Teststatistiken zur Bestimmung der Clusterzahl für Quick Cluster

Bacher, Johann

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Teststatistiken zur Bestimmung der Clusterzahl für QUICK CLUSTER

von Johann Bacher[1]

Zusammenfassung

Der SPSS-Anwender/die SPSS-Anwenderin, der/die eine Clusteranalyse rechnen will, sieht sicht einer grotesken Situation gegenübergestellt. Auf der einen Seite kann mit SPSS zwar eine Clusteranalyse durchgeführt werden, andererseits fehlen aber Teststatistiken zur Bestimmung der Clusterzahl. Ziel dieses Beitrages ist es, Hilfestellungen für Interessierte anzubieten. Dargestellt wird, wie mit Hilfe eines Syntaxprogramms Teststatistiken zur Festlegung der Clusterzahl berechnet werden können. Da diese häufig keine eindeutige Entscheidung ermöglichen, werden weitere Beurteilungskriterien erörtert.

Abstract

Users of SPSS who want to perform a cluster analysis are confronted with a grotesque situation. They can calculate clusters with SPSS, but SPSS provides no test criteria to determine the number of clusters. The paper intends to help interested users. It is shown how test criteria can be computed with a syntax programme. These tests frequently don't offer definite decisions. Hence further criteria are discussed.

1 Clusteranalyseverfahren in SPSS

Ziel der Clusteranalyse ist das *Auffinden von homogenen Gruppen bzw. Clustern* für eine Menge von Objekten. Damit ist gemeint, dass Objekte auf der Grundlage von bestimmten Merkmalen, den sogenannten Klassifikationsmerkmalen, so zu Clustern zusammengefasst werden, dass die Objekte eines Clusters einander sehr ähnlich sind, während sich Objekte aus unterschiedlichen Clustern deutlich voneinander unterscheiden. Objekte können Individuen (Personen), bestimmte Produkte, Dienstleistungen, Berufe, Räume, Orte oder andere beliebige Aggregate oder Verhaltensweisen sein. Geclustert werden können auch Vari-

1 Anschrift des Autors: Prof. Dr. ***Johann Bacher***, Lehrstuhl für Soziologie, WISO-Fakultät der Universität Erlangen-Nürnberg, Findelgasse 7-9, D-90402 Nürnberg, e-mail: j.bacher@demut.at.

ablen. Die Clusteranalyse ist eine explorative Methoden, die mit wenigen Annahmen auskommt. Sie kann aber auch konfirmatorisch eingesetzt werden (***Bacher*** 1996, S. 348-352). In diesem Fall wird vergleichbar den linearen Strukturgleichungsmodellen vorab eine Clusterstruktur definiert. Daran anschließend wird bei der Datenanalyse überprüft, wie gut sich die untersuchten Daten mit der angenommenen Typologie reproduzieren lassen.

In den *Sozialwissenschaften* wurde die Clusteranalyse für verschiedene Aufgabenstellungen eingesetzt (für einen Überblick siehe ***Bacher*** 2000). Als eine häufige Anwendung von hierarchisch agglomerativen Verfahren (siehe unten) lässt sich die Klassifikation von Ländern auf der Basis unterschiedlicher Merkmale anführen. So z.B. haben ***Obinger*** und ***Wagschal*** (1998) Esping Andersens Typologie von Wohlfahrtsstaaten (***Esping-Andersen*** 1990) mit Hilfe der Clusteranalyse empirisch überprüft. ***Haller*** (1996) untersuchte die Einstellungen zur sozialen Ungleichheit auf Länderebene und konnte unterschiedliche Ländertypen ermitteln. ***Vogel*** (1993) und ***Bacher*** (1999) - um ein weiteres Beispiel zu nennen - gehen der Frage nach, ob sich Länder hinsichtlich ihrer sozialen und ökonomischen Entwicklung zu Clustern zusammenfassen lassen.

Ein weiteres *typischen und intuitiv naheliegendes Anwendungsgebiet* ist die Lebensstil- und Milieuforschung, bei der angenommen wird, dass es Gruppen (Typen, Cluster) von Personen mit ähnlichen Ressourcen, Wertorientierungen, Besitztümern bzw. Verhaltensweisen gibt. Da hierbei Personen geclustert werden, kommen partionierende Verfahren, wie das K-Means-Verfahren (siehe unten), zum Einsatz, da i.d.R. eine größere Fallzahl vorliegt. Als Beispiele zur empirischen Bestimmung von Lebensstilen und sozialen Milieus mit Hilfe der Clusteranalyse lassen sich die Arbeiten von ***Lüdtke*** (1989), ***Giegler*** (1994) und ***Gutsche*** (2000) anführen. Auch das in dieser Arbeit verwendete empirische Beispiel (siehe unten) lässt sich diesem Anwendungsfeld zurechnen.

Verfahren der Clusteranalyse unterscheiden sich dadurch, wie die Vorstellung der Homogenität innerhalb der Cluster operationalisiert wird, wie die Cluster berechnet und wie die Objekte zu den Clustern zugeordnet werden. Nach der Zuordnung der Objekte zu den Clustern lassen sich *drei sehr allgemeine Verfahrenstypen* unterscheiden (***Bacher*** 1996, S. 4-6): *unvollständige Verfahren*, bei denen der Anwender/die Anwenderin auf der Grundlage einer graphischen Darstellung der Objekte in einem niedrigdimensionalen Raum die Clusterbildung selbst vornimmt, *deterministische Verfahren*, bei denen jedes Objekt eindeutig einem und nur einem Cluster zugeordnet wird, und *probabilistische Verfahren*, bei denen jedes Objekt mit einer bestimmten Wahrscheinlichkeit jedem Cluster angehört. Die deterministischen Verfahren stellen die eigentlichen Clusteranalyseverfahren dar.

In SPSS für Windows stehen zwei (deterministische) Verfahren zur Verfügung[2]: die Prozeduren CLUSTER und QUICK CLUSTER. CLUSTER enthält die sogenannten *hierarchisch agglomerativen Verfahren* (***Bacher*** 1996, S. 238-278, 297-307). Bei diesen Verfahren wird zunächst von der Annahme ausgegangen, dass jedes Objekt ein selbständiges Cluster bildet. Daran anschießend werden die Cluster (Objekte) schrittweise verschmolzen (agglomeriert, daher die Bezeichnung "agglomerativ"). Bei n Objekten gibt es im ersten Schritt n-1 Cluster, im zweiten Schritt n-2 Cluster, im dritten n-3 usw. In jedem Schritt werden dabei jene zwei Cluster verschmolzen, die sich am ähnlichsten sind. Die Fusionierung wird solange vorgenommen, bis alle Objekte ein einziges Cluster bilden. Eine einmal vorgenommene Verschmelzung wird nicht mehr aufgelöst. Zwischen den einzelnen Lösungen entsteht dadurch eine Hierarchie (daher die Bezeichnung "hierarchisch"). Die Lösung mit (k-1) Clustern enthält die k-Clusterlösung, wobei zwei Cluster der k-Clusterlösung zu einem Cluster fusioniert wurden. Hierarchisch agglomerative Verfahren haben den Nachteil, dass sie sich nur für kleine Fallzahlen eignen, wie z.B. für die Klassifikation von Berufen, Gütern, Stadtteilen, Städten oder Ländern

Für Umfragedaten mit 1000 oder mehr Befragten sind sie verfahrenstechnisch (es entsteht ein unübersichtlicher Output) und rechentechnisch (der Arbeitsspeicher kann zu klein werden) ungeeignet. Bei *Datensätzen mit einer größeren Fallzahl* ist der SPSS-Anwender/die SPSS-Anwenderin auf die Prozedur QUICK CLUSTER angewiesen. *QUICK CLUSTER* enthält das sogenannte *K-Means-Verfahren*. Es handelt sich dabei ebenfalls um ein deterministisches Verfahren, das der Gruppe der partionierenden Methoden angehört. Für eine bestimmte Clusterzahl k werden die Clusterzentren (=Mittelwerte der k Cluster in den Klassifikationsvariablen, daher die Bezeichnung "k-means") so bestimmt (***Bacher*** 1996, S. 308-316; ***SPSS*** 2001, S. 449), dass die Streuungsquadratsumme in den Clustern ein Minimum wird. Homogenität in den Clustern ist bei dem K-Means-Verfahren also über die Streuungsquadratsumme in den Clustern definiert. Letztere soll minimiert werden, was gewährleisten soll, dass in Summe die Objekte eines Clusters vom Clusterzentrum nur geringfügig abweichen und sich folglich auch untereinander nur geringfügig unterscheiden[3].

Für *eine gegebene Clusterzahl k* werden die Clusterzentren iterativ bestimmt. SPSS berechnet zunächst mit Hilfe eines speziellen Algorithmus eine Startpartition, bei der k Objekte als Clusterzentren so ausgewählt werden, dass sie voneinander maximal entfernt sind (***Bacher*** 1996, S. 340-341; ***SPSS*** 2001, S. 448-449). Die gefundene Ausgangslösung wird schrittweise durch Neuzuordnung der Objekte verbessert. SPSS bricht die Iteration

2 Daneben gibt es noch weitere deterministische Verfahren, wie z.B. Repräsentantenverfahren (***Bacher*** 1996, S. 279-297).

