Workshop Clusteranalyse

Clusteranalyse – K-Means-Verfahren

Graz, 8. – 9. Oktober 2009

Johann Bacher

Johannes Kepler Universität Linz

Linz 2009

1. Fragestellung und Algorithmus

Bestimmung von Wertetypen (Bacher 1996)

n=221 Angaben von SchülerInnen liegen vor

=> möglich: Anwendung eines hierarchischen Verfahrens zur Konstruktion von Mittelwerten, insbes. Ward-Verfahren (bei n=221 treten aber Bindungen auf!) oder: K-Means-Verfahren, synonyme Bezeichnungen: Forgys Methode, Verfahren zur Verbesserung einer Ausgangspartition, Minimaldistanzverfahren für Varianzkriterium, Square-Error-Clustering usw.

Schritt 1: Berechnung oder Eingabe von Startwerten für die Clusterzentren.

Schritt 2: Zuordnung der Klassifikationsobjekte: Die Klassifikationsobjekte g werden jenem Clusterzentrum k zugeordnet, zu dem die quadrierte euklidische Distanz minimal ist. Formal ausgedrückt: $g \in k \Leftrightarrow k = \min_{k^*=1,2,\ldots,K} \left(d^2_{g,k^*}\right)$

Dies führt dazu, dass die Streuungsquadratsumme in den Clustern $SQ_{in}(K) = \sum_k \sum_{g \in k} d^2_{g,k} = \sum_g \min_{k^*=1,2,\ldots K} (d^2_{g,k^*})$

in jedem Iterationszyklus minimiert wird.

Schritt 3: Neuberechnung der Clusterzentren: Nach der Zuordnung aller Objekte zu den Clustern werden die Clusterzentren neu berechnet mit: $\bar{x}_{kj} = \sum_{g \in k} x_{gj} \Big/ n_{kj}$

mit nkj = Zahl der Objekte des Clusters k mit gültigen Angaben in der Variablen j. In die Summenbildung werden nur Ausprägungen mit gültigen Angaben einbezogen.

Schritt 4: Iteration: Es wird geprüft, ob sich im Schritt 2 die Zuordnung der Objekte geändert hat. Ist dies der Fall, werden die Schritte 2 und 3 erneut durchgeführt. Bei nein wird der Algorithmus beendet.

ALMO verwendet verallgemeinerte Distanzfunktion (->Mahalanobis-Distanz)

$$d_{g,g^*} = \left(\mathbf{x_g} - \mathbf{x_{g^*}}\right) \cdot \mathbf{W} \cdot \left(\mathbf{x_g} - \mathbf{x_{g^*}}\right)^T$$

Für $W = I$ ergibt sich das Varianzkriterum.

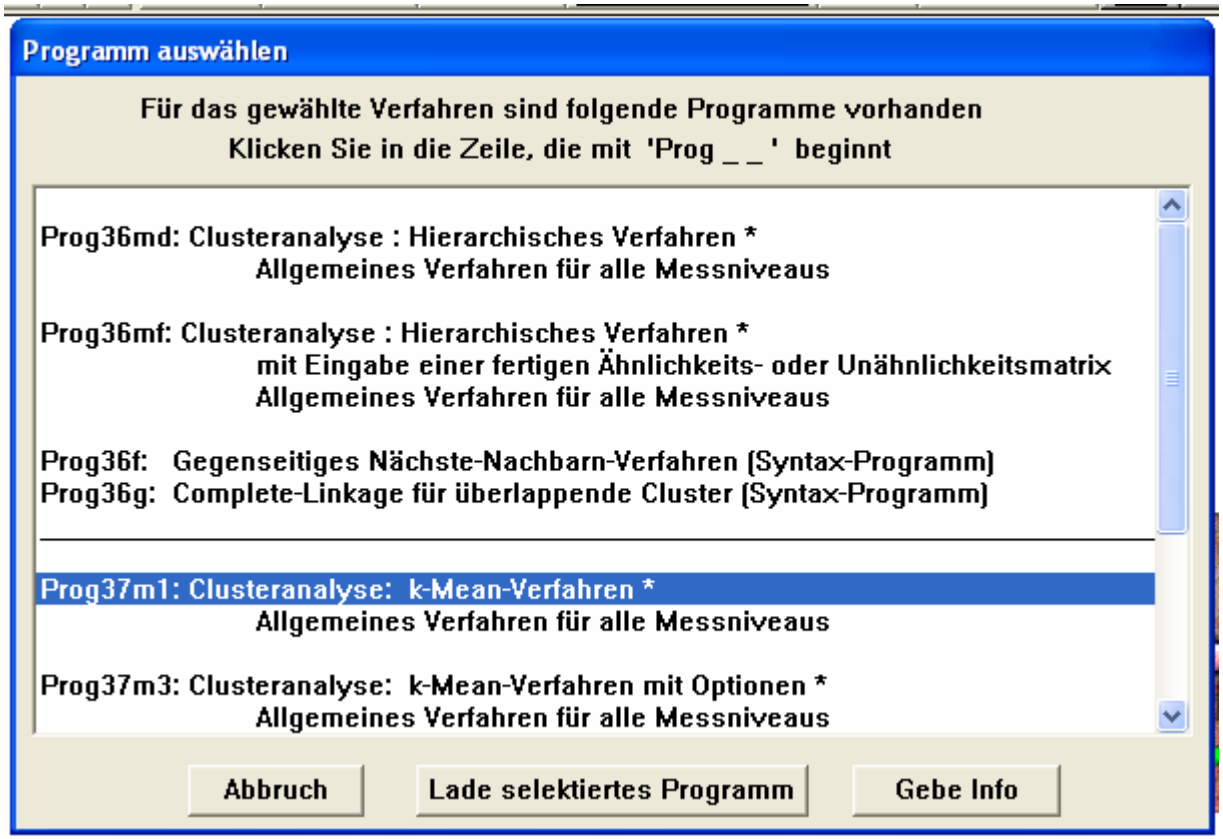
Für $W = \begin{bmatrix} 1/s_1^2 & 0 & 0 & & 0 \\ 0 & 1/s_2^2 & 0 & & 0 \\ 0 & 0 & 1/s_3^2 & & \\ & & & \ddots & \\ 0 & 0 & & & 1/s_4^2 \end{bmatrix}$

