Workshop Clusteranalyse

Clusteranalyse – Hierarchische Verfahren

Graz, 8. – 9. Oktober 2009

Johann Bacher

Johannes Kepler Universität Linz

Linz 2009

1. Programmsystem ALMO

vollständiges Statistikprogramm, von Prof. Holm (JKU Linz) entwickelt, Clusteranalyseteile von Bacher (1996, 1999).

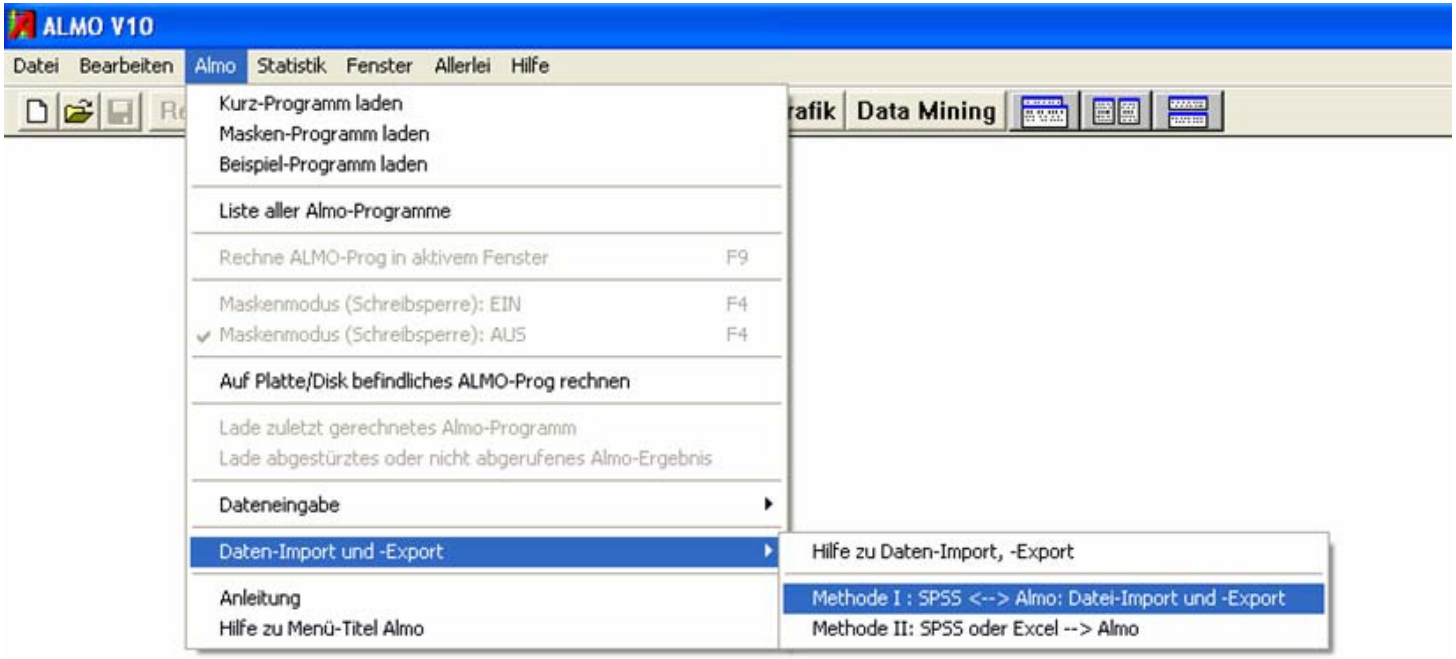
Clusteranalyse in ALMO umfasst:

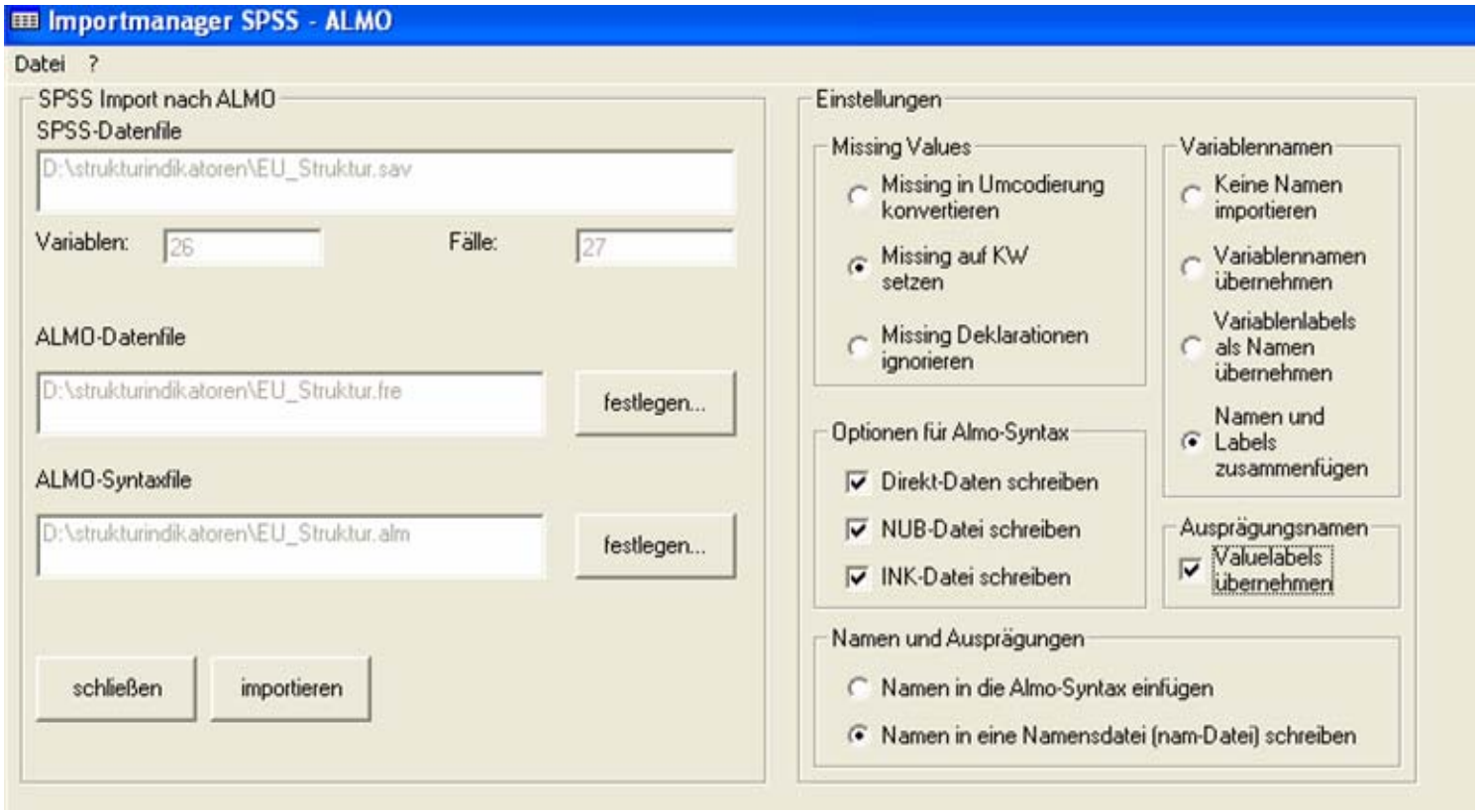
- hierarchische Verfahren
- partitionierende Verfahren
- probabilistische Verfahren (→ elaboriertes Modell LatentGOLD)