3 Bei den hierarchisch agglomerativen Verfahren gibt es dagegen unterschiedliche Formalisierungen der Homogenitätsvorstellungen. (***Bacher*** 1996, S. 142-148)

ab, wenn die vorgegebene Höchstzahl überschritten wird oder die maximale Änderung der Clusterzentren in zwei aufeinanderfolgenden Iterationen kleiner einem bestimmten Schwellenwert ist. Die Voreinstellungen für diese Werte sind "maximale Iterationszahl = 10" und "Schwellenwert = 0,02∗kleinste. Distanz zwischen den Startclusterzentren". Beide Voreinstellungen sollten – außer es entsteht dadurch eine unzumutbare hohe Rechenzeit – geändert werden, da das Abbruchskriterium im Output nicht eindeutig nachvollziehbar ist. In den nachfolgenden Analysen wurde der Schwellenwert (Voreinstellung 0,02) gleich 0,0001 gesetzt. Dies garantiert in den meisten Fällen absolute Konvergenz. Die berechneten Clusterzentren ändern sich nicht mehr. Da hierfür i.d.R. mehr als 10 Rechenschritte erforderlich sind, sollte die Zahl der zulässigen Iterationen erhöht werden (siehe dazu das Syntaxprogramm im Anhang).

Die *zentrale und gleichzeitig schwierigste Frage der (explorativen) Clusteranalyse* ist jene nach der Zahl der Cluster. Für ihre Beantwortung wurden Teststatistiken entwickelt, die den Anwender/die Anwenderin bei der Entscheidung unterstützen sollen. In den meisten Fällen ermöglichen diese zwar keine eindeutige Festlegung der Clusterzahl, sie schränken aber die Zahl der aus formalen Gesichtspunkten zulässigen Clusterlösungen ein und sind daher in der Praxis unentbehrlich. Von daher ist es unverständlich, dass in QUICK-CLUSTER (und auch in CLUSTER) *keine Teststatistiken zur Bestimmung der Clusterzahl* zur Verfügung stehen. Dadurch entsteht eine groteske Situation: Es kann zwar mit SPSS eine Clusteranalyse gerechnet werden, es kann aber nicht beurteilt werden, wie gut die gefundene Clusterlösung ist. Daher soll nachfolgend dargestellt werden, wie sich einige Teststatistiken – interpretiert durch ein Syntax-Programm in SPSS – berechnen lassen. Dadurch soll es Interessierten ermöglicht werden, mit SPSS eine Clusteranalyse auch tatsächlich durchzuführen. Da die Teststatistiken häufig keine eindeutige Entscheidung ermöglichen, werden weitere Beurteilungskriterien behandelt. Damit dies alles nicht zu abstrakt und praxisfern wird, werden die Verfahrensschritte anhand realer Daten aus einer empirischen Untersuchung erörtert.

2. Beispiel

Im Jahr 1999 wurden 620 *Berufsschüler/Berufsschülerinnen in Nürnberg* zu ihrer Lebenssituation, ihren Werthaltungen, politischen Einstellungen und Freizeitpräferenzen befragt.[4] Die *bevorzugten Freizeitaktivitäten* wurden durch eine Reihe von Fragen erfasst. Zu Beginn des entsprechenden Fragebogenteils wurden die Schüler/Schülerinnen gebeten, ihre

4 Eine Kurzbeschreibung der Studie befindet sich in ***Lechner*** (2000, S. 37-40). Die Nürnberger Berufsschülerbefragung wurde finanziell von der Staedtler-Stiftung und dem Schul- und Kulturreferat der Stadt Nürnberg unterstützt.

Übersicht 1: Freizeitpräferenzen von Berufsschülern/Berufschülerinnen in Nürnberg

Frage: Was sind Deine liebsten Freizeitaktivitäten? (Mehrfachantworten möglich, n=617*, prozentuale Anteile der Nennungen)

Musik hören	86%
faulenzen	68%
Auf Partys gehen	68%
Fernsehen / Videos anschauen	66%
Ins Kino gehen	62%
In die Disco gehen/tanzen	61%
Sport treiben/Fitnessstudio/Sauna	54%
Einkaufsbummel/Schaufensterbummel machen	53%
Auto/Motorrad/Moped/Fahrrad reparieren und damit in der Gegend herumfahren	41%
lesen	35%
Am Computer/an Spielautomaten spielen	31%
Zeichnen/Malen/Fotografieren/Filmen	29%
Rock-/Pop-Konzerte besuchen	26%
Verbotenes tun, nämlich ...	19%
etwas anderes, nämlich ...	19%
Ein Instrument spielen / Musik machen	15%
Ins Jugendzentrum gehen	14%
Theater, Museen, Kunstausstellungen, klassische Musikkonzerte besuchen	13%
Sich beruflich weiterbilden	13%

* 3 Fälle wurden wegen Antwortverweigerung eliminiert

liebsten Freizeitaktivitäten zu nennen. Sie konnten aus einer Liste von 19 Aktivitäten eine beliebige Zahl von Freizeitbeschäftigungen auswählen (siehe Übersicht 1).

An erster Stelle der beliebtesten Freizeitaktivitäten steht mit 86 % Musik hören. Es folgen: Faulenzen (68 %), auf Partys gehen (68 %), Fernsehen (66 %), das Kino und die Disco (62 bzw. 61 %). Ein Jugendzentrum ist immerhin für 14 % noch attraktiv und rangiert knapp vor dem Besuch eines Theaters, eines Museums oder einer Kunstausstellung (13 %).

Führt man eine dimensionale Analyse mit Hilfe der Hauptkomponentenmethode und anschließender Varimax-Rotation[5] durch, ergeben sich *sechs inhaltlich gut interpretierbare Faktoren/Dimensionen*[6]. Faktor 1 wird von den Items "Zeichnen/Malen/Fotografieren/ Filmen" (0,58)[7], "lesen" (0,54), "Ein Instrument spielen/Musik machen" (0,51) und "Theater, Museen, Kunstausstellungen, klassische Musikkonzerte besuchen" (0,73) gebildet und lässt zwei Interpretationen zu: traditionelle Freizeitorientierung oder Allgemeinbildung. Auch die Validitätsprüfung der Clusteranalyse (s.u.) erbringt kein klareres Bild, welche Bezeichnung angemessener ist. Die anderen Faktoren - mit Ausnahme des letzten - lassen sich eindeutig benennen. Faktor 2 bildet eine weitgehend passive Konsumorientierung ab, die sich aus den Items "Einkaufsbummel/Schaufensterbummel machen" (0,52) , "Musik hören" (0,54), "Faulenzen" (0,67) und "Fernsehen/Video ansehen" (0,58) zusammensetzt. Faktor 3 präsentiert mit den Items "In die Disco gehen/tanzen" (0,69) und "auf Partys gehen" (0,77) eine Orientierung in Richtung Partys/Events (Partyorientierung). Faktor 4 vereint den Besuch eines Jugendzentrums (0,54), den Besuch von Konzerten (0,53; dieses Item lädt aber auch auf dem dritten Faktor mit 0,41 hoch) und Computer- bzw. Spielautomatenspiele (0,60). Da in Jugendzentren im Regelfall Spielautomaten bzw. Computer stehen und in ihnen Konzerte stattfinden, wurde dieser Faktor als Jugendzentrumsorientierung bezeichnet. Auf dem Faktor 5 lädt das Item "Sport treiben/Fitnessstudio/Sauna" (0,56) stark positiv, stark negativ lädt das Item "etwas anderes tun" (-0,57), ebenfalls eine negative Ladung besitzt das Item "Verbotenes tun" (-0,30), das Item "ins Kino gehen" lädt dagegen auf dem Faktor positiv (0,46), es weist aber auch auf dem zweiten Faktor der passiven Konsumorientierung eine hohe Ladung auf (0,48). Dem Faktor wurde daher der Name "Sportorientierung" gegeben. Den letzten Faktor schließlich bilden zwei Items: "Auto/Motorrad/Moped/Fahrrad reparieren und damit in der Gegend herumfahren" (0,73) und "sich beruflich weiterbilden" (0,59). Wegen der stärkeren Ladung des ersten Items wurde der Faktor als "Auto-/Motorradorientierung" bezeichnet. Im Unterschied zu den anderen Faktoren besitzt dieser Faktor eine geringe Plausibilität, da es auf den ersten Blick schwer fällt, Gemeinsamkeiten der beiden Items (Weiterbildung einerseits und ein Fahrzeug reparieren

5 SPSS-Standardeinstellung für die Faktorenanalyse. Die Hauptkomponentenanalyse ist strenggenommen keine Faktorenanalyse, da sie nur gemeinsame Faktoren annimmt. Bei der Faktorenanalyse wird dagegen von der Annahme ausgegangen, dass neben gemeinsamen Faktoren spezifische Faktoren und Messfehler vorliegen (***Arminger*** 1979, S. 27-28). Eine "echte" Faktorenanalyse kann in SPSS z.B. mit der Anweisung "FACTOR ... /EXTRACTION = PAF/ ..." gerechnet werden. Bei dieser Spezifikation wird die sogenannte Hauptachsen- oder Hauptfaktorenmethode (***Arminger*** 1979, S. 40) durchgeführt. Sie resultiert i.d.R. in numerisch kleineren Faktorladungen. Nach ***Holm*** (1998, S. 16) unterscheiden sich ab 15 Items Hauptkomponentenmethode und "echte" Verfahren der Faktorenanalyse nur geringfügig. Dies entspricht auch den Erfahrungen des Autors. In dem hier untersuchten Beispiel ergeben sich aber trotz einer Variablenzahl von 19 Items inhaltlich bedeutsame Unterschiede (siehe dazu unten).

