es ist zu untersuchen, ob sich die Gruppen in ihrem Mittel signi-fikant voneinander unterschei-den. Gibt es nur eine Grup-pierungsvariable, so spricht man von einer **einfaktoriellen** Anova.

- Der bekannte t-Test für Mittelwertunterschiede läßt sich für mehr als zwei Variablen **nicht** anwenden.

- Beispiel: Ermittlung der Kaufhäufigkeit für Befragte aus unterschiedlichen Ländern:

lfd. Nr.	**Herkunftsland**	Kaufhäufigkeit
1	Portugal	2
2	Portugal	4
3	Italien	7
4	Frankreich	4
5	Großbritannien	8
6	Belgien	5
7	Österreich	3
8	Großbritannien	8
9	Belgien	5
10	Frankreich	6
11	Schweiz	4
12	Italien	4
13	Italien	5
14	Italien	4
15	Großbritannien	2
16	Deutschland	6
17	Italien	6
18	Deutschland	2
19	Deutschland	6
20	Belgien	2
21	Belgien	2
22	Großbritannien	6
23	Portugal	3
24	Großbritannien	2
25	Spanien	6
...	...	...

Varianzanalyse

- Die **Mittelwerte** für die einzelnen Länder sind:

Herkunftsland	Anzahl	Mittelwert
Belgien	135	4,07
Deutschland	1223	5,16
Frankreich	157	3,89
Großbritannien	252	5,03
Italien	197	4,3
Niederlande	103	4,82
Österreich	285	4,09
Portugal	65	3,63
Schweden	82	4,32
Schweiz	257	4,8
Spanien	164	4,16

Die Idee der Varianzanalyse läßt sich wie folgt charakterisieren:

- Alle Werte haben eine bestimmte Varianz (Gesamtvarianz). Diese Varianz läßt sich zerlegen in eine Varianz zwischen den einzelnen Gruppen (hier: Ländern) und eine Varianz innerhalb der Gruppen (Fehlervarianz).

- Das Verhältnis aus Varianz zwischen den Gruppen und Varianz innerhalb der Gruppen ist F-verteilt. Wie aus der schließenden Statistik bekannt, ist nun zu untersuchen, ob die ermittelte Prüfgröße F signifikant wird. Wird F signifikant, dann sind die Unterschiede zwischen den Gruppen nicht mehr auf den Zufall zurückzuführen.

Varianzanalyse

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	715,738	10	71,574	16,267	0,000
Innerhalb der Gruppen	12799,248	2909	4,4		
Gesamt	13514,986	2919			

Man errechnet die Prüfgröße F als Quotienten aus der Summme der Abweichungsquadrate zwischen den Gruppen, geteilt durch die Zahl der Freiheitsgrade, und innerhalb der Gruppen, geteilt durch die Zahl der Freiheitsgrade:

$$F = \frac{s_B^2}{s_W^2}$$

- Im vorliegenden Beispiel ist die ermittelte Prüfgröße F = (716/10)/(12799/2909) = 16,3. Damit wird F signifikant (Irrtumswahrscheinlichkeit < 1%), die Unterschiede sind also nicht auf den Zufall zurückzuführen.
- Die Beurteilung des Produkts hängt folglich davon ab, aus welchem Land die Testperson kommt.
- Werden zwei Merkmale gemessen, liegen also zwei Gruppierungsvariablen vor, so spricht man von einer **zweifaktoriellen Anova.**
- Mit einem entsprechenden Verfahren lassen sich die Unterscheide anschließend spezifizieren.

Multiple Regression

Multiple Regression

- Die multiple Regression dient der Untersuchung von Beziehungen zwischen einer **abhängigen** und mindestens einer **unabhängigen** Variablen. Im Rahmen der deskriptiven Statistik wurde bereits die lineare Regression vorgestellt, in der es um die Ermittlung der Beziehung zwischen einer abhängigen und einer unabhängigen Variablen geht.

- Mit Hilfe der multiplen Regression lassen sich drei Gruppen von Analysen durchführen
  - Ursachenanalysen: Wie stark ist der Einfluß der unabhängigen auf die abhängige Variable?
  - Wirkungsprognosen: Wie verändert sich die abhängige Variable, wenn eine der unabhängigen verändert wird?
  - Zeitreihenanalyse: Wie verändert sich die abhängige Variable im Zeitverlauf (→ Zeitreihenanalyse)?

Multiple Regression

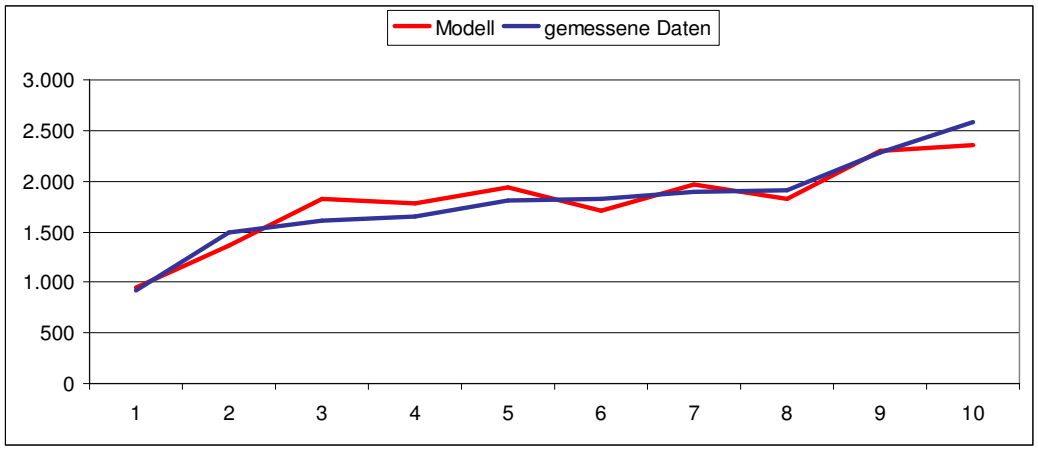
- Beispiel: Verkaufszahlen als abhängige Variable, VK-Preis, Werbeausgaben, Vertreterbesuche als unabhängige Variablen:

Verkauf	Preis	Werbeausgaben	Vertreterbesuche
921	12,00 €	0	81
1.496	11,50 €	800	70
1.612	9,50 €	1.100	87
1.647	9,30 €	800	100
1.810	8,00 €	800	110
1.819	8,50 €	550	107
1.897	8,50 €	1.200	92
1.913	12,50 €	1.300	79
2.278	10,00 €	1.500	102
2.585	7,80 €	1.200	120