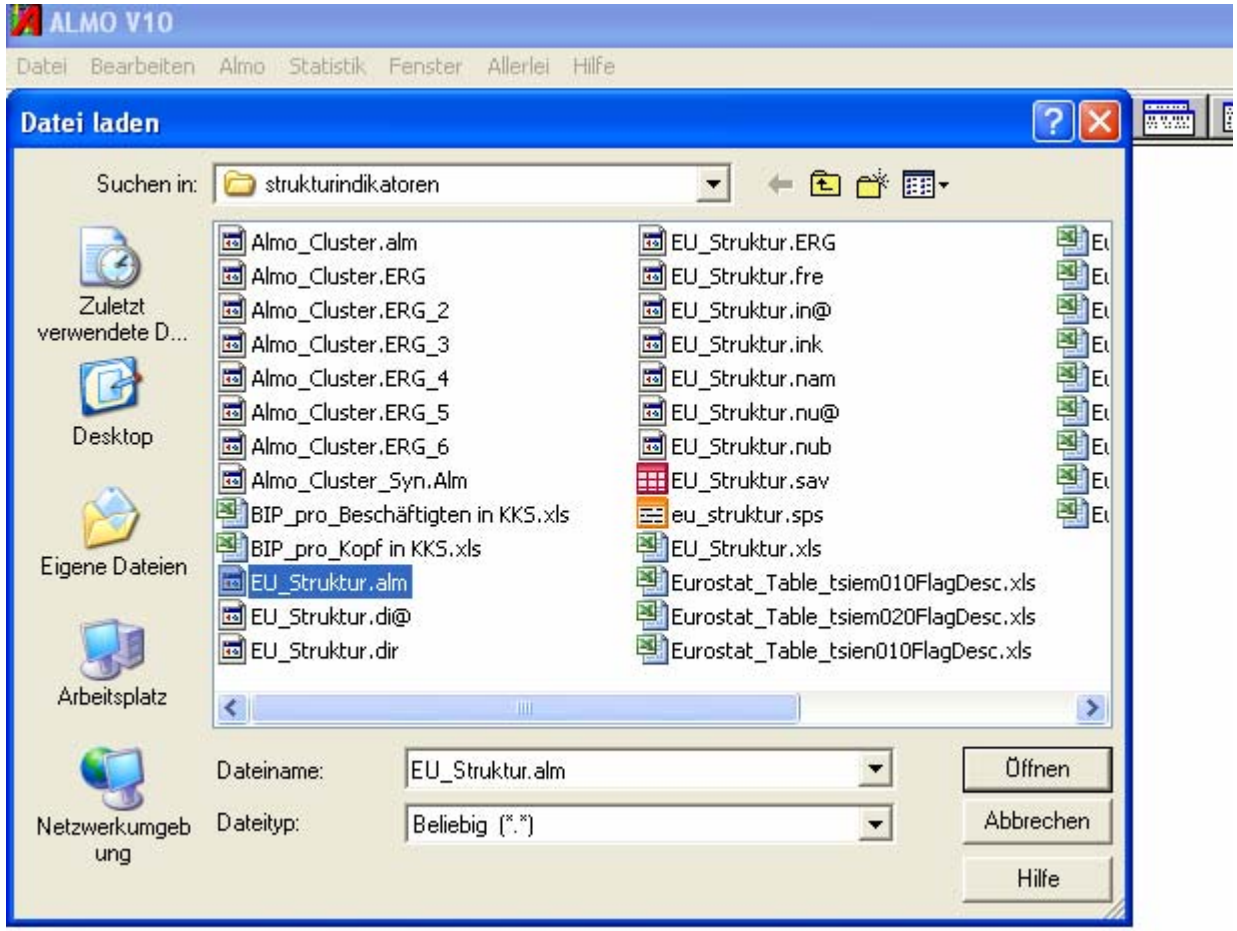
zusätzlich unvollständige (geometrische) Methoden (Korrespondenzanalyse, MDS, Faktorenanalyse)

Vorgehen bei gelegentlicher Nutzung:

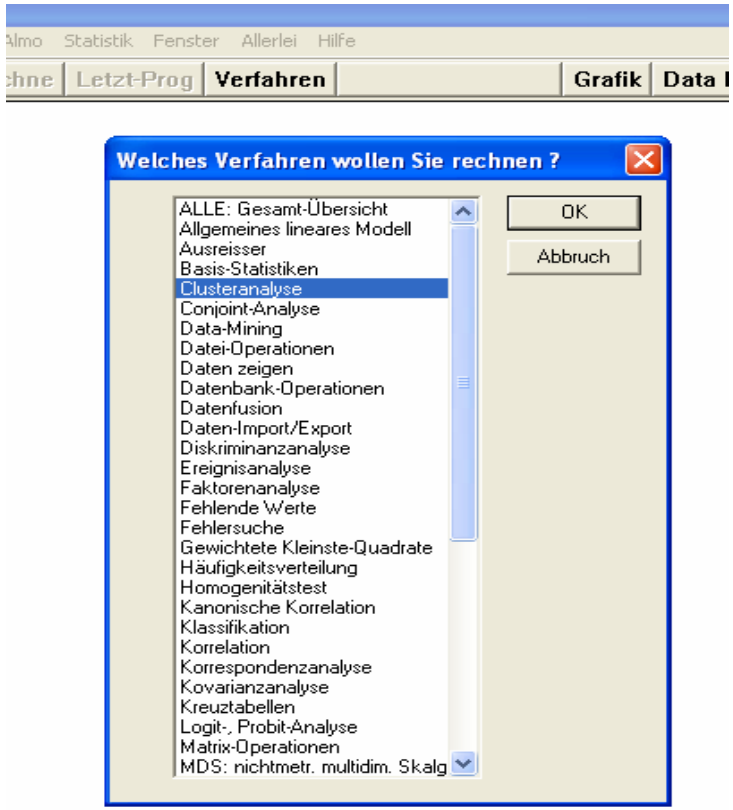
- vorhandene Daten im Standardstatistikprogramm aufbereiten und als SAV-Datei abspeichern
- Datentransferprogramm in ALMO ausführen (Optionen NUB-Datei, INK-Datei usw. anklicken)
- Erzeugtes ALMO-Programm mit Erweiterung „*.ALM“ rechnen → ALMO-Systemdatei und Namensdatei liegt vor, mit der gerechnet werden kann. (Mitunter sind kleine Nachbearbeitungen erforderlich)


ALMO V10
Datei Bearbeiten Almo Statistik Fenster Allerlei Hilfe
Data Mining
Kurz-Programm laden
Masken-Programm laden
Beispiel-Programm laden
Liste aller Almo-Programme
Rechne ALMO-Prog in aktivem Fenster F9
Maskenmodus (Schreibsperre): EIN F4
Maskenmodus (Schreibsperre): AUS F4
Auf Platte/Disk befindliches ALMO-Prog rechnen
Lade zuletzt gerechnetes Almo-Programm
Lade abgestürztes oder nicht abgerufenes Almo-Ergebnis
Dateneingabe
Daten-Import und -Export
Anleitung
Hilfe zu Menü-Titel Almo
Hilfe zu Daten-Import, -Export
Methode I : SPSS <--> Almo: Datei-Import und -Export
Methode II: SPSS oder Excel --> Almo