6 Da dichotome Items faktorisiert wurden, wurde untersucht, ob die Ergebnisse durch unterschiedliche Schwierigkeitsgrade verzerrt sind. (***Bacher*** 1996, S. 126-132) Dies war nicht der Fall.

7 Faktorladungen in Klammern.

und mit diesem herumfahren andererseits) zu erkennen. Trotz dieses Schönheitsfehlers soll der Faktor in die weiteren Analysen aufgenommen werden.

Mit Hilfe einer Clusteranalyse (QUICK-CLUSTER) soll nun untersucht werden, ob sich bestimmte Präferenztypen hinsichtlich des Freizeitverhaltens identifizieren lassen. Dafür stehen zwei Möglichkeiten zur Auswahl: Die Analyse kann für die Ursprungsvariablen oder die Faktorwerte durchgeführt werden. Zwei Argumente sprechen für die Verwendung von Faktorwerten: Faktorwerte sind Summenvariablen und daher weniger von zufälligen Messfehlern behaftet. Aus Simulationsstudien ist bekannt, dass *Clusteranalyseverfahren gegenüber irrelevanten Merkmalen*, wie z.B. zufälligen Messfehlern, sensibel sind (***Bacher***, 1996, S. 164-172). Bei der Analyse mit Faktorwerten ist daher in einem geringeren Ausmaß mit Verzerrungen durch Zufallsfehler zu rechnen. Ein weiterer Vorteil kann darin gesehen werden, dass der Interpretationsaufwand geringer ist. In unserem Beispiel müssten an Stelle von 19 Items nur 6 Faktoren analysiert werden. Die Verwendung von Faktorwerten hat aber auch Nachteile. Sie sind darin zu sehen, dass sich der Forscher/die Forscherin von den Ausgangsdaten entfernt. Bei der Interpretation der Faktorwerte ist ein Rückgriff auf die Einzelitems, die eine Dimension bilden, strenggenommen nicht mehr zulässig.[8] Diesem Nachteil kann man insofern begegnen, dass zur Kontrolle für eine gefundene Clusterlösung die Anteils- bzw. Mittelwerte in den Ursprungsvariablen berechnet werden. Ein weiterer Nachteil besteht allgemein darin, dass bei Verwendung von Faktoren von der Annahme ausgegangen wird, dass in jedem Cluster dieselbe dimensionale Struktur vorliegt. In unserem Beispiel schließlich kommt als negativer Sachverhalt hinzu, dass die Ergebnisse der dimensionalen Analyse nicht eindeutig sind. Rechnet man an Stelle der Hauptkomponentenanalyse eine Hauptfaktorenmethode ergeben sich z.T. andere inhaltliche Dimensionen[9].

Für die nachfolgende Analyse fiel die Entscheidung dennoch auf die Verwendung der Hauptkomponenten, da methodische Aspekte im Vordergrund stehen und der Interpretationsaufwand aus Platzgründen gering gehalten werden sollte.

8 Diesen Nachteil kann man dadurch umgehen, dass für die Ursprungsvariablen Clustermittelwerte berechnet werden.

9 Die Faktoren "traditionelle Aktivitäten" (F1), "passive Konsumorientierung" (F2) und "Party-/Eventorientierung" (F3) und "Sport" (F5) bleiben, z.T. werden einzelne Items etwas anderes zugeordnet. Der Faktor "Jugendzentrumsorientierung" (F4) verschwindet. An seine Stelle tritt ein Faktor, der als Medienorientierung bezeichnet werden könnte und die Items "fernsehen", "Computerspiele" und "Kino" umfasst. Der letzte Faktor schließlich gewinnt klarere Konturen und wird nur mehr von dem Item ""Auto/Motorrad/Moped/Fahrrad reparieren und damit in der Gegend herumfahren" gebildet.

3. Teststatistiken zur Bestimmung der Clusterzahl

Ziel der nachfolgenden Clusteranalyse war die Untersuchung der Frage, ob sich Präferenztypen hinsichtlich des Freizeitverhaltens identifizieren lassen. Liegen keine theoretischen Vorstellungen über die Zahl der Typen (Cluster) vor, bedeutet dies, dass zunächst die Zahl der Cluster bestimmt werden muss. Als formale Testgrößen können dazu verwendet werden: die erklärte Streuung (ETA^2_k), die relative Verbesserung gegenüber der vorausgehenden Lösung (PREk) und die F-MAX-Teststatistik (F-MAXk). Die Größen sind wie folgt definiert (***Bacher*** 1996, S. 316-322):

$$ETA_k^2 = 1 - \frac{SQ_{in}(k)}{SQ_{ges}} = 1 - \frac{SQ_{in}(k)}{SQ_{in}(1)},$$

$$PRE_k = 1 - \frac{SQ_{in}(k)}{SQ_{in}(k-1)} \quad \text{und}$$

$$F - MAX_k = \frac{SQ_{zw}(k)/k - 1}{SQ_{ges}/n - k},$$

wobei SQ_{ges} die Gesamtstreuungsquadratsumme ist. $SQ_{in}(k)$ ist die Streuungsquadratsumme innerhalb von k Clustern (Fehlerstreuung bei k Cluster) und $SQ_{zw}(k)$ die Streuungsquadratsumme zwischen k Clustern ("erklärte" Streuung bei k Clustern). Es gilt:

SQ_{ges} = konstant,

$$SQ_{ges} = (n-1) \cdot \sum_{j=1}^{p} s_j^2,$$

$SQ_{ges} = SQ_{in}(k) + SQ_{zw}(k)$ und

$SQ_{ges} = SQ_{in}(1)$.

Die Gesamtstreuungsquadratsumme ist eine konstante Größe. Sie setzt sich zusammen aus den Varianzen (s^2_j) der Klassifikationsmerkmale (j=1, 2, ..,p), die aufaddiert und anschließend mit der Fallzahl abzüglich 1 (n-1) multipliziert werden.[1] Die Gesamtstreuungsquadratsumme ist gleich der Streuungsquadratsumme zwischen k Clustern plus der Fehlerstreuung bei k Clustern. Die letzten beiden Größen sind nicht konstant, sondern hängen von der Clusterzahl k ab. Für die Lösung mit einem Cluster ist die Streuung zwischen den Clustern gleich Null, da es nur ein Cluster gibt. Die Gesamtstreuungsquadratsumme ist daher gleich der Streuung innerhalb der Cluster für die 1-Clusterlösung.

1 Diese Berechnungsformel gilt nur, wenn fehlende Werte fallweise ausgeschieden werden. Die beiden nachfolgenden Formel gelten allgemein.

In unserem Beispiel wird mit standardisierten Faktorwerten gerechnet. Die Varianzen sind daher gleich 1. Von den 620 Personen werden 617 in die Analyse einbezogen, n ist gleich 617 (Fehlende Werte wurden fallweise ausgeschieden!). Da sechs Faktorwerte in die Analyse eingehen, ist die Gesamtstreuungsquadratsumme gleich 3696 (= (617-1) ∗ 6)

Die Gesamtamtstreuungsquadratsumme kann in SPSS wie folgt berechnet werden (siehe Syntaxprogramm im Anhang):

- Es wird eine Analyse mit der Clusterzahl 1 gerechnet (=Lösung 1). Für jeden Fall wird die euklidische Distanz zum Clusterzentrum zwischengespeichert. In dem Syntaxprogramm wird dazu die Variable D1 verwendet.
- Die euklidischen Distanzen werden quadriert. In dem Syntaxprogramm geschieht dies durch die Anweisung COMPUE D1Q = D1*D1.
- Durch eine anschließende MEANS-Anweisung wird die Gesamtstreuungsquadratsumme berechnet. Sie ist gleich der Summe der Variablen D1Q.
- Schließlich wird für weitere Berechnungen die Gesamtstreuungsquadratsumme als neue Variable abgespeichert. Dazu werden die Daten mit AGGREGATE aufsummiert (aggregiert).

Der Berechnungsmodus im Syntaxprogramm gilt allgemein für unterschiedliche Methoden der Behandlung von fehlenden Werten.

Nach diesem Vorgriff auf das Beispiel und die Syntax zurück zu den Teststatistiken! ETA^2_k gibt die durch k Cluster *erklärte Streuung* an, PRE_k die durch die k Cluster erzielte *relative Verbesserung* gegenüber der vorausgehenden Lösung. PRE_k ist ein "proportional reduction of error"-Koeffizient. Es gilt:

$ETA_{k=1}^2=0$,

$PRE_{k=1}$ nicht definiert und

$PRE_{k=2} = ETA_{k=2}^2$.

