Clusteranalyse mit gemischtskalierten Merkmalen: SPSS-Makropaket „Paare“[1]

Norman Fickel

Friedrich-Alexander-Universität Erlangen-Nürnberg
Wirtschafts- und Sozialwissenschaftliche Fakultät
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Lange Gasse 20
D-90 403 Nürnberg

Abstract

Die Statistiksoftware SPSS unterstützt auch in der Version 6 nur Clusteranalysen, wenn alle Merkmale dasselbe Skalenniveau besitzen. Mit dem hier vorgestellten Makropaket „Paare“ kann SPSS auch gemischtskalierte Daten verarbeiten. Der zugrundeliegende Algorithmus abstrahiert dazu durch Abstandsmaße von den unterschiedlichen Skalenniveaus der gegebenen Merkmale. Die Güte der gefundenen Klassifikation läßt sich mit einer Streuungszerlegung sowohl insgesamt als auch merkmalsspezifisch beurteilen.

1 Clusteranalysemethode

Das Makropaket „Paare“ benützt die Methode der Abstrahierung durch Abstandsmaße. Dazu wird für jedes Merkmal ein vom Skalenniveau abhängiges Abstandsmaß definiert, so daß der Klassifizierungsalgorithmus in diesem Sinne vom Skalenniveau unabhängig mit Abständen rechnen kann. Charakteristisch für diese Methode sind das Konzept der Gleichgewichtung der Merkmale und die Idee der verwendeten Streuungszerlegung.

Die folgende Darstellung beschränkt sich im wesentlichen auf die Rechenschritte der Methode, da der theoretische Hintergrund bereits ausführlich von Buttler und Fickel (1995) beschrieben ist. Andere Ansätze zur Clusteranalyse mit gemischtskalierten Merkmalen findet man im Überblick etwa bei Anderberg (1973, S. 127-130), Bacher (1994, S. 186-191) oder Everitt und Merette (1990).

Für die drei Skalenniveaus metrisch, ordinal und nominal werden untenstehende Abstandsmaße als Standard vorgeschlagen. Dabei bezeichnet x_{ih} die Ausprägung des *h*ten Merkmals beim *i*ten Objekt. Die Gesamtzahl der Merkmale ist dabei *m* und die der Objekte *n*.

---

[1] Diese Beschreibung wurde im November 1995 verfaßt. Der Autor ist mit Email im Internet erreichbar unter wss120@wsrz2.wiso.uni-erlangen.de.

- *Nominales Niveau:* Der Abstand zwischen dem *i*ten und *j*ten Objekt ist gleich Eins, falls die Ausprägungen verschieden sind, und null, falls beide Objekte dieselbe Ausprägung haben. Also $d_{ij;h} = 1$, falls $x_{ih} \neq x_{jh}$ und $d_{ij;h} = 0$, sonst.

- *Ordinales Niveau:* Der Abstand ist die Anzahl der zwischen dem *i*ten und dem *j*ten Objekt liegenden anderen Objekte plus Eins, also $d_{ij;h} = |R_{ih} - R_{jh}|$. Dabei sind R_{1h}, R_{2h}, ..., R_{nh} die Ränge der *n* Objekte. Treten Bindungen auf, haben also Objekte dieselbe Ausprägung, so werden mittlere Ränge vergeben. Die Rangzahl des Objekts ist dann das arithmetische Mittel der Ränge bei durchlaufender Numerierung aller Objekte mit den Zahlen 1, 2, ..., *n*.

- *Metrisches Niveau:* Der Abstand je zweier Objekte ist der Absolutbetrag der Ausprägungsdifferenz, also $d_{ij;h} = |x_{ih} - x_{jh}|$.

Es ist sinnvoll, von diesem Vorschlag abzuweichen, falls aus sachlichen Gründen ein anderes Abstandsmaß angemessener erscheint. Etwa könnte bei ordinalem Niveau die Differenz der numerierten möglichen Ausprägungen verwendet werden, oder bei metrischem Niveau das Quadrat der Ausprägungsdifferenz. Von formaler Bedeutung ist nur, daß der Abstand immer eine nichtnegative reelle Zahl ist ($d_{ij;h} \geq 0$) und der Abstand eines Objekts zu sich selbst stets null wird ($d_{ii;h} = 0$). Zudem muß als Nichttrivialitätsbedingung jedes Merkmal echt variabel sein, also wenigstens für ein Objektpaar $d_{ij;h} > 0$ gelten.