Importmanager SPSS - ALMO
Datei ?
SPSS Import nach ALMO
SPSS-Datenfile
D:\strukturindikatoren\EU_Struktur.sav
Variablen: 26
Fälle: 27
ALMO-Datenfile
D:\strukturindikatoren\EU_Struktur.fre
festlegen...
ALMO-Syntaxfile
D:\strukturindikatoren\EU_Struktur.alm
festlegen...
schließen
importieren
Einstellungen
Missing Values
Missing in Umcodierung konvertieren
Missing auf KW setzen
Missing Deklarationen ignorieren
Optionen für Almo-Syntax
Direkt-Daten schreiben
NUB-Datei schreiben
INK-Datei schreiben
Variablennamen
Keine Namen importieren
Variablennamen übernehmen
Variablenlabels als Namen übernehmen
Namen und Labels zusammenfügen
Ausprägungsnamen
Valuelabels übernehmen
Namen und Ausprägungen
Namen in die Almo-Syntax einfügen
Namen in eine Namensdatei (nam-Datei) schreiben


ALMO V10
Datei Bearbeiten Almo Statistik Fenster Allerlei Hilfe
Datei laden
Suchen in: strukturindikatoren
Almo_Cluster.alm
Almo_Cluster.ERG
Almo_Cluster.ERG_2
Almo_Cluster.ERG_3
Almo_Cluster.ERG_4
Almo_Cluster.ERG_5
Almo_Cluster.ERG_6
Almo_Cluster_Syn.Alm
BIP_pro_Beschäftigten in KKS.xls
BIP_pro_Kopf in KKS.xls
EU_Struktur.alm
EU_Struktur.di@
EU_Struktur.dir
EU_Struktur.ERG
EU_Struktur.fre
EU_Struktur.in@
EU_Struktur.ink
EU_Struktur.nam
EU_Struktur.nu@
EU_Struktur.nub
EU_Struktur.sav
eu_struktur.sps
EU_Struktur.xls
Eurostat_Table_tsiem010FlagDesc.xls
Eurostat_Table_tsiem020FlagDesc.xls
Eurostat_Table_tsien010FlagDesc.xls
Zuletzt verwendete D...
Desktop
Eigene Dateien
Arbeitsplatz
Netzwerkumgebung
Dateiname: EU_Struktur.alm
Dateityp: Beliebig (*.*)
Öffnen
Abbrechen
Hilfe