$F\text{-}MAX_k$ wird analog der F-Statistik der Varianzanalyse aus dem Verhältnis von erklärter zur nicht erklärten Streuung berechnet. Da das K-Means-Verfahren die erklärte Streuung maximiert, besitzt F-MAX keine F-Verteilung (***Bacher*** 1996, S. 317). Eine statistische Signifikanzprüfung ist anders als bei der Varianzanalyse daher nicht möglich. Der *Vorteil* von $F\text{-}MAX_k$ gegenüber der erklärten Streuung ETA_k^2 ist darin zu sehen, dass die Abhängigkeit von der Clusterzahl, die bei ETA_k^2 dazu führt, dass bei einer größeren Clusterzahl automatisch mehr Varianz erklärt wird, beseitigt wird.

Zur Bestimmung der Clusterzahl werden die Testgrößen für k=1 bis k=k_{max} (z.B. 12) berechnet. Die Clusterzahl k kann wie folgt festgelegt werden.

- ETA_k^2: Die Clusterzahl ist gleich der Lösung mit k Clustern, wo nachfolgende Lösungen mit k+1 und mehr Clustern keine "wesentlichen" Verbesserungen der erklärten Streuung erbringen. Ergeben sich beispielsweise folgende Werte für ETA_k^2: $ETA_{k=1}^2$=0, $ETA_{k=2}^2$=0,60, $ETA_{k=3}^2$=0,70, $ETA_{k=4}^2$=0,80, $ETA_{k=5}^2$ =0,81, $ETA_{k=6}^2$=0,82, $ETA_{k=7}^2$=0,83 ..., so lassen sich ab 4 Cluster keine nennenswerten Verbesserungen mehr beobachten (ETA^2 für die 5-Clusterlösung ist 0,81 im Vergleich zu 0,80 usw.). Man wird sich daher für vier Cluster entscheiden.

- PRE_k: Die Clusterzahl wird ebenfalls dort festgelegt, wo nachfolgende Lösungen zu keinen wesentlichen Verbesserungen führen. Dies ist durch kleine PRE-Koeffizienten ersichtlich. Werden beispielsweise für PRE_k folgende Werte berechnet: $PRE_{k=1}$=nicht definiert, $PRE_{k=2}$=0,60, $PRE_{k=3}$=0,25, $PRE_{k=4}$=0,33, $PRE_{k=5}$=0,05, $PRE_{k=6}$=0,05, $PRE_{k=7}$=0,06 ..., so ist zwischen vier und fünf Clustern wiederum ein deutlicher Abfall erkennbar. Ab vier Cluster werden deutlich geringere Verbesserungen erzielt. In dem Beispiel wird man daher die Clusterzahl bei vier festlegen.

- $F\text{-}MAX_k$: Die Clusterzahl ist gleich der Lösung mit dem maximalen F-MAX-Wert. Für k=2 liefert das Kriterium keine eindeutige Entscheidung. Bei k=k_{max} muss die Clusterzahl um mindestens eins - am besten aber um eine größere Zahl - erhöht und eine erneute Analyse gerechnet werden. Ergeben sich folgende $F\text{-}MAX_k$ Werte: $F\text{-}MAX_{k=2}$=40, $F\text{-}MAX_{k=3}$=60, $F\text{-}MAX_{k=4}$=30, $MAX_{k=5}$=30 usw., so ist – abweichend von den vorausgehenden fiktiven Beispielen – die Clusterzahl gleich drei.

Die Clusterzahl k kann auch graphisch bestimmt werden. Dazu wird ein Screediagramm erstellt. In die X-Achse wird die Clusterzahl eingezeichnet, in die Y-Achse werden die entsprechenden erklärten Streuungen eingetragen. Tritt keine wesentliche Verbesserung mehr auf, erkennt man dies durch einen Knickpunkt und einen anschließenden flacheren, im Idealfall zur X-Achse parallelen Verlauf.

Das im Anhang wiedergegebene Syntaxprogramm berechnet die behandelten Testgrößen. Es werden zehn Clusteranalysen durchgeführt, beginnend mit der Clusterzahl 1. Für jede Clusterlösung wird für jeden Fall mit der Anweisung /SAVE ... DISTANCE(D<k>) die euklidische Distanz zum Clusterzentrum abgespeichert. Für die 1-Clusterlösung stehen diese in der Variablen D1, für die 2-Clusterlösung in der Variablen D2 usw. Die Variablen D1, D2, .. werden anschließend mit Hilfe von COMPUTE-Befehlen quadriert. Es entstehen die neuen Variablen D1Q, D2Q, D3Q usw. Für diese Variablen werden mit MEANS die Summen berechnet. Diese sind gleich den entsprechenden Fehlerstreuungen. Die Summe von D1Q ist gleich der Fehlerstreuungsquadratsumme (=Streuungsquadratsumme inner-

halb der Cluster) der 1-Clusterlösung, jene von D2Q gleich der Fehlerstreuungsquadratsumme der 2-Clusterlösung usw. Für die Berechnung der Testgrößen wird mit AGGREGATE eine neue Datenmatrix erzeugt, die nur aus einer Zeile besteht und die Variablen C1, NN und DD1 bis DD10 enthält. C1 ist für alle Fälle gleich 1. In NN steht die Fallzahl, in DD1 die Fehlerstreuungsquadratsumme der 1-Clusterlösung usw. Aus diesen Variablen werden im ersten Schritt entsprechend der oben angeführten Definition mit den Anweisungen COMPUTE ETA<k> = 1 - DD<k> / DD1 mit k = 1, 2, 3, .. die erklärten Streuungen berechnet. Dabei wird auf die Tatsache zurückgegriffen, dass die Fehlerstreuungsquadratsumme der 1-Clusterlösung (=DD1) gleich der Gesamtstreuung ist. Mit LIST VAR=ETA1 TO ETA10 werden die erklärten Streuungen im Anschluss an ihre Berechnung ausgegeben. Der PRE-Koeffizient für k Cluster ist definiert als 1 - $SQ_{in}(k) / SQ_{in}(k-1)$. Diese Formel lässt sich durch folgenden COMPUTE-Befehle realisieren: COMPUTE PRE<k> = 1 - DD<k> / DD<k-1>. Die Ausgabe erfolgt wiederum durch einen nachfolgenden LIST-Befehl. Als letztes berechnet das Syntaxprogramm mit der Anweisung COMPUTE FMAX<k> = ((DD1-DD<k>) / (<k>-1)) / (DD<k> / (nn - <k>)) die F-MAX-Statistik und gibt diese aus.

4 Anwendung der Teststatistiken für das Beispiel

Für die Berufsschüleruntersuchung ergeben sich folgende erklärte Streuungen:

ETA1	ETA2	ETA3	ETA4	ETA5	ETA6	ETA7	ETA8	ETA9	ETA10
,00	,12	,20	,28	,36	,40	,44	,48	,49	,52

Per Definition erklärt die 1-Clusterlösung 0 %. Die 2-Clusterlösung erklärt 12 %, die 3-Clusterlösung 20 % usw. Die erklärte Streuung nimmt bis zur 5-Clusterlösung deutlich zu (jeweils absolut um 8%). Nach fünf Clustern fällt der Zuwachs geringer aus. Dieser Befund spricht für eine 5-Clusterlösung. Nach acht Clustern ergibt sich eine weitere Abnahme. Die Zunahme der erklärten Streuung beträgt zwischen 5 und 8 Clustern jeweils 4% und fällt anschließend auf 1%. Auch die 8-Clusterlösung könnte somit formal noch als Lösung akzeptiert werden.

Betrachtet man die durch die PRE-Koeffizienten erfasste relative Verbesserung gegenüber der jeweils vorausgehenden Lösung, so ergeben sich folgende Werte:

PRE1	PRE2	PRE3	PRE4	PRE5	PRE6	PRE7	PRE8	PRE9	PRE10
-99(a)	,12	,10	,10	,10	,07	,07	,06	,02	,07

(a) nicht definiert

Bis zur 5-Clusterlösung haben die PRE-Koeffizienten Werte von 10%. Daran anschließend gehen sie zurück. Ein weiterer Rückgang ist bei 9-Clustern, so dass man wiederum auch 8 Cluster annehmen könnte. Bei 10 Clustern ergibt sich ein erneuter Anstieg. Die nachfolgenden Werte gehen erneut zurück (PRE11=0,04 und PRE12=0,03). Daraus folgt, dass formal auch 10 Cluster akzeptierbar wären.