Die Streuung eines Merkmals wird gemessen durch die Summe der Abstände aller möglichen Objektpaare, konkret $D_h = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij;h}$. Der normierte Abstand zweier Objekte ist der Anteil des gegebenen Abstandes an der Streuung, also $d_{ij;h}^{\text{norm}} = d_{ij;h} / D_h$. Diese Streuungsanteile werden zum Gesamtabstand aufsummiert, damit $d_{ij} = \sum_{h=1}^{m} d_{ij;h}^{\text{norm}}$. Dann sind die verschiedenen Merkmale insofern gleichgewichtet, als die Streuung der normierten Abstände für jedes Merkmal gleich Eins ist.

Mit dem Gesamtabstand lassen sich alle abstandsbasierten Klassifikationsalgorithmen zur Klassenbildung verwenden. In SPSS sind davon Single-linkage, Complete-linkage und zwei Varianten von Average-linkage implementiert.

Die Gesamtstreuung wird dadurch zerlegt, daß zum einen alle Abstände d_{ij}, bei denen beide Objekte in derselben Klasse sind, zu einer Innerklassenstreuung aufsummiert werden. Zum anderen sind alle Abstände d_{ij}, deren Objekte verschiedenen Klassen angehören, zur Zwischenklassenstreuung addiert. Die Gesamtstreuung ist dann die Summe aus Inner- und Zwischenklassenstreuung, was sich, wenn C_1, C_2, ..., C_s die Klassen bezeichnet, wie folgt schreiben läßt:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} = \sum_{k=1}^{s}\sum_{i,j\in C_k} d_{ij} + \sum_{k=1}^{s}\sum_{\substack{i\in C_k \\ j\notin C_k}} d_{ij}$$

Auf dieselbe Weise könnten auch die merkmalsspezifischen Streuungen zerlegt werden, wozu man in obiger Formel den Ausdruck d_{ij} für das *h*te Merkmal jeweils durch $d_{ij;h}$ ersetzen muß.

2 SPSS-Makropaket „Paare“

Mit dem Makropaket „Paare“ kann eine Clusteranalyse in den folgenden acht Schritten durchgeführt werden:

(1) Einlesen der Daten
(2) gegebenenfalls Umwandeln ordinaler Ausprägungen
(3) Erzeugen der Paare
(4) Berechnen der Abstände
(5) Erzeugen der Distanzmatrix
(6) Bilden der Klassen
(7) Auflisten der Klassengrößen
(8) Auflisten der Streuungszerlegung

Beispiel:
Das Makro !beisp führt die beschriebenen acht Schritte für ein kleines Beispiel durch. Dazu öffnet man in SPSS für Windows die zugehörige SPSS-Syntaxdatei (S. 8 ff.), markiert alles und klickt auf die Schaltfläche „Ausführen“.

2.1 Einlesen der Daten (Befehl GET TRANSLATE)

```
GET TRANSLATE FILE=Dateiname /TYPE=TAB.
```

Der angegebene SPSS-Befehl liest aus der Datei die Objekte mit ihren Merkmalen ein. Jede Zeile enthält die Daten eines Objekts, die Spalten müssen durch Tabulatoren getrennt sein. Der Dateiname bezieht sich auf das aktuelle Arbeitsverzeichnis von SPSS. Nach Ablauf des Makros stehen die Daten im Arbeitsbereich von SPSS, wobei die Merkmale die Bezeichnungen VAR1, VAR2, VAR3, ... erhalten haben.

Weitere Möglichkeiten, Daten auch aus anderen Formaten einzulesen, sind im Syntax-Handbuch der SPSS Inc. (1992, S. 279 ff.) dokumentiert.

Beispiel:

```
GET TRANSLATE FILE='beispiel.dat' /TYPE=TAB.
```

Die Datei beispiel.dat wird in den Arbeitsbereich geladen. Sie besteht aus folgenden fünf Zeilen, wobei die drei Spalten durch Tabulatoren getrennt sind:

```
8          4,3          2
10         2,3          1
8          2,7          2
11         2,7          3
9          1,0          2
```

Es handelt sich hier um fünf Studierende, deren Semesterzahl (erste Spalte), Prüfungsnote (zweite Spalte) und Studienfach (dritte Spalte) angegeben ist. Die Semesterzahl ist hier ein metrisches, die Prüfungsnote ein ordinales und das Studienfach ein nominales Merkmal.

