

# Einführung in die Cluster-Analyse mit SAS

---

Benutzertreffen am URZ  
Carina Ortseifen  
4. Juli 2003

---

---

---

---

---

---

---

---

## Inhalt

---

1. Clusteranalyse im allgemeinen  
Definition, Distanzmaße, Gruppierung, Kriterien
2. Clusteranalyse mit SAS
  - a) Hierarchische Clusteranalyse  
Prozedur Cluster
  - b) Partitionierende Clusteranalyse  
Prozedur Fastclus
3. Literatur

---

---

---

---

---

---

---

---

## 1. Cluster (dt.: Traube, Haufen)

---

- *heuristisches Verfahren* zur systematischen Klassifizierung von Beobachtungen, z.B. Personen, Autos, Schallplatten)
- *Ziel*: Auffinden von Gruppen, in denen sich Beobachtungen befinden, die innerhalb der Gruppe möglichst ähnlich sind und extern (zwischen den Gruppen) verschieden.
- *Anwendungsgebiete*: Sozialwissenschaften, Biologie, Wirtschaftswissenschaften, Marktforschung

---

---

---

---

---

---

---

---

## Ähnlichkeit / Unähnlichkeit

- Die Ähnlichkeit bzw. Unähnlichkeit wird auf der Basis von Merkmalen definiert.
- Z.B. gleiches Alter, gleiche Haarfarbe.
- Andere Begriffe für Unähnlichkeit: Distanz  
Ähnlichkeit: Proximität

---

---

---

---

---

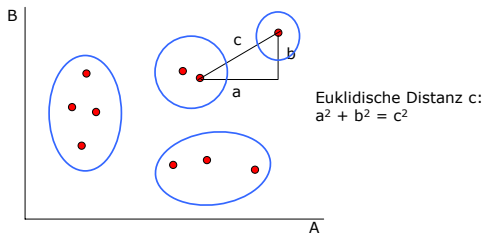
---

---

---

## Beispiel

- Zehn Fälle, zwei stetige Merkmale A/B



---

---

---

---

---

---

---

---

## Euklidische Distanz allgemein

P Merkmale: 
$$d_{ii'} = \left[ \sum_{i=1}^p (X_{ij} - X_{i'j})^2 \right]^{1/2}$$

Beispiel für 2 Fälle A und B, 7 Merkmale:

A: 5 7 8 1 3 2 5

B: 9 5 8 2 7 8 2

$$d_{ab} = \sqrt{(5-9)^2 + (7-5)^2 + \dots + (5-2)^2} = 9,055$$

---

---

---

---

---

---

---

---

## Distanzmaße für metrische Variabl.

- *Euklidische Distanz*
- *City Block-Distanz*
  - Summe der absoluten Differenzen
- = Spezialfälle der *Minkowski-Distanz*
  - Hohe Unterschiede werden stark gewichtet.
  - Maße sind translationsinvariant, aber nicht skaleninvariant. (Einkommen in Dollar oder Euro)

---

---

---

---

---

---

---

---

## Distanzmaße für metrische Var. (2)

- *Mahalanobis-Distanz*

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

Wurzel ?

wobei  $S^{-1}$  die Inverse der Stichproben-Varianz-Kovarianzmatrix der  $p$  Merkmale ist.

Translations- und Skaleninvariant

---

---

---

---

---

---

---

---

## Dichotome Merkmale (Beispiel)

Zwei Beobachtungen, A und B, 9 Merkmale, die angeben, ob ein Sachverhalt gegeben ist oder nicht

A: 0 0 1 1 0 0 1 1 1  
B: 1 0 1 0 0 0 0 1 1

	B	
A	1	0
1	a	b
0	c	d

---

---

---

---

---

---

---

---

## Ähnlichkeitskoeffizient von Jaccard

- $p_{ij} = a / (a+b+c)$  (d spielt keine Rolle)
- Das entsprechende Distanzmaß ist:  
 $d_{ij} = 1 - p_{ij} = (b+c) / (a+b+c)$
- $p_{ij}$  nimmt Werte zwischen 0 und 1 an.

Für das Beispiel:  $p_{AB} = 3/6 = 0.5$ .