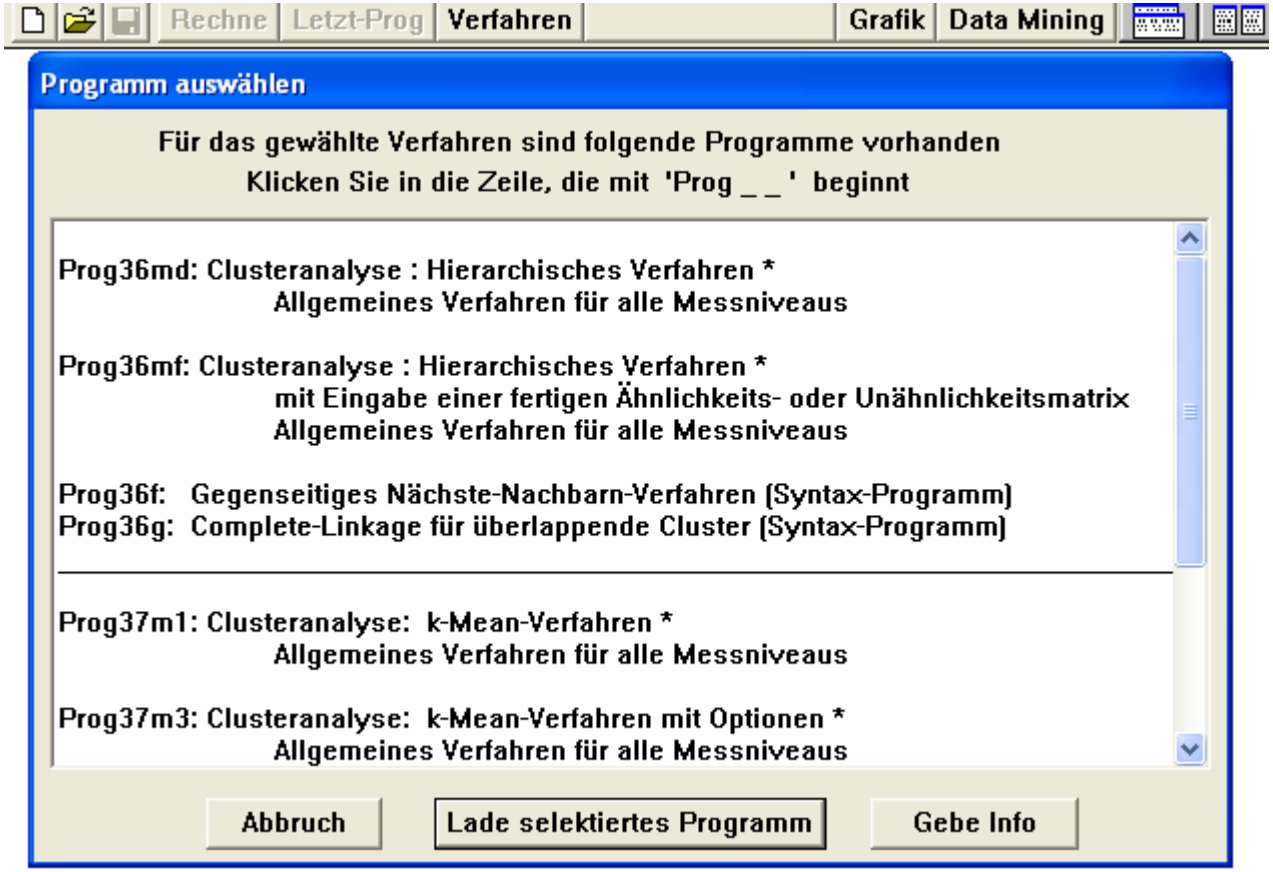
2. Durchführen einer Clusteranalyse



Menu-Knopf „Verfahren“ auswählen →

Verfahren Clusteranalyse auswählen

Innerhalb der Clusteranalyseverfahren entsprechendes Verfahren auswählen



3. Hierarchische Clusteranalyse

3.1. Beispiel

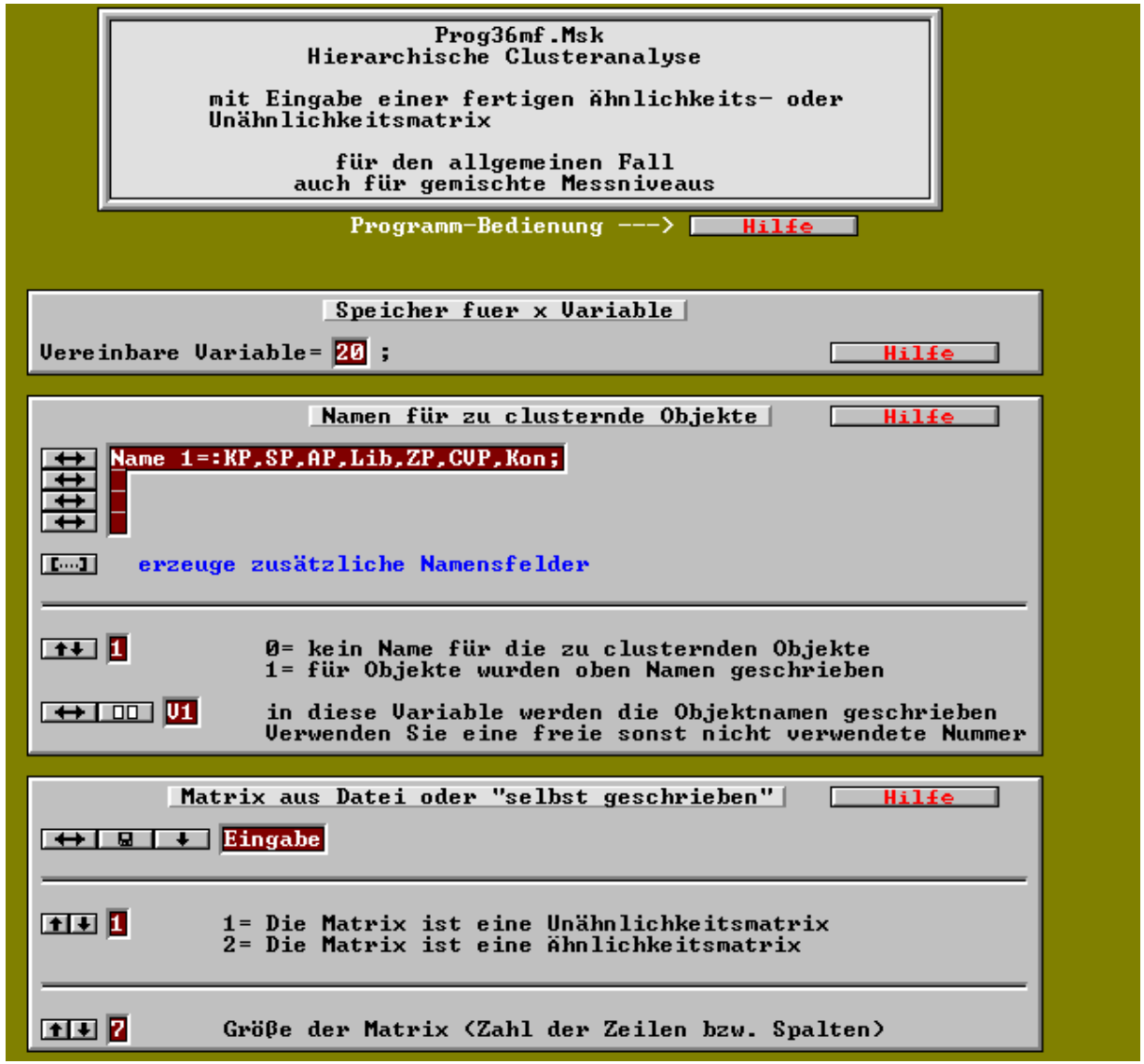
Ähnlichkeit von politischen Parteien.

	KP	SP	AP	Lib	ZP	CVP	Kon
Kommunistische Partei (KP)	0	8.7	25.3	33.7	37.9	49.3	50.2
Sozialistische Partei (SP)	8.7	0	14.8	19.0	33.2	50.5	40.0
Arbeiterpartei (AP)	25.3	14.8	0	10.0	17.8	21.3	24.3
Liberale (Lib)	33.7	19.0	10.0	0	10.5	18.9	12.9
Zentrumspartei (ZP)	37.9	33.2	17.8	10.5	0	7.6	8.1
Christliche Volkspartei (CVP)	49.3	50.5	21.3	18.9	7.6	0	7.3
Konservative (Kon)	50.2	40.0	24.3	12.9	8.1	7.3	0

entnommen aus: Lund (1974; zit. in Bacher 1996: 239)

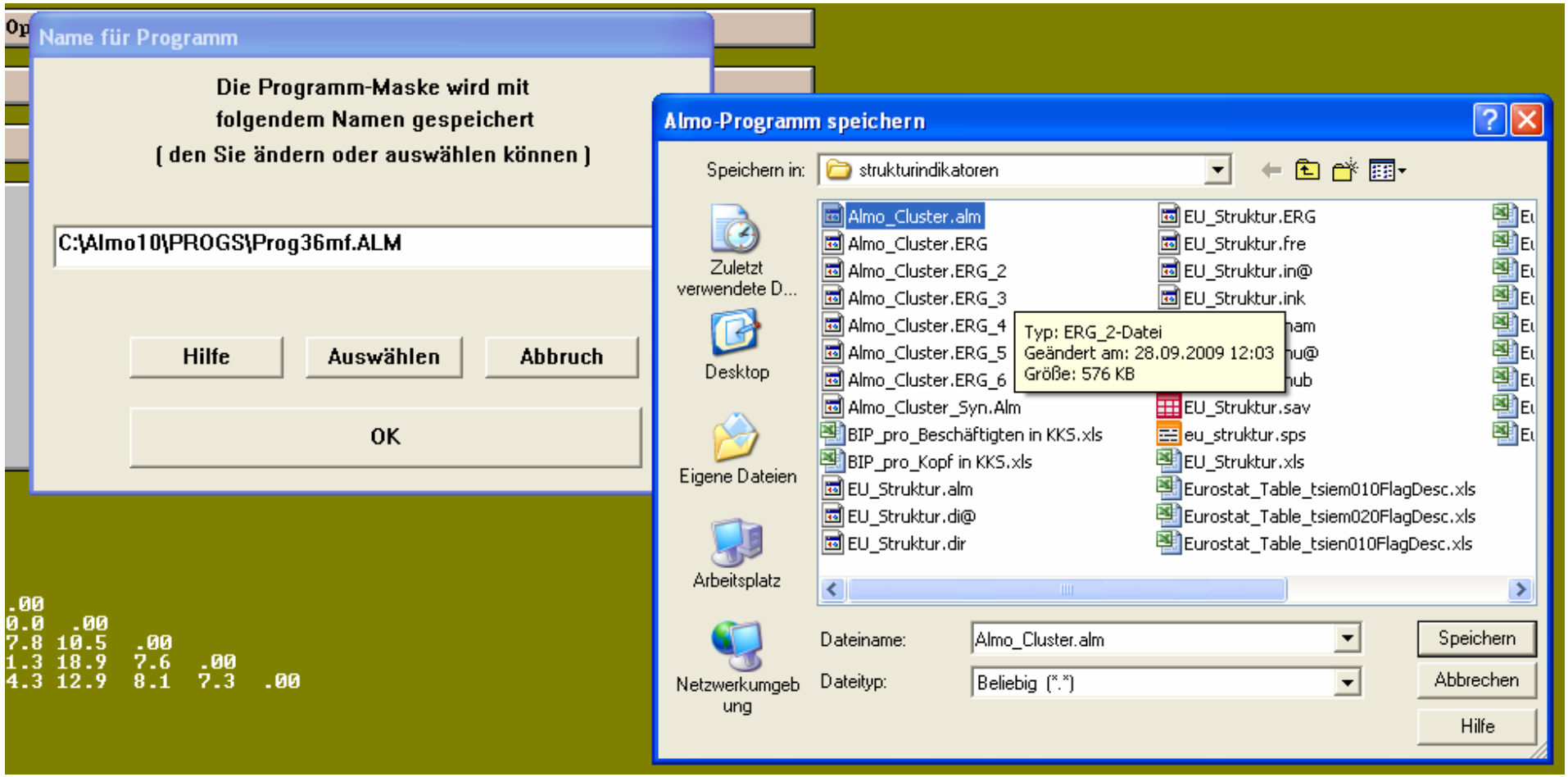
Fragestellungen:

- Besteht eine hierarchische Ähnlichkeitsbeziehung?
- Lassen sich die Parteien zu Clustern zusammenfassen?


