**Institut für Angewandtes Markt-Management**
**Prof. Dr. Wolfgang Müller**
**Reihe Forschungspapier**
**Band 9**

# Multivariate Statistik im Quantitativen Marketing - Konzeption und Anwendungsbereiche der Clusteranalyse -

*Wolfgang Müller*

Dortmund, September 2004

**Fachhochschule Dortmund**
Fachbereich Wirtschaft
Emil-Figge-Straße 44
44047 Dortmund
Telefon 0231 / 755 - 6796
Telefax 0231/ 755 - 4957
E-Mail: marktmanagement@t-online.de
www.iamm.de

# Inhaltsverzeichnis

**1. Gegenstand der Clusteranalyse.....3**

1.1. Problemstellung der Clusteranalyse.....3

1.2. Charakteristika der Clusteranalyse.....3

1.3. Einsatzfelder der Clusteranalyse im Marketing.....5

**2. Methodische Grundlagen der Clusteranalyse..... 6**

2.1. Struktur der Datenmatrix.....7

2.2. Proximitätsmaße..... 7

2.3. Fusionierungsverfahren.....11

2.4. Festlegung der Clusterzahl.....15

2.5. Clusterdiagnose.....17

**3. Hierarchische Clusteranalyse mit SPSS..... 18**

3.1. Die Datenmatrix des Demonstrationsbeispiels..... 18

3.2. Die SPSS-Auswertungsmethodik..... 20

3.3. Interpretation der Distanzmatrix.....23

3.4. Darstellung des Agglomerationsprozesses..... 24

3.5. Bestimmung der Clusterzahl.....27

3.6. Clusterdiagnose ..... 29

**4. Partitionierende Clusterzentrenanalyse mit SPSS..... 35**

4.1. Verfahrensbesonderheiten der Clusterzentrenanalyse.....35

4.2. SPSS-Methodik mit Vorinformationen über Clusterzentren.....36

4.3. SPSS-Methodik ohne Vorinformationen über Clusterzentren..... 43

**5. Fallbeispiele aus der Marketingpraxis.....50**

5.1. Serviceanalyse im Automobilhandel.....50

5.2. Strategische Wettbewerbergruppen im Großhandel..... 53

**Literaturverzeichnis..... 57**

**Dokumentation der Forschungsreihe..... 59**

# 1. Gegenstand der Clusteranalyse

## 1.1. Problemstellung der Clusteranalyse

Eines der Hauptprobleme in den empirischen Sozialwissenschaften besteht darin, umfangreiche Gesamtheiten von Objekten anhand von relevanten Merkmalen zu erfassen und in beschreibbare bzw. sachlich interpretierbare Gruppen aufzuteilen. Solche Teilgruppen können sowohl natürliche Gruppierungen darstellen (z.B. Käufer, Nichtkäufer einer Produktart) als auch das Resultat eines statistischen Klassifikationsverfahrens bilden. Gegenstand der Clusteranalyse bildet eine heterogene Menge von Untersuchungsobjekten (z.B. Personen, Produkte, Unternehmen, Regionen), die auf Basis von untersuchungsrelevanten Objektmerkmalen und mittels spezieller Fusionierungsalgorithmen zu möglichst homogenen Teilgruppen (Cluster, Klassen) zusammengefasst werden (vgl. Aaker/Kumar/Day 2001, S. 566 ff.; Bacher 1996; Backhaus et. al. 2003, S. 480 ff.; Böhler 2004, S. 230 ff.; Bortz 1993, S. 522 ff.; Büschken/von Thaden 2000; Churchill/Iacobucci 2005, S. 585 ff.; Eckey/ Kosfeld/Rengers 2002; S. 203 ff.; Hammann/Erichson 2000, S. 270 ff.; Hüttner 1997, S. 319 ff.; Litz 2000; S. 384 ff.; Malhotra 1999, S. 610 ff.; Rudolf/Müller 2004, S. 151 ff.; Sudman/Blair 1998, S. 558 ff.; Voß 2004, S. 565 ff.).

## 1.2. Charakteristika der Clusteranalyse

Kennzeichnend für die Clusteranalyse sind vorrangig vier Charakteristika:

**(1) Zielsetzungen**: Die grundsätzliche Aufgabe einer Clusteranalyse besteht darin, Objekte entsprechend ihrer Ähnlichkeit bezüglich untersuchungsrelevanter Klassifizierungsmerkmale zu gruppieren. Hiermit sind drei Teilzeile verbunden: Zum einen wird angestrebt, dass die in einer einzelnen Teilgruppe zusammengefassten Objekte einander möglichst ähnlich bzw. homogen sind. Zusätzlich soll zum anderen gewährleistet sein, dass die Unterschiede zwischen den Teilgruppen möglichst groß bzw. die Teilgruppen einander möglichst unähnlich bzw. heterogen sind. Schließlich sollen die solcherart gebildeten Cluster eine tragfähige Grundlage für gruppenspezifische Maßnahmen des Anwenders bilden (z.B. den clusterspezifischen Einsatz von Marketing-Instrumenten).

Das **clusteranalytische Grundprinzip** verdeutlicht Abbildung 1, in welcher die exemplarisch betrachtete Gesamtheit von 13 Nachfragern anhand der beiden Merkmale „Markenzufriedenheit“ und „Wiederkaufabsicht“ in zwei Cluster unterteilt ist. Aus dem Streudiagramm ist ersichtlich, dass Gruppe 1 aus Nachfragern besteht, die bezüglich beider Merkmale jeweils eine unterdurchschnittliche Ausprägung aufweisen und somit als die Gruppe der Markenwechsler beschrieben werden kann. Demgegenüber befinden sich im Cluster 2 solche Nachfrager, die hinsichtlich beider Merkmale eine jeweils überdurchschnittliche Ausprägung besitzen und sich damit als die Gruppe der Stammkunden kennzeichnen lässt.

Für Nachfrager 13 hingegen, der trotz einer geringen Markenzufriedenheit eine hohe Wiederkaufabsicht gegenüber der betreffenden Marke äußert, erweist sich eine

Gruppenzuordnung als nicht zweckmäßig. Denn wenngleich das scheinbar widersprüchliche Verhalten beispielsweise durch das Konzept der sog. Wechselbarrieren erklärt werden könnte, so bleibt im Rahmen einer Clusteranalyse dennoch zu prüfen, ob dieser nicht als ein Ausreißer aufzufassen und damit möglicherweise vom Untersuchungsprozeß auszuschließen wäre.

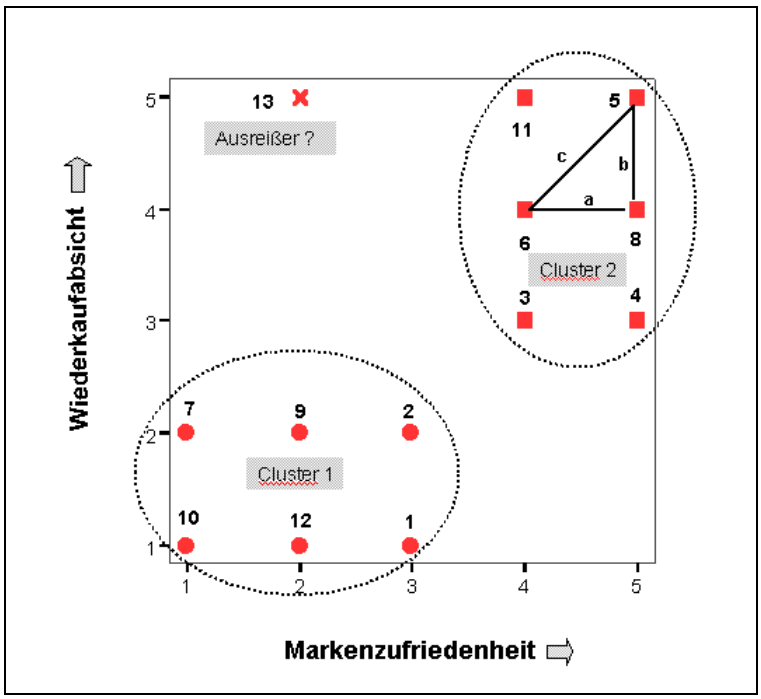



*Abbildung 1:* Grundprinzip der Clusterbildung

**(2) Klassifikationsansatz**: Die Clusteranalyse lässt sich als ein interdependenz-analytisches Klassifikationsverfahren kennzeichnen. Sie bildet zum einen ein Verfahren der Inderdependenzanalyse, das wechselseitige Beziehungen zwischen Objekten zur Aufdeckung von Strukturen untersucht. Ihr explorativer Einsatz ist insbesondere dann zweckmäßig, wenn zu Beginn der Analyse keinerlei Informationen über die Anzahl sowie Charakteristika von Teilgruppen vorliegen. Die Clusteranalyse stellt zum anderen ein spezielles objektzentriertes Klassifikationsverfahren dar, das eine Zusammenfassung der in einer Datenmatrix enthaltenen Objekte bzw. eine Gruppierung der Matrixzeilen anstrebt. Im Gegensatz zur Clusteranalyse knüpft eine **Faktorenanalyse** an den Variablen der Datenmatrix an und beinhaltet daher eine Gruppierung von Variablen bzw. eine Reduktion der Matrixspalten. Eine Gruppierung von Objekten kann auch mit Hilfe von dependenzanalytischen Klassifikationsverfahren vorgenommen werden. Während mit der **Diskriminanzanalyse** die Zugehörigkeit von Objekten zu vorab bereits unterteilten Gruppen erklärt und prognostiziert werden kann, strebt die **Kontrastgruppenanalyse** an, eine abhängige Variable dadurch zu erklären, dass die Ausgangsmenge von Objekten sukzessive in Gruppen aufgeteilt wird..

**(3) Verfahrensflexibilität:** Die Clusteranalyse bildet kein standardisiertes Verfahren, sondern bietet vielfältige Optionen zur Druchführung des Analyseprocederes. Hinsichtlich des Dateninputs ermöglicht die Clusteranalyse eine Objektgruppierung auf Basis von nominalen, ordinalen, metrischen oder gemischt-skalierten Klassifizierungsvariablen. Im clusteranalytischen Untersuchungsprozeß stehen dem

Anwender ferner zahlreiche Wahlmöglichkeiten zur Verfügung, die einerseits die Messung der Unterschiedlichkeit von Objekten betreffen und sich andererseits auf die Auswahl von Fusionierungsalgorithmen beziehen. Hinzu kommt, dass die Clusteranalyse zwar prinzipbedingt zur Gruppenbildung führt, jedoch keine aussagefähigen Gütekriterien, wie z.B. Teststatistiken zur Ergebnis-bewertung bereitstellt. Daher ist dem Anwender stets eine zweifache Aufgabe gestellt: Er hat zum einen über die Auswahl der statistischen Teilprozeduren der Cluster-analysen zu entscheiden. Und zum anderen bedarf es einer Beurteilung darüber, ob die gewonnenen Cluster sachlich interpretierbar, hinreichend stabil und intern ausreichend homogen bzw. extern heterogen sind.

**(4) Integrierter Verfahrenseinsatz:** Die Gütebeurteilung kann teilweise im Verbund mit anderen Analyseverfahren erfolgen. Hierbei vermittelt z.B. eine Varianzanalyse Aufschluß darüber, ob die erzeugten Cluster sich signifikant voneinander unterscheiden. Daneben lassen sich mit Hilfe einer Diskriminanzanalyse diejenigen Objektmerkmale aufdecken, die eine besonders starke Trennung der ermittelten Gruppen herbeiführen. Darüber hinaus kann der Einsatzverbund aber auch den Dateninput einer Clusteranalyse betreffen. Vielfach bietet es sich an, eine große Zahl von Klassifizierungsvariablen mittels einer vorgeschalteten Faktorenanalyse auf wenige Faktoren zu verdichten und diese als Gruppierungsmerkmale einer Clusteranalyse zu verwenden. Schließlich können die ermittelten Nutzenwerte einer vorgeschalteten Conjoint-Analyse als Dateninput einer Clusteranalyse dienen, um sog. Nutzensegmente zu identifizieren.

### 1.3. Einsatzfelder der Clusteranalyse im Marketing

Das Aufgabenfeld der Clusteranalyse erstreckt sich im Marketing primär auf sechs Entscheidungsbereiche (vgl. hierzu u.a. Malhotra 1999, S. 612 ff.):

- **Marktsegmentierung**: Ein zentrales marketingrelevantes Einsatzfeld der Clusteranalyse bildet die Abgrenzung und Beschreibung von Käufersegmenten, um konkrete Anhaltspunkte zur zielgruppengerechten Gestaltung der Marketing-Instrumente zu erhalten. Das Spektrum von Segmentierungsvariablen des Käuferverhaltens ist breit gesteckt und umfasst sozio-demographische Merkmale (z.B. Alter, Geschlecht von Personen), psychographische Merkmale (z.B. Lebensstile, Einstellungen von Personen) und Merkmale des beobachtbaren Käuferverhaltens (z.B. Markenwahl, Kaufintensität von Nachfragern).
- **Strategische Wettbewerbergruppen**: Daneben kann die Clusteranalyse zur Erfassung und Beschreibung von sog. strategischen Wettbewerbergruppen herangezogen werden. Strategische Wettbewerbergruppen setzen sich aus einer Teilmenge von Unternehmen einer bestimmten Branche zusammen, die sich im Hinblick auf den Einsatz von strategischen Unternehmensaktivitäten (z.B. Art des Produktionssystems, Maßnahmen der Personalentwicklung, Instrumente des Marketing-Mix) ähnlich sind. Aus solcherart abgegrenzten Wettbewerbergruppen sollen Anhaltspunkte über Erfolgspotentiale (z.B. Gewinnpotentiale) und die Wettbewerbsdynamik einer Branche gewonnen werden.
- **Markenpositionierung**: Ferner erweist sich die Clusteranalyse im Rahmen der sog. Markenpositionierung als überaus hilfreich. Hierbei wird auf Basis von

Produkteigenschaften der in einem bestimmten Produktmarkt (z.B. Markt für Fruchtsaftgetränke) angebotenen Marken angestrebt, zunächst Markengruppen voneinander abzugrenzen und anschließend die Frage zu beantworten, ob Marken mit spezifischen Wettbewerbsvorteilen oder –nachteilen ausgestattet sind.

