Haben
Sie
schon mal
geclustert ?
clusteranalyse
@gmx.de

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Zentrale Fragestellungen:

Was ist eine Clusteranalyse?

Wie wird eine Clusteranalyse angewendet?

Wann wird eine Clusteranalyse angewendet?

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Clusteranalyse =

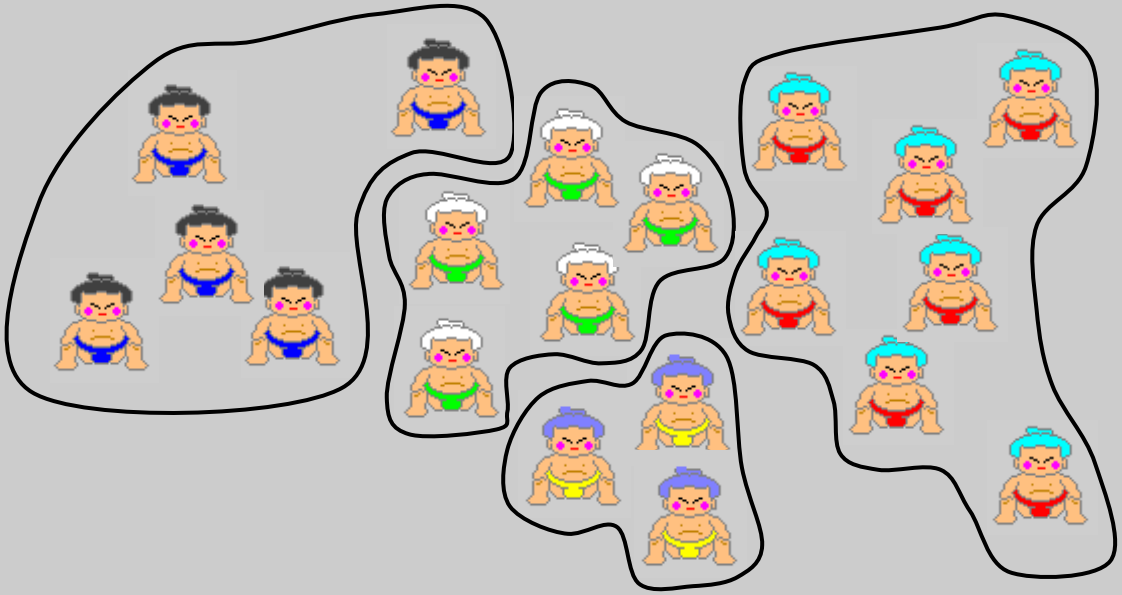
Gruppenbildungsverfahren =

eine Vielzahl von Objekten werden zu Gruppen zusammengefasst

zur Etymologie:
engl. Cluster = Haufen, Menge, Ballung
altdt. Kluster = „was dicht und dick zusammensitzt“
(Grimm‘sches Wörterbuch)

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

„was dicht und dick zusammensitzt“ - Sumo-Ringer



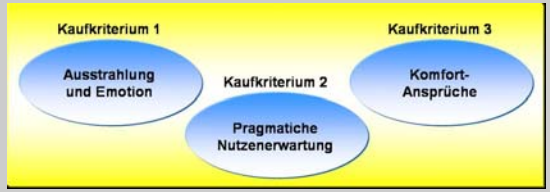
Objekte: Beschreibung durch verschiedene Merkmale unterschdl. Ausprägung

Cluster: Bildung durch Objekte mit ähnlichen Ausprägungen

Ojekte *innerhalb* einer Gruppe sollen *homogen* sein
Objekte *zwischen* den Gruppen sollen *heterogen* sein

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

- Marketing: Zusammenhang zw. Selbstbild und Wahl einer Automarke



- Archäologie: Kultureller Fingerabdrücke in der Kategorie „Schmuck“ in hallstattzeitlichen Siedlungen im Mittelrheingebiet

- Botanik: Ein pflanzensoziologisches Modell der Schattentoleranz von Baumarten in den Bayerischen Alpen

- Städtestatistiker: Sozialraumanalyse, Analyse zu Luftverschmutzung u. Lärmbelastung, Bürgerumfragen, Wirtschaftsräumen, Wahlanalysen etc.