Prog36mf.Msk
Hierarchische Clusteranalyse
mit Eingabe einer fertigen Ähnlichkeits- oder Unähnlichkeitsmatrix
für den allgemeinen Fall
auch für gemischte Messniveaus
Programm-Bedienung ---> Hilfe
Speicher fuer x Variable
Vereinbare Variable= 20 ;
Hilfe
Namen für zu clusternde Objekte
Hilfe
Name 1=:KP,SP,AP,Lib,ZP,CVP,Kon;
erzeuge zusätzliche Namensfelder
1
0= kein Name für die zu clusternden Objekte
1= für Objekte wurden oben Namen geschrieben
V1
in diese Variable werden die Objektnamen geschrieben
Verwenden Sie eine freie sonst nicht verwendete Nummer
Matrix aus Datei oder "selbst geschrieben"
Hilfe
Eingabe
1
1= Die Matrix ist eine Unähnlichkeitsmatrix
2= Die Matrix ist eine Ähnlichkeitsmatrix
7
Größe der Matrix (Zahl der Zeilen bzw. Spalten)


Klassifikations-Variable für die Clusterung
quantitative Variable für die Clusterung
V11:17
ordinale Variable für die Clusterung
nominale Variable für die Clusterung
Hilfe
Clusterzahl
Minimale Zahl von Clustern
Maximale Zahl von Clustern
X Loesche wieder diese Box
Verfahren
Voreinstellung: weighted average
ACHTUNG: ALLES GEHT NICHT !!!!!
BEACHTE: Im Prinzip ist es möglich, eine Analyse mit allen Verfahren und allen Distanzmaßen zu rechnen. Dabei kann aber ein so umfang-reicher Output erzeugt werden, dass er im Extremfall von ALMO nicht mehr geladen werden kann. Abhilfe: Verwendung nur einiger Verfahren und nur einiger Distanzmaße
complete_linkage
complete linkage
single linkage
average linkage
weighted average linkage
Median-Verfahren
Zentroid-Verfahren
Ward-Verfahren
within avarage linkage
gegenseitiges-nächstes-Nachbarn-Verfahren
complete linkage für überlappende Cluster


Option: Distanzmaß (Voreinstellung: city_block)
Option: Teststatistiken
Grafik-Optionen
Schreiben der Matrixwerte
Schreiben Sie hier dahinter das untere
Dreieck der Matrix inklusive Diagonale
BEACHTE:
Vor der Dreiecksmatrix stehen 3 Sterne
Hinter der Dreiecksmatrix stehen 2 Sterne
Schalten Sie dazu die Schreibsperre aus
[Schreibsperre] <--- EIN : rot
AUS : grau
*
*
*
.00
8.7 .00
25.3 14.8 .00
33.7 19.0 10.0 .00
37.9 33.2 17.8 10.5 .00
49.3 50.5 21.3 18.9 7.6 .00
50.2 40.0 24.3 12.9 8.1 7.3 .00
*
*


Name für Programm
Die Programm-Maske wird mit
folgendem Namen gespeichert
(den Sie ändern oder auswählen können)
C:\Almo10\PROGS\Prog36mf.ALM
Hilfe
Auswählen
Abbruch
OK
.00
0.0 .00
7.8 10.5 .00
1.3 18.9 7.6 .00
4.3 12.9 8.1 7.3 .00
Almo-Programm speichern
Speichern in:
strukturindikatoren
Zuletzt verwendete D...
Desktop
Eigene Dateien
Arbeitsplatz
Netzwerkumgebung
Almo_Cluster.alm
Almo_Cluster.ERG
Almo_Cluster.ERG_2
Almo_Cluster.ERG_3
Almo_Cluster.ERG_4
Almo_Cluster.ERG_5
Almo_Cluster.ERG_6
Almo_Cluster_Syn.Alm
BIP_pro_Beschäftigten in KKS.xls
BIP_pro_Kopf in KKS.xls
EU_Struktur.alm
EU_Struktur.di@
EU_Struktur.dir
EU_Struktur.ERG
EU_Struktur.fre
EU_Struktur.in@
EU_Struktur.ink
EU_Struktur.sav
eu_struktur.sps
EU_Struktur.xls
Eurostat_Table_tsiem010FlagDesc.xls
Eurostat_Table_tsiem020FlagDesc.xls
Eurostat_Table_tsien010FlagDesc.xls
Typ: ERG_2-Datei
Geändert am: 28.09.2009 12:03
Größe: 576 KB
Dateiname:
Almo_Cluster.alm
Dateityp:
Beliebig (*.*)
Speichern
Abbrechen
Hilfe

3.2. Modellspezifikationen

- Auswahl der Variablen = entfällt, da Unähnlichkeitsmatrix eingelesen wird
- Auswahl der Objekte = alle Objekte sollen einbezogen werden
- Auswahl eines geeigneten Unähnlichkeitsmaßes = entfällt, da Unähnlichkeitsmatrix eingelesen wird
- Auswahl eines geeigneten Verfahrens = abhängig von den Anforderungen, die an die Klassifikation und die Hierarchie gestellt werden

Anforderungen an die Hierarchie und Klassifikation

- invariant gegenüber monotonen Transformationen → Complete oder Single-Linkage (Complete-Linkage = sehr strenge Homogenitätsvorstellungen / Single-Linkage = sehr schwache Homogenitätsvorstellungen → Verkettungen)
- Invarianz kein wichtiges Merkmal, aber kein bestimmtes Distanzmaß → Mittelwertverfahren (Weighted-Average-Linkage; noch möglich: Average-Linkage, Within-Average-Linkage)
- Invarianz keine Rolle, Datenmatrix liegt vor und quadrierte euklidische Distanzen sinnvolles Distanzmaß → Ward-Verfahren, noch möglich: Zentroid- und Median-Verfahren

Vorgehen: Basisverfahren auswählen, anschließend Stabilität durch Verwendung weiterer Modelle untersuchen

3.3. Erster Durchlauf

dient der Bestimmung der Clusterzahl und der Hierarchie

Algorithmus der hierarchischen Verfahren:

Schritt 1: Jedes Klassifikationsobjekt bildet zu Beginn ein selbständiges Cluster. Setze daher die Clusterzahl K gleich der Klassifikationsobjektzahl n.

Schritt 2: Suche das Clusterpaar ({p}, {q}) mit der größten Ähnlichkeit bzw. der geringsten Unähnlichkeit, verschmelze das Clusterpaar zu einem neuen Cluster {p,q} und reduziere Clusterzahl K um 1 (K=K-1).

Schritt 3: Prüfe, ob K gleich 1 ist. Ist dies der Fall, beende den Algorithmus, da alle Klassifikationsobjekte einem einzigen Cluster angehören. Bei nein fahre mit Schritt 4 fort.