Die F-MAX-Statistik schließlich nimmt folgende Werte an:

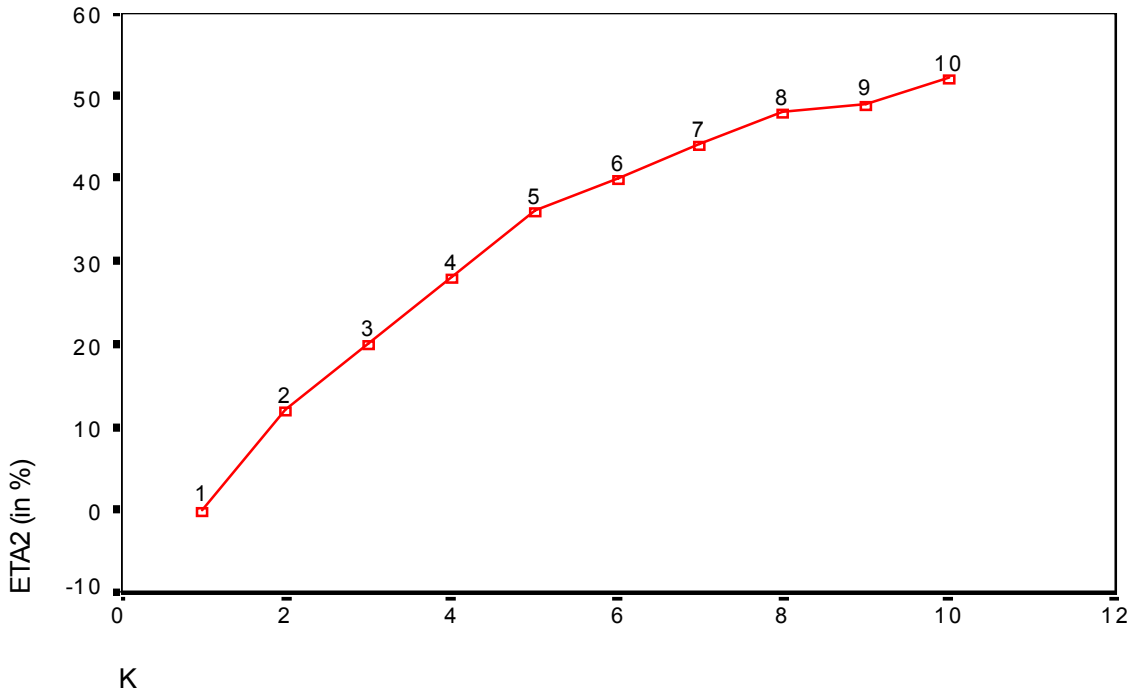
FMAX1	FMAX2	FMAX3	FMAX4	FMAX5	FMAX6	FMAX7	FMAX8	FMAX9	FMAX10
-99(a)	83,26	78,91	80,89	84,97	82,23	80,59	79,59	71,85	73,03

(a) nicht definiert

Das Maximum liegt bei fünf Clustern. In dem Beispiel weisen somit alle Teststatistiken auf eine 5-Clusterlösung hin. Weitere formale Hinweise gibt es für eine 8- und 10-Clusterlösung, die aber nicht mehr durch alle Teststatistiken abgesichert sind.

Auch graphisch lässt sich bei 5 Cluster ein leichter Knickpunkt erkennen (siehe Abbildung 1). Nach 5 Clustern verläuft die Kurve flacher. Das Ideal eines zur X-Achse parallelen Verlaufs wird nicht erreicht. Bei 8 Clustern ist eine weiterer Knickpunkt erkennbar, bei 9 ein erneuter Anstieg.

Abbildung 1: Screediagramm zur Bestimmung der Clusterzahl



5 Weitere Beurteilungskriterien

5.1 Inhaltliche Interpretierbarkeit

Die Clusterzahl sollte nicht alleine durch formale Teststatistiken bestimmt werden. In der Praxis ist dies oft auch gar nicht möglich, da formal mehrere Lösungen zulässig sind. *Eine Clusterlösung sollte auf jeden Fall inhaltlich interpretierbar sein,* d.h. allen Clustern müssen inhaltlich und theoretisch sinnvolle Namen gegeben werden können (***Bacher*** 1996, S. 151-172, 336-344). Sind für eine bestimmte Clusterlösung die formalen Kriterien erfüllt, ist sie aber nicht interpretierbar, so ist die Lösung unbrauchbar.

Übersicht 2: Ergebnisse von QUICK-Cluster

Final Cluster Centers.

Cluster	F1	F2	F3	F4	F5	F6
1	-,3676	,5655	,5443	-,5777	-,1461	-,2748
2	-,1702	-1,3612	-,0502	-,2734	,5198	-,2459
3	-,1576	,2497	-1,3825	,0518	-,4373	-,1354
4	,8967	,1627	-,1022	-,4259	,3783	1,6873
5	,3676	,0938	,5958	1,3405	-,0656	-,1743

unterstrichene Werte = für jeweiliges Cluster charakteristisch.

F1 bis F6 sind die Faktorwerte: F1 = Allgemeinbildung (oder traditionelle Freizeitorientierung); F2 = Konsumorientierung; F3 = Partyorientierung; F4 = Jugendzentrumsorientierung; F5 = Sportorientierung; F6 = Auto-/Motorradorientierung

Die fünf Cluster lassen sich aufgrund der Clusterzentren (siehe Übersicht 2) wie folgt benennen:

- *Cluster 1: konsumorientierte PartygeherInnen.* In ihrer Freizeit sind die Jugendlichen dieses Clusters zum einen sehr aktiv, sie gehen gerne auf Parties, ins Kino, bummeln durch Einkaufsstraßen, gehen Shopping usw. Zum anderen sind sie sehr passiv, sie faulenzen, sehen fern oder Video, hören Musik. Gemeinsam ist allen diesen Merkmalen eine ausgeprägte Konsumorientierung, die im Cluster gleichzeitig mit einer negativen Einstellungen zu allen nicht privatwirtschaftlichen und nicht kommerziellen Angeboten, wie z.B. dem Besuch eines Jugendzentrums, auftritt. Mit 189 Fällen (31 %) ist dies das größte Cluster.

- *Cluster 2: SportlerInnen.* Personen, die diesem Cluster angehören, betreiben in ihrer Freizeit am liebsten Sport. Passiven Konsum lehnen sie ab. Man kann sich unter diesem Cluster sportbegeisterte Jugendliche vorstellen, die mit Ausnahme des Besuchs eines Kinos für Shopping und Konsum wenig übrig haben. Zum Sport gehört dabei auch der Besuch eines Fitnessstudios. Dieses Cluster umfasst 117 Personen (19 %).

- *Cluster 3: Event- und SportablehnerInnen.* Dieses Cluster ist durch eine ablehnende Haltung von Sport und mit Einschränkungen von Events (Parties) gekennzeichnet. Es wird von 118 Befragten (19 %) gebildet.

- *Cluster 4: TraditionalistInnen.* Diese bevorzugen in der Freizeit traditionelle Tätigkeiten. Zum einen spielen sie gerne ein Musikinstrument, besuchen bevorzugt eine Theateraufführung, eine Kunstausstellung oder gehen in ein klassisches Konzert. Zum anderen gehört zu ihren Lieblingsbeschäftigungen das Herumfahren mit Fahrrad, Moped, Motorrad oder Auto und Weiterbildung. Man kann sich unter diesem Cluster Jugendliche vom Land vorstellen, die in einer Blasmusik mitspielen, gerne ein Konzert, Theater oder eine Kunstausstellung in einer größeren Stadt besuchen, und baldmöglichst ein billiges und daher reparaturanfälliges Moped, Motorrad oder Auto erwerben, um mobil zu sein. Vorstellbar ist aber auch, dass es sich bei diesem Cluster um ein städtisches (alternatives) Bildungsbürgertum handelt, das sich durch ein Lehre ganzheitlich bilden möchte. Dafür wird auch ein eigenes Instrument gespielt und an Theatervorstellungen, Kunstausstellungen und klassischen Konzerten teilgenommen. Das Fahrrad ist entsprechend der alternativen Orientierung das bevorzugte Fortbewegungsmittel. Schließlich könnte es sich bei dem Cluster auch um ein vorstädtisches Aufstiegsmilieu handeln, das einerseits den bereits erzielten Aufstieg durch den Besitz materieller Güter (hier ein Auto, ein Motorrad oder ein Moped) demonstriert, andererseits weiß, dass nur Bildung den Aufstieg absichern kann. Aufgrund der Klassifikationsmerkmale ist eine Entscheidung zwischen den Interpretationen nicht möglich. Mit 70 Fällen (11 %) ist dies das kleinste der fünf Cluster.

- *Cluster 5: JugendzentrumsbesucherInnen.* Die Jugendlichen dieses Clusters besuchen in ihrer Freizeit am liebsten ein Jugendzentrum, das ihnen kostengünstige Möglichkeiten für Partys und andere Veranstaltungen und Aktivitäten bietet (hoher Wert in F3). Diesem Cluster gehören 20 % (123 Fälle) an.[10]

In dem Beispiel erfüllt die 5-Clusterlösung das Kriterium der inhaltlichen Interpretierbarkeit großteils. Für Cluster 4 sind drei Interpretationen (typische Landjugend, städtisches

10 Problematisch an dieser Interpretation ist, dass nur 14% der befragten Jugendlichen angaben, gerne ins Jugendzentrum zu gehen.

alternatives Bildungsbürgertum, vorstädtisches Aufstiegsmilieu) denkbar. Jede der drei Interpretationen weist einige Schwächen auf. Bei der ersten Interpretation könnte als erklärungsbedürftig betrachtet werden, warum Landjugendliche Allgemeinbildung einschließlich Theateraufführungen, klassische Konzerte und Kunstausstellungen zu ihren liebsten Freizeittätigkeiten zählen. Bei der zweiten Interpretationsmöglichkeit könnte die Wahl des Fortbewegungsmittel, das den sechsten Faktor F6 bildet, zu Dissonanzen führen, wenn sich herausstellen würde, dass nicht ein Fahrrad, sondern ein Moped, Motorrad oder Auto zur Fortbewegung genützt wird. Gegen die dritte Interpretation lässt sich einwenden, warum überhaupt eine Lehre gemacht wird, wenn Allgemeinbildung als wichtiges Kriterium zur Absicherung des Aufstiegs gilt. Weitere Analysen können möglicherweise Anhaltpunkte liefern, welche Interpretation plausibler ist.