- **Selektion von Ländermärkten**: Die Clusteranalyse kann Entscheidungen bezüglich des internationalen Marketing dadurch unterstützen, indem diese dazu beiträgt, die Gesamtheit der relevanten Ländermärkte bezüglich von Merkmalen der Marktumwelt (z.B. politische Risiken) zu gruppieren und eine zieladäquate Ländermarktselektion vorzunehmen .
- **Standortanalyse**: Im Rahmen von betrieblichen Standortanalysen seitens z.B. von Handelsunternehmen ist es vielfach zweckmäßig, die Menge der Standort-alternativen nach Maßgabe ihrer Attraktivität (z.B. verkehrstechnische Infra-struktur) zu Standortclustern zusammenzufassen und hieran anschließend vertiefende Detailanalysen der Standortbewertung und -auswahl durchzuführen.
- **Abgrenzung von Testmärkten**: Ein weiteres, gleichwohl spezielles Aufgabenfeld der Clusteranalyse besteht darin, im Zuge der experimentellen Untersuchung der Erfolgswirksamkeit von Marketingmaßnahmen (z.B. bei der Markteinführung von Neuprodukten) jene Testmärkte abzugrenzen, in denen der Einsatz bzw. die Variation der testrelevanten Instrumente erfolgen und gemessen werden soll.

## 2. Methodische Grundlagen der Clusteranalyse

Der clusteranalytische Verfahrensablauf umfasst fünf Phasen, wobei die Erstellung der Proximitätsmatrix und die Auswahl des Fusionierungsverfahrens den methodischen Kern der Analyseprozedur darstellen (vgl. Abb. 2):

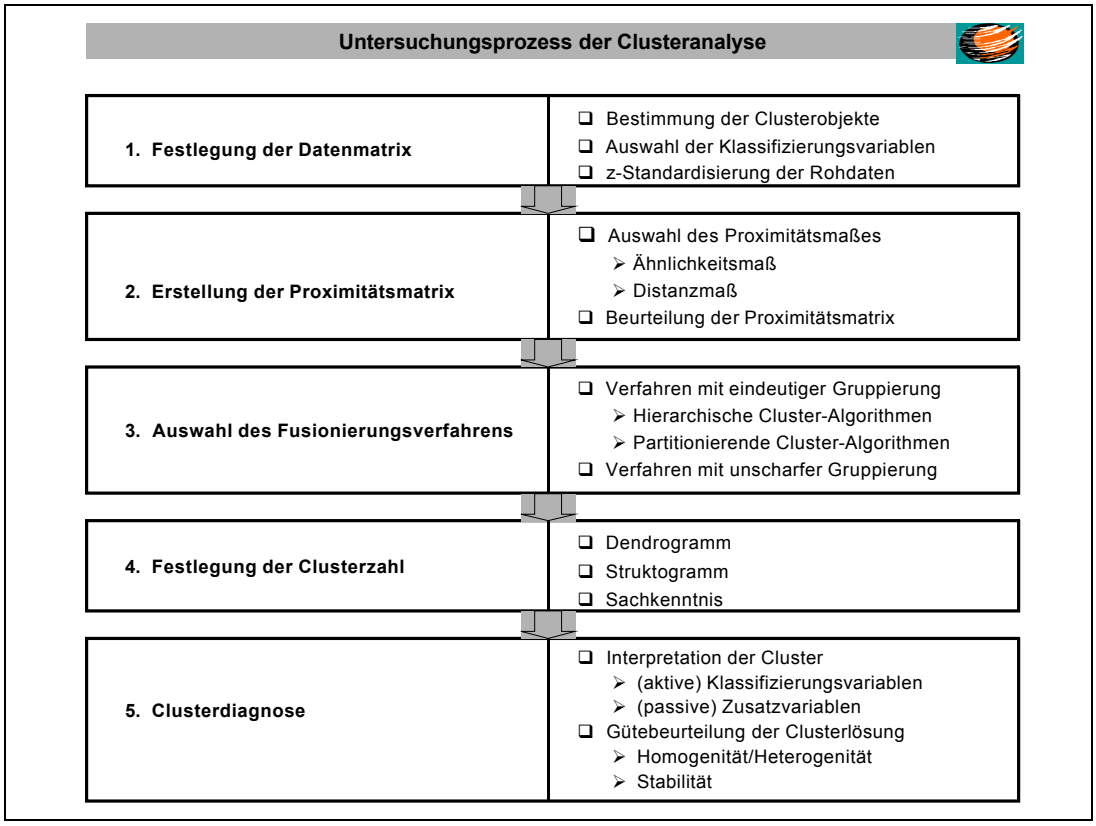


*Abbildung 2*: Verfahrensablauf der Clusteranalyse

## 2.1. Struktur der Datenmatrix

Ausgangspunkt einer Clusteranalyse bildet eine Datenmatrix, in der die Merkmalsausprägungen der betrachteten Objekte enthalten sind. Die Auswahl jener Objekte und Merkmale, die als Dateninput des clusteranalytischen Vorgehens dienen sollen, ist aus der marketingspezifischen Aufgabenstellung abzuleiten. Je nachdem, ob beispielsweise eine Marktsegmentierung, eine branchenbezogene Unternehmensklassifikation oder eine Gruppierung von Ländermärkten angestrebt wird, handelt es sich bei den relevanten **Objekten** um Käufer eines bestimmten Produktmarktes, Unternehmen einer spezifischen Branche oder eine Menge von Ländermärkten.

Vergleichbare Überlegungen betreffen die Festlegung der Art sowie der Anzahl untersuchungsrelevanter Klassifizierungsvariablen. Es ist demnach in den zuvor genannten Beispielen zu überlegen, welche **Merkmale** bei der Marktsegmentierung (z.B. Markenwahl, Einkaufsstättenwahl von Nachfragern), der Unternehmensgruppierung (z.B. Jahresgewinn, Rechtsform, Innovationsrate) oder der Zusammenfassung von Ländermärkten (z.B. Einwohnerzahl, Arbeitslosenquote) zweckmäßig sind. Daneben kann es im Einzelfall sinnvoll sein, die Ausgangsmerkmale mittels einer Faktorenanalyse zu verdichten und die daraus resultierenden Faktorwerte der Objekte als Klassifizierungsmerkmale heranzuziehen ( faktorielle Clusteranalyse).

Nach der **Zahl der Klassifizierungsvariablen** lassen sich eindimensionale (monothetische) und mehrdimensionale (polythetische) Clusteranalysen unterscheiden. In der Regel wird man im Marketing auf polythetische Ansätze zurückgreifen, bei denen die Clusterbildung auf Basis der simultanen Analyse mehrerer Merkmale erfolgt. Für metrische Klassifizierungsvariablen, die in unterschiedlichen Skaleneinheiten gemessen wurden, gilt es zu beachten, dass diese vor der Durchführung clusteranalytischer Prozeduren zu **standardisieren** sind (vgl. ausführlich Eckey/Kosfeld/Rengers 2002, S. 208 ff.; Voß 2004, S. 567 f.). Eine Merkmalsstandardisierung, (z.B. z-Standardisierung) beinhaltet die dimensionslose Vereinheitlichung der Variablen und führt dazu, dass bei der Berechnung von Distanz- bzw. Ähnlichkeitsmaßen die Variablen mit einem großem Mittelwert und großer Streuung (z.B. jährliches Haushaltseinkommen von Personen zwischen 15.000 € und 250.000 €) kein größeres Gewicht erhalten als Merkmale mit einer geringeren Streuung (z.B. Haushaltsgröße zwischen 1 und 6 Personen).

## 2.2. Proximitätsmaße

Um die betrachteten Objekte entsprechend ihrer Ähnlichkeit zu Teilgruppen zusammenfassen zu können, muß die Unterschiedlichkeit (Proximität) von Objekten anhand einer statistischen Maßzahl erfasst werden. Daher wird im zweiten Schritt einer Clusteranalyse die Datenmatrix der Rohdaten in eine Proximitätsmatrix überführt. In dieser wird die paarweise Unterschiedlichkeit zwischen den Objekten quantifiziert und dargestellt. Zur Messung der Unterschiedlichkeit von Objekten existiert eine Vielzahl von Proximitätsmaßen (vgl. Abb. 3), die auf dem Ähnlichkeits- oder dem Distanzkonzept beruhen (vgl. ausführlich Backhaus et. al. 2003; S. 482 ff.; Bortz 1993, S. 523 ff; Brosius 2002, S. 609 ff.; Büschgen/Thaden 2000, S. 344 ff.; Eckey/Kosfeld/Rengers 2002, S. 205 ff.; Litz 2000, S. 387 ff.).

**Ähnlichkeitsmaße**, die gewöhnlich auf das Intervall [0,1] normiert sind, reflektieren die Ähnlichkeit zwischen zwei Objekten: Je größer der Wert eines Ähnlichkeitsmaßes, desto ähnlicher sind sich zwei Objekte. Demgegenüber bringen **Distanzmaße** die Unähnlichkeit zwischen zwei Objekten zum Ausdruck: Je größer die Distanz zwischen zwei Objekten, desto unähnlicher sind sie sich. Gleichwohl ist darauf hinzuweisen, dass sich jedes Ähnlichkeitsmaß in ein Distanzmaß transformieren lässt und umgekehrt (vgl. Bortz 1993, S. 523 ff.). Proximitätsmaße können entweder direkt erhoben (z.B. durch Befragung) oder aber indirekt, durch einen merkmalsbezogenen Vergleich der Objekte berechnet werden Im zweiten, weitaus häufigeren Fall hängt die Wahl des Proximitätsmaßes entscheidend vom Skalenniveau der Klassifizierungsvariablen ab (vgl. Abb. 3.):

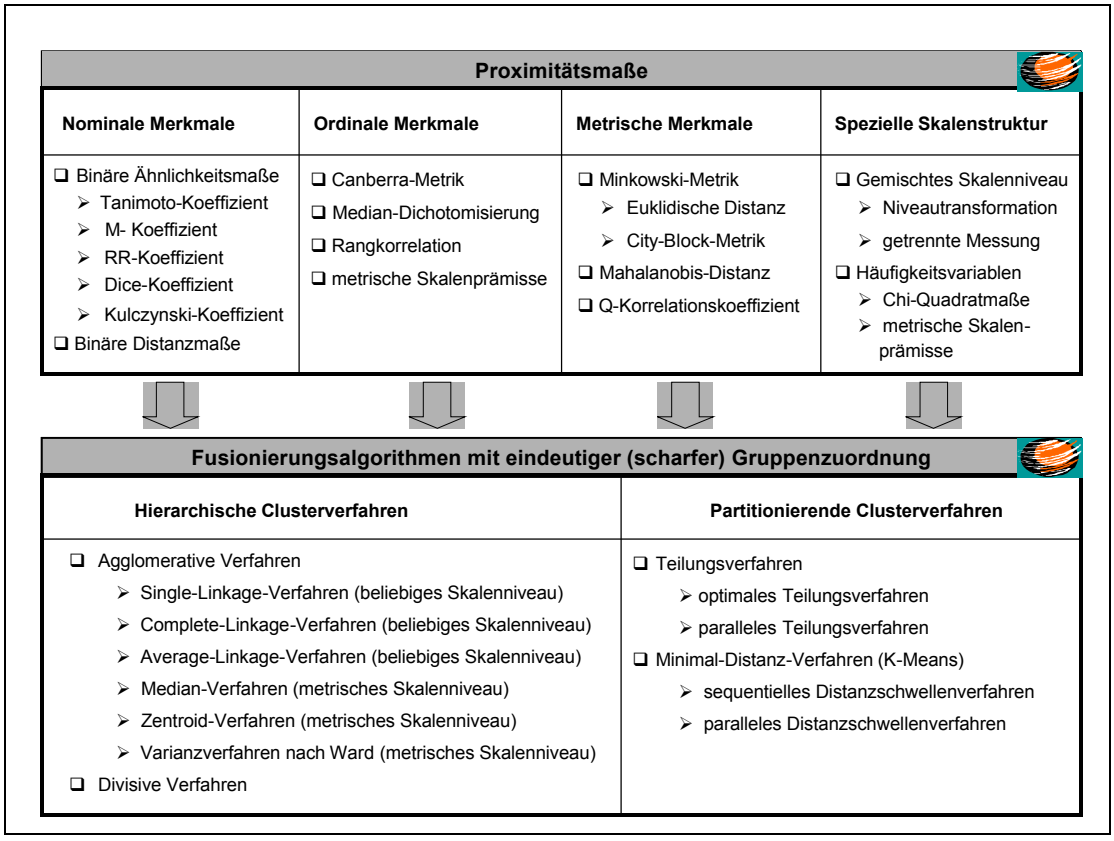



*Abbildung 3:* Überblick über ausgewählte Proximitätsmaße und Fusionierungsverfahren

**(1)** Für **nominalskalierte Klassifizierungsvariablen** kommen Ähnlichkeitsmaße sowie binäre Distanzmaße in Betracht. Hierbei wird stets von einem Paarvergleich hinsichtlich des Vorhandenseins bzw. Nichtvorhandenseins von Merkmalsausprägungen ausgegangen. Dies setzt generell voraus, dass die Merkmale in zweifach gestufter (dichotomer) Ausprägung vorliegen, d.h. binär kodiert sind. Hierbei wird jeder Merkmalsausprägung entweder der Wert 1 ( = Eigenschaft vorhanden) oder der Wert 0 ( = Eigenschaft nicht vorhanden) zugeordnet. Mehrfach gestufte (polytome) Merkmale sind hingegen zunächst in binäre Merkmale zu transformieren, bevor diese zur Ähnlichkeitsmessung der Objekte herangezogen werden können. Bei binären Merkmalen kann man sämtliche für einen Vergleich zweier Objekte möglichen Fälle in einer Vierfelder-Tafel darstellen. Eine solche Tafel ist exemplarisch in Tabelle 1 dargestellt, bei der zwei Objekte anhand von 15 binären Merkmalen miteinander verglichen werden.