Schritt 4: Berechne die Ähnlichkeit bzw. Unähnlichkeit des neu gebildeten Clusters{p,q} zu den verbleibenden Clustern i.

Schritt 5: Gehe zu Schritt 2.

KP						
SP	8.7					
AP	25.3	14.8				
Lib	33.7	19.0	10.0			
ZP	37.9	33.2	17.8	10.5		
CVP	49.3	50.5	21.3	18.9	7.7	
Kon	50.2	40.0	24.3	12.9	8.1	(7.3)

Unähnlichkeitsmatrix (nur unteres Dreieck)

Schritt 1: Jedes Objekt bildet ein Cluster.
Es liegen somit folgende Cluster vor:
{KP}, {SP}, {AP}, {Lib}, {ZP}, {CVP}, {Kon}

1. Verschmelzungsschritt:

Clusterpaar mit geringster Unähnlichkeit:
{CVP}, {Kon} wird verschmolzen.
Folgende Cluster liegen vor:
{KP}, {SP}, {AP}, {Lib}, {ZP}, {CVP, Kon}
Neuberechnung der Unähnlichkeiten:

verbl. Cluster	{CVP}	{Kon}	neue Unähnl. (=Max)
KP	49.3	50.2	50.2
SP	50.5	40.0	50.5
AP	21.3	24.3	24.3
Lib	18.9	12.9	18.9
ZP	7.7	8.1	8.1

KP					
SP	8.7				
AP	25.3	14.8			
Lib	33.7	19.0	10.0		
ZP	37.9	33.2	17.8	10.5	
CVP	50.2	50.5	24.3	18.9	(8.1)

neue Unähnlichkeitsmatrix nach 1. Verschmelzungsschritt

2. Verschmelzungsschritt:

Clusterpaar mit geringster Unähnlichkeit:
{ZP}, {CVP} wird verschmolzen.
Folgende Cluster liegen vor:
{KP}, {SP}, {AP}, {Lib}, {ZP, CVP, Kon}
Neuberechnung der Unähnlichkeiten

KP				
SP	(8.7)			
AP	25.3	14.8		
Lib	33.7	19.0	10.0	
ZP	50.2	50.5	24.3	18.9

neue Unähnlichkeitsmatrix nach 2. Verschmelzungsschritt

3. Verschmelzungsschritt:

Clusterpaar mit geringster Unähnlichkeit:
{KP}, {SP} wird verschmolzen.
Folgende Cluster liegen vor:
{KP, SP}, {AP}, {Lib}, {ZP, CVP, Kon}
Neuberechnung der Unähnlichkeiten

KP			
AP	25.3		
Lib	33.7	(10.0)	
ZP	50.5	24.3	18.9

neue Unähnlichkeitsmatrix nach 3. Verschmelzungsschritt

4. Verschmelzungsschritt:

Clusterpaar mit geringster Unähnlichkeit:
{AP}, {Lib} wird verschmolzen.
Folgende Cluster liegen vor:
{KP, SP}, {AP, Lib}, {ZP, CVP, Kon}
Neuberechnung der Unähnlichkeiten

KP		
AP	33.7	
ZP	50.5	(24.3)

neue Unähnlichkeitsmatrix nach 4. Verschmelzungsschritt

5. Verschmelzungsschritt:

Clusterpaar mit geringster Unähnlichkeit:
{AP}, {ZP}
Folgende Cluster liegen vor:
{KP, SP}, {AP, Lib, ZP, CVP, Kon}
Neuberechnung der Unähnlichkeiten

KP	
ZP	(50.5)

neue Unähnlichkeitsmatrix nach 5. Verschmelzungsschritt

6. Verschmelzungsschritt:

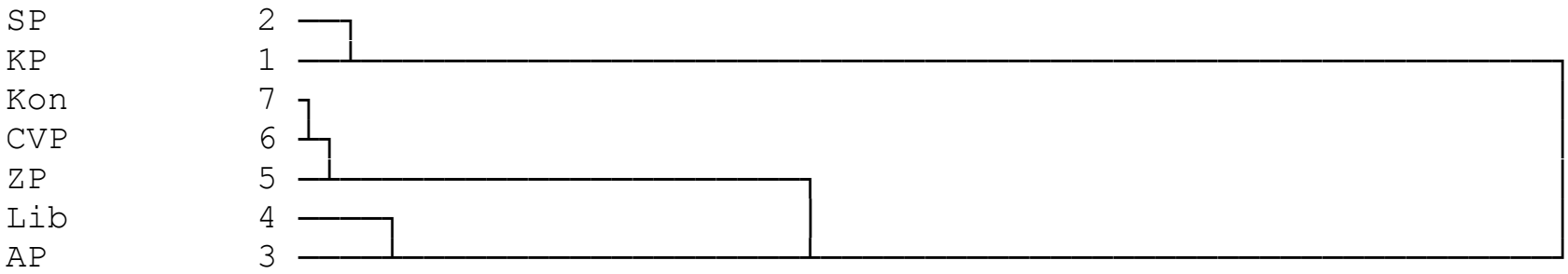
Clusterpaar mit geringster Unähnlichkeit:
{KP}, {AP}
Folgendes Cluster liegt vor:
{KP, SP, AP, Lib, ZP, CVP, Kon}
Neuberechnung der Unähnlichkeiten

Ende, da nur mehr ein Cluster vorliegt.

Ausgabe des Verschmelzungsschemas (siehe Tabelle 3.4.2)

```
*************************************************************************
Clusterverknuepfung   Clusterzahl       Distanzniveau       Zuwachs     Bindungen
      6       7              6              7.300            0.000           0
      5       6              5              8.100            0.800           0
      1       2              4              8.700            0.600           0
      3       4              3             10.000            1.300           0
      3       5              2             24.300           14.300           0
      1       3              1             50.500           26.200           0

Beachte: Ausgegeben werden alle Schritte (OPTION37=0;).
         Mit (OPTION37=x;) kann die Zahl reduziert werden.

Beachte: Bindungen koennen abhaengig von ihrer Behandlung zu unterschiedlichen
         Ergebnissen fuehren.

Behandlung der Bindungen (OPTION39=2;)    letztes Objektpaar wird ausgewaehlt.
Zahl der Bedingungen = 0

Minimum=   7.3, Maximum=  50.5
```

Dendrogramm:



3.4. Hierarchische Ähnlichkeitsstruktur

Aus dem Dendrogramm bzw. dem Verschmelzungsschema kann eine theoretische (Un-)Ähnlichkeitsmatrix erzeugt werden. Durch einen Vergleich mit der empirischen (Un-)Ähnlichkeitsmatrix kann geprüft werden, wie gut die hierarchische Struktur den Daten angepasst ist.