Häufig ist das Kriterium der inhaltlichen Interpretierbarkeit nicht für alle Cluster erfüllt und es bleiben - wie in unserem Beispiel - ein oder gelegentlich auch zwei oder mehr inhaltlich schwer interpretierbare Cluster übrig. In diesem Fall existieren mehrere Möglichkeiten: (a) die schlechte Interpretierbarkeit einiger Cluster könnte akzeptiert werden, (b) es könnte eine andere Clusterlösung verwendet werden (in Frage käme die 8-oder die 10-Clusterlösung; siehe oben), (c) weiterer Klassifikationsvariablen könnten hinzugenommen werden oder (d) es könnte die Vorstellung verworfen werden, dass es Cluster gäbe. Die Entscheidung für eine der Möglichkeiten hängt von der Anwendungssituation und dem Gesamtbefund einschließlich der Stabilitäts- und Validitätsprüfung (siehe unten) ab. Eine allgemeine Regel lässt sich nicht benennen.

5.2 Stabilität

Eine Clusterlösung sollte nicht nur formal durch Teststatistiken zur Bestimmung der Clusterzahl abgesichert und inhaltlich gut interpretierbar sein, sondern sie sollte darüber hinaus die Kriterien Stabilität und Validität erfüllen.

Eine *Clusterlösung* wird dann als *stabil* bezeichnet, wenn geringfügige, nicht substantielle Modifikationen zu keinen oder vernachlässigbaren Änderungen der Ergebnisse führen. *Zwei Arten der Stabilität* sind zu unterscheiden: (a) *Stabilität der Zuordnung der Fälle* zu den Clustern und (b) *Stabilität der Clusterzentren*. Wir wollen hier nur das Vorgehen für den Fall (a) erörtern, da es sich "leicht" realisieren lässt und das strengere Kriterium der beiden ist. Eine Methode der Stabilitätsprüfung von Clusterzentren wird in ***Bacher*** (1996, S. 343-344) behandelt.

Inhaltlich irrelevant sind beim K-Means-Verfahren i.d.R. das Startwertverfahren und die zufällige Elimination von einigen Personen. Änderungen hinsichtlich dieser Kriterien sollten keinen Einfluss auf die Ergebnisse haben[11].

Das allgemeine Prinzip der Stabilitätsprüfung besteht deshalb darin, dass mehrere zulässige Lösungen berechnet werden und durch geeignete Teststatistik geprüft wird, wie gut die Lösungen übereinstimmen. Gewünscht ist eine perfekt Übereinstimmung.

Die Prozedur *QUICK CLUSTER* bietet *keine Verfahren zur Stabilitätsprüfung* an. Man ist daher wiederum auf die Verwendung von Syntaxprogrammen[12] angewiesen. Eine einfache Möglichkeit, um z.B. die Stabilität gegenüber dem Startwertverfahren zu testen, besteht darin, dass eine zufällige Ausgangspartition berechnet und als Startkonfiguration für eine Clusteranalyse eingelesen wird. Nach deren Durchführung liegen als Ergebnis zwei Clusterlösungen vor: die ursprüngliche Lösung mit SPSS-internen Startwerten und eine Lösung mit zufälligen Startwerten. Die *Stabilität der Zuordnungen der Fälle* lässt sich in diesem Fall einfach durch Kreuztabellierungen der unterschiedlichen Lösungen prüfen, vorausgesetzt, die Clusterzugehörigkeit wurde für beide Varianten als neue Variable abgespeichert. Wenn die beiden Lösungen, die einem Vergleich unterzogen werden, dieselbe Clusterzahl haben, kann ***Cohens*** Kappa (***SPSS*** 2001, S. 110; ***Bacher*** 1996, S. 206) verwendet werden. Bei ungleicher Clusterzahl ist Kappa nicht anwendbar[13]. Mitunter ist eine Umkodierung der Nummerierung der Cluster erforderlich, damit in der Hauptdiagonalen der Tabelle die maximalen Werte stehen. Kappa als Übereinstimmungskoeffizient prüft, ob die Übereinstimmungen in der Hauptdiagonalen überzufällig auftreten. Nach ***Fleiss*** (zit. in ***Bacher*** 1996, S. 206) sollte Kappa größer 0,45 sein. Werte zwischen 0,45 und 0,75 können als befriedigende bzw. gute Übereinstimmung interpretiert werden, Werte größer 0,75 als sehr gute oder ausgezeichnete Übereinstimmung.

Für unser Beispiel (siehe Übersicht 3) errechnet sich für Kappa mit 0,66 eine befriedigende Übereinstimmung, die mit einem Fehlerniveau von $p < 0{,}1\%$ signifikant von Null verschieden ist. *Die Zuordnungen der Fälle* zu den Clustern kann als stabil betrachtet werden. Wünschenswert wäre natürlich ein noch deutlich bessere Übereinstimmung.

11 Bei hierarchisch agglomerativen Verfahren können darüber hinaus für eine Klassifikationsaufgabe mehrere Distanzmaße und Verschmelzungsalgorithmen zulässig sein. Die Ergebnisse sollten dann auch gegenüber diesen Optionen stabil sein.

12 Das Syntaxprogramm ist auf Anfrage beim Autor erhältlich.

13 In diesem Fall kann z.B. der Rand-Index (***Bacher*** 1996, S. 278) verwendet werden, der in SPSS leider nicht verfügbar ist.

Übersicht 3: Stabilität der Zuordnung der Fälle

C5 * C5A Crosstabulation

			C5A					
			1	2	3	4	5	Total
C5	1	Count	158	4		12	15	189
		% within C5	83,6%	2,1%		6,3%	7,9%	100,0%
		% of Total	25,6%	,6%		1,9%	2,4%	30,6%
	2	Count	2	84	12	9	10	117
		% within C5	1,7%	71,8%	10,3%	7,7%	8,5%	100,0%
		% of Total	,3%	13,6%	1,9%	1,5%	1,6%	19,0%
	3	Count	5	2	100	7	4	118
		% within C5	4,2%	1,7%	84,7%	5,9%	3,4%	100,0%
		% of Total	,8%	,3%	16,2%	1,1%	,6%	19,1%
	4	Count	7	11	10	30	12	70
		% within C5	10,0%	15,7%	14,3%	42,9%	17,1%	100,0%
		% of Total	1,1%	1,8%	1,6%	4,9%	1,9%	11,3%
	5	Count		5		37	81	123
		% within C5		4,1%		30,1%	65,9%	100,0%
		% of Total		,8%		6,0%	13,1%	19,9%
Total		Count	172	106	122	95	122	617
		% within C5	27,9%	17,2%	19,8%	15,4%	19,8%	100,0%
		% of Total	27,9%	17,2%	19,8%	15,4%	19,8%	100,0%

C5 = Clusterlösung mit SPSS-internen Startwerten, C5A = Clusterlösung bei zufälligen Startwerten

Symmetric Measures

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Measure of Agreement	Kappa	,662	,022	32,272	,000
N of Valid Cases		617			

Eine genauere Betrachtung der Ergebnisse zeigt, dass das schwer zu interpretierende und kleinste Cluster C4 die Stabilität beeinträchtigt. Von den 70 Fällen dieses Clusters bei SPSS-internen Startwerten werden nur 30 (=42,9 %) bei zufälligen Startwerten wieder richtig zugeordnet. Für die anderen Cluster variieren die Werte zwischen 65,9 % (Cluster C5) und 84,7 % (Cluster C3). Auch eine Stabilitätsprüfung der Clusterzentren führte zu einem ähnlichen Ergebnis.[14] [15] Zusammenfassend ist Stabilität somit großteils gegeben. Mit

14 Ergebnisse auf Anfrage beim Autor erhältlich.

dem hier erörterten Verfahren lassen sich beliebig viele Lösungen vergleichen, vorausgesetzt die Clusterzahl ist gleich. In diesem Fall werden einfach mehrere Kreuztabellen erstellt.

5.3 Validitätsprüfung

Bei der *Clusteranalyse* wird von einem *kriterienbezogenen Konzept von Gültigkeit* ausgegangen. Eine Clusterlösung wird dann als valide bezeichnet, wenn vermutete Zusammenhänge zwischen den Clustern und Außenkriterien empirisch zutreffen. Bei der Validitätsprüfung werden Hypothesen über die Cluster getestet, in denen auf Variablen Bezug genommen wird, die nicht in die Clusterbildung eingingen. Die zu analysierenden Hypothesen ergeben sich häufig bei der Interpretation der Cluster, d.h. sie sind explorativ und ihnen liegt i.d.R. keine empirisch bewährte Theorie zugrunde.