2.2 Umwandeln ordinaler Ausprägungen (Makro !rangord)

```
!rangord von=erstes_ordinale_Merkmal
         bis=letztes_ordinale_Merkmal.
```

Das Makro wandelt die Ausprägungen der angegebenen ordinalen Merkmale in Ränge um. Es kann mehrmals hintereinander aufgerufen werden, so daß im Arbeitsbereich die ordinalen Merkmale nicht alle aufeinander folgen müssen.

Beispiel:

```
!rangord von=2 bis=2.
```

Im Beispiel werden die Ausprägungen des Merkmals „Prüfungsnote“ wie folgt umgewandelt:

vorher:	4,3	2,3	2,7	2,7	1,0
nachher:	5	2	3,5	3,5	1

2.3 Erzeugen der Paare (Makro !erzpaar)

```
!erzpaar m=Merkmalsanzahl n=Objektanzahl.
```

Die n Objekte im Arbeitsbereich werden zunächst in der Datei objekte.sav mit ihrer Objektnummer (Variable objekt_i) zur späteren Verwendung gesichert. Dann erzeugt das Makro n^2 Paare, die durch die Variablen objekt_i und objekt_j identifizierbar sind.

Beispiel:

```
!erzpaar m=3 n=5.
```

Im Beispiel stehen nach Ausführung des Makros als Fälle 25 Paare im Arbeitsbereich.

2.4 Berechnen der Abstände (Makros !metri, !ordi und !nomi)

```
!metri von=erstes_metrische_Merkmal
       bis=letztes_metrische_Merkmal.
!ordi  von=erstes_ordinale_Merkmal
       bis=letztes_ordinale_Merkmal.
!nomi  von=erstes_nominale_Merkmal
       bis=letztes_nominale_Merkmal.
```

Jedes der drei Makros berechnet das dem Skalenniveau adäquate Abstandsmaß und fügt das Ergebnis als Variable d*k* dem Arbeitsbereich hinzu, wobei *k* für die jeweilige Nummer des Merkmals steht. Für alle ordinalen Merkmale müssen vor dem Erzeugen der Paare (Makro !erzpaar) die Ausprägungen in Ränge umgewandelt worden sein (vgl. Makro !rangord).

Beispiel:

```
!metri von=1 bis=1.
!ordi  von=2 bis=2.
!nomi  von=3 bis=3.
```

Im Beispiel befinden sich dann im Arbeitsbereich die Variablen objekt_i, objekt_j, d1, d2 und d3.

2.5 Erzeugen der Distanzmatrix (Makro !erzmat)

```
!erzmat m=Merkmalsanzahl n=Objektanzahl
        mat=Dateiname.
```

Die Ausführung setzt voraus, daß sich im Arbeitsbereich die n^2 Paare (objekt_i, objekt_j) mit unnormierten Abständen d1, d2, ..., d*m* befinden. Die Abstände werden zunächst normiert, wodurch als Zwischenergebnis die Datei mittel.sav alle mittleren Summen der paarweisen Abstände enthält. Anschließend werden die normierten Abstände zum aggregierten Gesamtabstand (Variable distanz) aufsummiert. Der aktuelle Arbeitsbereich wird dann in der Datei paare.sav abgespeichert, um später zur Verfügung zu stehen. Der Gesamtabstand wird benutzt, um eine Distanzmatrix im SPSS-Format zu erzeugen, die unter dem nach „mat=“ angegebenen Dateinamen gesichert wird. Zuletzt lädt das Makro die Datei objekte.sav in den Arbeitsbereich.

Beispiel:

```
!erzmat m=3 n=5 mat='matrix.sav'.
```

Nach Ablauf des Makros ist in der Datei matrix.sav eine SPSS-Distanzmatrix abgespeichert. Der Arbeitsbereich enthält als Fälle die fünf Studierenden.