Multiple Regression

- Ziel: Angabe einer linearen Funktion, die **alle** Variablen berücksichtigt:
  - $Y = \beta_1 X_1 + ... + \beta_J X_J$
- also hier:
  - Verkäufe = β_1·Stückpreis + β_2·Werbeausgaben + β_3·Vertreterbesuche
- Ergebnis:
  - Verkauf (St.) = -10,54·VK-Preis + 0,71·Werbeausgaben + 13,18·Vertreterbesuche

Multiple Regression



$R^2 = 0{,}90$

Preis	Werbung	Vertreterb.	VK-Prognose	VK real	Differenz	Differenz (%)
12,00	0	81	**941**	921	20	2,2%
11,50	800	70	**1.367**	1.496	-129	-8,6%
9,50	1.100	87	**1.824**	1.612	212	13,1%
9,30	800	100	**1.785**	1.647	138	8,4%
8,00	800	110	**1.931**	1.810	121	6,7%
8,50	550	107	**1.709**	1.819	-110	-6,0%
8,50	1.200	92	**1.971**	1.897	74	3,9%
12,50	1.300	79	**1.828**	1.913	-85	-4,4%
10,00	1.500	102	**2.299**	2.278	21	0,9%
7,80	1.200	120	**2.347**	2.585	-238	-9,2%
10,00	1.000	100	**1.919**			
9,50	2.000	128	**3.000**			

Diskriminanzanalyse

Diskriminanzanalyse

- Ziel ist die Erklärung einer abhängigen Variablen durch die Werte einer oder mehrerer unabhängiger Variablen wie auch in der multiplen Regression.
- Die Diskriminanzanalyse nimmt jedoch eine Zuordnung von Fällen zu einer von zwei oder mehreren alternativen Gruppen vor. Die Werte der abhängigen Variablen geben also lediglich eine Gruppenzugehörigkeit an und besitzen damit Nominalskalenniveau.
- Typische Anwendung: Kreditwürdigkeitsprüfung
  - Angabe von Personenmerkmalen (Alter, Einkommen, ...) und Verhaltensmerkmalen für bekannte kreditwürdige bzw. nicht kreditwürdige Personen.
  - Ermittlung einer Diskriminanzfunktion, die die Zuordnung zu einer entsprechenden Gruppe („kreditwürdig“ bzw. „nicht kreditwürdig“) auf Basis der Variablenwerte erlaubt.

Clusteranalyse

Clusteranalyse

- Ziel der Clusteranalyse ist die Gruppierung einer Anzahl von Objekten nach ihrer Ähnlichkeit bzw. Unterschiedlichkeit.
- Die Objekte sollen dabei den unterschiedlichen Gruppen (Clustern) so zugeordnet werden, daß die Ähnlichkeit der Objekte innerhalb der Cluster so groß wie möglich wird und die Ähnlichkeit zwischen den Clustern so klein wie möglich wird.
- Die Clusteranalyse kann sowohl für Objekte (z.B. Personen) als auch Merkmale durchgeführt werden.
- Bei der Bezeichnung „Clusteranalyse“ handelt es sich um einen Sammelbegriff, hinter dem sich eine Vielzahl von Verfahren verbergen.

Clusteranalyse - Beispiel

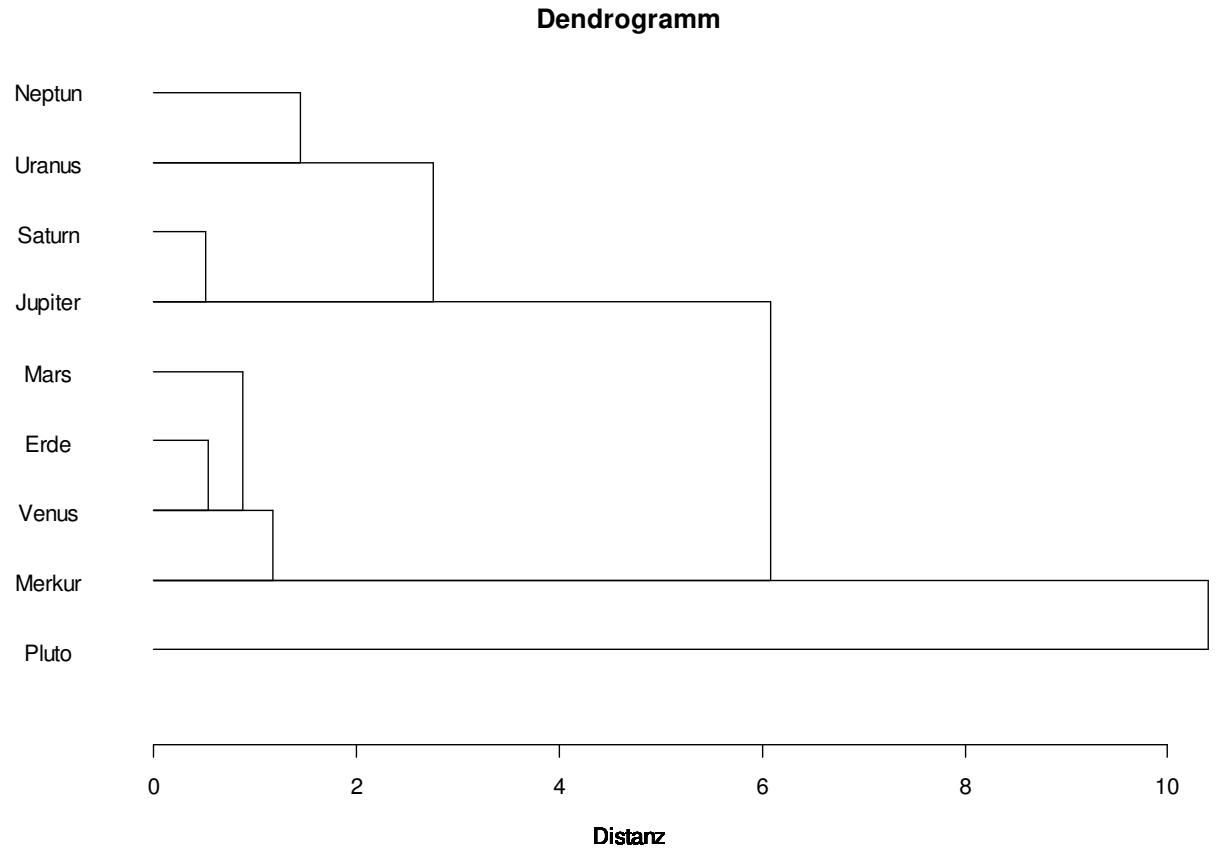
- Daten über Planeten unseres Sonnensystems (inkl. Pluto)