---

---

---

---

---

---

---

---

## Distanzmaße für binäre Merkmale

(Simple) *Matching Koeffizient*

$$p_{ij} = a+d / (a+b+c+d)$$

Jaccard- (Tanimoto-) *Koeffizient*

$$p_{ij} = a / (a+b+c)$$

*RR-Koeffizient*

$$p_{ij} = a / (a+b+c+d)$$

*Dice-Koeffizient*

$$p_{ij} = 2a / (2a+b+c)$$

---

---

---

---

---

---

---

---

## Mögliche Probleme

- Ungleiche Skala  
→ *Standardisierung*
- Ungleiches Skalenniveau der Merkmale  
→ binäre Merkmale als metrische betrachten  
→ metrische Merkmale binär kodieren  
→ Aggregation der verschiedenen Distanzmaße
- Merkmale sind korreliert  
→ *Berechnung von Faktorwerten*  
→ *Mahalanobis-Distanz*
- Ordinalskalierte Merkmale  
→ Merkmale am Median dichotomisieren  
→ Merkmale als metrische Daten behandeln

---

---

---

---

---

---

---

---

## Cluster-Analyse-Verfahren

	<b>Hierarchische Verfahren</b>	<b>Nicht-hierarchische Verfahren(*)</b>
<b>Start</b>	feinste Partionierung, jedes Objekt bildet ein eigenes Cluster	Vorgabe einer Startgruppierung
<b>Clusterbildung</b>	Fusionierung von Clustern	Verschieben der Objekte
<b>Ziel</b>	Das zuvor festgelegte Kriterium ist erfüllt.	Das zuvor festgelegte Kriterium ist erfüllt.

\* Auch: Partitionierendes Cluster-Analyse Verfahren

---

---

---

---

---

---

---

---

## Nicht-Hierarchische Verfahren

- Objekte werden solange in verschiedene Gruppen sortiert, bis die beste Lösung im Sinne des Kriteriums gefunden ist.
- *Problem*: enormer Arbeits- und Zeitaufwand (bei 10 Objekten gibt es schon 115 975 verschiedene Möglichkeiten), deshalb sind meist nur Annäherungen möglich.

---

---

---

---

---

---

---

---

## Hierarchische Verfahren

1. Berechnung der Distanzen zwischen den Clustern
2. Fusionierung der beiden Cluster, die die geringste Distanz zueinander haben
3. Berechnung des Ende-Kriteriums  
Wenn erfüllt, dann Ende; sonst weiter.
4. Berechnung der neuen Distanzen
5. Zurück zu Punkt 2

---

---

---

---

---

---

---

---

## Distanzen zwischen den Clustern

### *Single Linkage*

Kleinste Distanz zwischen einem Objekt des einen Clusters und einem Objekt des anderen Clusters

- Ketten-Tendenz

### *Complete Linkage*

größte Distanz zwischen einem Objekt des einen Clusters und einem Objekt des anderen Clusters

- anfällig für Ausreißer

---

---

---

---

---

---

---

---

## Distanzen zwischen Clustern (2)

### *Average Linkage*

Durchschnitt aller Distanz zwischen den Objekten der beiden betrachteten Cluster

- tendiert dazu Cluster mit kleinen Varianzen zu verbinden, neigt zu Clustern mit gleicher Varianz

### *Zentroid*

Quadrierte Euklidische Distanz zwischen Cluster-Mittelwerten

- nur für metrische Merkmale, robust gegenüber Ausreißern)

---

---

---

---

---

---

---

---

## Distanzen zwischen Clustern: Ward

- Distanz ist die Anova-Quadratsumme zwischen zwei Clustern (nur für intervallskalierte normalverteilte Daten)
- vereinigt diejenigen Elemente, deren Fusion die Gesamtvarianz innerhalb der Cluster am geringsten erhöht
- findet Cluster mit annähernd gleicher Besetzungszahl
- anfällig für Ausreißer

---

---

---

---

---

---

---

---

## Bewertungskriterien

---

### *Bestimmtheitsmaß $r^2$ (RSQ)*

- Anteil der Abweichungsquadratsumme zwischen den Clustern an der gesamten Abweichungsquadratsumme

### *Semipartielles Bestimmtheitsmaß*

- Zunahme der Abweichungsquadratsumme innerhalb der Cluster durch die Fusionierung der beiden zuletzt vereinigten Cluster dividiert durch die gesamte Abweichungsquadratsumme (Homogenitätsverlust innerhalb der Cl.)

---

---

---

---

---

---

---

---

## Bewertungskriterien (2)

---

### *Pseudo-F*

- Abweichungsquadratsumme zwischen den Clustern durch Abweichungsquadrate innerhalb der Cluster
- Ist nicht F-verteilt !!!