2.6 Bilden der Klassen (Befehl CLUSTER)

```
CLUSTER /MATRIX=IN(Dateiname)
    /METHOD=Methodenname
    /SAVE=CLUSTERS(Klassenanzahl).
```

Die Fälle im Arbeitsbereich werden in die als *Klassenanzahl* angegebene Zahl von Klassen aufgeteilt. Den Fällen zugrundegelegt sind dabei die in der Distanzmatrix aus der Datei *Dateiname* festgelegten Abstände. Als Klassifizierungsmethode ist Single-linkage (*Methodenname* SINGLE), Complete-linkage (COMPLETE), Average-linkage zwischen (BAVERAGE) oder innerhalb der Gruppen (WAVERAGE) angemessen. Nach Ausführung des Befehls be-

findet sich eine zusätzliche Variable im Arbeitsbereich, welche die Klassenzugehörigkeit angibt. Der Variablenname wird von SPSS in der Form CLU*s*_1 gebildet, wobei *s* die Klassenanzahl bezeichnet.

Der Befehl CLUSTER wird ausführlich im SPSS-Handbuch von Norušis (1993, S. 267 ff.) beschrieben.

Beispiel:

```
CLUSTER /MATRIX=IN('matrix.sav')
    /METHOD=COMPLETE
    /SAVE=CLUSTERS(2).
```

Mit den Abständen aus der Datei matrix.sav bildet die Methode Complete-linkage zwei Klassen. Der Arbeitsbereich enthält dann die zusätzliche Variable CLU2_1.

2.7 Auflisten der Klassengrößen (Befehl FREQUENCIES)

`FREQUENCIES` *`Variablenname`*`.`

Der Befehl listet die absoluten und relativen Häufigkeiten für die angegebene Variable, also werden insbesondere die Klassengrößen ausgegeben, wenn man den von CLUSTER erzeugten Variablennamen der Form CLU*s*_1 wählt.

Beispiel:

```
FREQUENCIES CLU2_1.
```

Damit liefert SPSS folgende Ausgabe:

Value	Frequency	Percent	Valid Percent	Cum Percent
1	2	40,0	40,0	40,0
2	3	60,0	60,0	100,0
Total	5	100,0	100,0	

Die eine Klasse enthält also zwei Studierende und die andere die restlichen drei.

2.8 Auflisten der Streuungszerlegung (Makro !zerleg)

`!zerleg var=`*`Variablenname`* `m=`*`Merkmalsanzahl`*`.`

Das Makro listet die Streuungszerlegung sowohl insgesamt als auch für jedes einzelne Merkmal auf. Die Klassenzugehörigkeit wird durch die nach „var=“ angegebene Variable festgestellt, welche in der Liste mit D0 benannt ist. Nach seinem Ablauf befinden sich als Fälle alle Merkmale und als Variablen die prozentuale Inner- und Zwischenklassenstreuung im Arbeitsbereich.

Beispiel:

```
!zerleg var=CLU2_1 m=3.
```

Das Makro liefert folgende Ausgabe:

```
CASE_LBL    INNER ZWISCHEN

D0            ,00   100,00
D1          25,00    75,00
D2          34,21    65,79
D3          42,86    57,14
DISTANZ     34,02    65,98
```

Insgesamt werden durch die Klassifikation rund 66 % der Streuung erklärt. Speziell sind das bei der Semesterzahl (Merkmal Nr. 1) 75 %, bei der Prüfungsnote (Merkmal Nr. 2) etwa 66 % und beim Studienfach (Merkmal Nr. 3) rund 57 %.

3 Technische Probleme

Die Anzahl der verarbeitbaren Objekte ist im Makropaket durch den verfügbaren Arbeitsspeicher des Computers beschränkt, da SPSS bei der hierarchisch-agglomerativen Clusteranalyse die gesamte Distanzmatrix dort zugriffsbereit halten muß. Für einen PC mit acht Megabyte Arbeitsspeicher sind maximal etwa 200 bis 300 Objekte möglich.

Zusätzlich ist bei größeren Objektzahlen die Laufzeit einer Clusteranalyse mit dem Makropaket „Paare“ ganz erheblich höher als bei einer der in SPSS vorgesehenen Clusteranalysen mit einheitlich skalierten Merkmalen. Beispielsweise benötigt für einen Datensatz mit 112 Objekten und 57 Merkmalen die Ausführung der acht Schritte rund zehn Minuten Rechenzeit. Dabei wird die Distanzmatrix (Makro !erzmat) in etwa sechs Minuten erzeugt, wohingegen der SPSS-Befehl CLUSTER nach nur knapp 15 Sekunden abgelaufen ist. Diese Zeiten wurden auf einem PC mit Pentium-Prozessor bei einer Taktfrequenz von 100 Megahertz gemessen.