| | | Nachfrager II | | |
|---|---|---|---|---|
| | | **1** | **0** | *Summe* |
| Nachfrager I | **1** | a = 3 | c = 5 | *8* |
| | **0** | b = 4 | d = 3 | *7* |
| | *Summe* | *7* | *8* | *15* |

a = Anzahl der Merkmale, die bei beiden Personen vorliegen (1;1)
b = Anzahl der Merkmale, die bei Person I nicht vorhanden sind und bei Person II vorliegen (0;1)
c = Anzahl der Merkmale, die bei Person I vorliegen und bei Person II nicht vorhanden sind (1;0)
d = Anzahl der Merkmale, die bei beiden Personen nicht vorliegen (0,0)

*Tabelle 1:* Vierfelder-Tafel zum Vergleich zweier Objekte

Durch unterschiedliche Kombination der Größen a, b, c und d, können zahlreiche Proximitätsmaße gebildet werden. Diese unterscheiden sich primär darin, in welcher Weise positive Übereinstimmungen (a), Nicht-Übereinstimmungen (b und c) sowie negative Übereinstimmungen (d) zwischen den Merkmalen beider Objekte gewichtet werden. Weite Verbreitung haben u.a. die nachstehenden Maße gefunden (vgl. Tabelle 2):

| **Proximitätsmaß** | **Formal** | **Beispiel** |
|---|---|---|
| Simple-Matching-Koeffizient | $M_{ik} = \frac{a+d}{a+b+c+d}$ | $M_{ik} = \frac{3+3}{3+4+5+3} = 0,40$ |
| Tanimoto-Koeffizient | $J_{ik} = \frac{a}{a+b+c}$ | $J_{ik} = \frac{3}{3+4+5} = 0,25$ |
| RR-Koeffizient | $RR_{ik} = \frac{a}{a+b+c+d}$ | $RR_{ik} = \frac{3}{3+4+5+3} = 0,20$ |
| Euklidische Distanz | $d_{ik} = \sqrt{b+c}$ | $d_{ik} = \sqrt{4+5} = 3$ |

*Tabelle 2:* Ausgewählte Proximitätsmaße für nominalskalierte Merkmale

- Der Simple-Matching-Ähnlichkeitskoeffizient (M-Koeffizient) setzt die Zahl der positiven und negativen Übereinstimmungen bei den beiden Objekten j und k zur Gesamtzahl der Merkmale in Beziehung. Dieser weist für die Beispieldaten der Tabelle 1 den Wert 0,40 auf.
- Demgegenüber verzichtet der Tanimoto- bzw. Jaccard-Ähnlichkeitskoeffizient (J) auf die Einbeziehung der negativen Übereinstimmungen und berücksichtigt nur jene Merkmale, die bei beiden Objekten tatsächlich vorkommen. Für die Beispieldaten der Tabelle 1 erhält man hierfür einen Wert von 0,25.
- Im Unterschied zum Tanimoto-Koeffizienten werden beim RR-Koeffizienten nach Russel & Rao im Nenner auch diejenigen Fälle erfasst, bei denen beide Objekte das interessierende Merkmal nicht aufweisen (d). Der RR-Koeffizient stellt daher die Anzahl der Wertepaare, bei denen das interessierende Merkmal bei beiden Objekten vorliegt (a), der Anzahl aller Wertepaare gegenüber und ergibt für unsere Beispieldaten einen Wert von 0,20.

❒ Ein gebräuchliches Distanzmaß für einen Objektvergleich bei binären Merkmalen bildet die Euklidische Distanz d, die sich als Quadratwurzel aus der Anzahl ungleicher Wertepaare errechnet und in unserem Beispiel einen Wert von 3 aufweist.

**(2)** Liegen **ordinalskalierte Merkmale** vor, so werden diese zumeist auf ein nominales Skalenniveau zurückgeführt, um die Proximitätsmaße für Nominaldaten verwenden zu können. Weiterhin besteht die Möglichkeit, auf Basis einer Rangziffernbildung spezielle Ähnlichkeitsmaße für Ordinaldaten, wie z.B. die sog. Canberra-Metrik oder den Rangkorrelationskoeffizienten heranzuziehen. Schließlich werden in der Praxis ordinale Daten recht häufig auch als metrisch skalierte Größen aufgefasst, für die spezielle Distanzmaße bestimmbar sind.

**(3)** Bei **metrischen Merkmalen** erfolgt die Proximitätsmessung gewöhnlich mit den aus der allgemeinen Minkowski-Metrik ableitbaren Distanzmaßen, wie z.B. der Euklidischen Distanz. Daneben gelangen vereinzelt auch die sog. Mahalanobis-Distanz und der Q-Korrelationskoeffizient zur Anwendung. Grundlage der Distanzmessung mittels der Minkowski-Metrik bildet die statistische Entfernung zwischen den Objekten (vgl. Tabelle 3):

| Distanzmaß | Formal | Geometrisch |
|---|---|---|
| Euklidische Distanz | $d_{j,k} = \sqrt{\sum_{i}^{n} \left(x_{ij} - x_{ik}\right)^2}$ | $c = \sqrt{a^2 + b^2}$ |
| Quadrierte Euklidische Distanz | $d_{j,k} = \sum_{i}^{n} \left(x_{ij} - x_{ik}\right)^2$ | $c^2 = a^2 + b^2$ |
| City-Block-Distanz | $d_{j,k} = \sum_{i}^{n} \left\lvert x_{ij} - x_{ik} \right\rvert$ | $c = \lvert a \rvert + \lvert b \rvert$ |

*Tabelle 3*: Ausgewählte Distanzmaße der Minkowski-Metrik für metrische Merkmale

❒ Eines der gebräuchlichsten Distanzmaße für metrische Daten ist die Euklidische Distanz, die sich als Quadratwurzel aus der Summe der quadrierten Differenzen zwischen den Merkmalsausprägungen zweier Objekte ergibt. Durch die Quadrierung werden große Differenzwerte stärker gewichtet als kleinere Differenzwerte. Die Euklidische Distanz ist vergleichsweise robust gegenüber Datentransformationen und entspricht unserer räumlichen Anschauung. Im Beispiel der Abbildung 1 lässt sich die Euklidische Distanz zwischen z.B. den beiden Nachfragern 6 und 5 berechnen als: $d_{(6,5)} = \sqrt{(4-5)^2 + (4-5)^2} = 1,41$. Geometrisch entspricht die Euklidische Distanz der direkten Distanz zwischen den Objekten. Diese lässt sich nach dem Satz von Pythagoras als Hypothenuse eines „gedachten“ rechtwinkligen Dreiecks berechnen. Hiernach ist im Merkmalsraum der Abbildung 1 die Euklidische Distanz c zwischen den beiden Nachfragern 6 und 5 durch die Wurzel der Summe der beiden Kathetenquadrate $a^2$ und $b^2$ gegeben.

❒ Demgegenüber errechnet sich die quadrierte euklidische Distanz im Beispiel der Abbildung 1 als: $d_{(6,5)} = (4\text{-}5)^2 + (4\text{-}5)^2 = 2$ bzw. $c^2$ als Summe der beiden Kathetenquadrate $a^2$ und $b^2$. Bei Verwendung der quadrierten euklidischen Distanz

anstelle der euklidischen Distanz bleibt die Rangfolge der ähnlichen Objektpaaare unverändert. Es verändern sich jedoch die Abstandsverhältnisse, was mitunter einen Einfluss auf die zu ermittelnde Clusterzahl ausüben kann.

- Sollen alle Differenzen gleichgewichtig berücksichtigt werden, so kann man statt der vorstehenden Distanzmaße die sog. City-Block-Distanz berechnen. Diese ist als die Summe der absoluten Merkmalsdifferenzen definiert und entspricht graphisch einer rechtwinkligen Verbindung zwischen zwei Punkten. Diese beträgt in unserem Beispiel $d_{(6,5)} = |4-5| + |4-5| = 2$ bzw. $c = |a| + |b|$.

**(4)** Neben den zuvor angeführten Datenstrukturen sind spezielle Skalensituationen denkbar. Bei Vorliegen **gemischt-skalierter Merkmale** bieten sich im wesentlichen zwei Ansätze zur Proximitätsmessung an (vgl. Bortz 1993, S. 527). Eine erste Vorgehensweise besteht darin, Merkmale mit unterschiedlichen Skalenniveaus auf ein einheitliches Meßniveau zu transformieren. Hierbei werden Merkmale mit einem höheren Skalenniveau zunächst in Merkmale mit einem niedrigeren Skalenniveau umgewandelt, auf deren Grundlage anschließend eine gemeinsame Proximitäts-messung durchgeführt werden kann. Demgegenüber werden im Rahmen eines zweiten Ansatzes für die unterschiedlichen Skalenarten von Merkmalen jeweils getrennt die adäquaten Proximitätsmaße berechnet und anschließend zu einem „gewogenen, mittleren Distanzwert“ zusammengefasst. Die Gewichtungsstruktur kann dabei dem relativen Anteil der Anzahl einer Skalierungsart an der Gesamtzahl der Merkmale entsprechen. Hierbei ist darauf zu achten, dass die Koeffizienten für nominal und metrisch skalierte Merkmale dieselbe Richtung aufweisen ( je ähnlicher, desto kleiner etc.). Hierzu müssen die ermittelten Ähnlichkeitswerte nominaler Merkmale ($s_{jk}$) in binäre Distanzmaße $d_{jk}$ transfomiert werden, indem diese von Eins subtrahiert werden, d.h. $d_{jk} = 1 - s_{jk}$. Enthalten die Klassifizierungsvariablen sog. **Häufigkeitswerte**, so können diese zum einen als metrische Variablen behandelt werden. Zum anderen kann es jedoch auch zweckmäßig sein, für Häufigkeitsvariablen spezielle $\chi^2$-basierte Ähnlichkeitsmaße, wie z.B. das Chi-Quadrat-Maß zu verwenden, bei welchem die Distanz zwischen zwei Objekten der Quadratwurzel der $\chi^2$–Statistik entspricht (vgl. ausführlich Brosius 2002, S. 615 ff.).

### 2.3. Fusionierungsverfahren

Die Ähnlichkeits- bzw. Distanzwerte der Proximitätsmatrix bilden die Datenbasis zur Gruppierung von Objekten. Das Gruppierungsprinzip besteht allgemein darin, Objekte mit geringen Distanzen bzw. großer Ähnlichkeit zu einem Cluster zusammenzufassen und solche Objekte, zwischen denen große Distanzen bzw. geringe Ähnlichkeiten bestehen, unterschiedlichen Clustern zuzuordnen. Zur Gruppierung einer gegebenen Objektmenge ist der Einsatz eines Fusionierungsverfahrens erforderlich, wozu dem Anwender ein breites Methodenspektrum zur Verfügung steht (vgl. ausführlich Aaker/Kumar/Day 2001, S. 568 ff.; Büschken/Thaden 2000, S. 352 ff.; Eckey/ Kosfeld/Rengers 2002, S. 229 ff.; Litz 2000, S. 401 ff.; Malhotra 1999, S. 617 ff.; Voß 2004, S. 571 ff.).

Von grundsätzlicher Bedeutung ist in diesem Zusammenhang zunächst die Unterscheidung zwischen Verfahren mit eindeutiger (scharfer) Gruppenzuordnung und Verfahren mit uneindeutiger (unscharfer) Gruppenbildung. Bei **scharfen**

**Methoden** werden die einzelnen Objekte eindeutig bzw. überlappungsfrei den verschiedenen Clustern zugeordnet. Werden dabei sämtliche Elemente den Gruppen zugewiesen, spricht man von einer exhaustiven (erschöpfenden) Gruppierung; bleiben demgegenüber einzelne Elemente clusterfrei, so handelt es sich um eine nicht-exhaustive Zuordnung. Die **Verfahren mit unscharfer Gruppenbildung** ermöglichen es, dass ein Objekt entweder vollständig mehreren Gruppen angehören kann (überlappende Gruppierung) oder aber anteilsmäßig verschiedenen Gruppen zugeordnet wird (Fuzzy Clustering). In der Marketingforschung besitzen scharfe Methoden den weitaus größeren Stellenwert, so dass sich die nachfolgenden Ausführungen auf diesen Verfahrenstyp beschränken. Hierzu vermittelt Abbildung 3 einen Methodenüberblick

**(1) Hierarchische Fusionierungsverfahren**: Der Einsatz von hierarchischen Clusterverfahren bietet sich dann an, wenn im Rahmen der Marketinganalyse keinerlei Vorkenntnisse über die Anzahl der zu bildenden Gruppen vorhanden sind (z.B. bei einer Marktsegmentierung für ein Innovationsprodukt). Im Zuge der Gruppierung wird eine hierarchische Verschachtelung von Ober- und Untergruppen gebildet, die mit Hilfe agglomerativer oder divisiver Fusionierungsalgorithmen vorgenommen werden kann.

**Divisive Verfahren**, die in der empirischen Statistik überaus selten zur Anwendung gelangen, starten den Gruppierungsprozeß mit einer Gruppe, die alle Objekte enthält, und spalten dann von dieser schrittweise solange Untergruppen ab, bis alle Objekte jeweils eine Gruppe bilden. **Agglomerative Verfahren** beschreiten den umgekehrten Weg, indem Objekte ausgehend von der feinsten Gruppierung ein-elementiger Cluster sukzessive solange zusammengefasst werden, bis alle Objekte in einem gemeinsamen Cluster enthalten sind. Auf diese Weise entsteht eine Hierarchie von Clustern innerhalb des mehrstufigen Gruppierungsprozesses, wobei jene Objekte, die bereits einmal zu einem Cluster zusammengefasst wurden, auf den späteren Stufen der Clusterbildung nicht mehr getrennt werden und die Cluster auf jeder Stufe disjunkt sind. Bei der agglomerativen Gruppierung werden somit die Klassen von Stufe zu Stufe heterogener, da immer „entferntere“ Objekte hinzukommen, bis sich schließlich alle Objekte in einem Cluster befinden.