Planet	Abstand zur Sonne in AE	Neigung der Bahnebene in Grad	Umlaufzeit um die Sonne in Jahren	mittlere Bahnge-schwindigkeit in km/s	Äquator-durch-messer in km	Mittlere Dichte in g/cm³
Markur	0,387	7,00	0,24	47,87	4.879	5,427
Venus	0,723	3,40	0,62	35,02	12.104	5,243
Erde	1,000	0,00	1,00	29,78	12.756	5,515
Mars	1,524	1,85	1,88	24,14	6.794	3,933
Jupiter	5,203	1,31	11,86	13,07	142.984	1,326
Saturn	9,582	2,48	29,46	9,67	120.536	0,687
Uranus	19,201	0,77	84,01	6,84	51.118	1,270
Neptun	30,047	1,77	164,79	5,48	49.528	1,638
Pluto	39,482	17,16	248,20	4,75	2.390	1,750

Clusteranalyse – Durchführung
Beispiel: Hierarchische Clusteranalyse

- Ermittlung der Distanz der Objekte zueinander (üblicherweise wird die euklidische Distanz ermittelt, eine Standardisierung der Meßwerte kann/sollte zuvor vorgenommen werden, sofern es sich um metrisch skalierte Daten handelt).
- Ermittlung des Paares von Objekten (bzw. Merkmalen) mit der kleinsten Distanz; Zusammenfassung dieses Paares.
- Neuberechnung der Distanzen: dem zusammengefassten Paar wird jeweils die kleinere der beiden Distanzen zu jedem anderen Objekt (bzw. Merkmal) zugewiesen.
- Wiederholung der Schritte zwei und drei, bis alle Objekte (bzw. Merkmale) zusammengefasst sind.
- Das Ergebnis lässt sich am besten in Form eines sog. Dendrogramms darstellen: hier wird angegeben, welche Objekte (bzw. Merkmale) bei welcher Distanz zusammengefasst werden. Je ähnlicher die Objekte (bzw. Merkmale), desto eher erfolgt die Zusammenfassung.

Clusteranalse - Ergebnis



- Man kann erkennen, daß die inneren Planeten (Merkur, Venus, Erde, Mars) sowie die Gasplaneten (Jupiter, Saturn, Uranus und Neptun) jeweils ein Cluster bilden. Pluto passt nicht zu diesen beiden Clustern und wird entsprechend spät zugeordnet.

Hauptkomponentenanalyse (PCA)

Hauptkomponentenanalyse (PCA)

- Ziel der PCA ist die Visualisierung von Objekten bzw. Merkmalen einer Zahlentafel.
- Ähnliche Objekte (Merkmale) sollen möglichst nah bei einander liegen, unähnliche Objekte (Merkmale) möglichst weit entfernt.
- Weiterhin soll die Lage der Objekte (Merkmale) Aufschluss über die wichtigsten trennenden Variablen liefern – dies führt auf die Ermittlung „verdeckter“ Variablen, die hinter den Objekten (Merkmalen) stehen.
- Bei mehr als drei Merkmalen (Objekten) ist eine entsprechende Visualisierung im allgemeinen nicht ohne weiteres möglich, da ein mehr als dreidimensionaler Raum erforderlich wäre.
- Die PCA bildet die „Information“ aus einem höher dimensionalen Raum in einem Raum geringer Dimensionalität – üblicherweise in einer Ebene – ab und erlaubt so die Interpretation der Objekte nach Ähnlichkeit und die Ermittlung verdeckter Variablen.

Hauptkomponentenanalyse (PCA)

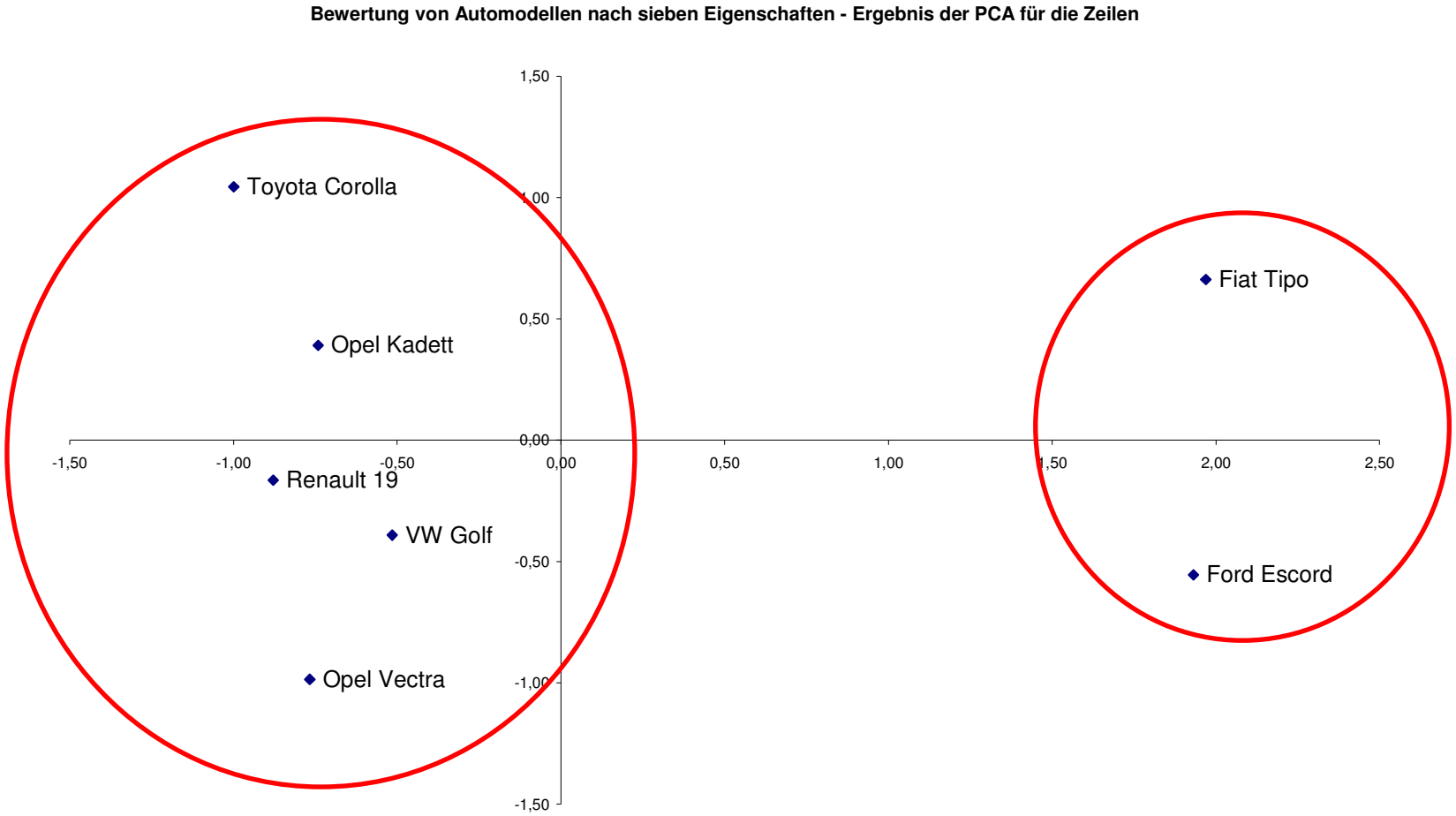
- Die zugrunde liegenden Daten haben metrisches Niveau.
- Beispiel: Verschiedenen Automodellen werden bezüglich verschiedener Eigenschaften Noten gegeben (in diesem Fall werden Noten als metrische Werte betrachtet):