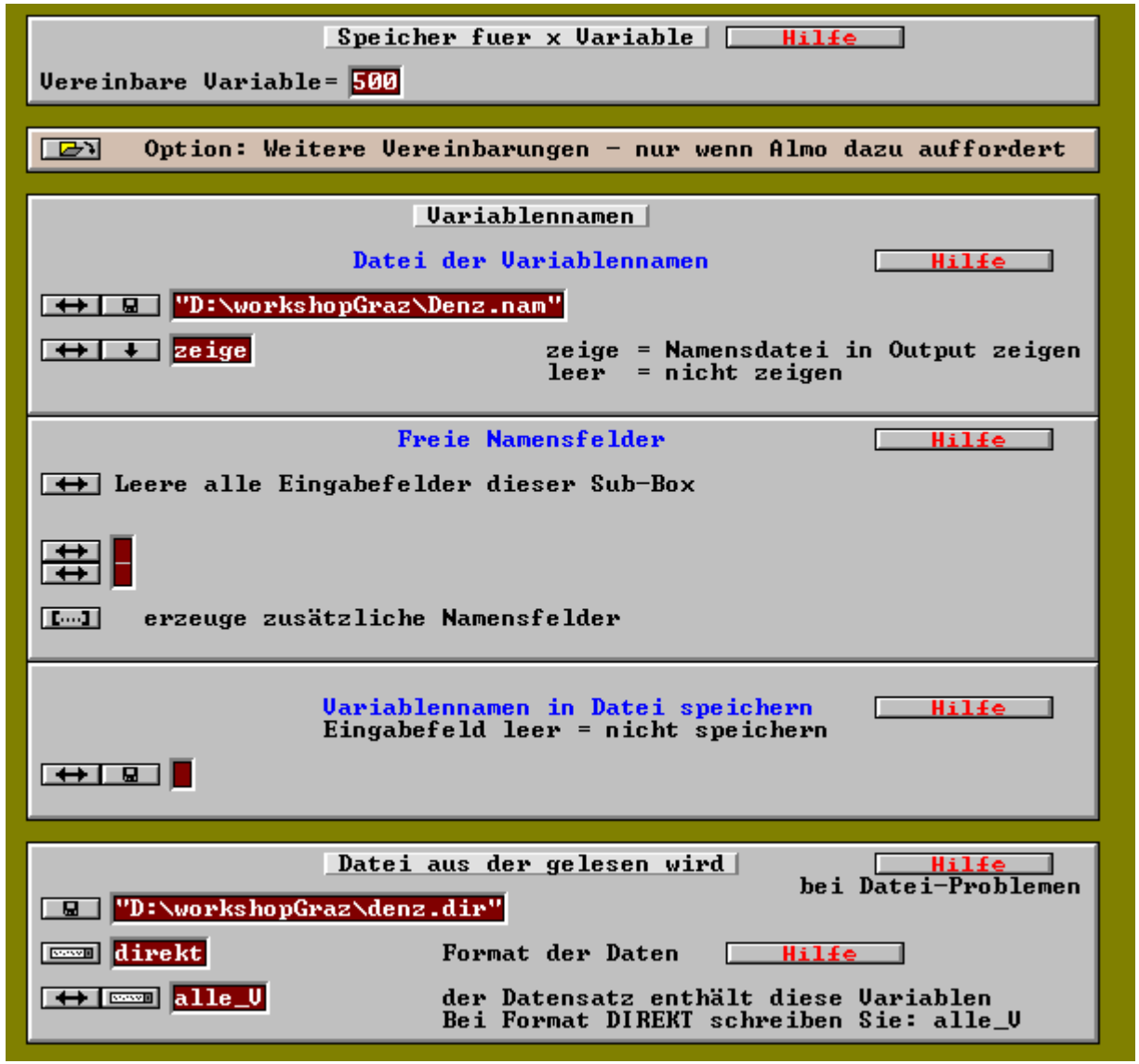
erfolgt eine Gewichtung mit 1/Varianz des Merkmals x. Dies entspricht einer Standardisierung