Exemplarisch sollen zwei bei der Interpretation der Cluster entwickelte Hypothesen untersucht werden:

- H1: Das Cluster der TraditionalistInnen (Cluster 4) rekrutiert sich häufiger aus BerufsschülerInnen, die außerhalb von Nürnberg auf dem Land leben.
- H2: Das Cluster der JugendzentrumsbesucherInnen (Cluster 5) rekrutiert sich häufiger aus BerufsschülerInnen mit geringen finanziellen Möglichkeiten.

Diese Hypothesen lassen sich durch einfache Tabellierungen prüfen. Dazu ist es erforderlich, dass die Clusterzugehörigkeit der Fälle als neue Variable mit der Anweisung .../SAVE CLUSTER(<Variablenname>) ... (siehe Syntaxprogramm) abgespeichert wird. Daran anschließend wird die Variable "Clusterzugehörigkeit" mit den Variablen Wohnort (Hypothese H1) und finanzielle Möglichkeiten (Hypothese H2) kreuztabelliert. Zur Überprüfung der Hypothesen muss die Clusterzugehörigkeit jeweils in TraditionalistInnen mit den Ausprägungen "ja" und "nein" (H1) und in JugendzentrumsbesucherInnen ebenfalls mit den Ausprägungen "ja" und "nein" (H2) dichotomisiert werden.

Die Kreuztabellenanalyse kann in unserem Beispiel die beiden Hypothesen nicht bestätigen (siehe Übersicht 4 und 5). Zwischen der Zugehörigkeit zum Cluster der TraditionalistInnen und dem Wohnort besteht bei einem Fehlerniveau von 5% kein signifikanter Zusammenhang ($Chi^2 = 3{,}32$; $p = 0{,}068$). In der Tendenz zeigt sich sogar eine umgekehrte Beziehung.

15 Leider ist es nicht immer der Fall, dass schwer zu interpretierende Cluster auch mangelnde Stabilität und Validität aufweisen. Eine mathematische Beziehung zwischen den drei Kriterien Interpretierbarkeit, Stabilität und Validität ist noch nicht nachgewiesen

Übersicht 4: Ergebnisse aus der Validitätsprüfung

Crosstab

			TRAD Traditionalist.		
			,00 nein	1,00 ja	Total
V50 Wohnort	1 Nünberg	Count	237	37	274
		% within TRAD Traditionalist.	45,0%	56,9%	46,3%
	2 außerhalb	Count	290	28	318
		% within TRAD Traditionalist.	55,0%	43,1%	53,7%
Total		Count	527	65	592
		% within TRAD Traditionalist.	100,0%	100,0%	100,0%

Übersicht 5: Weiteres Ergebnis aus der Validitätsprüfung

Crosstab

			JUGZ Jugendzentrumsbes.		
			,00 nein	1,00 ja	Total
V60 verfügbares Einkommen	1 -200 DM	Count	109	27	136
		% within JUGZ Jugendzentrumsbes.	23,5%	23,5%	23,5%
	2 201-400 DM	Count	108	21	129
		% within JUGZ Jugendzentrumsbes.	23,3%	18,3%	22,3%
	3 401 u.m. DM	Count	246	67	313
		% within JUGZ Jugendzentrumsbes.	53,1%	58,3%	54,2%
Total		Count	463	115	578
		% within JUGZ Jugendzentrumsbes.	100,0%	100,0%	100,0%

TraditionalistInnen kommen häufiger aus Nürnberg.[16] Auch Hypothese H2 ist nicht zutreffend. JugenzentrumsbesucherInnen verfügen nicht über weniger finanzielle Mittel. Beide Variablen stehen in keiner signifikanten Beziehung zueinander ($Chi^2 = 1{,}51$; $p = 0{,}471$).

16 Dies legt die Vermutung nahe, dass die alternativen Interpretationen für dieses Cluster, es handle sich um ein alternatives Bildungsbürgertum oder um ein aufstiegsorientiertes Milieu, zutreffender seien. Eine Kreuztabellierung mit dem Bildungsabschluss zeigt aber, dass Personen aus dem Cluster keine höhere schulische Vorbildung haben. Die Interpretation, dass ein alternatives Bildungsbürgertum mit hoher Schulbildung vorliegt, besitzt somit ebenfalls eine geringe Wahrscheinlichkeit.

Die Validitätsprüfung konnte die Ausgangshypothesen nicht bestätigen. Dennoch wäre in dem vorliegenden Beispiel der Schluss, dass die Cluster nicht valide seien, voreilig, da sich eine Validitätsprüfung auf mehr Hypothesen stützen muss[17]. Dabei ist zu beachten, dass auch nicht signifikante Ergebnisse valide sein können, wenn in der Ausgangshypothese kein Zusammenhang vermutet wird, wenn also z.B. die Hypothese "Cluster x unterscheidet sich nicht von Cluster x* hinsichtlich der Variablen y" untersucht wird. Hinzu kommt, dass die Hypothesen häufig durch keine empirisch bewährten Theorien abgesichert sind. Kann ein vermuteter Zusammenhang nicht bestätigt werden, kann dies mehrere Ursachen haben: Die Cluster sind invalide oder der angenommene theoretische Zusammenhang ist falsch.[18]

Neben der Tabellenanalyse können in der Validitätsprüfung beliebige multivariate Verfahren eingesetzt werden, in welche die Variable "Clusterzugehörigkeit" in ihrer ursprünglichen Form oder zusammengefasst einbezogen wird. Für unser Beispiel könnte z.B. eine logistische Regression mit der abhängigen dichotomen Variablen "TraditionalistInnen" mit den Ausprägungen "ja" und "nein" und den unabhängigen Variablen Alter, schulische Bildung, soziale Herkunft, Wohnort, finanzielle Möglichkeiten und Nationalität gerechnet werden. Die Clusterzugehörigkeit kann aber auch als unabhängige Variable in eine multivariate Analyse eingehen. In unserem Beispiel könnte z.B. ein allgemeines lineares Modell gerechnet werden, bei dem angenommen wird, dass der Alkoholkonsum von der Clusterzugehörigkeit und weiteren Kontrollvariablen abhängt.

6 Zusammenfassung

Primäres Ziel dieses Beitrages war die Darstellung der Berechnung von Teststatistiken zur Bestimmung der Clusterzahl mit Hilfe eines SPSS-Syntaxprogramms anhand eines Beispiels aus der Forschung. Dies erfordert vom Benutzer bzw. von der Benutzerin einen bestimmten Programmieraufwand. Abzuwägen ist daher, ob es sich nicht langfristig lohnen würde, auf ein anderes Statistikprogramm, das entsprechende Teststatistiken anbietet, wie z.B. ALMO (***Holm*** 1999)[19], umzusteigen.

Die behandelten Teststatistiken sind wichtige formale Hilfskriterien zur Bestimmung der Clusterzahl. Eine Clusteranalyse sollte des Weiteren inhaltlich gut interpretierbar, stabil und valide sein. Daher wurden auch diese Kriterien behandelt. Bei der Prüfung dieser Kriterien tritt häufig das Problem auf, dass sie nur teilweise erfüllt sind. So auch im vorlie-

17 Umfangreiche Validitätsstudien hat ***Lechner*** (2000) durchgeführt.

18 Möglich ist des Weiteren, dass die Kriterienvariable nicht valide gemessen wurde.

19 Die dort enthaltenen Clusteranalyseverfahren wurden vom Autor (***Bacher*** 1999) programmiert. Informationen über ALMO können auf der Homepage der Abteilung für empirische Sozialforschung des Instituts für Soziologie der Universität Linz abgerufen werden.

genden Fall. Eines der fünf Cluster erwies sich als schwer interpretierbar und beeinträchtigte die Stabilität.

Auch ***Giegler*** (1994) kam in seiner Lebensstilanalyse zu ähnlichen Ergebnissen. Er konnte nachweisen, dass unterschiedliche Lebensstiltypologien starke Überschneidungen und Ähnlichkeiten aufweisen, sich aber in Teilaspekten unterscheiden.

Eine Patentlösung für diese Situationen (die Ergebnisse erfüllen mit Ausnahme eines oder zwei Clusters die geforderten Eigenschaften gut) gibt es nicht, sondern es hängt von der Forschungsfrage ab, welche Strategie weiter verfolgt wird. Formal bieten sich vier Möglichkeiten an: (a) Die Ergebnisse werden akzeptiert, (b) es wird eine andere Clusterlösung ausgewählt, (c) weitere Variablen werden in die Analyse aufgenommen oder (d) die Vorstellung, dass den Daten Cluster zugrunde liegen, wird verworfen. Die letzte Forderung erscheint für unser Beispiel angesichts der in vielen Teilaspekten erfüllten Anforderungen an eine brauchbare Clusterlösung zu radikal. Notwendig ist aber die Durchführung weiterer Validitätstests und die Analyse anderer Clusterlösungen, bevor die 5-Clusterlösung endgültig akzeptiert oder verworfen wird.[20]

Literatur[21]**:**

Arminger, G. 1979: Faktorenanalyse. Stuttgart.

Bacher, J. 1996:
Clusteranalyse. 2. Auflage. München-Wien.

Bacher, J. 1999:
ALMO Statistik-System. P36, P37 - Clusteranalyse. Linz.

Bacher, J. 2000:
Auffinden komplexer Zusammenhänge? - Ein Erfahrungsbericht über Erkenntnisstand und Forschungsbedarf der Clusteranalyse. ÖZS, 25. Jg., Heft 4, S. 48-60.