	Fahrgast-zelle	Lenkrad	Armat.-brett	Fußraum	Seiten-verkl.	Vorder-sitze	Rücksitze
Renault 19	2,1	2,6	2	2,5	3,8	2,3	3,3
Toyota Corolla	1,6	3	2,3	3,5	3	2,8	2,2
VW Golf	2,6	2,3	2,8	2,8	3	2,2	2,7
Opel Vectra	2,9	2	2,1	2,5	3,5	2,2	2,3
Fiat Tipo	3,8	4,1	3,5	3,5	3,8	3,3	3,3
Ford Escord	4,6	3,6	3,1	3	3,5	2,7	3,2
Opel Kadett	2	2,8	2,1	2,8	3,3	2,8	3,3

Hauptkomponentenanalyse (PCA)

- Ansatz: Die Objekte (bzw. Merkmale) werden durch Vektoren im J-dimensionalen Raum repräsentiert, hier: J = 7, denn jedes Auto wurde nach sieben Kriterien bewertet.
- Ziel: Reproduktion der Information in einem Raum geringerer Dimensionalität (vorzugsweise im zweidimensionalen Raum)
- Vorgehen: Projektion der Vektorenspitzen in diesen Raum (auf eine Ebene)
- Interpretation: Beurteilung der Ähnlichkeit/Unähnlichkeit von Objekten im reproduzierten Raum. Interpretation der Achsen als wichtigste – verdeckte – Variablen.

Ergebnis der Hauptkomponentenanalyse



Im Ergebnis lassen sich zwei Gruppen (Cluster) erkennen. Das Hauptmerkmal, das die beiden Gruppen trennt, wird durch die waagrechte Achse repräsentiert, das zweitwichtigste Merkmal durch die senkrechte Achse. Inhaltlich könnten diese beiden Merkmale z.B. Preis und Sicherheit, etc. sein.

Faktorenanalyse

Faktorenanalyse

- Ziel der Faktorenanalyse ist es, mehrere Variablen durch möglichst wenige gemeinsame, hinter ihnen stehende „Faktoren“ zu beschreiben.
- Hinter der Faktorenanalyse verbirgt sich kein bestimmtes Rechenverfahren, sondern es handelt sich vielmehr um eine Sammelbezeichnung für eine Vielzahl von Verfahren, die Lösungen für das o.g. Problem liefern. Es gibt jedoch ein gemeinsames Modell.
- Grundsätzlich führt die Faktorenanalyse auf vier Teilaufgaben:
  - Ermittlung der Anzahl der Faktoren, die benötigt werden, um die Variablen zu erklären.
  - Bestimmung der Faktorenladungen: Die Faktorenladungen geben an, wie stark die Faktoren auf eine Variable wirken.
  - Ermittlung der Kommunalitäten: Die Kommunalität gibt an, welcher Anteil der Varianz einer Variablen durch die „gemeinsamen“ Faktoren aufgeklärt wird.
  - Rotation der Faktorenladungen: Zur besseren Interpretation der Faktoren ist es im Allgemeinen erforderlich, eine „Rotation“ der Faktorenladungsmatrix durchzuführen.

Faktorenanalyse - Beispiel

- Durchführung der Faktorenanalyse nach der Hauptkomponentenmethode.
- Ausgangsdaten: Schulnoten (metrisch)

	Deutsch	Philosophie	Mathematik	Physik
Schüler 1	2	3	1	2
Schüler 2	4	4	3	2
Schüler 3	1	1	2	2
Schüler 4	3	2	4	5
Schüler 5	2	2	1	2
Schüler 6	4	3	2	2
Schüler 7	4	3	2	3
Schüler 8	1	2	4	4
Schüler 9	2	2	3	3
Schüler 10	3	2	1	3

- Ausgangsfrage: Stehen hinter den Leistungen in den Fächern gemeinsame Faktoren, z.B. „sprachliche Begabung“ und „naturwissenschaftliche Begabung“?

Faktorenanalyse – Durchführung (1)

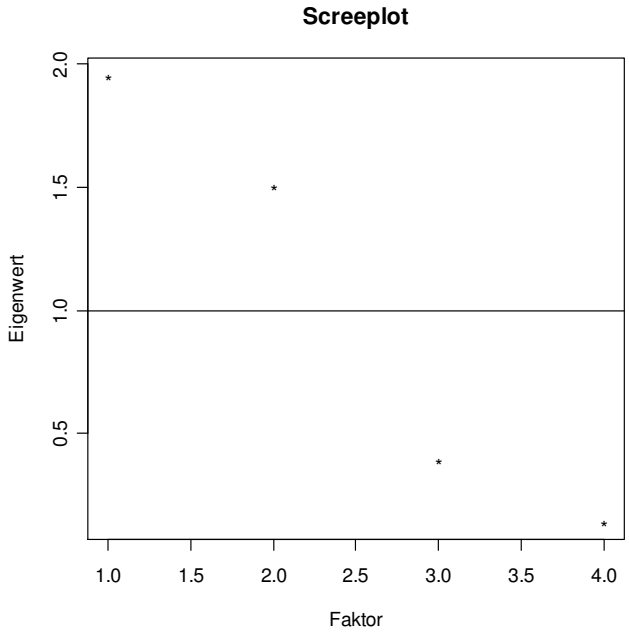
- Angabe der Korrelationsmatrix

	Deutsch	Philosophie	Mathematik	Physik
Deutsch	1	0,74	-0,07	-0,07
Philosophie	0,74	1	-0,02	-0,28
Mathematik	-0,07	-0,02	1	0,71
Physik	-0,07	-0,28	0,71	1

- Die Korrelationsmatrix soll aus möglichst wenigen „Faktoren“ und deren „Ladungen“ so gut wie möglich reproduziert werden.