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Sozialraumanalyse:

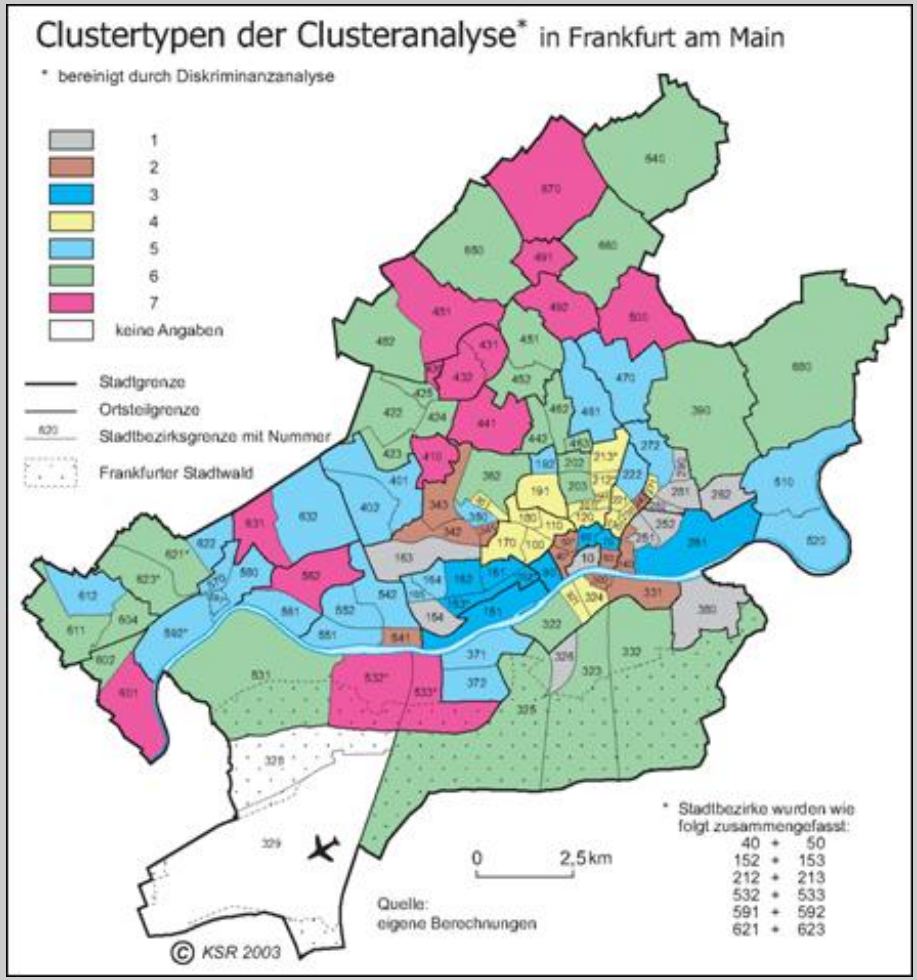
- Cluster = ähnliche Bezirke

- Merkmale sozio-ökonomische Variablen*:

- unter 6-Jährige
- über 65-Jährige
- Zu- u. Wegzüge
- Ausländer/-innen
- Einpersonenhaushalte
- Arbeitslosendichte
- Sozialhilfeempfänger/innen
- Mehrfamilienhäuser
- Wahlbeteiligung

* jeweils Anteile

=> Variablenauswahl extrem wichtig



Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Datenvoraussetzungen - Empfehlungen:

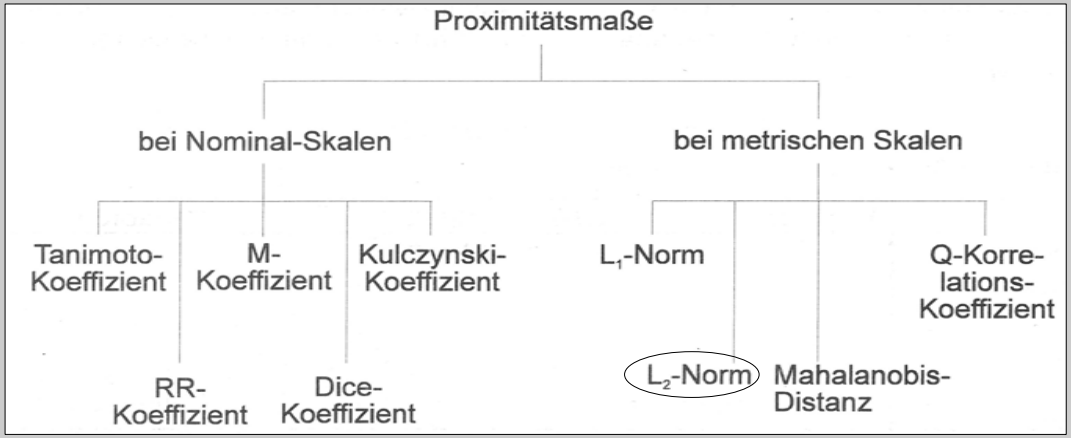
- **kein spezielles Skalenniveau**
- **Standardisierte Merkmale** (z-Transformation)
- **Ausreißer ausschließen** (Verzerrungen)
- **Anzahl der Merkmale**
  Keine Begrenzung, aber:
  - nur relevante Variablen einbeziehen - Vorüberlegungen
  - möglichst hoch korrelierenden Variablen ausschließen
  - keine Variablen mit konstanten Ausprägungen bei allen Objekten
- **Anzahl der Objekte**
  Keine Begrenzung

Drei Ablaufschritte:

1. Bestimmung der Distanz (Abstand-Differenz) durch Proximitätsmaße
2. Auswahl des Fusionierungsalgorithmuses
3. Bestimmung der Clusteranzahl

Proximitätsmaße: Maßzahl zur Quantifizierung des Abstandes der Objekte durch Merkmalswerte

Vielzahl von Maßberechnungen - abhängig vom Skalenniveau



Standard bei Distanzmaßen bei metrischen Skalen:
L2-Norm = Quadrierte Euklidische Distanz

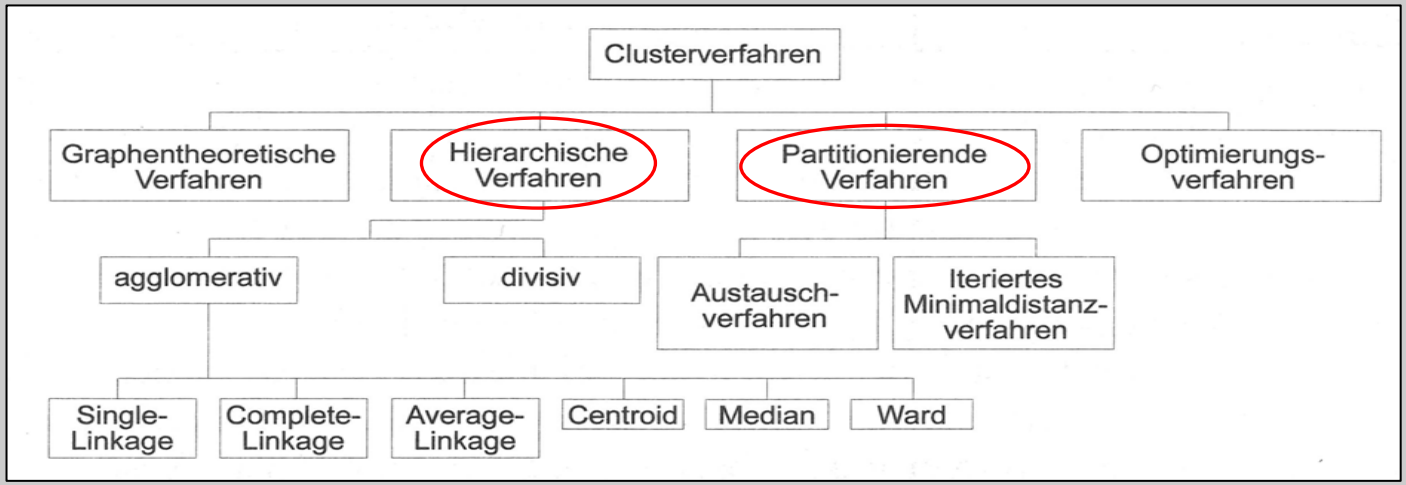
Berechnung: absoluten Differenzwerte werden quadriert und addiert