Zudem belegen die während des Ablaufs erzeugten Dateien unter Umständen nennenswerten Festplattenspeicher. Im oben erwähnten Datensatz mit 112 Objekten und 57 Merkmalen hat die größte Datei paare.sav einen Umfang von rund sechs Megabyte.

Die SPSS-Makrosprache ermöglicht keinen direkten Zugriff auf die internen Datenstrukturen von SPSS und wird zudem erst während der Laufzeit interpretiert. Eine Compilierung ist vorab nicht möglich. Bei einer direkten Implementation des in Abschnitt 1 beschriebenen Algorithmus (S. 1 ff.) dürften die Laufzeit und der Speicherbedarf nicht wesentlich größer sein als bei den in SPSS vorhandenen Clusteranalysen.

4 Literatur

Anderberg, Michael R. *Cluster Analysis for Applications.* New York [u.a.] 1973.

Bacher, Johann. *Clusteranalyse: anwendungsorientierte Einführung.* München [u.a.] 1994.

Buttler, Günter, u. Norman Fickel. *Clusteranalyse mit gemischtskalierten Merkmalen*. Diskussionspapier der Lehrstühle für Statistik. Nürnberg 1995.

Everitt, Brian S., u. Chantal Merette. „The clustering of mixed-mode data: a comparison of possible approaches". *Journal of Applied Statistics* 17.3 (1990): 283-297.

Norušis, Marija J. *SPSS for Windows, Professional Statistics, Release 6.0.* Chicago 1993.

SPSS Inc., Hg. *SPSS Base System, Syntax Reference Guide, Release 5.0.* Chicago 1992.

5 Anhang: Vollständige SPSS-Syntax des Makropakets „Paare"

Das Makro !beisp liefert die Clusteranalyse für das in Abschnitt 2 (S. 3 ff.) benützte Demonstrationsbeispiel. Für eine eigene Clusteranalyse kann es entsprechend abgeändert werden. Möchte man andere Abstandsmaße verwenden, als die in Abschnitt 1 (S. 1 ff.) vorgeschlagenen, so läßt sich der COMPUTE-Befehl in den Makros !metri, !ordi und !nomi modifizieren.

Zusätzlich zur SPSS-Syntax der in Abschnitt 2 beschriebenen Makros ist unten auch die der verwendeten Hilfsmakros angegeben:

- *Makro !hinzu1:* Die Variable mit der beim Aufruf angegebenen Nummer wird den Objektpaaren im Arbeitsbereich hinzugefügt. Dies wird zur Streuungszerlegung benötigt (Makro !zerleg).
- *Makro !hinzu:* Jede der Variablen der angegebenen Nummern wird den Objektpaare im Arbeitsbereich hinzugefügt. Im Ergebnis ist dies gleichwertig mit mehreren entsprechenden Aufrufen des Makros !hinzu1, jedoch ist die Abarbeitung effizienter. Dies wird beim Erzeugen der Distanzmatrix benötigt (Makro !erzmat).
- *Makro !norme:* Die Abstände d1, d2, ..., d*m* der Objektpaare im Arbeitsbereich werden normiert. Dies gehört zum Erzeugen der Distanzmatrix (Makro !erzmat).
- *Makro !matrix:* Die n^2 Werte für den Gesamtabstand werden in das vom SPSS-Befehl CLUSTER benötigte Format umsortiert. Auch dies gehört zum Erzeugen der Distanzmatrix (Makro !erzmat).