### *Pseudo-t2*

- Zunahme der Abweichungsquadratsumme innerhalb der Cluster  $k$ ,  $l$  und  $m$  dividiert durch Abweichungsquadratsumme in den Clustern  $k$  und  $l$

---

---

---

---

---

---

---

---

## Überprüfung der Cluster-Lösung

---

- Inhaltliche Interpretation
  - Deskriptive Unterschiede zwischen den Clustern auf weiteren Variablen
- Diskriminanzanalytische Überprüfung
  - Clustervariable als Gruppenvariable
- Graphische Veranschaulichung
  - Eiszapfen, Dendogramm, Plot (auch von den Kriteriumswerten wie Pseudo-F)

---

---

---

---

---

---

---

---

## 2. Cluster-Analyse im SAS

Prozeduren für Cluster-Analysen:  
CLUSTER (hierarchische Methoden)  
FASTCLUS (besonders für große Datensätze, nicht-hierarchisch, k-means)  
VARCLUS (Clusteranalyse auf der Basis von Korrelations- oder Kovarianzmatrizen)  
MODECLUS, OVERCLUS, ACECLUS

---

---

---

---

---

---

---

---

### a. Hierarchische Clusteranalyse

5 Probanden wurden gefragt, wie viele Stunden pro Woche sie für Sport, Medien, Hobbies aufbringen.

```
DATA clusdat;  
  INPUT nr sport medien hobbies @@;  
  LINES;  
  1 1 5 3 2 0 6 3 3 2 2 8  
  4 5 3 1 5 5 4 0  
  ;
```

---

---

---

---

---

---

---

---

### Single Linkage Verfahren

```
PROC CLUSTER DATA=clusdat  
  METHOD=SINGLE  
  PSEUDO RSQUARE NONORM;  
  VAR sport medien hobbies;  
RUN;
```

---

---

---

---

---

---

---

---

## Ergebnis

NCL	-Clusters	Joined-	FREQ	SPRSQ	RSQ	PSF	PST2	Min Dist
4	OB1	OB2	2	0.01445	0.986	22.7	.	1.4142
3	OB4	OB5	2	0.01445	0.971	33.6	.	1.4142
2	CL4	CL3	4	0.44075	0.530	3.4	30.5	4.8990
1	CL2	OB3	5	0.53035	0.000	.	3.4	5.9161

Entscheidung fällt für 3 Cluster, aufgrund des  $R^2$  und des Pseudo-F-Wertes.

---

---

---

---

---

---

---

---

## Dendrogramm

```
PROC CLUSTER ... OUTTREE=clustree . . ;  
PROC TREE DATA=clustree NCLUSTERS=3  
HORIZONTAL GRAPHICS;  
ID nr;  
RUN;
```

---

---

---

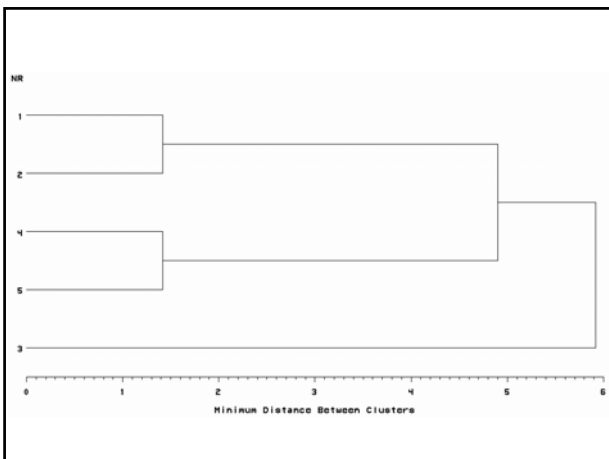
---

---

---

---

---



---

---

---

---

---

---

---

---

## Plot der Cluster-Lösung

```
PROC TREE DATA=clustree OUT=outtree  
  NCLUSTERS=3;  
  COPY nr sport medien hobbies;  
RUN;  
  
SYMBOL1 v=dot;  
PROC GPLOT DATA=outtree;  
  PLOT sport*medien=cluster;  
RUN;
```