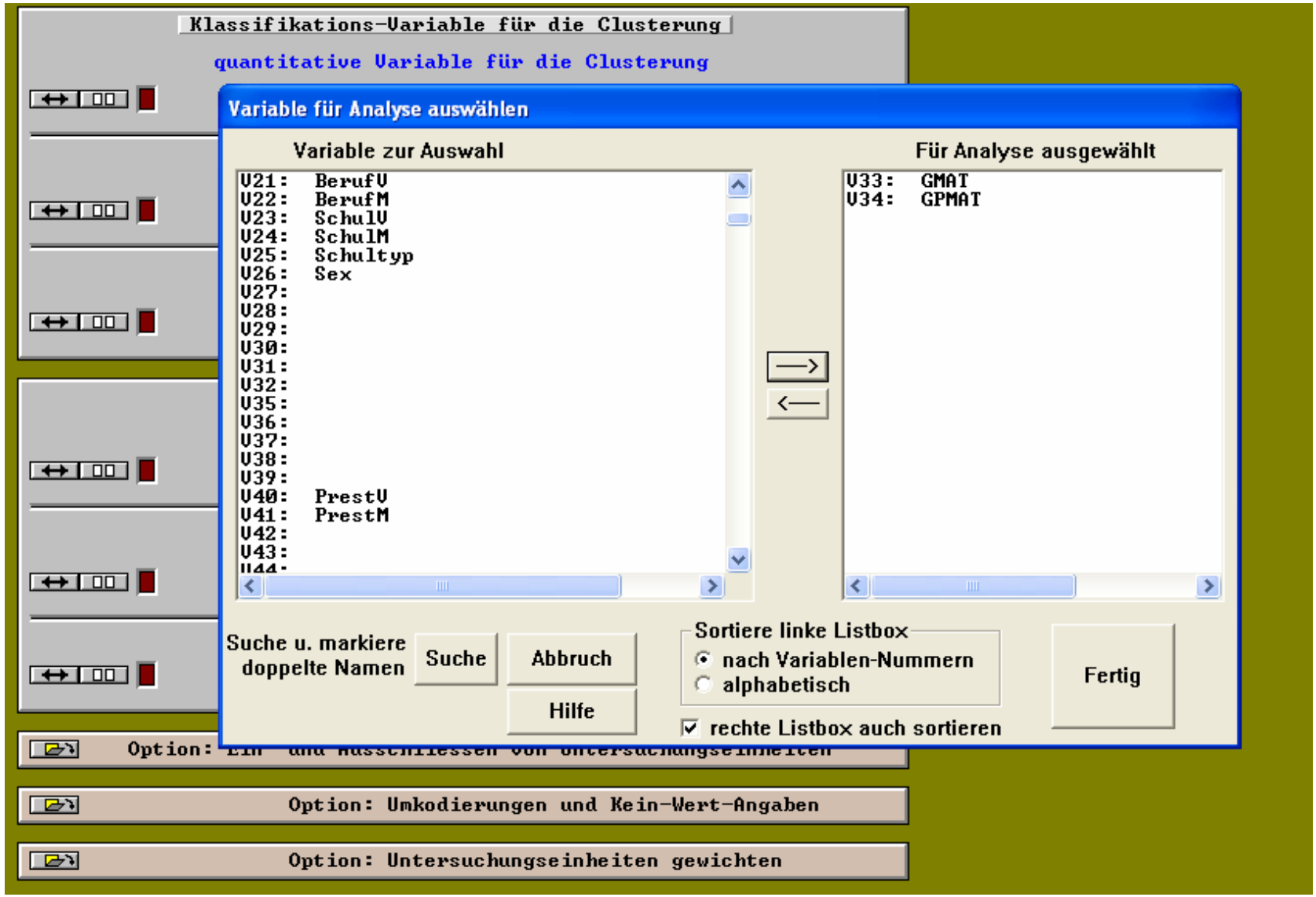
Konvergenz des Algorithmus nachgewiesen, es kann sich aber um ein lokales Minimum handeln → sinnvoll, mit unterschiedlichen Startwerten zu rechnen

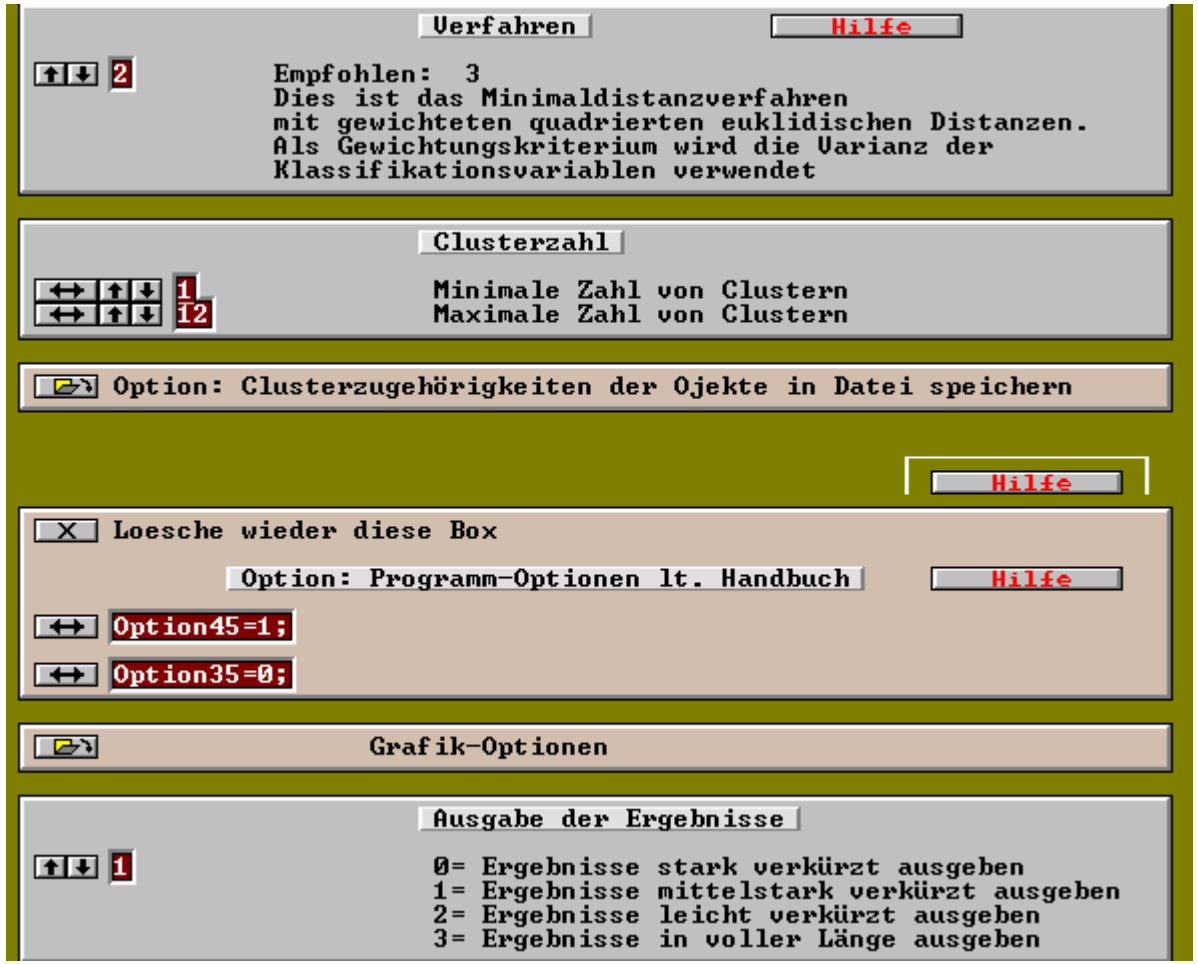
	Versuche je Clusterzahl				
Cluster	1	10	100	1000	2000
1	148.805	148.805	148.805	148.805	148.805
2	85.545	85.545	85.545	85.545	85.545
3	60.616	60.613	60.613	60.613	60.613
4	45.633	44.996	44.971	44.960	44.960
5	40.752	37.423	36.882	36.856	36.856
6	40.659	33.059	31.220	30.708	30.708
7	28.740	28.740	25.720	24.684	24.684
8	25.173	22.841	22.389	20.798	20.798
9	28.746	25.024	19.292	18.593	18.593
10	25.356	17.993	17.993	16.280	16.280
11	24.725	16.677	16.677	14.800	14.800
12	19.428	15.497	14.602	13.942	13.557