Esping-Andersen, G., 1990:
The Three Worlds of Welfare Capitalism. Cambridge.

Giegler, H. 1994:
Lebensstile in Hamburg. In: ***Dangschat, J.; Blasius, J.*** (Hg.): Lebensstile in Städten. Opladen. S. 255-272.

Gutsche, G. 2000:
Kriminalitätseinstellungen im Kontext von Wertorientierungen und gesellschaftlichen Leitbildern am Beispiel sozialer Milieus in den neuen Bundesländern. In: ***Ludwig-Mayerhofer, W.*** (Hg.): Soziale Ungleichheit, Kriminalität und Kriminalisierung. Opladen. S. 119-145.

20 Weiterführende Analysen wurden von ***Lechner*** (2000) durchgeführt. Neben den allgemeinen Freizeitpräferenzen wurden als weitere Präferenzen der Musik- und Filmgeschmack einbezogen. Für die Gruppe der Lehrlinge konnte sie sieben gut interpretierbare Cluster ermittelt werden (***Lechner*** 2000, S. 65-109).

21 Eine umfangreiche Bibliographie zur Clusteranalyse enthält ***Bacher*** (1996). Neuere Trends werden beschrieben in ***Bacher*** (2000).

Haller, M. 1996:
Einstellungen zur sozialen Ungleichheit im internationalen Vergleich. In: ***Haller, M.*** u.a. (Hg.): Österreich im Wandel. Wien-München, S. 188-220.

Holm, K. 1998:
ALMO Statistik-System. Handbuch. P30 - Faktorenanalyse, Nominale Faktorenanalyse, Multiple Korrespondenzanalyse. Linz.

Holm, K. 1999:
ALMO Statistik-System. Handbuch. Teil 1 - Bedienungsanleitung. Linz.

Lechner, B. 2000:
Freizeitverhalten von BerufsschülerInnen im Rahmen der Lebensstilforschung und Subkulturtheorie. Nürnberg: Bericht 2001-1 des Lehrstuhls für Soziologie.

Lüdtke, H. 1989:
Expressive Ungleichheit. Zur Soziologie der Lebensstile. Opladen.

Obinger, H.; Wagschal, U. 1998: Drei Welten des Wohlfahrtsstaates? Das Stratifizierungskonzept in der clusteranalytischen Überprüfung. In: ***Lessenich, S.; Ostner, I.*** (Hg.): Welten des Wohlfahrtskapitalismus. Frankfurt a.M., S. 109-136.

SPSS Inc. 2001:
SPSS 7.5. Statistical Algorithms. Chicago, Illinois 60606 (http://www.spss.com/tech/stat/Algorithms.htm; 1.3.2001).

Vogel, F. 1993:
Some Remarks on a Classification of the Countries of the World According to their Stage of Development. Jahrbuch f. Nationalökonomie und Statistik. S. 306-323, Stuttgart.

Anhang:

SPSS-Syntaxprogramm zur Berechnung von Teststatistiken für die Bestimmung der Clusterzahl

```
* Einlesen der Daten.
get file"c:\texte\quickcluster\lehrlinge.sav".

recode v31.01 to v31.19 (0=sysmis) (2=0).

* Faktorenanalyse der Freizeitaktivitäten.
* Die Faktorwerte werden in den Variablen F1 bis
* F6 abgespeichert.
factor var=v31.01 to v31.19/save (6 F).

* Namensgebung für die Faktorwerte.
var labels
  F1 "Allgemeinbildung"
  F2 "Konsumorientierung"
  F3 "Partyorientierung"
  F4 "Jugendzentrum-/Konzertorientierung"
  F5 "Sport"
  F6 "Auto".

* Kontrollberechnung. Mittelwerte müssen 0 sein,
* Standardabweichungen gleich 1.
means var=f1 to f6.

* Lösung 1: Es wird die 1-Clusterlösung berechnet.
* Die Clusterzugehörigkeit der Objekte und ihre Distanzen zum
* Clusterzentrum werden in die Variablen C1 und D1 geschrieben.
* D1 wird zur Berechnung der Gesamtstreuung benötigt.
QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(1) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c1) DISTANCE (d1)
  /PRINT INITIAL ANOVA.
```

```
* Lösung 2: Es wird die 2-Clusterlösung berechnet.
* Die Clusterzugehörigkeit der Objekte und ihre Distanzen zum
* Clusterzentrum werden in die Variablen C2 und D2 geschrieben.
* D2 wird zur Berechnung der Fehlerstreuung der 2-Clusterlösung
* benötigt.
QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(2) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c2) DISTANCE (d2)
  /PRINT INITIAL ANOVA.

* Lösungen 3 bis 10: Analog werden die 3- bis 10-Clusterlösung berechnet.
QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(3) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c3) DISTANCE (d3)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(4) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c4) DISTANCE (d4)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(5) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c5) DISTANCE (d5)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
```

```
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(6) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c6) DISTANCE (d6)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(7) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c7) DISTANCE (d7)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(8) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c8) DISTANCE (d8)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(9) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c9) DISTANCE (d9)
  /PRINT INITIAL ANOVA.

QUICK CLUSTER
  f1 to f6
  /MISSING=LISTWISE
  /CRITERIA= CLUSTER(10) MXITER(100) CONVERGE(.0001)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER (c10) DISTANCE (d10)
  /PRINT INITIAL ANOVA.
```

```
* Die Variablen D1 bis D10 müssen quadriert werden.
compute d1q=d1*d1.
compute d2q=d2*d2.
compute d3q=d3*d3.
compute d4q=d4*d4.
compute d5q=d5*d5.
compute d6q=d6*d6.
compute d7q=d7*d7.
compute d8q=d8*d8.
compute d9q=d9*d9.
compute d10q=d10*d10.

* Die Streuungsquadratsummen werden berechnet.
* D1Q ist die Gesamtstreuung, D2Q ist die Fehlerstreuung
* der 2-Clusterlösung, D3Q die Fehlerstreuung
* der 3-Clusterlösung usw.
means var=d1q to d10q/cells=sum.

* Elimination von Objekten, die wegen fehlender Werte
* nicht in die Clusteranalyse einbezogen wurden.
select if (c1=1).

aggregate outfile=*    /* Die Daten werden aggregiert.
    /break=c1          /* Aggregierungsvariable
    /nn = sum(c1)      /* Fallzahl zur Berechnung von F-MAX
    /dd1 to dd10 = sum(d1q to d10q)
                      /* Die quadrierten Distanzen werden
                      /* aufsummiert.

* Eine neue Datenmatrix befindet sich im Arbeitsspeicher.
* Sie besteht nur aus einem "Fall" und enthält die Variablen
* c1, dd1 bis dd8 und nn. In dd(i) steht die Fehlerstreuung
* der i-Clusterlösung. dd1 ist gleich der Gesamtstreuung.

compute eta1=1-dd1/dd1. /* Berechnung der erklärten Streuungen
compute eta2=1-dd2/dd1. /* entsprechend der im Text wieder-
compute eta3=1-dd3/dd1. /* gegebenen Formel.
compute eta4=1-dd4/dd1.
compute eta5=1-dd5/dd1.
compute eta6=1-dd6/dd1.
```

```
compute eta7=1-dd7/dd1.
compute eta8=1-dd8/dd1.
compute eta9=1-dd9/dd1.
compute eta10=1-dd10/dd1.

list var=eta1 to eta10. /* Ausgabe der erklärten Streuungen.

compute pre1=-99.       /* Berechnung der PRE-Koeffizienten.
compute pre2=1-dd2/dd1. /* PRE für die 1-Clusterlösung ist nicht
compute pre3=1-dd3/dd2. /* nicht definiert. Er erhält daher den
compute pre4=1-dd4/dd3. /* Wert -99.
compute pre5=1-dd5/dd4.
compute pre6=1-dd6/dd5.
compute pre7=1-dd7/dd6.
compute pre8=1-dd8/dd7.
compute pre9=1-dd9/dd8.
compute pre10=1-dd10/dd9.

list var=pre1 to pre10. /* Ausgabe der PRE-Koeffizienten.

compute fmax1=-99.  /* Berechnung der F-MAX-Statistiken.
compute fmax2=( (dd1-dd2)/(2-1) ) / (dd2 / (nn-2) ).
compute fmax3=( (dd1-dd3)/(3-1) ) / (dd3 / (nn-3) ).
compute fmax4=( (dd1-dd4)/(4-1) ) / (dd4 / (nn-4) ).
compute fmax5=( (dd1-dd5)/(5-1) ) / (dd5 / (nn-5) ).
compute fmax6=( (dd1-dd6)/(6-1) ) / (dd6 / (nn-6) ).
compute fmax7=( (dd1-dd7)/(7-1) ) / (dd7 / (nn-7) ).
compute fmax8=( (dd1-dd8)/(8-1) ) / (dd8 / (nn-8) ).
compute fmax9=( (dd1-dd9)/(9-1) ) / (dd9 / (nn-9) ).
compute fmax10=( (dd1-dd10)/(10-1) ) / (dd10 / (nn-10) ).

list var=fmax1 to fmax10. /* Ausgabe der F-MAX-Statistiken.
```