- Praktisch kommt hierfür die Anwendung der PCA auf die Korrelations-. matrix zur Anwendung

- Ergebnis der PCA: Ermittlung der Zahl der Faktoren (hier: zwei)



Faktorenanalyse – Durchführung (2)

- Die Faktorenladungsmatrix (ebenfalls ein Ergebnis der PCA) gibt an, in welcher Stärke die einzelnen Variablen auf die (hier: zwei) Faktoren wirken:

	Faktor 1	Faktor 2
Deutsch	-0,69	-0,62
Philosophie	-0,77	-0,54
Mathematik	0,60	-0,70
Physik	0,72	-0,58

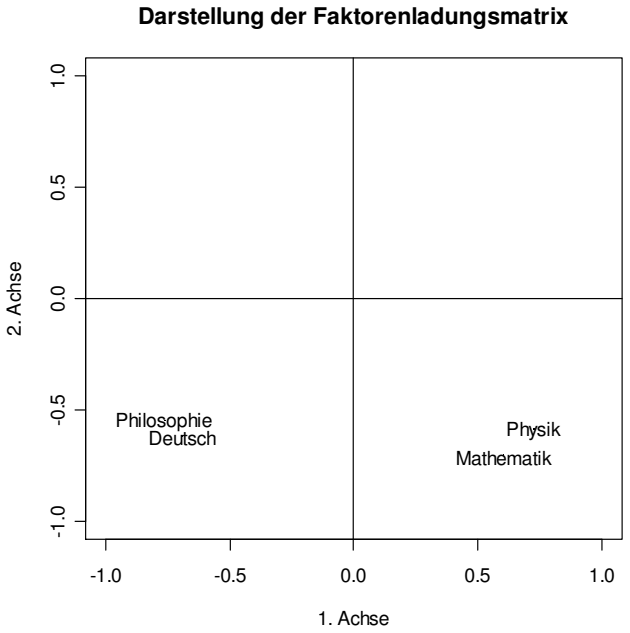
- Die Kommunalitäten geben an, welcher Anteil der Varianz der Variablen durch die (hier: zwei) Faktoren erklärt wird (in der Diagonalen dargestellt):

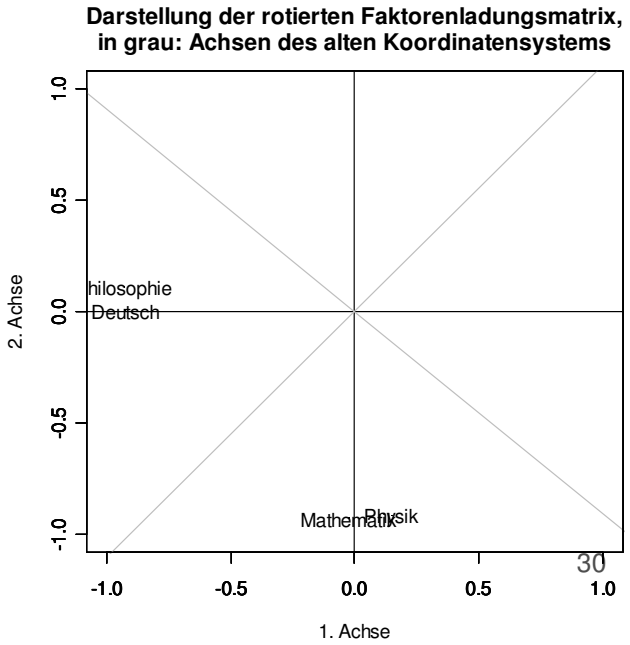
	Deutsch	Philosophie	Mathematik	Physik
Deutsch	**0,86**	0,87	0,02	-0,14
Philosophie	0,87	**0,88**	-0,08	-0,24
Mathematik	0,02	-0,08	**0,86**	0,84
Physik	-0,14	-0,24	0,84	**0,86**

Faktorenanalyse – Durchführung (3)

- Die Werte der Faktorenladungsmatrix können (bei zwei Faktoren) in einem Diagramm dargestellt werden. Man erkennt im vorliegenden Beispiel sofort, daß Philosophie und Deutsch sowie Mathematik und Physik jeweils ein Cluster bilden.

- Zur besseren Darstellung wird jedoch häufig eine Rotation der Faktoren-ladungsmatrix durchgeführt. Durch die Rotation ist es häufig einfacher, die Faktoren inhaltlich zu interpretieren. Im Ergebnis würde man hier den ersten Faktor als „sprachliche Begabung“, den zweiten als „naturwissenschaftliche Begabung“ interpretieren.





Metrische Mehrdimensionale Skalierung (MDS)

Metrische Mehrdimensionale Skalierung (MDS)

- Für die MDS werden Distanzmaße zugrunde gelegt.
- Beispiel: Verbraucher werden befragt, wie ähnlich sich je zwei Sorten Weichspüler im paarweisen Vergleich sind (z.B. von 1 = sehr ähnlich bis 10 = völlig unähnlich).
- Aus den angegebenen Distanzen wird die Lage der entsprechenden Punkte im J-dimensionalen Raum (also z.B. bei zehn Variablen im zehndimensionalen Raum) rekonstruiert. Die Lage der Punkte wird anschließend im zweidimensionalen Raum so gut wie möglich reproduziert.
- „Gefühlte Distanzen“ lassen sich nicht immer exakt reproduzieren, Beispiel: A – B = 6, B – C = 3, A – C = 2. Auf Basis dieser Distanzen können die Punkte A, B und C nicht in der Ebene dargestellt werden, denn die Dreiecksungleichung (eine Dreiecksseite ist höchstens so lang wie die Summe der beiden anderen Seiten) ist hier nicht erfüllt.