2. Durchführung




Speicher fuer x Variable
Hilfe
Vereinbare Variable= 500
Option: Weitere Vereinbarungen - nur wenn Almo dazu auffordert
Variablennamen
Datei der Variablennamen
Hilfe
"D:\workshopGraz\Denz.nam"
zeige
zeige = Namensdatei in Output zeigen
leer = nicht zeigen
Freie Namensfelder
Hilfe
Leere alle Eingabefelder dieser Sub-Box
erzeuge zusätzliche Namensfelder
Variablennamen in Datei speichern
Hilfe
Eingabefeld leer = nicht speichern
Datei aus der gelesen wird
Hilfe
bei Datei-Problemen
"D:\workshopGraz\denz.dir"
direkt
Format der Daten
Hilfe
alle_V
der Datensatz enthält diese Variablen
Bei Format DIREKT schreiben Sie: alle_V


Klassifikations-Variable für die Clusterung
quantitative Variable für die Clusterung
Variable für Analyse auswählen
Variable zur Auswahl
V21: BerufV
V22: BerufM
V23: SchulV
V24: SchulM
V25: Schultyp
V26: Sex
V27:
V28:
V29:
V30:
V31:
V32:
V35:
V36:
V37:
V38:
V39:
V40: PrestV
V41: PrestM
V42:
V43:
Für Analyse ausgewählt
V33: GMAT
V34: GPMAT
—>
<—
Suche u. markiere doppelte Namen
Suche
Abbruch
Hilfe
Sortiere linke Listbox
nach Variablen-Nummern
alphabetisch
rechte Listbox auch sortieren
Fertig
Option: Umkodierungen und Kein-Wert-Angaben
Option: Untersuchungseinheiten gewichten


Verfahren
Hilfe
2
Empfohlen: 3
Dies ist das Minimaldistanzverfahren
mit gewichteten quadrierten euklidischen Distanzen.
Als Gewichtungskriterium wird die Varianz der
Klassifikationsvariablen verwendet
Clusterzahl
1
12
Minimale Zahl von Clustern
Maximale Zahl von Clustern
Option: Clusterzugehörigkeiten der Ojekte in Datei speichern
Hilfe
Loesche wieder diese Box
Option: Programm-Optionen lt. Handbuch
Hilfe
Option45=1;
Option35=0;
Grafik-Optionen
Ausgabe der Ergebnisse
1
0= Ergebnisse stark verkürzt ausgeben
1= Ergebnisse mittelstark verkürzt ausgeben
2= Ergebnisse leicht verkürzt ausgeben
3= Ergebnisse in voller Länge ausgeben

3. Ergebnisse

3.1. Bestimmung der Clusterzahl

- Mindestwert für ETA^2: schwer definierbar, Berechnung:

$$ETA_K^2 = 1 - \frac{SQ_{in}(K)}{SQ_{ges}} = 1 - \frac{SQ_{in}(K)}{SQ_{in}(1)}$$

- PRE-Maßzahl: hat sich sehr gut bewährt, Formel: $PRE_K^2 = 1 - \frac{SQ_{in}(K)}{SQ_{in}(K-1)}$
- Scree-Test: Lösung mit dem Knickpunkt, oft schwer erkennbar
- F-Max-Statistik: Lösung mit dem maximalen F-Wert:

$$F - MAX_K = \frac{SQ_{zw}(K)/K-1}{SQ_{in}(K)/n-K} = \frac{\left(SQ_{ges} - SQ_{in}(K)\right)/K-1}{SQ_{in}(K)/n-K}$$

- Bealsche F-Werte: Lösung, die vorausgehende Lösungen mit geringerer Clusterzahl signifikant verbessert, während nachfolgende Lösungen mit höherer Clusterzahl keine signifikante Verbesserung erbringen

$$F-WERT_{K_2,K_1}(Beale)=\left(\frac{SQ_{in}(K_1)-SQ_{in}(K_2)}{SQ_{in}(K_2)}\right)\Bigg/\left(\frac{n-K_1}{n-K_2}\cdot\left(\frac{K_2}{K_1}\right)^{2/m}-1\right)$$

Ergebnisse der Iteration

Cluster-zahl	Itera-tionen	Kriterium	prozentuelle Verbesserung gegenueber H-0
1	3	148.805	0.000
2	5	85.545	42.512
3	4	60.613	59.267
4	6	44.960	69.786
5	12	36.856	75.232
6	7	30.708	79.364
7	19	24.684	83.412
8	11	20.798	86.024
9	10	18.593	87.505
10	16	16.280	89.059
11	13	14.800	90.054
12	18	13.942	90.631