Fragestellung | Definition | Voraussetzung | **Methodik** | Interpretation

Fusionierung

Ausgangsdatenmatrix → **Distanzmaß** (Quadrierte Euklidische Distanz) → **Distanzmatrix**

Distanzmatrix → **Fusionierung** → **mit Hilfe von Cluster-Algorithmen:**



Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

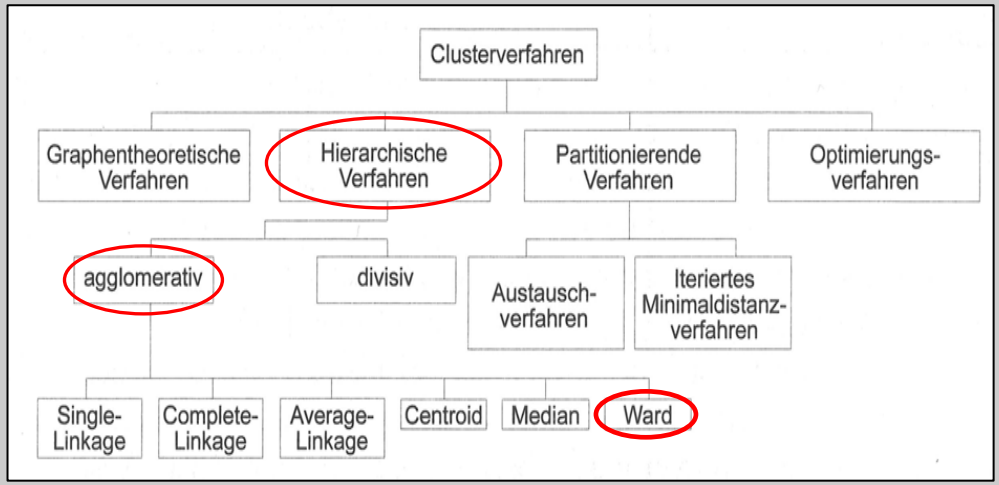
Partitionierendes Verfahren:

1. Vorgabe einer Anfangspartition
2. jedes Objekt kann im Prozess jederzeit verschoben werden
3. Festlegung der Clusteranzahl im vorhinein

Hierarchisch-agglomerativ Verfahren:

1. Keine Vorgabe – Start mit feinster Partition (jedes Objekt ist ein Cluster)
2. Objekte mit der geringsten Distanz werden verbunden, später Gruppen
3. „Durchlaufen“ zu einem Großcluster

Häufigste Anwendung:



Ward Verfahren:

Bildung von homogeneren Clustern –

„Vereinigt diejenigen Objekte, die die Fehlerquadratsumme (Varianz/Streuung) am wenigsten erhöhen“

Möglichkeiten zur Bestimmung der Clusteranzahl:

1. Fehlerquadratsumme
2. Elbow-Kriterium
3. Dendrogramm

Zur Erinnerung! : aggl. Verf. nach Ward geht von der kleinsten Partition aus und endet bei einem Großcluster!