Metrische Mehrdimensionale Skalierung (MDS)

- Ermittlung der Distanzen:

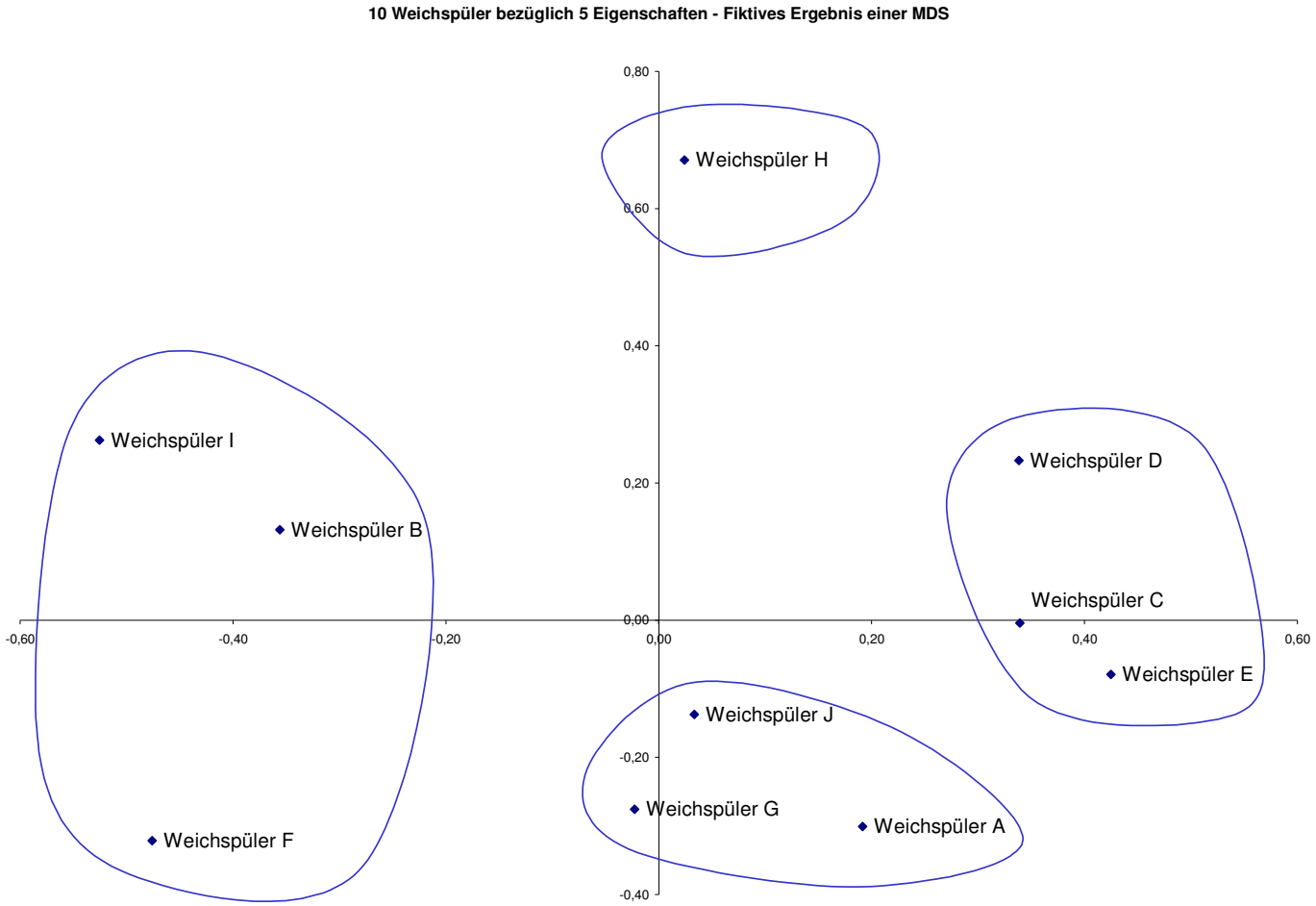
Person	A - B	A - C	A - D	...	A - J	...	I - J
1	5	1	5	4	10	1	9
2	2	1	5	8	1	6	9
3	7	4	10	2	5	7	1
4	4	3	5	1	7	9	5
5	7	2	7	8	3	8	3
6	6	8	7	10	8	1	3
7	3	1	8	3	10	3	1
8	10	6	7	5	5	9	7
9	3	1	10	1	4	4	8
10	5	4	3	1	10	8	2
...							

- Die Distanzen werden anschließend gemittelt und in einer Distanzmatrix angeordnet.

Weichspüler A	Weichspüler A	Weichspüler B	Weichspüler C	Weichspüler D	Weichspüler E	Weichspüler F	Weichspüler G	Weichspüler H	Weichspüler I	Weichspüler J
Weichspüler B	1	7,15	4,02	3,4	6,01	3,17	4,14	5,62	9,35	5,18
Weichspüler C		1	7,07	5,22	2,72	2,33	3,87	0,75	1,93	5,44
Weichspüler D			1	6,14	2,04	7,95	3,57	1,52	4,35	2,59
Weichspüler E				1	0,59	8,21	6,83	0,36	4,54	6
Weichspüler F					1	8,48	6,08	9,84	7,59	1,33
Weichspüler G						1	2,06	9,68	0,51	4,88
Weichspüler H							1	2,61	9,37	4,33
Weichspüler I								1	1,62	5,18
Weichspüler J									1	0,81
										1
...										

- Die Objekte werden auf Basis dieser Distanzmatrix in einem Rechenverfahren ähnlich der PCA in einem Raum geringer Dimensionalität abgebildet.

Ergebnis der MDS



Die Interpretation erfolgt analog zur Interpretation der Hauptkomponentenanalyse.