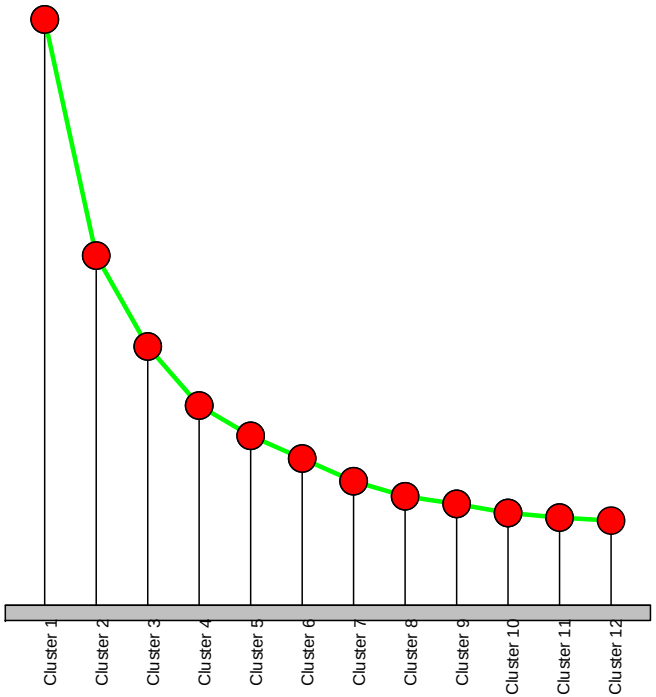
Bei Modellen 1 bis 6: Kriterium = Wert des Varianzkriteriums
Bei Modell 7: : Kriterium = Wert der Log-Likelihood-Funktion
Bei Modell 8: : Kriterium = Ueberlapp. + Nichtklass.

Ergebnisse fuer Reproduktionsindex

Cluster-	Kriterium	Versuche	Startzahl	Reproduktionen	Rep.Index.
1	148.805	1000	123123	1000.000	1.000
2	85.545	1000	123123	477.000	0.477
3	60.613	1000	16965405	274.000	0.274
4	44.960	1000	9323078	10.000	0.010
5	36.856	1000	27797712	9.000	0.009
6	30.708	1000	25819777	1.000	0.001
7	24.684	1000	12874299	1.000	0.001
8	20.798	1000	21162909	1.000	0.001
9	18.593	1000	14539713	1.000	0.001
10	16.280	1000	17512316	1.000	0.001
11	14.800	1000	29752758	1.000	0.001
12	13.942	1000	1450757	1.000	0.001

Reproduktionen= Anzahl der Versuche, bei deinen Mimimum gefunden wurde.
Rep.Index = Reproduktionen in Porzent, Rep.Index sollte im Idealfall 1 sein

Kriterium



Knickpunkt bei 4 Clustern???

Cluster-zahl	Streuungsquadratsummen innerhalb	Streuungsquadratsummen zwischen	F-Wert	ETA**2	PRE
1	148.805	-0.000	0.000	0.000	KW
2	85.545	63.260	161.949	0.425	***0.425***
3	60.613	88.193	158.598	0.593	0.291
4	44.960	103.846	167.073	0.698	***0.258***
5	36.856	111.950	164.025	0.752	0.180
6	30.708	118.097	165.369	0.794	0.167
7	24.684	124.121	179.345	0.834	0.196
8	20.798	128.008	187.287	0.860	***0.157***
9	18.593	130.213	185.588	0.875	0.106
10	16.280	132.525	***190.840***	0.891	***0.124***
11	14.800	134.006	190.145	0.901	0.091
12	13.942	134.863	183.786	0.906	0.058

maximaler F-Wert → 10 Cluster (= (132.525/(10-1)) / (16.280/(221-10))
PRE-Koeffizient → 4 Cluster (= 1 – 44.960/60.613))
Mindestanforderung an ETA^2, z.B. 70% → 5 Cluster, 4 Clusterlösung nahe dran
(ETA^2 für 4-Clusterlösung: (=1 – 44.960/148.805))
Bealsche F-Werte → 1 Cluster, da keine signifikanten Zuwächse

Bealsche F-Werte:
(Spalte1..1-Clusterloesung, Spalte2..2-Clusterloesung usw.;
 unteres Dreieck = F-Werte;
 oberes Dreieck = Signifikanzen der F-Werte)

Spalte 1	Spalte 2	Spalte 3	Spalte 4	Spalte 5	Spalte 6	Spalte 7	Spalte 8
0	51.4442	41.7075	39.3178	34.4120	31.9776	36.0279	38.2176
0.7328	0	55.1461	52.6162	47.5104	45.6234	51.6977	55.1838
0.7176	0.8115	0	63.9336	56.0844	53.8162	60.7533	64.5271
0.7560	0.8864	1.0256	0	57.2814	53.6141	61.4257	65.1901
0.7422	0.8608	0.9450	0.8596	0	61.9766	68.7716	71.6415
0.7483	0.8686	0.9474	0.9030	0.9738	0	75.8044	76.0438
0.8115	0.9550	1.0571	1.0605	1.1937	1.4179	0	71.6215
0.8475	1.0002	1.1071	1.1197	1.2403	1.3778	1.2608	0
0.8398	0.9869	1.0840	1.0883	1.1778	1.2501	1.1000	0.9099
0.8635	1.0155	1.1142	1.1212	1.2066	1.2691	1.1499	1.0596
0.8604	1.0094	1.1030	1.1065	1.1801	1.2257	1.1106	1.0269
0.8316	0.9714	1.0552	1.0519	1.1102	1.1372	1.0201	0.9300