Maßzahlen zur Beurteilung:

- kophenetische Korrelationen (Gamma, r)
- Stresskoeffizienten
- weitere Tests

-> neue Datenmatrix

emp. Unähnl.	(2,1)	(3,1)	(3,2)	(4,1)	(4,2)	usw.
theoret. Unähnl.	(2,1)	(3,1)	(3,2)	(4,1)	(4,2)	usw.

reproduzierte Unaehnlichkeitsmatrix (Distanzmatrix)

		KP V1-1	SP V1-2	AP V1-3	Lib V1-4
KP	V1-1	0	8.7000	50.5000	50.5000
SP	V1-2	8.7000	0	50.5000	50.5000
AP	V1-3	50.5000	50.5000	0	10.0000
Lib	V1-4	50.5000	50.5000	10.0000	0
ZP	V1-5	50.5000	50.5000	24.3000	24.3000
CVP	V1-6	50.5000	50.5000	24.3000	24.3000
Kon	V1-7	50.5000	50.5000	24.3000	24.3000

		ZP V1-5	CVP V1-6	Kon V1-7
KP	V1-1	50.5000	50.5000	50.5000
SP	V1-2	50.5000	50.5000	50.5000
AP	V1-3	24.3000	24.3000	24.3000
Lib	V1-4	24.3000	24.3000	24.3000
ZP	V1-5	0	8.1000	8.1000
CVP	V1-6	8.1000	0	7.3000
Kon	V1-7	8.1000	7.3000	0

```
************************************************************************
kophenetischer Korrelationskoeffizient =    0.795
------------------------------------------------------------------------
Zahl der vorgeg.  Simulationen = 100
Zahl der erfolgr. Simulationen = 100
Erwartungswert         =    -0.04
Standardabweichung     =     0.20
Teststatistik          =     4.13
Schwellwert fuer       =    90.00 Prozent
                       =     0.23
------------------------------------------------------------------------
************************************************************************
Gamma  =    0.919
------------------------------------------------------------------------
Zahl der vorgeg.  Simulationen = 100
Zahl der erfolgr. Simulationen = 100
Erwartungswert         =    -0.04
Standardabweichung     =     0.22
Teststatistik          =     4.31
Schwellwert fuer       =    90.00 Prozent
                       =     0.27
------------------------------------------------------------------------
```

Nullmodell für Simulation: emp.Unähnl. = $NV(\bar{d}, s_d)$

3.5. Zahl der Cluster

gute Erfahrungen mit Zuwachs im Agglomerationsschema

Clusterzahl = Zeile vor Zuwachs → 3 Cluster im Beispiel

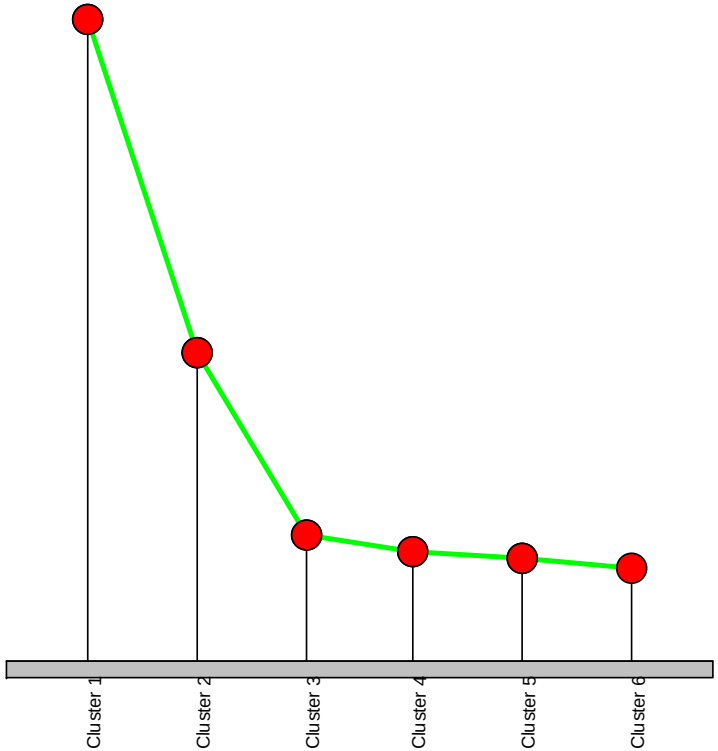
Clusterverknuepfung		Clusterzahl	Distanzniveau	Zuwachs	Bindungen
6	7	6	7.300	0.000	0
5	6	5	8.100	0.800	0
1	2	4	8.700	0.600	0
3	4	3	10.000	1.300	0
3	5	2	24.300	***14.300***	0
1	3	1	50.500	26.200	0

Weitere Maßzahlen

- inverser Screetest
- Teststatistiken von Mojena
- Zufallstestung des Verschmelzungsschemas

Inverser Screetest

Kriterium



Teststatistiken von Mojena

Teststatstik zur Bestimmung der Clusterzahl
nach MOJENA (Regel 1) - analog zu CLUSTAN

$$M_{1k} = \frac{v_k - \bar{v}}{s_v}$$

Nullmodell: $v_k \approx NV(\bar{v}, s_v)$

```
Mittelwert          =   18.150
Standardabweichung  =   17.081
Freiheitsgrade      =        5

Clusterzahl    Teststatistik    Freiheitsgrade    Signifikanz
     6            -0.588               5                -
     5            -0.553               5                -
     4            -0.477               5             67.263
     3             0.360               5             63.561
     2             1.894               5             94.227
*************************************************************************
```

➔ kein signifikanter Zuwachs erkennbar, Schwellenwert 99,7%

➔ alternative Regel: Auswahl der Lösung mit der maximalen Teststatistik

Teststatstik zur Bestimmung der Clusterzahl
nach MOJENA (Regel 1) - modifiziert

$$M_{1k} = \frac{v_k - \bar{v}_{k-1}}{s_{v,k-1}}$$ Nullmodell: $v_k \approx NV(\bar{v}_{k-1}, s_{v,k-1})$

Clusterzahl	Teststatistik	Signifikanz
6	-	-
5	1.768	-
4	2.800	88.649
3	13.858	99.821
2	5.450	99.470

→ deutlicher Zuwachs bei 3 Clustern erkennbar, Signifikanzschwelle von 99.7% wird überschritten

→ bei 3 Clustern auch maximaler Wert

Teststatstik zur Bestimmung der Clusterzahl
nach MOJENA (Regel 2) - modifiziert

$$M_{1k} = \frac{v_k - \hat{v}_{k-1}}{s_{\hat{v},k-1}}$$

Nullmodell: $v_k = a_{k-1} + b_{k-1} \cdot k + e_k$

Clusterzahl	Teststatistik	Signifikanz
6	-	-
5	-0.354	-
4	0.807	71.707
3	11.947	99.755
2	3.938	98.632

→ deutlicher Zuwachs bei 3 Clustern erkennbar, Signifikanzschwelle von 99.7% wird überschritten

→ bei 3 Clustern auch maximaler Wert

Zufallstestung des Verschmelzungsschemas (OPTION17=100;)

nur möglich, wenn eine Datenmatrix vorliegt → Übungsaufgabe am Nachmittag

3.6. Interpretation der Cluster

im vorliegenden Fall auf der Basis der Zuordnung der Objekte zu den Clustern, liegt eine Datenmatrix vor, können zusätzlich Verteilungskennwerte der Variablen berechnet werden.

```
**************************************************************************
Clusterzugehoerigkeit der Elemente bei 3 Clustern
--------------------------------------------------------------------------
Cluster  1 (n=   2)      1 KP
                         2 SP
Cluster  2 (n=   2)      3 AP
                         4 Lib
Cluster  3 (n=   3)      5 ZP
                         6 CVP
                         7 Kon
**************************************************************************
```

inhaltlich sinnvolle Interpretation möglich: sozialistisches Cluster, liberales Cluster und konservatives Cluster