Agglomerativen Verfahren ist gemein, dass auf jeder Stufe die Distanzen bzw. Ähnlichkeiten aller Cluster zueinander ermittelt und jene beiden Gruppen fusioniert werden, welche die geringste Distanz oder größte Ähnlichkeit aufweisen. Auf der ersten Fusionierungsstufe unterscheiden sich die verschiedenen hierachischen Gruppierungsverfahren nicht voneinander, da in der Ausgangspartition jedes Cluster aus lediglich einem Objekt besteht und daher die Clusterdistanzen (Clusterähnlichkeiten) exakt den in der Proximitätsmatrix berechneten Distanzen der Einzelobjekte (Objektähnlichkeiten) entsprechen. Auf der Grundlage des verwendeten Distanzmaßes oder Ähnlichkeitsmaßes erhält man daher im ersten Schritt der Clusterbildung stets dasselbe Fusionsergebnis. In den darauffolgenden Fusionsstufen hingegen unterscheiden sich die Clusterlösungen der Agglomerationsverfahren gewöhnlich voneinander, da die Clusterbildung auf unterschiedlichen Fusionskriterien beruhen kann (vgl. Abb. 4):

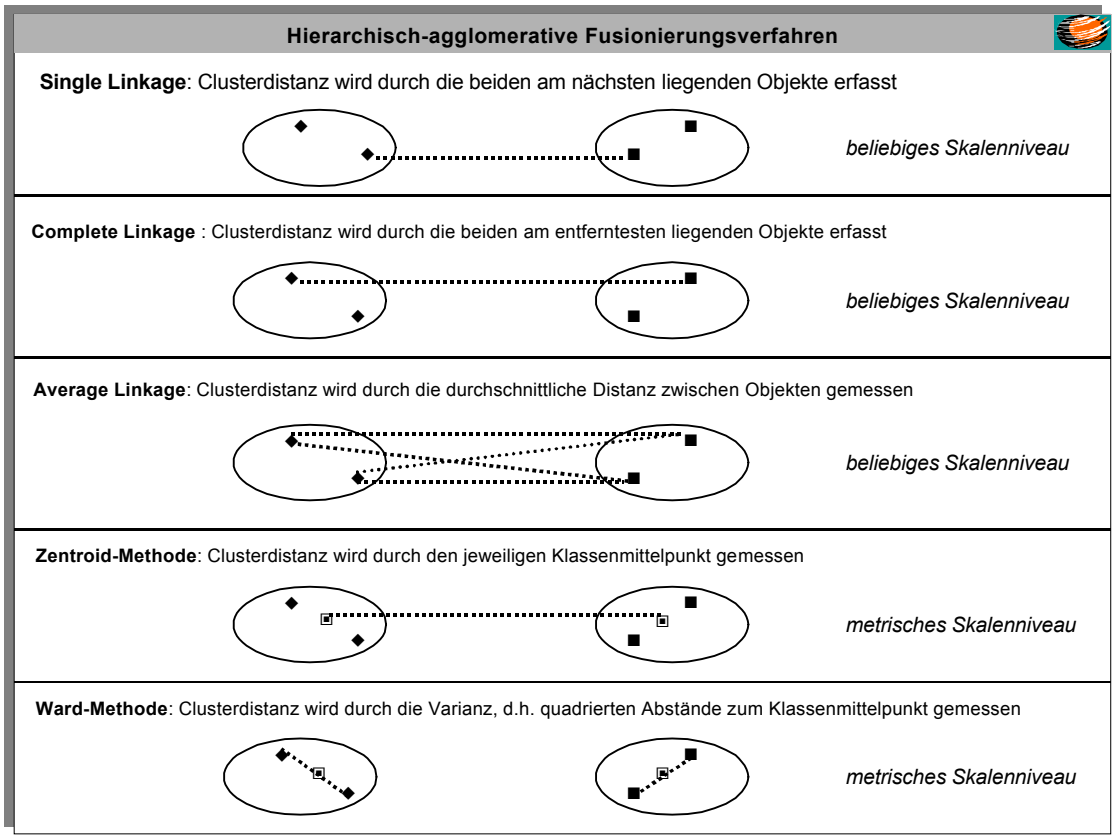



*Abbildung 4:* Hierarchisch-agglomerative Verfahren im Überblick

- Beim **Single-Linkage-Verfahren** (Minimummethode, nearest neighbor rule), das für beliebige Skalenniveaus geeignet ist, wird die Distanz zwischen zwei Clustern (bzw. einem Einzelobjekt und und einem Cluster) durch die Distanz der beiden am nächsten liegenden Einzelobjekte beider Gruppen gemessen. Es werden dann stets jene beiden Cluster zusammengefasst, deren Clusterdistanz minimal ist. Die Verbindung zweier Objekte wird demnach „brückenförmig“ durch je ein Objekt der beiden Objektmengen („single link“) hergestellt. Da hierbei als neue Distanz zwischen zwei Gruppen immer der kleinste Wert der Einzeldistanzen herangezogen wird, eignet sich das Single-Linkage-Verfahren in besonderer Weise dazu, Ausreißer in einer Objektgruppe zu identifizieren. Hiermit einher geht jedoch die Tendenz zur Kettenbildung (chainig effect), d.h. zur Zusammenfassung von wenigen großen und heterogenen Gruppen (kontrahierendes Verfahren).

- Das **Complete-Linkage-Verfahren** ( Maximummethode; furthest neighbor rule), welches gleichfalls bei einem beliebigen Skalenniveau einsetzbar ist, wird die Distanz zwischen zwei Clustern (bzw. einem Einzelobjekt und einem Cluster) durch die beiden am entferntesten liegenden Einzelobjekte beider Gruppen erfasst. Dabei werden auf jeder Fusionierungsstufe jene zwei Cluster verschmolzen, deren größte Einzeldistanz minimal ist. Im Gegensatz zum Single-Linkage-Verfahren ermöglicht das Complete-Linkage zwar keine Identifikation von Ausreißern; es tendiert jedoch zur Bildung vieler kleiner, kompakter und homogener Gruppen (dilatierendes Verfahren) und ist daher für praktische Fragestellungen besser geeignet.

- Beim **Average-Linkage-Verfahren,** das kein bestimmtes Skalenniveau erfordert, berechnet man für je zwei Cluster (bzw. einem Objekt und einem Cluster) das ungewichtete arithmetische Mittel aller Objektdistanzen und fusioniert diejenigen Objekte bzw. Cluster, welche die geringste Durchschnittsdistanz aufweisen. Hierbei sind zwei Ansätze zu unterscheiden: Während das sog. „Linkage zwischen den Gruppen“ ( average linkage between groups) bei der Distanzberechnung nur jene Objektpaare berücksichtigt, bei denen die Objekte aus verschiedenen Gruppen entstammen, werden beim sog. „Linkage innerhalb der Gruppen“ (average linkage within groups) sämtliche Objektpaare einbezogen, d.h. auch jene Objekte, die einem gemeinsamen Cluster angehören. Das Average-Linkage-Verfahren stellt einen vergleichsweise konservativen Fusionierungsalgorithmus dar, der eine Kompensation von größeren Objektdistanzen durch geringere Distanzen nahe beieinander liegender Objekte gestattet und somit zu Gruppen mit einer durchschnittlichen Besetzungszahl und Homogonität führt.

- Im Unterschied zu den Linkage-Verfahren setzen sowohl das Zentroid-Verfahren als auch das Medianverfahren metrisch skalierte Klassifizierungsmerkmale voraus. Beide Verfahren messen die Clusterdistanzen anhand der Abstände bzw. der quadrierten euklidischen Distanzen zwischen den Clusterschwerpunkten; sie untercheiden sich jedoch hinsichtlich der Gewichtung der Clusterschwerpunkte. Beim **Medianverfahren** werden diejenigen Cluster vereint, deren quadrierter euklidischer Centroidabstand minimal ist, wobei ein Clustercentroid (Gruppenschwerpunkt) den durchschnittlichen Merkmalsausprägungen aller Objekte eines Clusters entspricht. Dabei werden allerdings unterschiedliche Objekthäufigkeiten der zu fusionierenden Cluster vernachlässigt, so dass der Centroid eines neu gebildeten Clusters dem Mittelpunkt (Median) der Linie, welche die Centroide der zu fusionierenden Cluster verbindet, entspricht. Sollen unterschiedliche Objekthäufigkeiten der zu fusionierenden Cluster berücksichtigt werden, wählt man das gewichtete Median- oder das Zentroidverfahren. Das **Zentroid-Verfahren** berechnet die Clusterdistanz im Gegensatz zum Average-Linkage-Verfahren, welches auf durchschnittliche Objektdistanzen abstellt, als quadrierte euklidische Distanzen zwischen den Clustermittelwerten. Clustermittelwerte bilden ein Maß des Clusterschwerpunktes und dienen als fiktive Objektrepräsentanten der Cluster. Zur Ermittlung des Zentrums von zwei vereinigten Clustern wird das gewichtete arithmetische Mittel der betreffenden Cluster berechnet, wobei die Gewichtungsstruktur dem clusterspezifischen Anteil der Objekte an der Gesamtzahl der Objekte entspricht.

- Das **Ward-Verfahren** (Fehlerquadratsummen-Methode, minimum variance method), welches metrische Klassifizierungsmerkmale voraussetzt, unterscheidet sich von den vorhergehend besprochenen Clustermethoden dadurch, dass die Objektgruppierung nicht anhand der geringsten Clusterdistanz, sondern anhand eines vorgegebenen Heterogenitätsmaßes bzw. des Varianzkriteriums erfolgt. Ziel ist es, jeweils diejenigen Objekte bzw. Gruppen zusammenzufassen, welche die Gesamtstreuung in einer Gruppe (Varianz, Fehlerquadratsumme innerhalb der Gruppen, within-groups sum of squares) möglichst wenig erhöhen, so dass die Gruppen intern möglichst homogen sind. Die Varianz der Gruppen ist in der Ausgangspartition, in der noch keine Gruppenbildung erfolgt ist, für jede „Gruppe“

0. Im Verlauf der darauffolgenden Fusionsstufen verringert sich die Homogentiät in Form einer zunehmenden Streuung innerhalb der Gruppen, wobei das Verfahren anstrebt, die Objekte so zusammenzufassen, dass die Varianz der neu gebildeten Gruppen möglichst gering ist. Der Ward-Algorithmus tendiert zur Bildung von kompakten, in sich homogenen Clustern mit annähernd gleichgroßen Besetzungszahlen. Er gehört deshalb zu den am häufigsten eingesetzten Fusionierungsverfahren in der Marketingpraxis, wobei gewöhnlich auf die quadrierte euklidische Distanz als Proximitätsmaß zurückgegriffen wird.

**(2) Partitionierende Fusionierungsverfahren**: Im Gegensatz zu hierarchischen Verfahren gehen die partitionierenden Ansätze von einer bereits gegebenen Zerlegung der Objektmenge in Cluster aus. Diese wird jedoch nicht als „optimal" erachtet, so dass versucht wird, die Ausgangspartition durch eine Umgruppierung der Objekte sukzessive zu verbessern. Dabei bleibt jedoch – im Gegensatz zu den hierarchischen Methoden – die Anzahl der Cluster in jeder Verfahrensstufe konstant. Partitionierende Verfahren besitzen somit den Vorteil, dass die Zuordnung von Objekten im Gruppierungsprozeß revidierbar ist, um zu einer besseren Clusterlösung zu gelangen. Ihr Einsatz bietet sich insbesondere dann an, wenn der Anwender bereits über Vorkenntnisse bezüglich der relevanten Objektstrukturen (z.B. Zusammensetzung von Marktsegmenten) verfügt. In der Marketingpraxis erfolgt deshalb recht häufig eine kombinierte Anwendung von Fusionierungsmethoden ("sog. two-stage-clustering"), bei welcher z.B. zunächst mittels des Single-Linkage-Verfahrens einzelne Ausreißer identifziert bzw. aus dem Datensatz eliminiert werden, anschließend eine Clusterbildung auf Basis des Ward-Verfahrens erfolgt und die daraus resultierende Clusterlösung sodann mit Hilfe eines partitionierenden Verfahrens verfeinert wird (vgl. Homburg/Krohmer 2003, S. 324).

Partitionierende Verfahren können in zwei Methodengruppen unterteilt werden (vgl. Abb. 3): Jene Verfahren, die eine optimale Teilung der Objektmenge im Hinblick auf ein bestimmtes Kriterium anstreben, werden als **Teilungsverfahren** ( optimierende Austauschverfahren, hill climbing) bezeichnet. Demgegenüber beruhen iterative **Minimal-Distanzverfahren** (Distanzschwellenverfahren, Clusterzentren-Methode, K-Means-Methode) nicht auf einem expliziten Optimierungskriterium. Vielmehr wird hierbei - analog zum Zentroid-Verfahren – für jedes Cluster der Ausgangspartition zunächst der Gruppenschwerpunkt berechnet. Sodann wird für jedes Objekt die Euklidische Distanz zu allen Clusterschwerpunkten errechnet. Sofern ein Objekt eine - im Vergleich zur bislang zugehörigen Gruppe - geringere Distanz zu einem Clusterschwerpunkt besitzt, wird das Objekt dem betreffenden anderen Cluster zugewiesen. Nach dieser Umgruppierung werden die Clusterzentren erneut errechnet und die Objekte möglicherweise wiederum umgruppiert. Diese iterativen Schritte werden solange fortgesetzt, bis eine „optimale" Lösung gefunden ist.