Spalte 9	Spalte10	Spalte11	Spalte12
35.9208	37.5568	36.0928	31.3103
53.3948	56.1007	55.2123	50.3457
62.8441	65.7757	65.0932	60.3753
63.0614	65.9854	65.0866	59.9743
68.9788	71.6241	70.5775	65.4114
72.1308	74.2982	72.8528	67.2749
64.3892	66.8041	64.5627	57.4334
59.4150	62.3523	59.2493	50.7857
0	70.4320	64.1876	53.7394
1.2205	0	61.0640	46.7664
1.0959	0.9506	0	46.7240
0.9464	0.7930	0.6398	0

ALMO wählt automatisch die Lösung mit dem höchsten F-Wert aus, in dem Beispiel also die 10-Clusterlösung → Inspektion: zwei Ausreißercluster mit sehr hohen Werten in PMAT und GMAT

4-Clusterlösung soll weiter untersucht werden → erneuter Durchlauf mit Clusterzahl 4, Eingabe des Startwertes, der Minimum erzielt hat, also 9323078


Verfahren
Hilfe
2
Empfohlen: 3
Dies ist das Minimaldistanzverfahren
mit gewichteten quadrierten euklidischen Distanzen.
Als Gewichtungskriterium wird die Varianz der
Klassifikationsvariablen verwendet
Clusterzahl
4
4
Minimale Zahl von Clustern
Maximale Zahl von Clustern
Option: Clusterzugehörigkeiten der Ojekte in Datei speichern
Hilfe
Loesche wieder diese Box
Option: Programm-Optionen lt. Handbuch
Hilfe
Option45=1; Option15=9323078;
Option35=0;
Grafik-Optionen
Ausgabe der Ergebnisse
1
0= Ergebnisse stark verkürzt ausgeben
1= Ergebnisse mittelstark verkürzt ausgeben
2= Ergebnisse leicht verkürzt ausgeben
3= Ergebnisse in voller Länge ausgeben

3.2. Interpretation der Cluster

```
================================================================
Die 4-Clusterloesung wird weiter untersucht
================================================================

Clustergroessen:
   C1            26  (  11.765 %)
   C2            98  (  44.344 %)
   C3            22  (   9.955 %)
   C4            75  (  33.937 %)
 KW-Faelle (ungewichtet)=       0
```

```
============================================================
Zellenmittelwerte der Klassifikationsvariablen
(Mittelwerte bei quantitativen / ordinalen Variablen)
(Anteilswerte bei nominalen Variablen)

------------------------------------------------------------
Variable     C1       C2       C3       C4
------------------------------------------------------------
  V33      2.15     1.53     3.49     2.33                  GMAT
  V34      2.53     1.51     1.37     1.46                  GPMAT
------------------------------------------------------------
```

C1 = Mischtypus (Nicht-Orientierte) oder Materialisten
C2 = Mischtypus (Konsenstypus)
C3 = extremer Postmaterialist
C4 = gemäßigter Postmaterialist

Standardabweichungen:

Variable	C1	C2	C3	C4	
V33	0.37	0.24	0.53	0.28	GMAT
V34	0.54	0.30	0.42	0.23	GPMAT

Besetzungszahlen:

Variable	C1	C2	C3	C4	
V33	26	98	22	75	GMAT
V34	26	98	22	75	GPMAT

Z-Werte:

Variable	C1	C2	C3	C4	
V33	1.06	-22.23	12.36	7.97	GMAT
V34	8.72	-2.89	-2.47	-5.26	GPMAT

z-Werte wichtige Information, werden analog zur hierarchischen Clusteranalyse berechnet mit $z_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{s_{\bar{x}.ik}}$ (=dürften nicht mit den gewöhnlichen z-Werten verwechselt werden)
manchmal sinnvoll, die Interpretation auf diese zu stützen, ermöglichen aber nur relative Aussagen

Signifikanz (1-p)*100 der z-Werte:

Variable	C1	C2	C3	C4	
V33	69.94	100.00	100.00	100.00	GMAT
V34	100.00	99.54	97.78	100.00	GPMAT

Detailanalysen

Oft sollen bei der Interpretation folgende Hypothesen geprüft:

- Unterschiede zwischen den Clustern in einer Variablen x
- Unterschiede zwischen Variablen innerhalb eines Clusters

ALMO berechnet dazu folgende Werte:

```
Paarweise Clusterdifferenzen fuer Cluster=1   (n=  26)

Klassifikationsvariablen:
-------------------------------------------------------------------
Variable     C2       C3       C4
-------------------------------------------------------------------
  V33         >        <        =                          GMAT
  V34         >        >        >                          GPMAT
```

```
Paarweise Clusterdifferenzen fuer Cluster=2   (n=  98)
Klassifikationsvariablen:

-------------------------------------------------------------------
Variable     C1       C3       C4
-------------------------------------------------------------------
  V33         <        <        <                          GMAT
  V34         <        =        =                          GPMAT
```

usw.

```
Beziehung der Variablen im Cluster   1
====================================

keine Verschmelzung bei vorgegebenen Alpha_Niveau moeglich
```

Cluster 1 könnte auch als „Materialisten“ bezeichnet werden, da beide Mittelwerte nicht gleich sind

```
Beziehung der Variablen im Cluster   2
=====================================

Verschmelzungsprotokoll:
 Schritt   1  alpha=  0.597  V33=GMAT
                           + V34=GPMAT

Variablengruppen bei einem Fehlerniveau von alpha=  0.050
(Alpha-Niveau nach Bonferroni-Korrektur = 0.05)
Variablengruppe   1
         V33=GMAT
         V34=GPMAT
   (Mittelwert=    1.52; Varianz=    0.07)

Beziehung der Variablen im Cluster   3
=====================================
keine Verschmelzung bei vorgegebenen Alpha_Niveau moeglich
```

usw.