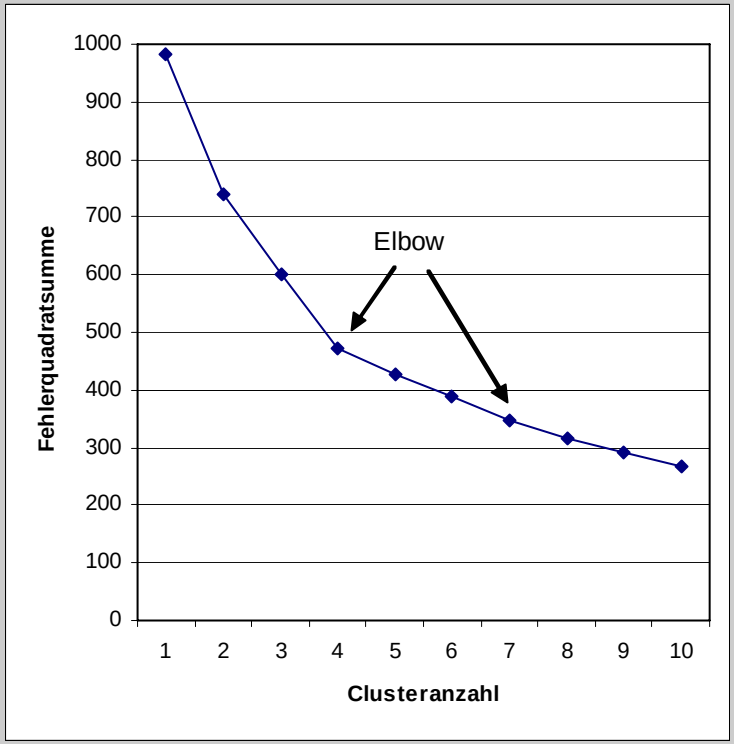
1. Fehlerquadratsumme

= gibt die Varianzveränderung an, bei großen Sprüngen werden heterogene Cluster zusammengeführt

Schritt	Zusammen-geführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt	Diff. CL1:CL2	Clusteranzahl
	Cluster 1	Cluster 2		Cluster 1	Cluster 2			
100	25	62	267,35	99	91	103	21,8	10
101	14	18	291,56	94	97	105	24,2	9
102	3	13	317,16	96	0	108	25,6	8
103	**25**	**47**	**346,46**	**100**	**74**	**107**	**29,3**	**7**
104	**2**	**7**	**387,24**	**93**	**89**	**106**	**40,8**	**6**
105	14	65	428,42	101	98	107	41,2	5
106	**1**	**2**	**473,85**	**95**	**104**	**108**	**45,4**	**4**
107	**14**	**25**	**599,89**	**105**	**103**	**109**	**126,0**	**3**
108	1	3	738,35	106	102	109	138,5	2
109	1	14	981,00	108	107	0	242,7	1