### 2.4. Festlegung der Clusterzahl

Während bei partitionierenden Verfahren die **Zahl der Cluster** vorzugeben ist, bedarf es bei hierarchischen Fusionierungsverfahren der Festlegung der „richtigen" Clusterzahl bzw. einer Entscheidung darüber, an welcher Stelle der Verschmelzungsprozess abgebrochen werden soll. Hierfür steht dem Anwender allerdings kein fest definiertes Abbruchkriterium zur Verfügung. Neben sachlichen Überlegungen erfolgt

daher zumeist ein Rückgriff auf graphische Darstellungen des Fusionierungsprozesses in Form eines Dendrogramms und/oder eines Struktogramms (vgl. Abb. 5).

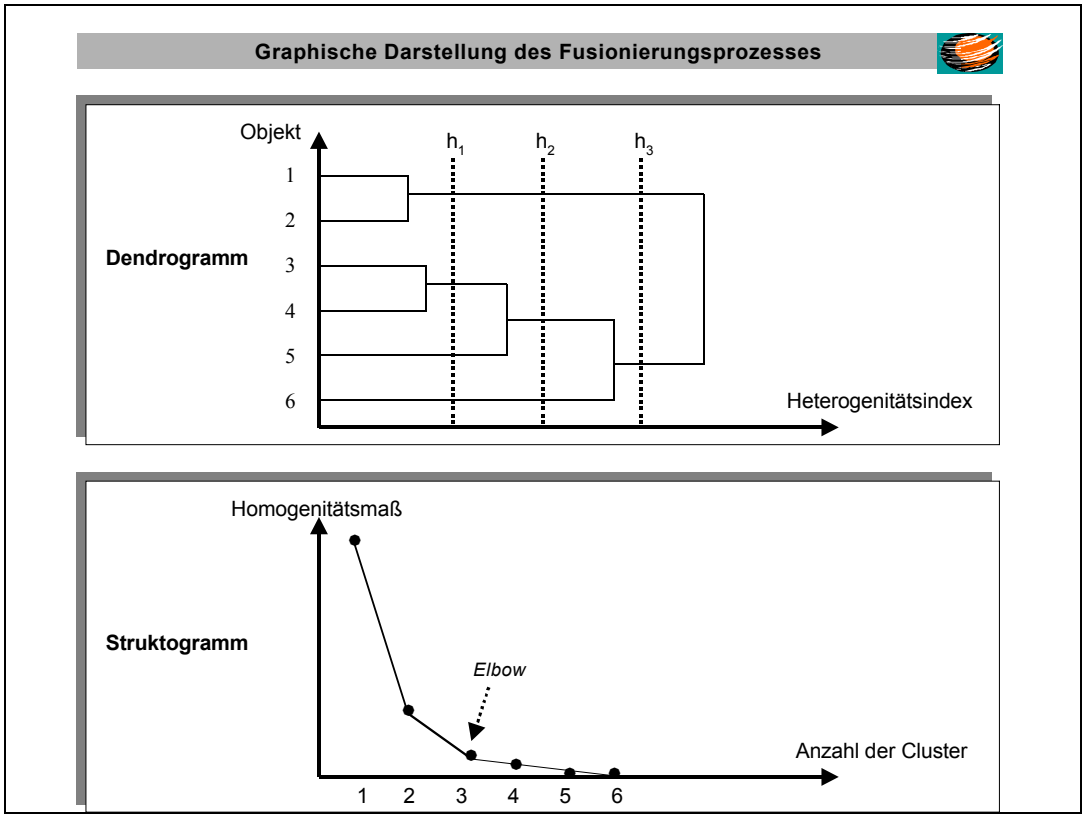



*Abbildung 5*: Dendrogramm und Struktogramm

Im **Dendrogramm** ist der hierarchische Verschmelzungsprozess schematisch dargestellt. Zusätzlich wird angezeigt, bei welchem Heterogenitätsgrad eine Fusion zweier Objektmengen, d.h. Objekte oder Cluster stattfindet. Die Messung der Heterogenität erfolgt anhand des dem betreffenden Fusionierungsverfahren zugrundeliegenden Vereinigungsmaßes, wie z.B. der Distanz der beiden zuletzt fusionierten Objektmengen oder der Fehlerquadratsumme. Im Extremfall bildet jedes Objekt einen eigenen Cluster ( Heterogenität = Max. bzw. Homogenität = 0) oder alle Objekte sind in einem Cluster vereinigt (Heterogenität = 0 bzw. Homogenität = Max). Im Beispiel der Abbildung 5 werden im ersten Fusionierungsschritt die Objekte 1 | 2 zu einer Gruppe vereinigt, während auf der letzten Fusionierungsstufe die aus den Objekten 1 | 2 zusammengesetzte Gruppe mit der aus den Objekten 3 | 4 | 5 | 6 bestehenden Gruppe zu einem Cluster verschmolzen werden.

Der Anwender kann anhand des Verlaufs des Heterogenitätsmaßes über die Zahl der „richtigen“ Cluster entscheiden. Legt man hierzu bei einem vorgegebenen Heterogenitätsindex eine vertikale Linie durch das Dendrogramm, dann entspricht die Anzahl der Cluster jener Anzahl der horizontalen Linien, die von dieser geschnitten werden. Legt der Anwender seinen Überlegungen beispielsweise den Heterogenitätsindex $h_3$ zugrunde, dann geht aus Abbildung 5 hervor, dass hieraus einerseits zwei Zweiergruppen mit den Objekten 1 | 2 sowie 3 | 4 und andererseits zwei einelementige Gruppen der Objekte 5 und 6 resultieren. Betrachtet man hingegen den Indexwert $h_2$

so wird ersichtlich, dass in diesem Fall nur noch zwei Gruppen vorliegen: Das erste Cluster besteht aus den Objekten 1 | 2, während das zweite Cluster aus den Objekten 3 | 4 | 5 | 6 zusammengesetzt ist. Als visuelle Entscheidungshilfe zur Festlegung der Clusterzahl kann jener Bereich des Dendrogramms dienen, bei welchem einer starker Anstieg des Heterogenitätsmaßes erfolgt. Im vorliegenden Beispiel zeigt sich eine sprunghafte Zunahme des Heterogenitätsmaßes beim Index $h_2$, die eine 3-Clusterlösung nahe legt.

Eine alternative Darstellungsweise bietet das **Struktogramm**, bei welchem an der Ordinate die Abnahme des Heterogenitätsmaßes und an der Abzisse die Anzahl der Cluster abgetragen ist. Analog zum sog. Scree-Test einer Faktorenanalyse kann die „richtige“ Anzahl von Clustern an derjenigen Stelle des Liniendiagrammes abgelesen werden, bei der dieses einen starken Knick (Elbow) aufweist. Bezogen auf das Beispiel der Abbildung 5 wäre demnach von einer 3-Clusterlösung auszugehen. Gleichwohl ist darauf hinzuweisen, dass es im praktischen Anwendungsfall durchaus vorkommen kann, dass man an mehr als einer Stelle einen „Ellenbogen“ vorfindet und dieses Kriterium somit keine eindeutige Lösung gestattet. Hiermit einher geht die generelle Empfehlung, die durch ein Dendrogramm oder Struktogramm identifizierte Clusterzahl stets auch vor dem Hintergrund sachlicher Überlegungen zu überprüfen.

### 2.5. Clusterdiagnose

Die abschließende Phase einer Clusteranalyse besteht in der Diagnose der ermittelten Clusterstruktur. Diese beinhaltet Gütebeurteilung der Clusterlösung sowie die Interpretation von Clustern.

Zur **Gütebeurteilung** der gewonnenen Clusterlösung kann der Anwender nicht auf verfahrensimmanente statistische Kriterien zurückgreifen. Ersatzweise bietet es sich daher an, zum einen mit Hilfe von Varianz- und/oder Diskriminanzanalysen zu untersuchen, ob signifikante Gruppenunterschiede hinsichtlich der Klassifizierungsvariablen vorliegen. Daneben kann zum die Stabilität der Clusterlösung bei Anwendung mehrerer Fusionierungsverfahren überprüft werden. Hierbei ist es zudem möglich, den sog. Rand-Index oder Kappa-Wert als Kenngrößen für den Grad der Übereinstimmung der Clusterlösungen zweier Fusionierungsverfahren zu bestimmen (vgl. Bortz 1993, S. 538 ff; Eckey/Kosfeld/Rengers 2002, S. 270 ff.).

Eine vergleichbare Vorgehensweise ist bei der **Interpretation der Cluster** hilfreich. Diese knüpft zum einen an den relevanten Klassifizierungsvariablen an, für die sich z.B. im Fall metrischer Variablen graphische Mittelwertprofile und/oder varainz- bzw. diskrimininzanalytische Mittelwertuntersuchungen durchführen lassen. Um einen differenzierten Einblick in die Clusterstrukturen zu erlangen, ist es darüber hinaus zweckmäßig, Cluster mit Hilfe sog. passiver Zusatzvariablen, d.h. Merkmalen, die nicht als Klassifizierungsvariablen herangezogen wurden, zu beschreiben.

# 3. Hierarchische Clusteranalyse mit SPSS

## 3.1. Die Datenmatrix des Demonstrationsbeispiels

In der Marketingpraxis ist es vielfach üblich, Informationen über die Leistungsangebote konkurrierender Anbieter auf dem Wege einer Durchsicht von Produktprospekten zu sammeln und in Form einer Leistungstabelle gegenüberzustellen. Ein derartiges Vorgehen unterliegt auch der Zusammenstellung der Leistungsausprägungen verschiedener Marken des bundesdeutschen Pkw-Marktes in der Tabelle 4.

In der Datenmatrix sind zwölf Pkw-Modelle anhand von jeweils neun technischen Produkteigenschaften und einem ökonomischen Merkmal bzw. dem Verkaufspreis gegenübergestellt. Ist man nun daran interessiert zu untersuchen, bezüglich welcher Leistungsmerkmale Unterschiede oder Gemeinsamkeiten zwischen den Marken vorliegen, so wird recht schnell deutlich, dass der Vergleich von 12 x 10 = 120 Eigenschaftsausprägungen eine komplexe Beurteilungsaufgabe darstellt, die nicht nur zeitaufwendig ist, sondern auch ein unübersichtliches Leistungsbild vermittelt.

| | **Preis** (DM) | **Länge** (mm) | **Breite** (mm) | **Höhe** (mm) | **Gewicht** (kg) | **PS** | **Hubraum** (ccm) | **Geschwindig-keit** (km/h) | **Beschleuni-gung** (Sek. für 0-100 km/h) | **Verbrauch** (l pro 100 km) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Audi 80** | 12655 | 4383 | 1682 | 1365 | 910 | 55 | 1273 | 145 | 17,5 | 8,9 |
| **BMW 320** | 19300 | 4355 | 1610 | 1380 | 1115 | 122 | 1990 | 181 | 10,7 | 9,5 |
| **Citroen GSX** | 14490 | 4120 | 1608 | 1349 | 935 | 55 | 1130 | 145 | 20,8 | 8,4 |
| **Fiat 131** | 12590 | 4264 | 1651 | 1381 | 1015 | 75 | 1585 | 160 | 12,8 | 9,2 |
| **Ford Taunus** | 11930 | 4340 | 1700 | 1362 | 1020 | 55 | 1285 | 137 | 20,3 | 9,5 |
| **Mercedes 200** | 20261 | 4725 | 1786 | 1438 | 1340 | 94 | 1988 | 160 | 15,2 | 11,1 |
| **Opel Rekord** | 14685 | 4593 | 1726 | 1420 | 1100 | 75 | 1875 | 155 | 16 | 10,2 |
| **Peugeot 244** | 14995 | 4490 | 1690 | 1460 | 1160 | 79 | 1796 | 154 | 15,8 | 10,5 |
| **Renault 20** | 18670 | 4520 | 1726 | 1435 | 1260 | 109 | 1994 | 173 | 12,7 | 10,2 |
| **Simca** | 13224 | 4245 | 1680 | 1390 | 1075 | 75 | 1442 | 154 | 13,9 | 9,7 |
| **VW Passat** | 14925 | 4290 | 1615 | 1360 | 885 | 75 | 1588 | 164 | 13 | 8,8 |
| **Volvo 244** | 17990 | 4898 | 1707 | 1435 | 1280 | 90 | 1986 | 155 | 15 | 11,5 |

*Tabelle 4:* Leistungsmerkmale von Automobilen (Quelle: Hammann/Erichson 2000, S. 257)

Im vorliegenden Beispiel läßt sich mittels einer Clusteranalyse die Frage beantworten, ob die betrachteten Pkw-Modelle anhand ausgewählter Leistungsmerkmale zu Leistungsgruppen zusammengefasst werden können. Bezogen auf die eingangs diskutierten Einsatzfelder der Cluster im Marketing, ist diese Problemstellung dem Bereich der Emittlung von strategischen Wettbewerbergruppen zuzuordnen.

Bei der Auswahl derjenigen Merkmale, anhand derer die relevanten Untersuchungs-objekte klassifiziert werden sollen, bieten sich für unser Beispiel die folgenden drei Alternativen an:

- Erstens könnte man versuchen, die Pkw-Modelle hinsichtlich eines einzelnen Leistungsmerkmals (z.B. dem Hubraum, der Länge oder dem Verkaufspreis) zu gruppieren ( = eindimensionale bzw. monothetische Clusteranalyse).
- Eine zweite Alternative besteht in der Markengruppierung auf der Basis von mehreren Leistungseigenschaften ( = mehrdimensionale bzw. polythetische Clusteranalyse).

❒ Schließlich kann man unmittelbar an den Ergebnissen einer vorgeschalteten Faktorenanalyse ansetzen und die dabei extrahierten Faktoren als Klassifizierungsmerkmale verwenden (= faktorielle Clusteranalyse).