3.3. Zufallstestung einer Clusterlösung

Nullmodell: $x_i \approx NV(\bar{x}, s_x)$ und paarweise unabhängig → Wie gut werden Zufallsdaten durch die Clusterlösung erklärt? Die Erklärungskraft sollte deutlich geringer sein.

```
Zahl der Simulationen = 100
Simulationswerte:
   0.608    0.565    0.557    0.597    0.568    0.566    0.574    0.538    0.570
   0.557    0.577    0.573    0.572    0.526    0.531    0.567    0.600    0.591
   0.606    0.594    0.562    0.542    0.613    0.579    0.595    0.567    0.569
   0.620    0.552    0.601    0.568    0.576    0.562    0.597    0.650    0.618
   0.567    0.552    0.578    0.634    0.573    0.582    0.570    0.611    0.594
   0.546    0.594    0.575    0.607    0.614    0.546    0.588    0.585    0.644
   0.615    0.575    0.558    0.577    0.614    0.596    0.556    0.568    0.578
   0.601    0.598    0.547    0.586    0.551    0.599    0.611    0.639    0.594
   0.605    0.612    0.575    0.575    0.620    0.591    0.529    0.622    0.596
   0.565    0.565    0.536    0.594    0.557    0.536    0.606    0.579    0.523
   0.596    0.563    0.564    0.600    0.554    0.595    0.559    0.553    0.608
   0.590
emp. Wert      =   0.698
Erwartungswert =   0.581
Standardabw.   =   0.027
z-Wert         =   4.371
Signifikanz    = 100.000
```

→ auch Zufallsdaten würden zu 58% erklärt werden

3.4. Validitätsprüfung

Formale Gültigkeitsprüfung: Zur formalen Gültigkeitsprüfung berechnet ALMO die Homogenität innerhalb der Cluster und die Entfernungen zwischen den Clusterzentren

Cluster	n=	Streuung innerhalb	Homogenitaet innerhalb
C1	26	0.210	0.378
C2	98	0.072	0.786
C3	22	0.227	0.329
C4	75	0.066	0.805

$$H_k = 1 - \frac{s(k)_{in}^2}{s^2}$$

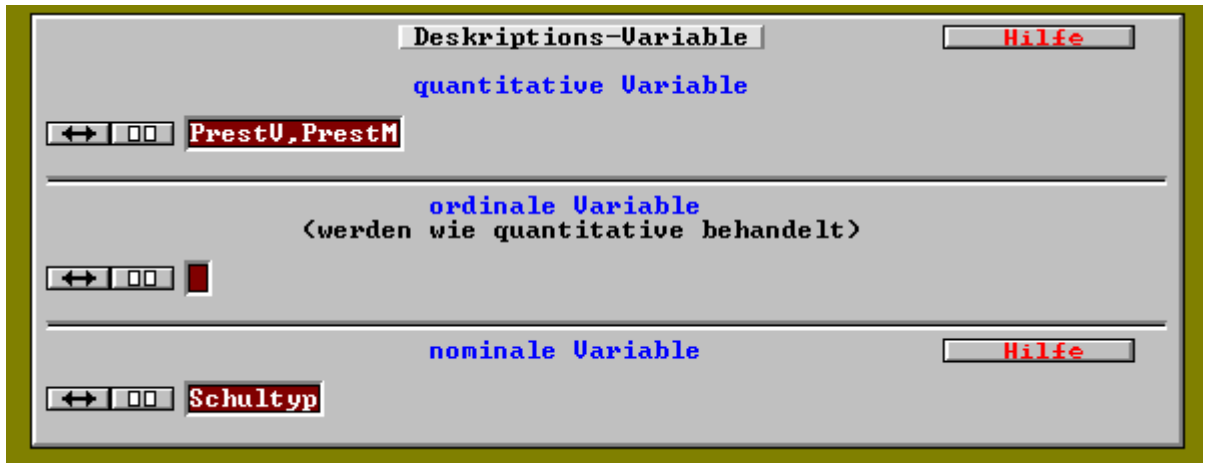
```
Entfernungen der Clusterzentren bzw. Repraesentanten zueinander:
quadrierte (gewichtete) euklidische Distanzen

Spalte 1 Spalte 2 Spalte 3 Spalte 4
   0       1.4175   3.1524   1.1902
  1.4175    0       3.8555   0.6327
  3.1524   3.8555    0       1.3645
  1.1902   0.6327   1.3645    0
```

Kriterienbezogene Gültigkeitsprüfung

Es werden Hypothesen über die Cluster formuliert und anschließend empirisch geprüft:

- Extreme Postmaterialisten und Postmaterialisten kommen aus höheren sozialen Schichten
- Extreme Postmaterialisten und Postmaterialisten kommen häufiger aus einer AHS



Zellenmittelwerte der Deskriptionsvariablen:
(Mittelwerte bei quantitativen Variablen, Anteilswerte bei nominalen)

C1=Nicht-Orientiert, C2=Konsenstyous, C3=etrem Postmaterialisten, C4=Postmaterialisten

Variable	C1	C2	C3	C4	
V25					Schultyp
1	0.42	0.27	0.23	0.52	BHS
2	0.23	0.14	0.68	0.32	AHS
3	0.35	0.59	0.09	0.16	BS
V40	3.54	3.58	4.09	3.91	PrestV
V41	2.83	3.02	4.09	3.59	PrestM

Z-Werte:

Variable	C1	C2	C3	C4	
V25					Schultyp
1	0.57	-2.26	-1.52	2.64	BHS
2	-0.43	-3.49	4.08	0.98	AHS
3	-0.21	4.52	-4.39	-4.85	BS
V40	-0.68	-1.28	1.14	1.07	PrestV
V41	-1.00	-1.26	2.28	1.18	PrestM

Signifikanz (1-p)*100 der z-Werte:

Variable	C1	C2	C3	C4	
V25					Schultyp
1	42.88	97.38	85.67	99.01	BHS
2	32.95	99.92	99.95	66.88	AHS
3	16.68	100.00	99.98	100.00	BS
V40	49.97	79.81	73.21	71.27	PrestV
V41	66.34	78.36	95.36	75.27	PrestM

Hypothesen werden für extreme Postmaterialisten bestätigt. Für Postmaterialisten zeigen sich die Zusammenhänge in der Tendenz.