Verteilungskennwerte bei Datenmatrix zur Beschreibung der Cluster

wichtige Maßzahl z-Werte: $z_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{s_{\bar{x}.ik}}$

```
Masszahlen fuer Klassifikationsvariablen im Clustern 1:
Cluster  1 (n=   4)      1 AUT
                         2 BEL
                         5 GER
                        15 SUI
gewichtete Fallzahl =4
```

Variable	n=	Min.	Max.	MA	SA	z-Wert
10 ZMATH	***4***	***504.00***	***530.00***	***514.75***	***10.85***	***2.71***
11 ZREAD	***4***	***490.00***	***501.00***	***496.25***	***4.21***	***2.81***
12 ZTOP	***4***	***0.17***	***0.23***	***0.20***	***0.03***	***3.97***
13 ZRISK	4	0.19	0.26	0.23	0.03	-1.58
14 BERUF_MA	4	0.29	0.37	0.33	0.03	-0.07
15 BERUF_RE	4	0.31	0.33	0.32	0.01	-0.28
16 BERUF_TO	4	0.19	0.27	0.24	0.03	0.41
17 BERUF_RI	4	-0.26	-0.23	-0.25	0.01	-0.13

18	BUB_MATH	4	0.06	0.15	0.11	0.04	1.45
19	BUB_READ	4	-0.20	-0.16	-0.18	0.02	1.24
20	***BUB_TOP***	***4***	***0.05***	***0.09***	***0.07***	***0.01***	***4.79***
21	BUB_RISK	4	0.04	0.09	0.06	0.02	0.03
22	***MIGRA_MA***	***4***	***-0.31***	***-0.21***	***-0.26***	***0.04***	***-5.53***
23	***MIGRA_RE***	***4***	***-0.28***	***-0.14***	***-0.23***	***0.05***	***-2.93***
24	***MIGRA_TO***	***4***	***-0.16***	***-0.08***	***-0.14***	***0.03***	***-4.06***
25	***MIGRA_RI***	***4***	***0.18***	***0.28***	***0.23***	***0.04***	***4.39***

Merkmale des Clusters:

- überdurchschnittliche Testleistungen
- überdurchschnittlicher Anteil an SpitzenschülerInnen
- überdurchschnittlicher Bubenanteil bei SpitzenschülerInnen
- unterdurchschnittliche Integration von Kindern mit Migrationshintergrund

3.7. Validitätsprüfung

- formale Validitätsprüfung (Wie gut sind die Modellvorstellungen erfüllt?)
- kriterienbezogene Validitätsprüfung (Wie gut sind Hypothesen über die Cluster erfüllt?) → setzt weitere Daten voraus, in unserem Beispiel nicht der Fall
- Expertenvalidierung (Wie gut stimmen die Ergebnisse mit ExpertInnenmeinungen überein?) → Vorlage der Ergebnisse ExpertInnen oder Prüfung mit der Literatur

formale Gültigkeitsmaße

zahlreiche Maßzahlen, oft fehlen aber Schwellenwerte, Almo berechnet:

- Homogenitätsindex: $h = \frac{g - E(g)}{s_g}$ mit $h = u_{between} - u_{within}$
- kophenetische Korrelation: analog zu oben, nur mit anderer theoretischer Unähnlichkeitsmatrix: $\delta_{ij} = \begin{cases} 0 \; wenn \; i \; und \; j \; demselben \; Cluster \; angehören \\ 1 sonst \end{cases}$
- weitere Indizes, wie z.B. W/B-Index (Within-Between-Index): $W / B = \frac{\bar{d}_{within}}{\bar{d}_{between}}$

```
************************************************************************
Clusterzugehoerigkeit der Elemente bei 3 Clustern
------------------------------------------------------------------------
Cluster  1 (n=   2)     1 KP
                        2 SP
Cluster  2 (n=   2)     3 AP
                        4 Lib
Cluster  3 (n=   3)     5 ZP
                        6 CVP
                        7 Kon
************************************************************************

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 Clusterkennwerte fuer die 3-Loesung
------------------------------------------------------------------------
```

Unaehnlichkeiten in den Clustern:

Cluster	Paare	Minimum	Maximum	arithm.M.	Standardabw.
1	1	8.70	8.70	8.70	0.00
2	1	10.00	10.00	10.00	0.00
3	3	7.30	8.10	7.67	0.33

Unaehnlichkeiten zwischen den Clustern:

Cluster	Cluster	Paare	Minimum	Maximum	arithm.M.	Standardabw.
1	2	4	14.80	33.70	23.20	7.12
1	3	6	33.20	50.50	43.52	6.80
2	3	6	10.50	24.30	17.62	4.70

W/B-Kriterium = 0.271
C-Index = 0.306

G1-Homogenitaetsmass = 19.322
Erwartungswert = 0.000
Varianz = 65.973
z-Wert = 2.379
Signifikanz = 98.247
Fehler (Chebychev) = 17.671 (=1/z**2)

```
Gamma  =    1.000

-------------------------------------------------------------------------
Zahl der vorgeg.  Simulationen = 100
Zahl der erfolgr. Simulationen = 100
Erwartungswert         =    -0.01
Standardabweichung     =     0.28
Teststatistik          =     3.64
Schwellwert fuer       =    90.00 Prozent
                       =     0.45
-------------------------------------------------------------------------

kophenetischer Korrelationskoeffizient =    0.601

-------------------------------------------------------------------------
Zahl der vorgeg.  Simulationen = 100
Zahl der erfolgr. Simulationen = 100
Erwartungswert         =    -0.01
Standardabweichung     =     0.21
Teststatistik          =     2.90
Schwellwert fuer       =    90.00 Prozent
                       =     0.37
-------------------------------------------------------------------------
```

3.8. Stabilitätsprüfung

Es werden mehrere CA-Methoden durchgerechnet. (Möglich auch: geringe Änderungen in der Unähnlichkeitsmatrix)

Übereinstimmung mehrer Lösung kann mittels Rand-Index gemessen werden:

$RAND = \frac{1}{(n-1)n/2} \sum_{g} \sum_{g^*>g} a_{gg^*}$, a_{gg^*} ist gleich 1, wenn g und g* in beiden Lösungen demselben Cluster angehören oder in beiden Lösungen unterschiedlichen Clustern angehören.

Aggregierte Randindizes fuer (Un)Aehnlichkeitsmasse:

(Un)Aehnlichkeitsmass	Rand-Index
	1.000

Aggregierte Randindizes fuer Modelle:

Modell	Rand-Index
Complete-Linkage	1.000
Single-Linkage	1.000
Average-Linkage	1.000
Weighted-Average	1.000
Within-Average-Linkage	1.000

Aggregierte Randindizes fuer Clusterzahl:

Clusterzahl	Rand-Index
Clusterzahl= 3	1.000

Berechnungsformel: r = (a + b) / (a + b + c + d) mit
a = Zahl der Objektpaare i und j in der Lösung 1 und in der Lösung 2 im selben Cluster
b = Zahl der Objektpaare i und j in der Lösung 1 und in der Lösung 2 nicht im selben Cluster
c = Zahl der Objektpaare i und j in der Lösung 1 nicht im selben Cluster und in der Lösung 2 im selben Cluster
d = Zahl der Objektpaare i und j in der Lösung 1 im selben Cluster und in der Lösung 2 nicht im selben Cluster

Literatur:

Bacher, J., 1996: Clusteranalyse. München.

Bacher, J., 1999: P36 P37 Clusteranalyse. Linz.