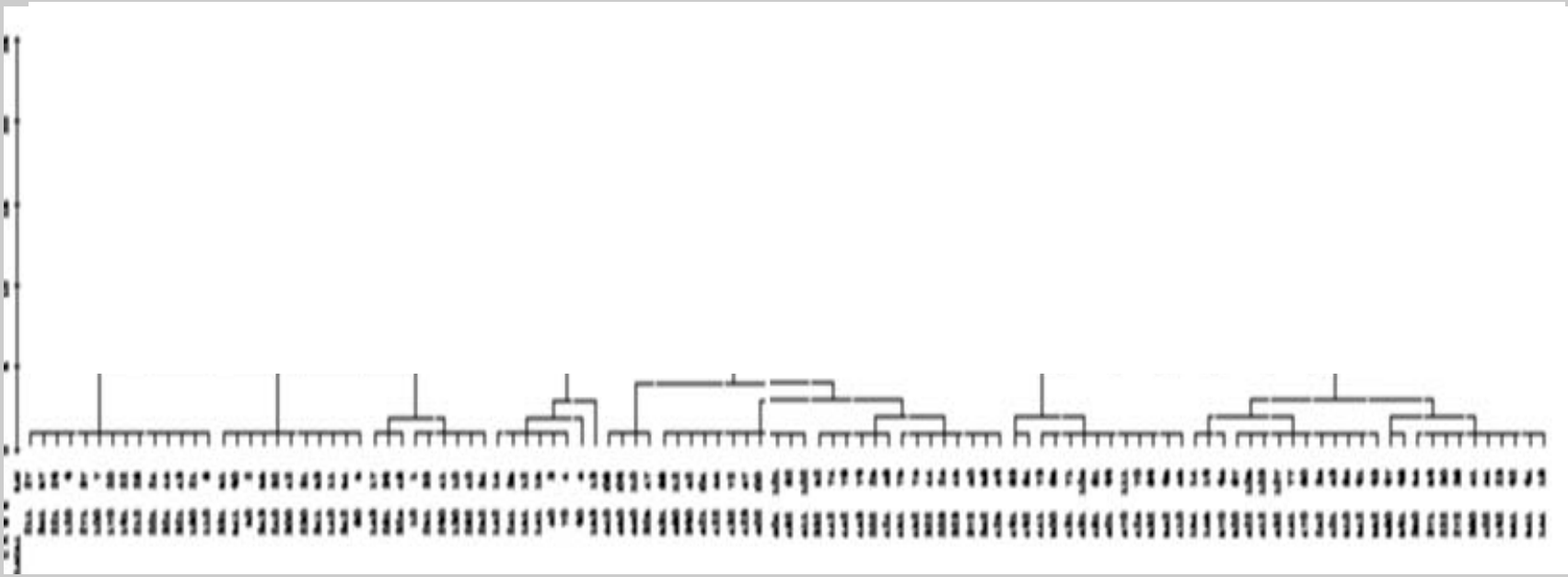
2. Elbow-Kriterium

= Abtragung der Fehlerquadratsumme in ein Diagramm



3. Dendrogramm

= Abtragung der Fusionsierungsschritte



Festlegung Clusteranzahl => Clustermatrix

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Drei Ablaufschritte:

1. Bestimmung der **Distanz** durch Proximitätsmaße:

   Quadrierte Euklidische Distanz

2. Auswahl des **Fusionierungsalgorithmuses**:

   hierarchisch, agglomerativ nach Ward

3. Bestimmung der **Clusteranzahl**:

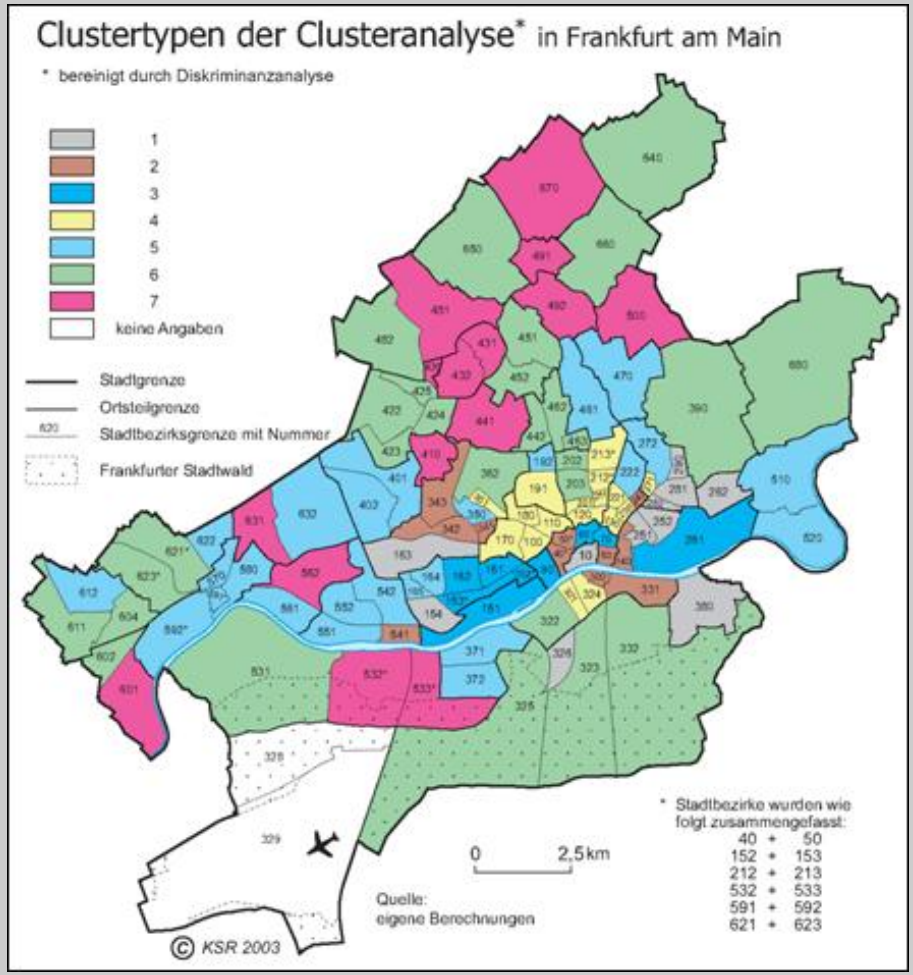
   Fehlerquadratsumme, Elbow und Dendrogramm