Zur besseren Lesbarkeit ist in der folgenden Syntax jeder DEFINE-Befehl fett gedruckt.

```
/*** Clusteranalyse mit       ***/
/*** gemischtskalierten       ***/
/*** Merkmalen:               ***/
/*** SPSS-Makropaket "Paare"  ***/
/*** Autor: Norman Fickel     ***/

DEFINE !rangord (von = !TOKENS(1)
  /bis = !TOKENS(1)).
* Werte ordinaler Variablen
  in Ränge umwandeln.
RANK !CONCAT('var',!von)
  TO !CONCAT('var',!bis)
  /PRINT=NO.
!DO !k = !von !TO !bis.
  COMPUTE !CONCAT('var',!k)
    = !CONCAT('rvar',!k).
```

```
!DOEND.
!ENDDEFINE.

DEFINE !hinzu1 (nr = !TOKENS(1)
  /table= !TOKENS(1)).
  * Variable var[nr] aus table den
    Paaren hinzufügen.
MATCH FILES FILE=*
  /TABLE=!table
  /RENAME=(!CONCAT('VAR',!nr)
    =!CONCAT('i',!nr))
  /BY objekt_i.
SORT CASES BY objekt_j.
MATCH FILES FILE=*
  /TABLE=!table
  /RENAME=(!CONCAT('VAR',!nr)
    =!CONCAT('j',!nr))
    (objekt_i=objekt_j)
  /BY objekt_j.
RENAME VARIABLES
  (objekt_i objekt_j
  =objekt_j objekt_i).
!ENDDEFINE.

DEFINE !hinzu (von = !TOKENS(1)
  /bis = !TOKENS(1)
  /table= !TOKENS(1)).
  * Variablen
    var[von] bis var[bis]
    aus table den
    Paaren hinzufügen.
MATCH FILES FILE=*
  /TABLE=!table
  /RENAME=
    (!CONCAT('var',!von)
    TO !CONCAT('var',!bis)
    =!CONCAT('i',!von)
    TO !CONCAT('i',!bis))
  /BY objekt_i.
SORT CASES BY objekt_j.
MATCH FILES FILE=*
  /TABLE=!table
  /RENAME=
    (!CONCAT('var',!von)
    TO !CONCAT('var',!bis)
    =!CONCAT('j',!von)
    TO !CONCAT('j',!bis))
    (objekt_i=objekt_j)
  /BY objekt_j.
RENAME VARIABLES
  (objekt_i objekt_j
  =objekt_j objekt_i).
!ENDDEFINE.

DEFINE !erzpaar (m = !TOKENS(1)
  /n = !TOKENS(1)).
  * Paare erzeugen.
COMPUTE objekt_i = $CASENUM.
SAVE OUTFILE='objekte.sav'
  /KEEP objekt_i
  var1 TO !CONCAT('var',!m).

/*** n-Quadrat Paare erzeugen ***/
NEW FILE.
INPUT PROGRAM.
LOOP #I = 1 TO !n.
  LOOP #J = 1 TO !n.
    COMPUTE objekt_i = #I.
    COMPUTE objekt_j = #J.
    END CASE.
  END LOOP.
END LOOP.
END FILE.
END INPUT PROGRAM.

!hinzu von=1 bis=!m
  table='objekte.sav'.
  * Variablenpaare zuordnen.
!ENDDEFINE.

DEFINE !metri (von = !TOKENS(1)
  /bis = !TOKENS(1)).
  * metrische Abstände berechnen.
!DO !k = !von !TO !bis.
  COMPUTE !CONCAT('d',!k) =
    ABS(!CONCAT('i',!k)
    - !CONCAT('j',!k)).
!DOEND.
!ENDDEFINE.

DEFINE !ordi (von = !TOKENS(1)
  /bis = !TOKENS(1)).
  * ordinale Abstände aus den
    vorgegebenen Rängen berechnen.
!DO !k = !von !TO !bis.
  COMPUTE !CONCAT('d',!k) =
    ABS(!CONCAT('i',!k)
    - !CONCAT('j',!k)).
!DOEND.
!ENDDEFINE.

DEFINE !nomi (von = !TOKENS(1)
  /bis = !TOKENS(1)).
  * nominale Abstände berechnen.
!DO !k = !von !TO !bis.
  DO IF (!CONCAT('i',!k)
          = !CONCAT('j',!k)).
    COMPUTE !CONCAT('d',!k) = 0.
  ELSE.
    COMPUTE !CONCAT('d',!k) = 1.
  END IF.
!DOEND.
!ENDDEFINE.

DEFINE !norme (m = !TOKENS(1)
  /table = !TOKENS(1)).
  * alle Abstände normieren.
COMPUTE e_1=1.
AGGREGATE OUTFILE=!table
  /PRESORTED
  /BREAK=e_1
  /m1 TO !CONCAT('m',!m)