Korrespondenzanalyse

Korrespondenzanalyse und Optimal Scaling

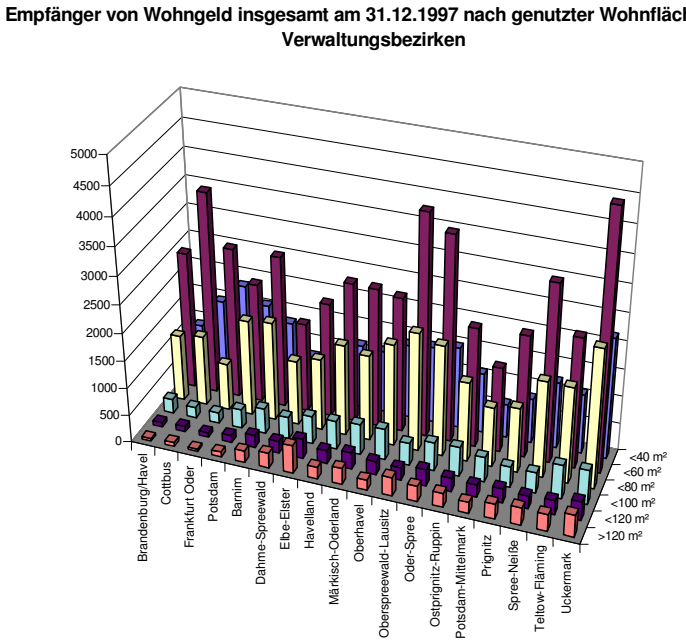
- Ausgangsdaten: Nominalskalierte Daten (Häufigkeits-daten), die in Form einer Kontingenztabelle vorliegen.
- Ziel: Darstellung der Ähnlichkeit der Zeilen , keine Darstellung einzelner Häufigkeiten, dadurch Aufdeckung der „inneren Struktur“: Welche Merkmale trennen die Zeilen am besten?

Beispiel: **Empfänger von Wohngeld insgesamt am 31.12.1997 nach genutzter Wohnfläche und Verwaltungsbezirken**

Verwaltungsbezirk/ Wohnungsgröße	<40 m²	<60 m²	<80 m²	<100 m²	<120 m²	>120 m²
Brandenburg/Havel	917	2.484	1.198	255	80	48
Cottbus	1.455	3.650	1.272	194	103	71
Frankfurt Oder	1.825	2.723	852	194	82	47
Potsdam	1.545	2.163	1.738	345	128	92
Barnim	1.302	2.743	1.793	460	228	213
Dahme-Spreewald	760	1.610	1.176	394	217	272
Elbe-Elster	605	2.067	1.298	511	356	512
Havelland	1.149	2.521	1.646	517	248	214
Märkisch-Oderland	1.128	2.510	1.550	556	306	299
Oberhavel	1.334	2.432	1.845	569	245	189
Oberspreewald-Lausitz	1.387	3.990	2.140	403	230	337
Oder-Spree	1.475	3.696	2.004	514	301	283
Ostprignitz-Ruppin	1.086	2.157	1.437	531	251	257
Potsdam-Mittelmark	609	1.537	1.080	446	224	205
Prignitz	805	2.206	1.168	377	265	278
Spree-Neiße	1.205	3.200	1.740	377	235	318
Teltow-Fläming	1.080	2.350	1.730	610	258	302
Uckermark	2.186	4.625	2.503	622	348	403

Korrespondenzanalyse und Optimal Scaling

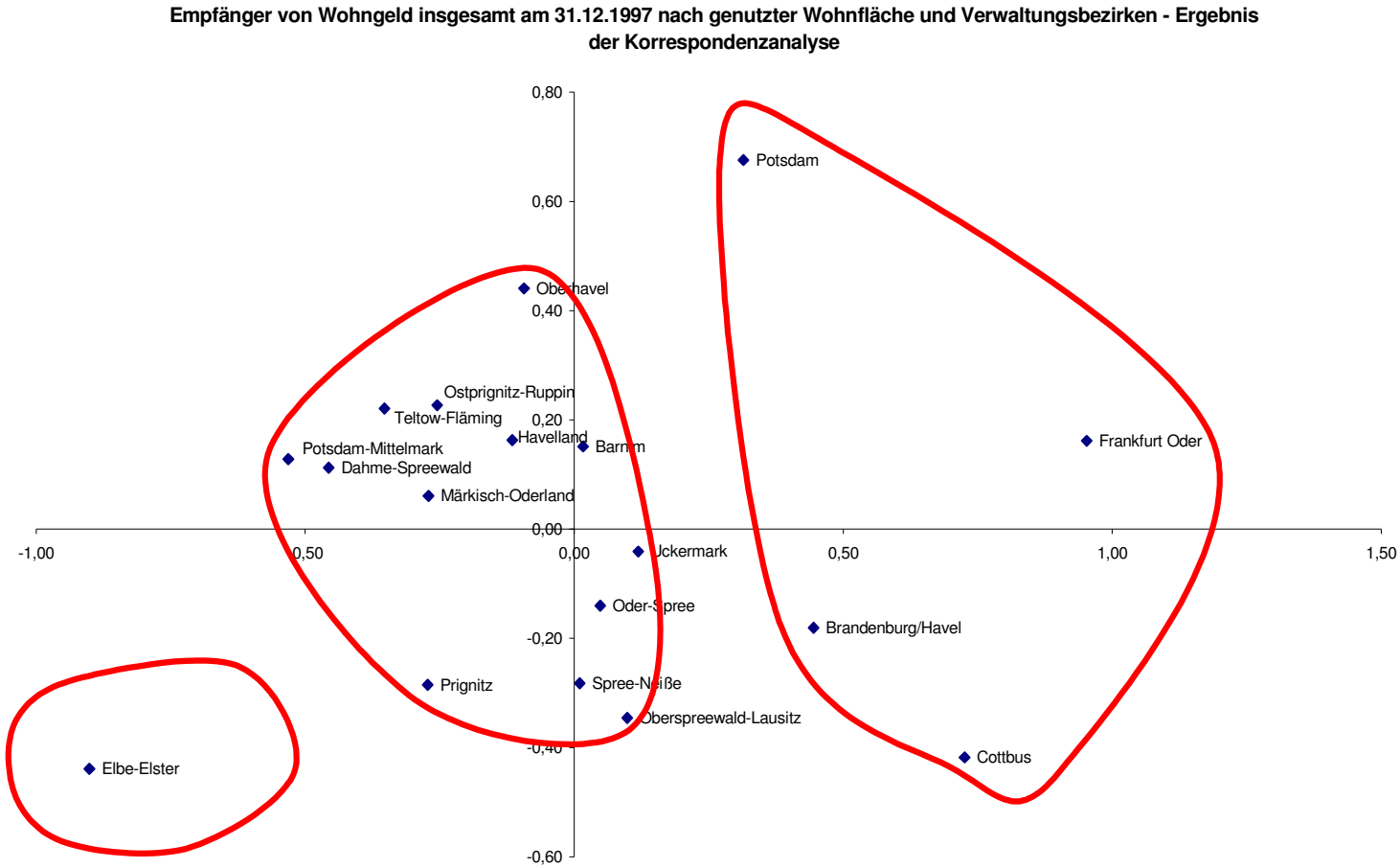
- „1. Versuch“: Darstellung mittels deskriptiver Statistik
  - Histogramme (bzw. andere Darstellungen) sind hervorragend geeignet, um Einzelwerte miteinander zu vergleichen
  - Ein Vergleich von Zeilen und/oder Spalten insgesamt ist jedoch nicht möglich