Hilfestellung für die Interpretation:

- absoluten Merkmals-ausprägungen
- Karte

Für Spezialisten:

- F- Werte (Varianz/Streuung)
- t-Wert (Standardabweichung)



Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

- **F-Werte (Varianz): Homogenität einer Gruppe;**
- **t-Werte (Standardabweichung): Überbewertung/Unterbewertung einer Variablen)**

Cluster	Arbeits-losendichte	Anteil d. Sozial-hilfeempfänger/ -innen an der Bevölkerung	Wahlbe-teiligung b.d. Bundes-tagswahl	Anteil d. Einpersonen-haushalte a.d. Privathaush.	Anteil der Bev. im Alter v. u. 6 Jahren a.d. Bevölkerung	Anteil d. Mehr-familienhäuser an den Wohn-gebäuden	Anteil d. Zu- u. Wegzüge an der Bevölkerung	Anteil der ausl. Bev. an der Bevölkerung	Anteil der Bev. v. 65 Jahren und älter a.d. Bevölkerung
1	0,35	0,40	0,40	0,13	0,19	0,32	0,15	0,27	0,76
2	0,42	0,35	0,35	0,34	0,18	0,28	0,43	0,10	0,58
3	1,38	0,60	0,60	0,34	1,88	0,07	1,12	0,28	0,66
4	0,38	0,15	0,15	0,06	0,09	0,07	0,11	0,08	0,21
5	0,41	0,34	0,34	0,30	0,49	0,35	0,33	0,21	0,40
6	0,50	0,39	0,39	0,71	0,42	0,50	0,31	0,19	0,89
7	0,23	0,18	0,18	0,12	0,48	0,23	0,09	0,06	0,28

Cluster	Arbeits-losendichte	Anteil d. Sozial-hilfeempfänger/ -innen an der Bevölkerung	Wahlbe-teiligung b.d. Bundes-tagswahl	Anteil d. Einpersonen-haushalte a.d. Privathaush.	Anteil der Bev. im Alter v. u. 6 Jahren a.d. Bevölkerung	Anteil d. Mehr-familienhäuser an den Wohn-gebäuden	Anteil d. Zu- u. Wegzüge an der Bevölkerung	Anteil der ausl. Bev. an der Bevölkerung	Anteil der Bev. v. 65 Jahren und älter a.d. Bevölkerung
1	0,55	0,37	0,02	0,44	-0,81	0,60	-0,50	-0,30	1,18
2	-0,09	0,25	-0,76	0,90	-0,86	0,77	1,21	0,97	-0,74
3	1,57	1,16	-2,19	1,07	-0,47	0,98	2,31	2,42	-1,18
4	-0,82	-0,83	0,64	1,15	-0,92	1,12	0,22	-0,22	-0,75
5	0,63	0,48	-0,46	-0,28	0,67	0,00	-0,13	0,45	-0,04
6	-0,80	-0,92	0,83	-0,61	0,09	-0,91	-0,61	-0,97	0,65
7	0,14	0,77	0,24	-1,25	1,25	-1,11	-0,65	-0,49	-0,06

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

- Vorüberlegungen sind wichtig:
  - Probleme bei der Auswahl der Variablen
  - Korrelationen
  - Ausreißer
- Festlegung von Grenzwerten – aber welche Grenzwerte sind gültig?
- Bestimmung der Clusteranzahl – ein Glücksspiel?
- Zuweisung bei hierarchisch-agglomerativen Verfahren
  (nicht revidierbar, dafür Tendenz, gleich große Gruppen zu bilden, Ausreißer erkennbarer)