Um im Folgenden den integrierten Einsatz von Datenanalyseverfahren demonstrieren zu können, legen wir unserer Markengruppierung eine faktorielle Clusteranalyse zugrunde. Hierzu dienen die Ergebnisse einer vorgeschalteten Faktorenanalyse als Dateninput. Im Zuge einer Faktorenanalyse wurden die neun technischen Leistungsmerkale des Datensatzes der Tabelle 4 zu zwei Faktoren verdichtet und mit den Bezeichnungen „Geräumigkeit“ und „Sportlichkeit“ versehen (vgl. ausführlich Müller 2004 b). Jedes der zwölf Pkw-Modelle lässt sich daher neben seinen Leistungsausprägungen zusätzlich durch Faktorwerte, d.h. den Ausprägungen hinsichtlich beider Faktoren beschreiben. Überträgt man die Faktorwerte in ein Streudiagramm, so erhält man den in Abbildung veranschaulichten Marktraum.

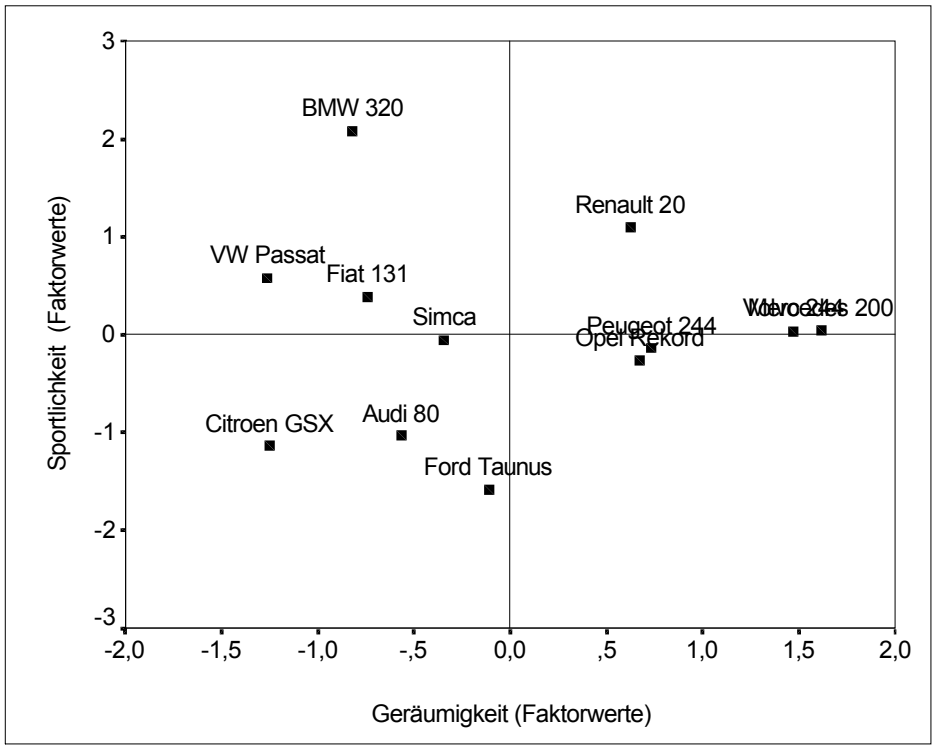



*Abbildung 6*: Streudiagramm auf Basis von Faktorwerten

Angesichts des Sachverhaltes, dass die betrachteten Leistungsmerkmale unterschiedliche Dimensionen aufweisen, wurden diese im Rahmen der Faktorenalyse z-standardisiert. Daher sind auch die Faktorwerte standardisiert und besitzen somit einen Mittelwert von 0 sowie eine Varianz von 1: Ein Faktorwert von 0 indiziert somit, dass das Objekt eine lediglich durchschnittliche Ausprägung besitzt. Demgegenüber zeigt ein positiver (negativer) Faktorwert an, dass das betreffende Objekt in bezug auf diesen Faktor eine im Vergleich zu allen anderen Objekten überdurchschnittliche (unterdurchschnittliche) Ausprägung aufweist. Aus Abbildung 6 ist darüber hinaus ersichtlich, dass die Markenpositionen in annähernd drei Klumpen gebündelt sind. Insofern legt bereits die visuelle Marktbetrachtung der Schluß nahe, dass im Zuge einer Clusteranalyse möglicherweise drei Markengruppen gebildet werden können.

### 3.2. SPSS-Auswertungsmethodik

Um nun die SPSS-Datenmatrix „Leistungspositionierung“ entsprechend der in Abbildung 2 dargestellten clusteranalytischen Untersuchungsmethodik auszuwerten, ist in SPSS die nachstehende Schrittfolge durchzuführen:

(1) Im ersten Schritt einer Clusteranalyse ist die auswertungsrelevante **Datenmatrix** zu erstellen. Führen Sie zunächst auf Basis des Datensatzes der Tabelle 4 eine Faktorenanalyse durch und speichern Sie die erzeugten Faktoren als Variablen „Fgeraeumigkeit“ sowie „Fsportlichkeit“. Speichern Sie die daraus resultierende Datendatei unter der Bezeichnung “Leistungspositionierung“ (vgl. Tabelle 5). Achten Sie bei der Definition der Fallvariablen „Marke“ darauf, dass diese als String-Variable definiert ist. Diese Konvention erleichtert im Rahmen der nachfolgenden hierarchischen Clusteranalyse die Interpretation der Clusterbesetzung.

| | marke | preis | länge | breite | höhe | gewicht | ps | hubraum | geschwin | beschleu | verbrauc | Fgeraeumigkeit | Fsportlichkeit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Audi 80 | 12655,0 | 4383,00 | 1682,00 | 1365,00 | 910,00 | 55,00 | 1273,00 | 145,00 | 17,50 | 8,90 | -,56546 | -1,03617 |
| 2 | BMW 320 | 19300,0 | 4355,00 | 1610,00 | 1380,00 | 1115,00 | 122,00 | 1990,00 | 181,00 | 10,70 | 9,50 | -,82159 | 2,08154 |
| 3 | Citroen GSX | 14490,0 | 4120,00 | 1608,00 | 1349,00 | 935,00 | 55,00 | 1130,00 | 145,00 | 20,80 | 8,40 | -1,25584 | -1,13894 |
| 4 | Fiat 131 | 12590,0 | 4264,00 | 1651,00 | 1381,00 | 1015,00 | 75,00 | 1585,00 | 160,00 | 12,80 | 9,20 | -,74302 | ,38629 |
| 5 | Ford Taunus | 11930,0 | 4340,00 | 1700,00 | 1362,00 | 1020,00 | 55,00 | 1285,00 | 137,00 | 20,30 | 9,50 | -,10682 | -1,58498 |
| 6 | Mercedes 200 | 20261,0 | 4725,00 | 1786,00 | 1438,00 | 1340,00 | 94,00 | 1988,00 | 160,00 | 15,20 | 11,10 | 1,61367 | ,04147 |
| 7 | Opel Rekord | 14685,0 | 4593,00 | 1726,00 | 1420,00 | 1100,00 | 75,00 | 1875,00 | 155,00 | 16,00 | 10,20 | ,66931 | -,26023 |
| 8 | Peugeot 244 | 14995,0 | 4490,00 | 1690,00 | 1460,00 | 1160,00 | 79,00 | 1796,00 | 154,00 | 15,80 | 10,50 | ,73323 | -,14065 |
| 9 | Renault 20 | 18670,0 | 4520,00 | 1726,00 | 1435,00 | 1260,00 | 109,00 | 1994,00 | 173,00 | 12,70 | 10,20 | ,62275 | 1,09832 |
| 10 | Simca | 13224,0 | 4245,00 | 1680,00 | 1390,00 | 1075,00 | 75,00 | 1442,00 | 154,00 | 13,90 | 9,70 | -,34790 | -,05964 |
| 11 | VW Passat | 14925,0 | 4290,00 | 1615,00 | 1360,00 | 885,00 | 75,00 | 1588,00 | 164,00 | 13,00 | 8,80 | -1,26772 | ,57850 |
| 12 | Volvo 244 | 17990,0 | 4898,00 | 1707,00 | 1435,00 | 1280,00 | 90,00 | 1986,00 | 155,00 | 15,00 | 11,50 | 1,46939 | ,03449 |

*Tabelle 5:* Datendatei „Leistungspositionierung“

(2) Fordern Sie nun das Dialogmenü der hierarchischen Clusteranalyse durch die Befehlsfolge „Analysieren/Klassieren/Hierarchische Cluster...“ an. Das Dialogfeld „ *Hierarchische Clusteranalyse“* wird geöffnet (vgl. Abb. 7).

(3) Zunächst wollen wir eine **Auswahl der untersuchungsrelevanten Variablen** vornehmen:

- ⌦ Markieren Sie daher die beiden Klassifizierungsvariablen „Fgeraeumigkeit“ sowie „Fsportlichkeit“ im linken Bereich des Dialogfeldes.
- ⌦ Überführen Sie anschließend die ausgewählten Merkmale aus diesem sog. Quellverzeichnis durch ein Anklicken des oberen Pfeils in die Liste „Variable(n)“
- ⌦ Überführen Sie die Variable „Marke“ in das Feld „Fallbeschriftung“.
- ⌦ Belassen Sie die Voreinstellungen in den Feldern „Cluster: Fälle“ sowie „Anzeigen: Statistik/Diagramme“.

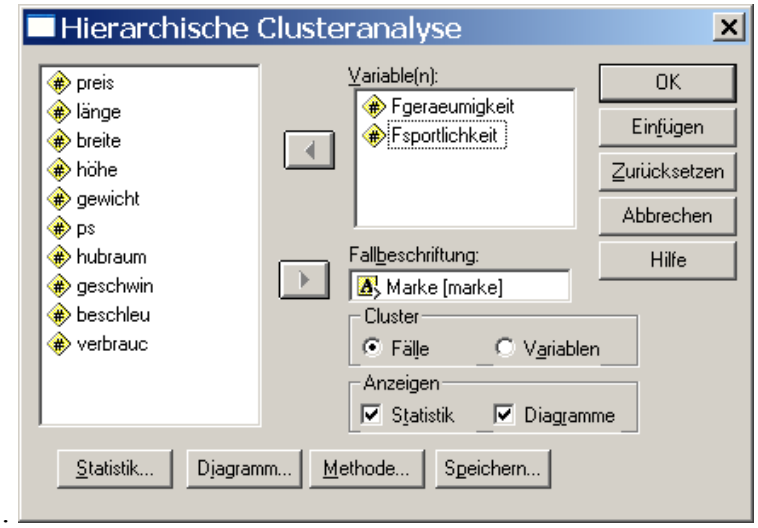



*Abbildung 7*: Dialogfeld „Hierarchische Clusteranalyse“

(4) Der nächste Schritt besteht in der Ermittlung der **Proximitätsmatrix**. Ein Klick auf die Schaltfläche „Statistik“ öffnet das Dialogfenster „*Hierarchische Clusteranalyse: Statistik*“ (vgl. Abb. 8).

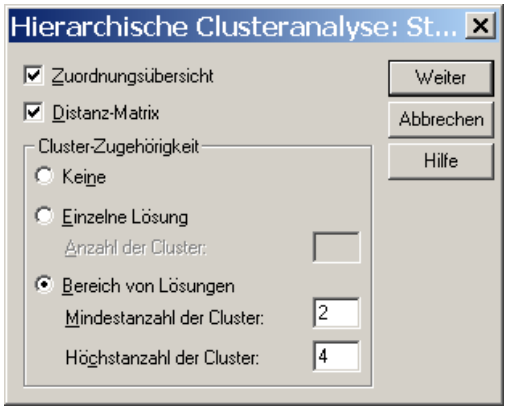



*Abbildung 8*: Dialogfeld „Hierarchische Clusteranalyse: Statistik“

⌦ Die Ausgabe der Distanzmatrix ist standardmäßig nicht voreingestellt. Daher fordern wir diese mit einem Klick auf das Kästchen „Distanz-Matrix“ an.

⌦ Die voreingestellte, von uns übernommene Option „Zuordnungsübersicht“ liefert ein tabellarisches Protokoll der fortlaufenden Fusionierung der Objekte.

⌦ Sofern Vorkenntnisse über die Zahl der zu bildenden Cluster vorhanden sind, kann im Feld „Cluster-Zugehörigkeit“ die Zahl der Cluster exakt vorgegeben, auf bestimmte Bereiche begrenzt oder auch unbestimmt gelassen werden. In unserem Beispiel hingegen wollen wir bereits an dieser Stelle aufgrund unserer Betrachtung des Faktorraumes der Abbildung 6 davon ausgehen, dass sich die Pkw-Modelle möglicherweise zu zwei, drei oder zu vier Markengruppen zusammenfassen lassen. Da wir jedoch keine Kenntnis über die exakte Clusterzahl verfügen, weisen wir SPSS im Feld „Bereich von Lösungen“ an, die Clusterzugehörigkeit der Fälle für ein Clusterergebnis von zwei bis vier Gruppen auszuweisen.

⌦ Bestätigen Sie Ihre Einstellungen mit „Weiter“. Hierauf öffnet sich erneut das Dialogeld „ *Hierarchische Clusteranalyse“*.

(5) Im nachfolgenden Schritt ist das **Fusionierungsverfahren** festzulegen. Klicken Sie daher auf die Schaltfläche „Methode...“, worauf sich die Dialogbox *„Hierarchische Clusteranalyse: Methode“* öffnet (vgl. Abb. 9):

⌦ Das Pull-down-Menü „Cluster-Methode“ ermöglicht die Wahl eines der zuvor besprochenen Fusionierungsverfahrens. Wählen Sie dort die Option „Ward-Methode“.

⌦ Im Feld „Maß“ finden sich für verschiedene Skalenniveaus der Variablen alternative Proximätsmaße. In unserem Beispiel liegen metrische Daten vor, so dass wir im Kästchen “Intervall“ ein Distanzmaß bzw. das Maß „Quadrierter Euklidischer Abstand“ wählen.