  =MEAN(d1 TO !CONCAT('d',!m)).
MATCH FILES FILE=*
  /TABLE=!table
  /BY e_1
  /KEEP objekt_i objekt_j
  d1 TO !CONCAT('d',!m)
  m1 TO !CONCAT('m',!m).
!DO !k = 1 !TO !m.
  COMPUTE !CONCAT('d',!k)
    = !CONCAT('d',!k)
    / !CONCAT('m',!k).
!DOEND.
```

```
!ENDDEFINE.

DEFINE !matrix (n = !TOKENS(1)).
/*** Variablen VAR1 bis VARn  ***/
/*** erzeugen                 ***/
!DO !j = 1 !TO !n.
  DO IF (objekt_j = !j).
    COMPUTE !CONCAT('VAR',!j)
      = distanz.
  ELSE.
    COMPUTE !CONCAT('VAR',!j)
      = -1.
  END IF.
!DOEND.

/*** die n-Quadrat Paare zu   ***/
/*** n Fällen aggregieren     ***/
MISSING VALUES
  VAR1 TO !CONCAT('VAR',!n) (-1).
AGGREGATE OUTFILE=*
  /PRESORTED
  /BREAK=objekt_i
  /VAR1 TO !CONCAT('VAR',!n)
    = FIRST(VAR1 TO
!CONCAT('VAR',!n)).

/*** Variablen ROWTYPE_ und   ***/
/*** VARNAME_ erzeugen        ***/
STRING ROWTYPE_ VARNAME_ (A8).
COMPUTE ROWTYPE_ = 'PROX'.
VALUE LABELS ROWTYPE_
  'PROX' 'DISSIMILARITY'.
COMPUTE VARNAME_ = CONCAT('VAR',
  LTRIM(STRING(objekt_i,F5),' ')).
!ENDDEFINE.

DEFINE !erzmat (m = !TOKENS(1)
  /n = !TOKENS(1)
  /mat=!TOKENS(1)).
  * Distanzmatrix erzeugen.
!norme m=!m table='mittel.sav'.
  * Abstände normieren.
COMPUTE distanz =
  SUM(d1 TO !CONCAT('d',!m)).
  * Gesamtabstand berechnen.
SAVE OUTFILE='paare.sav'
  /KEEP objekt_i objekt_j
   d1 TO !CONCAT('d',!m) distanz.
  * Paare sichern.
MATCH FILES FILE=*
  /KEEP objekt_i objekt_j
     distanz.
  * überflüssige Variablen aus
    dem Arbeitsbereich löschen.
!matrix n=!n.
  * Distanzmatrix erstellen.
SAVE OUTFILE=!mat
  /KEEP ROWTYPE_ VARNAME_
    VAR1 TO !CONCAT('VAR',!n).
GET FILE='objekte.sav'.
!ENDDEFINE.

DEFINE !zerleg (var = !TOKENS(1)
  /m = !TOKENS(1)).
SAVE OUTFILE='ergebnis.sav'
  /RENAME=(!var=VAR0)
  /KEEP objekt_i VAR0.
GET FILE='paare.sav'.
  * Paare wiederherstellen.
!hinzu1 nr=0
  table='ergebnis.sav'.
  * Zerlegungsvariable
    hinzufügen.
!nomi von=0 bis=0.
  * Zerlegungsvariable
    berechnen.
AGGREGATE
  /OUTFILE=*
  /BREAK=d0
  /d1 TO !CONCAT('d',!m)
    =SUM(d1 TO !CONCAT('d',!m))
  /distanz=SUM(distanz).
  * Streuung zerlegen.
FLIP.
COMPUTE summe =
  (VAR001 + VAR002) / 100.
COMPUTE inner  = VAR001  / summe.
COMPUTE zwischen = VAR002 / summe.
LIST CASE_LBL inner zwischen.
  * Zerlegung darstellen.
!ENDDEFINE.

DEFINE !beisp ().
GET TRANSLATE FILE='beispiel.dat'
  /TYPE=TAB.
  * Daten zu 5 Studierenden
    mit 3 Merkmalen einlesen.
!rangord von=2 bis=2.
  * Ausprägungen von Merkmal Nr. 2
    in Ränge umwandeln.
!erzpaar m=3 n=5.
  * 25 Paare erzeugen.
!metri von=1 bis=1.
!ordi  von=2 bis=2.
!nomi  von=3 bis=3.
  * Abstände berechnen.
!erzmat m=3 n=5 mat='matrix.sav'.
  * Distanzmatrix erzeugen.
CLUSTER /MATRIX=IN('matrix.sav')
  /METHOD=COMPLETE
  /SAVE=CLUSTERS(2).
  * mit Complete-Linkage 2 Klassen
    bilden.
FREQUENCIES CLU2_1.
  * Klassengrößen auflisten.
!zerleg var=CLU2_1 m=3.
  * Streuungszerlegung auflisten.
!ENDDEFINE.


/***      Hauptabschnitt      ***/
!beisp.
```