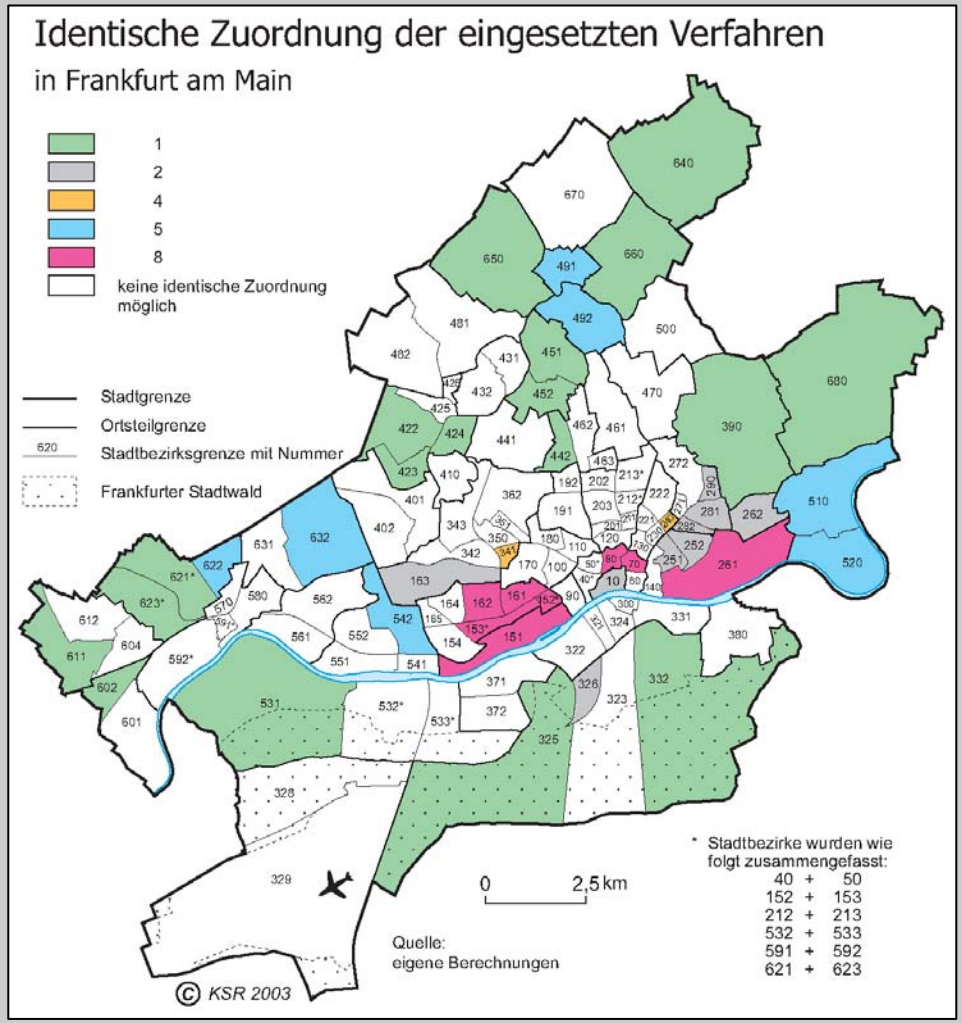
- **Zur Überprüfung der Klassenzugehörigkeit: Diskriminanzanalyse**

- **Vorüberlegungen sind wichtig:**
  Konkretisierung der Problemstellung der Untersuchung
  Verfahren:
  - **Faktorenanalyse (Reduzierung der Merkmale)**
  - **Klassische Raumanalyse (feste Grenzwerte)**

Ergebnisvergleich:

Konkruente Zuordnung – Klasssiche Raumanalyse:

- 43 Bezirke von 110 Bezirken waren identisch (39,1 %)
- Bezirke mit eindeutiger „inhaltlicher Aussage“



Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.

Ergebnis:

- Auswahl der Analyseverfahren muss sich an konkreten Fragestellungen orientieren
- methodische Annahmen, Voraussetzungen und Bedingungen müssen erläutert werden
- Clusteranalyse mit unterschiedlichen Programmen durchführbar

=> Clusteranalyse ist ein geeignetes Verfahren für Städtestatistiker

Sind Sie nun von der

Clusteranalyse

überzeugt ! ?

Dr. Ralf Gutfleisch, Stadt Frankfurt a.M.