Korrespondenzanalyse

- Prinzip der Korrespondenzanalyse
  - Ansatz: Die Zeilen werden durch Vektoren im J-dimensionalen Raum (hier: J = 6 bei sechs Wohnungstypen) repräsentiert
  - Ziel: Reproduktion der Information in einem Raum geringerer Dimensionalität (vorzugsweise im zweidimensionalen Raum)
  - Betrachtung des Raumes, der von den Vektorenspitzen aufgespannt wird
  - Projektion der Vektorenspitzen auf eine Ebene (analog zur PCA)
  - Einführung der sog. „χ^2-Metrik“ – nach der Korrespondenzanalyse sind die zuvor Ergebnisse erst nach Anwendung dieser Metrik geeignet, um die Abstände zwischen den Objekten zu ermitteln.
- Nachteil:
  - keine gleichzeitige Darstellung von Zeilen und Spalten möglich
  - „χ^2-Metrik“ nicht nachvollziehbar

Ergebnis der Korrespondenzanalyse



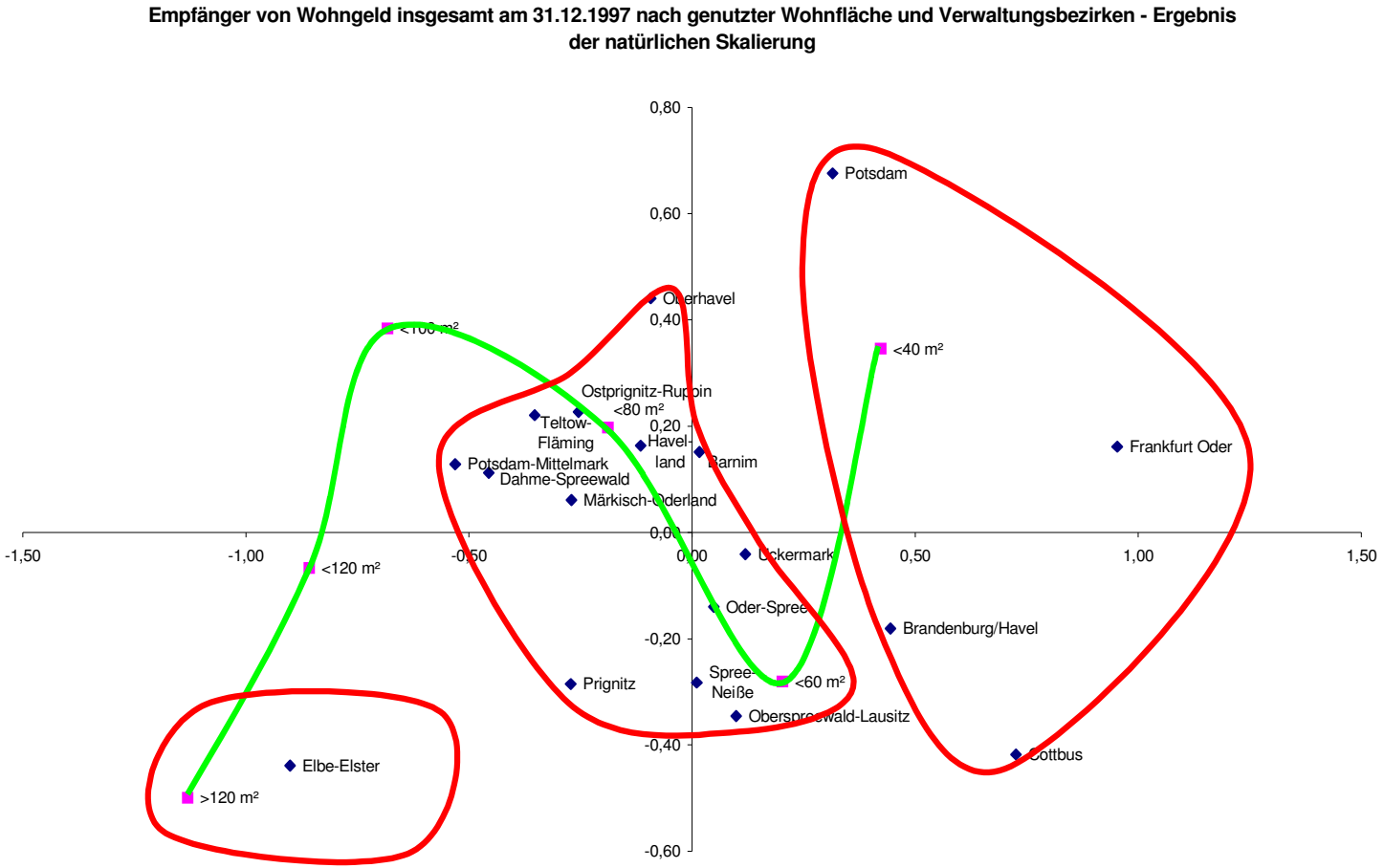
Interpretation → z.B. waagrechte Achse als Land-Stadt-Achse, etc.

Optimal Scaling

Optimal Scaling

- Aufgabenstellung und Datenanforderungen analog zur Korrespondenzanalyse
- Anwendung des Optimal Scalings
  - Ansatz: Korrelationsmaximierung (Alternative: Varianzmaximierung)
  - Ermittlung der sog. *Standardkoordinaten*
- Vorteile gegenüber der Korrespondenzanalyse:
  - Gleichzeitige graphische Darstellung von Zeilen und Spaltenmöglich, daher Zusammenhänge zwischen Zeilen und Spalten ersichtlich
  - Hierdurch Auffinden von Ansätzen für weitere Untersuchungen und Interpretationen
    - Wie genau beeinflussen die Merkmale die Objekte
    - Welches Merkmal ist für welches Objekt typisch
  - Weitere Ansätze möglich, z.B. Betrachtung der Veränderung eines Merkmals im Zeitverlauf

Ergebnis des Optimal Scalings



Zusätzliche Interpretation → Wohnungsgröße in Abhängigkeit von Stadt/Land; kleine Wohnungsgrößen typisch für Stadt, etc.