3.5. Stabilitätsprüfung

z.B. Auswirkungen von geringen Änderungen in den Daten oder anderes Verfahren, z.B. Ward-Linkage

→ Ward-Linkage

```
Modell = Ward-Verfahren
    fuer quadrierte euklidische Distanz
*************************************************************************
Clusterverknuepfung   Clusterzahl       Distanzniveau       Zuwachs     Bindungen
     42    173              20              1.197           0.240             0
     20     23              19              1.253           0.056             0
     14     58              18              1.419           0.166             0
     29     37              17              1.441           0.022             0
      4      5              16              1.716           0.275             0
      3      9              15              1.720           0.004             0
      1     25              14              1.977           0.256             0
```

81	82	13	2.105	0.128	0
29	35	12	3.100	0.995	0
21	42	11	3.217	0.116	0
3	96	10	3.458	0.241	0
14	20	9	4.022	0.564	0
3	4	8	6.967	2.945	0
21	216	7	9.531	2.564	0
1	81	6	13.222	3.691	0
21	29	5	17.109	3.887	0
3	14	4	18.582	1.474	0
1	164	3	24.350	5.768	0
3	21	2	57.900	33.550	0
1	3	1	109.304	51.404	0

Zuwachs bei 4-Clustern erkennbar, aber 4-Clusterlösugn weicht von K-Means, es wird ein „Ausreißer“-Cluster ermittelt

Masszahlen fuer Klassifikationsvariablen im Clustern 1:

gewichtete Fallzahl =52

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	52	2.33	3.80	2.78	0.38	13.37
34 GPMAT	52	1.00	1.83	1.36	0.25	-6.87

Masszahlen fuer Klassifikationsvariablen im Clustern 2:

gewichtete Fallzahl =123

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	123	1.00	2.33	1.68	0.33	-12.95
34 GPMAT	123	1.00	2.33	1.45	0.26	-6.41

Masszahlen fuer Klassifikationsvariablen im Clustern 3:

gewichtete Fallzahl =42

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	42	1.50	3.50	2.11	0.45	0.59
34 GPMAT	42	1.83	4.33	2.31	0.52	8.75

Masszahlen fuer Klassifikationsvariablen im Clustern 4:

gewichtete Fallzahl =4

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	4	4.33	4.83	4.50	0.20	20.63
34 GPMAT	4	1.67	2.33	1.92	0.25	2.20

****

Hinweis auf mögliche Ausreißer → Single Linkage → bestätigt diese Vermutung

Masszahlen fuer Klassifikationsvariablen im Clustern 1:

gewichtete Fallzahl =215

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	215	1.00	3.80	2.02	0.58	-1.25
34 GPMAT	215	1.00	3.00	1.57	0.41	-1.03

Masszahlen fuer Klassifikationsvariablen im Clustern 2:

gewichtete Fallzahl =4

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	4	4.33	4.83	4.50	0.20	20.63
34 GPMAT	4	1.67	2.33	1.92	0.25	2.20

Masszahlen fuer Klassifikationsvariablen im Clustern 3:
gewichtete Fallzahl =1

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	1	3.00	3.00	3.00	0.00	0.00
34 GPMAT	1	4.33	4.33	4.33	0.00	0.00

Masszahlen fuer Klassifikationsvariablen im Clustern 4:
gewichtete Fallzahl =1

Variable	n=	Min.	Max.	MA	SA	z-Wert
33 GMAT	1	2.17	2.17	2.17	0.00	0.00
34 GPMAT	1	3.83	3.83	3.83	0.00	0.00

4. Probabilsitische Clusteranalyse

in ALMO einfach durchführbar. Es muss nur Modell = 7 definiert werden. Auch hier ergeben sich Hinweise auf Ausreißer → 6 Cluster, davon zwei sehr schwach besetzt

Clusterzahl	Log(L)	df	PV(k-1)
1	-374.100	4.000	KW
2	-330.417	9.000	11.677
3	-316.099	14.000	4.333
4	-303.603	19.000	3.953
5	-292.559	24.000	3.638
6	-286.292	29.000	2.142
7	-283.439	34.000	0.997
8	-280.111	39.000	1.174
9	-278.203	44.000	0.681
10	-277.608	49.000	0.214
11	-275.014	54.000	0.934
12	-273.961	59.000	0.383

Clusterzahl	AIK	-2*AIK	BIC	CAIC
1	-378.100	756.201	769.793	773.793
2	-339.417	678.833	709.417	718.417
3	-330.099	660.198	707.773	721.773
4	-322.603	645.205	709.770	728.770
5	-316.559	633.118	714.674	738.674
6	-315.292	630.585	729.131	758.131
7	-317.439	634.878	750.415	784.415
8	-319.111	638.223	770.751	809.751
9	-322.203	644.406	793.926	837.926
10	-326.608	653.216	819.726	868.726
11	-329.014	658.029	841.529	895.529
12	-332.961	665.922	866.414	925.414

AIK=Informationskriterium von Akaike nach Bacher (1994: 366)
-2*AIK=Informationskriterium von Akaike nach Rost (1996: 443)
BIC=Best Information Criterion; Rost (1996: 443)
CAIC=Consistent Akaikes Information Criterion ; Rost (1996: 443)
beste Loesung = Loesung mit dem kleinsten Wert
Bacher, J., 1994: Clusteranalyse. München
Rost, J., 1996: Testtheorie. Testkonstruktion. Bern u.a.
PV(k-1)=prozentuelle Verb. gegen. vorausgeh. Loesung