⌦ Sofern eine Standardisierung der Klassifizierungsvariablen erforderlich ist, kann im Feld „Werte transformieren“ eine zweckmäßige Standardisierungsform, z.B. eine z-Standardisierung“ angefordert werden. Da unsere Klassfizierungsvariablen standardisierte Faktoren darstellen, belassen wir hier die Voreinstellung „ Keine“.

⌦ Bestätigen Sie Ihre Einstellungen „Weiter“.

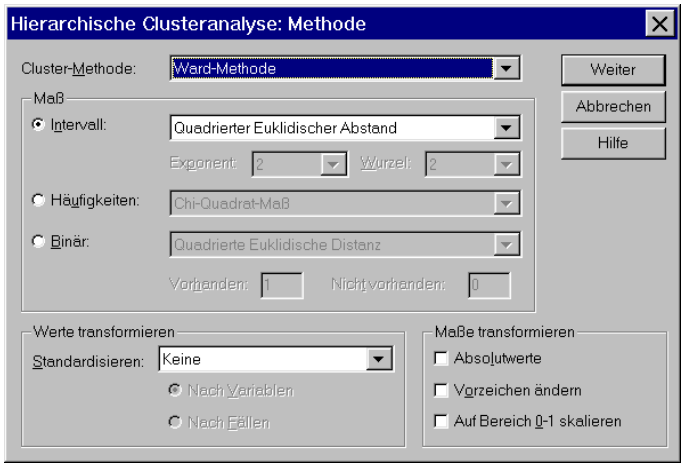



*Abbildung 9*: Dialogbox „Hierarchische Clusteranalyse: Methode“

(6) In einem weiteren Schritt sind Informationen zur **Clusterdiagnose** auszuwählen. Hierzu stellt SPSS im Dialogfenster „*Hierarchische Clusteranalyse: Diagramme*“ zwei graphische Darstellungsalternativen zum Verlauf des Fusionierungsprozesses zur Verfügung (vgl. Abb. 10):

⌦ Dort wählen wir zum einen die Option „Dendrogramm“.

⌦ Zum anderen möchten wir ein „Eiszapfen“-Diagramm betrachten.

⌦ Im Hinblick auf die Darstellungsform wählen wir unter „Orientierung“ die voreingestellte Option „vertikal“.

⌦ Anschließend bestätigen wir unsere Einestellungen mit „Weiter“.


*Abbildung 10*: Dialogbox „ Hierarchische Clusteranalyse: Diagramme“

(7) Schließlich bietet SPSS in die Möglichkeit, die **Clusterzugehörigkeit der Fälle** als zusätzliche Variablen in der Datendatei zu speichern. Hierzu klicken wir in der Dialogbox „Hierarchische Clusteranalyse“ auf die Schaltfläche „Speichern“. Es öffnet sich das Dialogfenster “*Hierarchische Clusteranalyse: Neue Variablen speichern*“ (vgl. Abb. 11):

- In der Analysesituation einer stringent explorativen Clusteranalyse, würde man die Clusterlösung erst im Anschluß an die Clusterprozedur und -diagnose festlegen und speichern.
- In unserem Beispiel haben wir allerdings im Schritt (4) SPSS dazu veranlasst, die Analyseprozedur für einen Lösungsbereich von zwei bis vier Clustern durchzuführen. Deshalb wählen wir an dieser Stelle nun die Option, die Clusterzugehörigkeit der Fälle für ein Clusterergebnis von zwei bis vier Gruppen zu speichern. Hiermit fügt SPSS der Datendatei drei neue Variablen unter den Bezeichnungen „CLU4_1; CLU3_1; CLU2_1“ an.

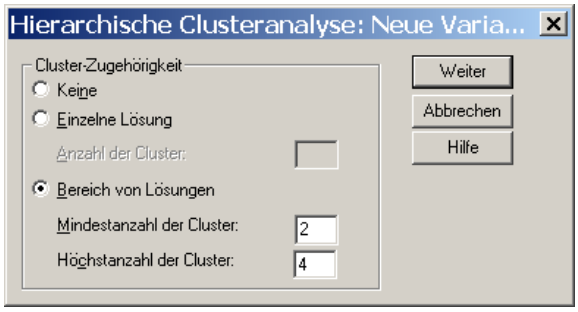



*Abbildung 11:* Dialogbox „Hierarchische Clusteranalyse: Neue Variablen speichern“

### 3.3. Interpretation der Distanzmatrix

Die SPSS-Ausgabe umfasst zunächst die von SPSS als Näherungsmatrix bezeichnete Distanzmatrix für die analysierten zwölf Fälle (vgl. Tabelle 6). In dieser wird die paarweise quadrierte euklidische Distanz zwischen den zwölf Pkw-Modellen ausgewiesen. Je kleiner der Distanzwert, desto ähnlicher sind die zwei Fälle bezüglich der Klassifizierungsvariablen Geräumigkeit und Sportlichkeit. Die Matrix ist symmetrisch strukturiert, so dass alle Distanzwerte zweimal angeführt sind. Es genügt

daher, entweder das obere rechte Dreieick oder das untere linke Dreieck der Matrix zu betrachten. Die Hauptdiagonalwerte der Matrix, welche die Eigendistanzen der Fälle beinhalten, sind jeweils 0,000.

**Näherungsmatrix**

| | Quadriertes euklidisches Distanzmaß | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fall | 1:Audi 80 | 2:BMW 320 | 3:Citroen GSX | 4:Fiat 131 | 5:Ford Taunus | 6: Mercedes 200 | 7:Opel Rekord | 8: Peugeot 244 | 9:Renault 20 | 10: Simca | 11:VW Passat | 12:Volvo 244 |
| 1:Audi 80 | ,000 | 9,786 | ,487 | 2,055 | ,512 | 5,910 | 2,127 | 2,489 | 5,968 | 1,001 | 3,100 | 5,287 |
| 2:BMW 320 | 9,786 | ,000 | 10,560 | 2,880 | 13,954 | 10,092 | 7,707 | 7,356 | 3,053 | 4,809 | 2,458 | 9,439 |
| 3:Citroen GSX | ,487 | 10,560 | ,000 | 2,589 | 1,519 | 9,627 | 4,478 | 4,953 | 8,534 | 1,989 | 2,950 | 8,804 |
| 4:Fiat 131 | 2,055 | 2,880 | 2,589 | ,000 | 4,291 | 5,673 | 2,413 | 2,457 | 2,372 | ,355 | ,312 | 5,019 |
| 5:Ford Taunus | ,512 | 13,954 | 1,519 | 4,291 | ,000 | 5,605 | 2,357 | 2,792 | 7,732 | 2,385 | 6,028 | 5,107 |
| 6:Mercedes 200 | 5,910 | 10,092 | 9,627 | 5,673 | 5,605 | ,000 | ,983 | ,808 | 2,099 | 3,858 | 8,591 | ,021 |
| 7:Opel Rekord | 2,127 | 7,707 | 4,478 | 2,413 | 2,357 | ,983 | ,000 | ,018 | 1,848 | 1,075 | 4,456 | ,727 |
| 8:Peugeot 244 | 2,489 | 7,356 | 4,953 | 2,457 | 2,792 | ,808 | ,018 | ,000 | 1,547 | 1,175 | 4,521 | ,573 |
| 9:Renault 20 | 5,968 | 3,053 | 8,534 | 2,372 | 7,732 | 2,099 | 1,848 | 1,547 | ,000 | 2,283 | 3,844 | 1,849 |
| 10:Simca | 1,001 | 4,809 | 1,989 | ,355 | 2,385 | 3,858 | 1,075 | 1,175 | 2,283 | ,000 | 1,253 | 3,311 |
| 11:VW Passat | 3,100 | 2,458 | 2,950 | ,312 | 6,028 | 8,591 | 4,456 | 4,521 | 3,844 | 1,253 | ,000 | 7,788 |
| 12:Volvo 244 | 5,287 | 9,439 | 8,804 | 5,019 | 5,107 | ,021 | ,727 | ,573 | 1,849 | 3,311 | 7,788 | ,000 |

Dies ist eine Unähnlichkeitsmatrix

*Tabelle 6:* Distanzmatrix

Die **Distanzberechnung** sei exemplarisch anhand des Objektpaares „BMW320 | VWPassat“ erläutert. In der Distanzmatrix wird für das betreffende Objektpaar ein Distanzwert von 2,458 ausgewiesen. Die quadrierte Euklidische Distanz errechnet sich als Summe der quadrierten Differenzen zwischen den Variablenwerten (hier: den Faktorwerten, vgl. Tabelle 5). Somit ergibt sich der Distanzwert von 2,458 als: $((- 0{,}82159) – (-1{,}26772))^2 + (2{,}08154 – 0{,}57850)^2$.

Mit 13,954 wird der größte Distanzwert für das Markenpaar „BMW | Ford Taunus“ ausgewiesen. Beide Pkw-Modelle sind sich demnach sehr unähnlich; was vornehmlich auf die starken Unterschiede in der Sportlichkeitsdimension zurückzuführen ist. Demgegenüber liegen für die Markenpaare „Volvo244 | Mercedes 200“ sowie „Opel Rekord | Peugeot 244“ vergleichsweise geringfügige Unterschiede vor. Diese Objektrelationen entsprechen dem optischen Eindruck des Faktorraumes der Abbildung 6.

### 3.4. Darstellung des Agglomerationsprozesses

Der Verlauf des hierarchischen Fusionierungprozesses kann in dreifacher Weise nachvollzogen werden, und zwar durch die Betrachtung

- der Agglomerationstabelle,
- des Eiszapfen-Diagramms sowie
- des Dendrogramms.

Die **Agglomerationstabelle** zeigt den Verlauf der Clusterbildung von der ersten Stufe, in der jedes Objekt einen eigenständigen Cluster bildet, bis zur letzten Stufe, in der alle Objekte zu einem Cluster zusammengefasst sind (vgl. Tabelle 8):

**Zuordnungsübersicht**

| Schritt | Zusammengeführte Cluster | | Koeffizienten | Erstes Vorkommen des Clusters | | Nächster Schritt |
|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | | Cluster 1 | Cluster 2 | |
| 1 | 7 | 8 | 9,193E-03 | 0 | 0 | 7 |
| 2 | 6 | 12 | 1,962E-02 | 0 | 0 | 7 |
| 3 | 4 | 11 | ,176 | 0 | 0 | 5 |
| 4 | 1 | 3 | ,419 | 0 | 0 | 6 |
| 5 | 4 | 10 | ,903 | 3 | 0 | 9 |
| 6 | 1 | 5 | 1,499 | 4 | 0 | 10 |
| 7 | 6 | 7 | 2,262 | 2 | 1 | 8 |
| 8 | 6 | 9 | 3,574 | 7 | 0 | 11 |
| 9 | 2 | 4 | 5,951 | 0 | 5 | 10 |
| 10 | 1 | 2 | 12,848 | 6 | 9 | 11 |
| 11 | 1 | 6 | 22,000 | 10 | 8 | 0 |

*Tabelle 7: Agglomerationstabelle*

(*Anmerkung*: Vielfach stört die von SPSS per Voreinstellung vorgenommene Notation für kleine Zahlen, wie z.B. der Wert 9,193E-03 in der vorstehenden Tabelle. In der SPSS Version 11.0 (oder höher) kann man diese jedoch mittels der Befehlsfolge „Bearbeiten/Optionen/Keine wissenschaftliche Notation für kleine Zahlen“ ausschalten. So entspricht beispielsweise die Notation 9.193E-03 dem Wert 0,009).

Jede **Zeile** der Tabelle beschreibt eine Stufe der Clusterbildung, die in der ersten Spalte **„Schritt“** angezeigt wird. Bei einer Gesamtzahl von n Objekten werden insgesamt (n–1)-Agglomerationsschritte durchgeführt; im vorliegenden Beispiel daher 11 Schritte. Gemäß der Spalte **„Zusammengeführte Cluster**“ werden im ersten Schritt die Objekte mit den Fallnummern 7 (Opel Rekord) und 8 (Peugeot 244) zu einem Cluster zusammengefasst. Die in der Spalte **„Koeffizienten“** ausgewiesenen Werte repräsentieren die jeweiligen Werte des verfahrensspezifischen Distanzmaßes. Da der Agglomerationssprozess des Ward-Verfahrens jedoch nicht auf Clusterdistanzen beruht, sondern das Varianzkriterium als Heterogentitätsmaß verwendet, zeigen die Koeffizienten der Tabelle 7 die Entwicklung der Fehlerquadratsumme an. Im zweiten Fusionierungsschritt werden die Fälle mit dem zweitgeringsten Varianzzuwachs und somit die Fallnummern 6 ( Mercedes) sowie 12 ( Volvo 244) zu einem Cluster vereinigt.

In der Spalte „**Nächster Schritt**“ wird für jede Agglomerationsstufe dargelegt, in welchem Schritt der gerade neu gebildete Cluster mit einem weiteren Cluster zusammengeführt wird. So wird z.B. angezeigt, dass die auf der ersten Stufe gebildete Gruppe (7 | 8) erst in Schritt 7 wieder mit einem anderen Cluster verschmolzen wird. Unter der Überschrift „**Erstes Vorkommen des Clusters“** wird angegeben, auf welcher Stufe der betreffende Cluster in dieser Form gebildet wurde. Da z.B. im Schritt 1 beide Cluster nur aus einem Objekt bestehen, finden sich in dieser Spalte nur Nullen. Demgegenüber erscheint z.B. im Schritt 5 unter „Erstes Vorkommen des Clusters“ eine 3, weil das betreffende Cluster (4 | 10) im dritten Schritt gebildet wurde. Dabei werden die Cluster stets durch das Objekt mit der niedrigsten Fallnummer gekenzeichnet, das ihm angehört. Dies bedeutet, dass sich im vorliegenden Beispiel die Nummer 4 auf das im dritten Schritt gebildete Cluster (4 | 11)

bezieht, während die Nummer 10 das nunmehr neu hinzugefügte Einzelobjekt 10 (Simca) anzeigt.