---

---

---

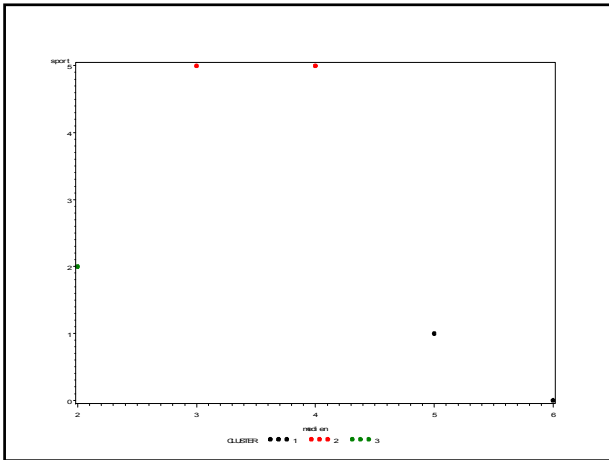
---

---

---

---

---



---

---

---

---

---

---

---

---

## Syntax der Proc Cluster

```
PROC CLUSTER METHOD=Name Optionen;  
  VAR Variablenliste;  
  ID Identifikations-Variable;  
  (z.B. nr)  
  COPY Variablenliste;  
  BY Variablenliste;
```

---

---

---

---

---

---

---

---

## Wichtige Optionen

DATA *Koordinaten oder Distanzen (type=distance)*  
NOEIGEN *Unterdrückung der Eigenwertberechnung*  
NOID *unterdrückt die Identifikation*  
NONORM *verhindert Normierung der Distanzen*  
OUTTREE=SAS-Datei *Herstellen einer Datei zum Zeichnen von Dendrogrammen*

---

---

---

---

---

---

---

---

## Wichtige Optionen (2)

PSEUDO *Berechnung von Pseudo F und Pseudo  $t^2$*   
RSQUARE *Berechnung des Bestimmtheitsmaßes*  
STANDARD *Standardisierung aller Merkmale vor Durchführung der Clusteranalyse*  
CCC *Cubic Cluster Criterion, vereinigt ERSQ und RSQ zu einem Maß*

---

---

---

---

---

---

---

---

## b. Partitionierende Clusteranalyse

- Prozedur FASTCLUS
- Nearest Centroid Sorting-Verfahren
  - Distanzmaß: Quadr. eukl. Distanz
  - 1. Schritt: Festlegung der Startwerte (n Beobachtungen mit Mindestdistanz)
  - 2. Schritt: Zuordnung der übrigen Beobachtungen zu den Startwerten
  - Evtl. 3. Schritt: Berechnung der Zentroide und erneute Zuordnung
  - 4. Schritt: Wiederholung der Schritte 1- 3 bis sich Zentroide nicht mehr ändern

---

---

---

---

---

---

---

---

## Beispiel

```
Proc Fastclus Data=clusdat
      Maxclusters=3
      Out=fastout;
Var sport medien hobbies;
Id nr;
Run;
```

The FASTCLUS Procedure  
Replace=FULL Radius=0 Maxclusters=3 Maxiter=1

### Initial Seeds

Cluster	hobbies	medien	sport
1	1.000000000	3.000000000	5.000000000
2	3.000000000	6.000000000	0.000000000
3	8.000000000	2.000000000	2.000000000

Criterion Based on Final Seeds = 0.3651

### Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	2	0.5774	0.7071	2	5.5227
2	2	0.5774	0.7071	1	5.5227
3	1	.	0	2	6.2849

### Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
hobbies	3.08221	0.50000	0.986842	75.000000
medien	1.58114	0.70711	0.900000	9.000000
sport	2.30217	0.50000	0.976415	41.400000
OVER-ALL	2.40139	0.57735	0.971098	33.600000

Pseudo F Statistic = 33.60

Approximate Expected Over-All R-Squared = .

Cubic Clustering Criterion = .

ARNING: The two values above are invalid for correlated variables.

```
Proc Print Data=fastout;
Run;
```

Obs	nr	sport	medien	hobbies	CLUSTER	DISTANCE
1	1	1	5	3	2	0.70711
2	2	0	6	3	2	0.70711
3	3	2	2	8	3	0.00000
4	4	5	3	1	1	0.70711
5	5	5	4	0	1	0.70711

Die Option OUT= liefert eine SAS-Tabelle, welche die Clustervariable automatisch enthält.

---

---

---

---

---

---

---

---

## Weitere Optionen

- Delete= n Zentroide mit weniger als n Objekten werden gelöscht (trifft nicht unbedingt für die Endlösung zu)
- Impute Ersetzung von fehlenden Werten
- Replace=Full|Part|None|Random Austausch der Startwerte

---

---

---

---

---

---

---

---

## Literatur

- SAS/STAT Users Guide Volume 1
- Oerthel, F. & Tuschl, S.: Statistische Datenanalyse mit dem Programmpaket SAS. Oldenbourg, 1995.
- Bortz, J. : Statistik für Sozialwissenschaftler Springer Lehrbuch, 4.Aufl. 1993.
- Späth, H. :Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. Oldenbourg, 1977.

---

---

---

---

---

---

---

---