Der Agglomerationstabelle ist ferner entnehmbar, dass die Zunahme der Fehlerquadratsumme zunächst vergleichsweise gering ist, während diese auf den späteren Stufen größer wird und insbesondere bei den Schritten 9 und 10 deutliche Varianzsprünge aufweist. Dies bedeutet, dass die auf den unteren Stufen gebildeten Cluster noch vergleichsweise homogen sind, während die darauffolgenden Schritte zunehmend heterogene Cluster erzeugen.

Das vertikale **Eiszapfen-Diagramm** beinhaltet die grafische Darstellung der in der Agglomerationstabelle enthaltenen Informationen (vgl. Tabelle 8). Das Diagramm beschreibt, von unten nach oben gelesen, den Ablauf der Clusterbildung.

**Vertikales Eiszapfendiagramm**

| Anzahl der Cluster | Fall | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 9:Renault 20 | | 8:Peugeot 244 | | 7:Opel Rekord | | 12:Volvo 244 | | 6:Mercedes 200 | | 10:Simca | | 11:VWPassat | | 4:Fiat 131 | | 2:BMW 320 | | 5:Ford Taunus | | 3:Citroen GSX | | 1:Audi 80 |
| 1 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 2 | X | X | X | X | X | X | X | X | X | | X | X | X | X | X | X | X | X | X | X | X | X | X |
| 3 | X | X | X | X | X | X | X | X | X | | X | X | X | X | X | X | X | | X | X | X | X | X |
| 4 | X | X | X | X | X | X | X | X | X | | X | X | X | X | X | | X | | X | X | X | X | X |
| 5 | X | | X | X | X | X | X | X | X | | X | X | X | X | X | | X | | X | X | X | X | X |
| 6 | X | | X | X | X | | X | X | X | | X | X | X | X | X | | X | | X | X | X | X | X |
| 7 | X | | X | X | X | | X | X | X | | X | X | X | X | X | | X | | X | | X | X | X |
| 8 | X | | X | X | X | | X | X | X | | X | | X | X | X | | X | | X | | X | X | X |
| 9 | X | | X | X | X | | X | X | X | | X | | X | X | X | | X | | X | | X | | X |
| 10 | X | | X | X | X | | X | X | X | | X | | X | | X | | X | | X | | X | | X |
| 11 | X | | X | X | X | | X | | X | | X | | X | | X | | X | | X | | X | | X |

*Tabelle 8: Eiszapfen-Diagramm*

Jede **Zeile** des Eiszapfen-Diagramms bezieht sich auf eine Stufe des Fusionierungsprozesses, wobei die unterste Zeile den ersten Agglomerationsschritt darstellt. Die mit den **Kreuzen** ausgefüllten Spalten repräsentieren jeweils ein einzelnes Objekt; so stellt z.B. die Spalte 1 das Pkw-Modell „Renault 20“ dar.

In der untersten Zeile sind z.B. die benachbarten Objekte Mercedes und Simca durch einen **Freiraum** voneinander getrennt. Dieser Freiraum symbolisiert, dass die beiden benachbarten Modelle unterschiedlichen Clustern angehören. Demgegenüber sind die benachbarten Objekte „Peugeot 244 und „Opel Rekord“ nicht durch einen Freiraum getrennt. Vielmehr wird der Raum zwischen beiden Spalten durch ein Kreuz ausgefüllt, so dass eine Verbindung zwischen beiden Spalten besteht.

Die auf jeder Stufe vorhandene Clusterzahl wird in der Spalte „**Anzahl der Cluster**“ angezeigt. Nach der ersten Agglomerationsstufe bestehen dementsprechend 11 Cluster, von denen 10 Cluster jeweils nur ein einzelnes Pkw-Modell umfassen und ein Cluster die beiden Objekte „ Peugeot 244“ und „Opel Rekord“ enthält. Im nächsten Schritt (Clusterzahl = 10) findet eine Zusammenlegung der Fälle 12 (Volvo 244) und

6 (Mercedes) statt. Im letzten Schritt der Agglomeration sind alle Fälle in einem Cluster vereint bzw. durch ein Kreuz verbunden.

Sowohl die Agglomerationstabelle als auch das Eiszapfen-Diagramm bieten einen recht unübersichtlichen Einblick in den Fusionierungsprozess. Eine grafisch ansprechendere und mit größerem Informationsgehalt versehene Darstellung liefert das im nächsten Abschnitt behandelte Dendrogramm, anhand dessen auch eine Bestimmung der Clusterzahl vorgenommen werden kann.

### 3.5. Bestimmung der Clusterzahl

Zur Bestimmung der „richtigen“ Clusterzahl stehen im Rahmen der Clusteranalyse keine mathematische Kriterien zur Verfügung. Ersatzweise bedarf es daher Überlegungen, die durch drei SPSS-Teilausgaben unterstützt werden können:

- den Distanzkoeffizienten in der Agglomerationstabelle
- dem Dendrogramm sowie
- dem Struktogramm..

Einen ersten Hinweis auf die sachgemäße Clusterzahl liefert die Analyse der Distanzkoeffizienten in der **Agglomerationstabelle** (vgl. Tabelle 7). Ín der Agglomerationstabelle ist der Verschmelzungsprozess generell an jener Stelle abzubrechen, bei der ein deutlicher Sprung des betreffenden Distanzkoeffizienten auftritt, da andernfalls unähnliche bzw. heterogene Cluster zusammengefasst werden. Im vorliegenden Beispiel ist mit Übergang vom Agglomerationsschritt 9 nach Schritt 10 eine deutliche Sprungstelle verbunden, denn hiermit geht ein Zuwachs der Fehlerquadratsumme von 5,95 nach 12,84 einher. Diese Sprungstelle kann als Indikator für eine sinnvolle Clusterlösung dienen. Hiernach entspricht die sachgemäße Clusterzahl der Differenz zwischen der Anzahl der zu clusternden Fälle (hier: 12) und der Schrittzahl, hinter der sich der Koeffizient sprunghaft erhöht (hier: 9). Im vorliegenden Beispiel erscheint demzufolge eine 3-Clusterlösung sinnvoll zu sein.

Ein **Struktogramm** beinhaltet die grafische Darstellung der Agglomerationstabelle**,** das in SPSS allerdings keinen Ausgabe-Bestandteil der Clusteranalyse darstellt, sondern mittels des SPSS-Grafikmenüs als Liniendiagramm erstellt werden kann. Bei diesem sind die Koeffizientenwerte der Agglomerationstabelle an der Ordinate und die Clusterzahl an der Abzsisse abgetragen. Für unser Beispiel resultiert daraus die Abbildung 12, die gleichfalls eine 3-Clusterlösung nahe liegt, da dort ein „Knick“ erkennbar ist.

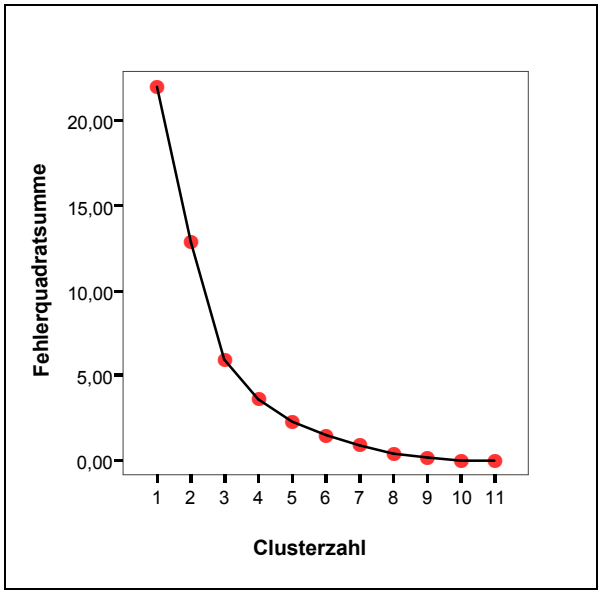



*Abbildung 12:* Struktogramm der Beispieldaten

Ein **Dendrogramm** vermittelt neben der Darstellung des Fusionierungsprozesses und dem Ausweis von Distanzen zwischen den jeweils gebildeten Clustern, auch einen Einblick in die Clusterzugehörigkeit von Fällen. Das Dendrogramm für das vorliegende Beispiel ist in Abbildung 13 veranschaulicht.

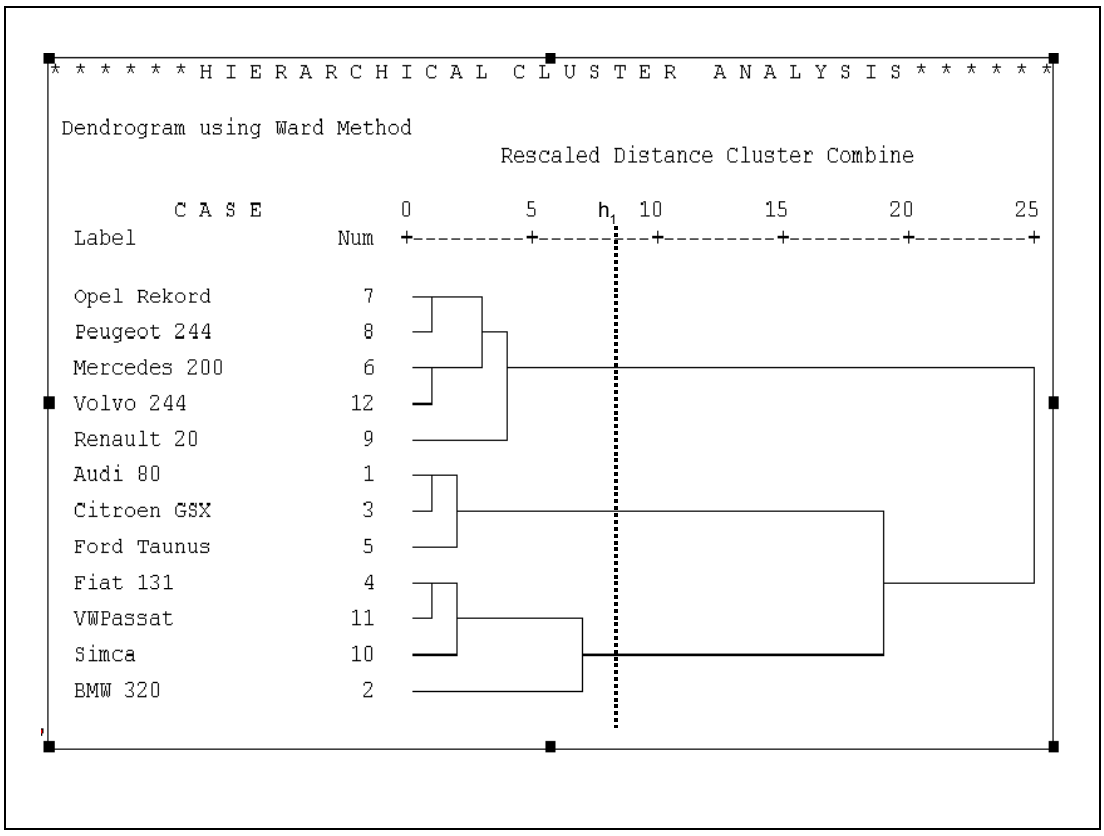



*Abbildung 13:* Dendrogramm für die Beispieldaten

Das Dendrogramm ist von links nach rechts zu lesen und beschreibt in dieser Richtung die einzelnen Stufen der Clusterbildung. Jede Zeile des Diagramms repräsentiert ein einzelnes Objekt. Jene Objekte, die durch eine senkrechte Linie miteinander

verbunden sind, gehören einem gemeinsamen Cluster an. Die Distanz der Cluster wird durch die Länge der horizontalen Linie angezeigt und mittels der Skala „Rescaled Distance Cluster Combine“ quantifiziert. Die in der Skala abgetragenen Distanzwerte entsprechen allerdings nicht den in der Agglomerationstabelle enthaltenen tatsächlichen Distanzwerten, sondern werden aus grafischen Darstellungsgründen in einer Skala mit dem Wertebereich von 0 bis 25 transformiert (vgl. zur Transformationsmethodik ausführlich Rudolf/Müller 2004, S. 160 f.). Beim Ward-Verfahren besteht hierbei allerdings eine Besonderheit darin, dass die Koeffizienten der Agglomerationstabelle und somit auch Werte der Distanzskala des Dendrogramms nicht die Clusterdistanzen, sondern die Werte des Heterogenitätsmaßes bzw. der Fehlerquadratsumme zum Ausdruck bringen. So entspricht z.B. die in der letzten Zeile der Agglomerationstabelle ausgewiesene Fehlerquadratsumme von 22 dem im Dendrogramm angezeigten Skalenwert von 25.

Der Fusionierungsprozess kann u.a. an den folgenden Schritten abgelesen werden:

- Im ersten Fusionierungsschritt wurden vier Cluster gebildet. Diese bestehen aus den Markenpaaren „Opel Rekord | Peugeot 244“, „Mercedes 200 | Volvo 244“, „Audi 80 | Citroen GSX“ und „Fiat 131 | VW Passat“.
- Im zweiten Schritt wurde einerseits das Objekt „Taunus“ dem bereits bestehenden Cluster „Audi 80 | Citroen GSX“ zugeordnet und andererseits das Objekt „Simca“ dem ebenfalls bereits gebildeten Cluster „Fiat 131 | VW Passat“ zugewiesen.
- In der dritten Fusionierungsstufe wurden die beiden Cluster „Opel Rekord | Peugeot 244“ und „Mercedes 200 | Volvo 244“ zu einem gemeinsamen Cluster verbunden.
- Im letzten Schritt wurde ein globales, aus sämtlichen 12 Pkw-Modellen zusammengesetztes Cluster gebildet.

Zur Festlegung der Clusterzahl ist der Fusionierungsprozeß an jener Stelle des Dendrogramms zu beenden, die durch einen großen Distanzsprung gekennzeichnet ist bzw. sich die Heterogenität sprunghaft erhöht. Denn ein überproportionaler Varianzzuwachs signalisiert, dass die weitere Verringerung der Clusterzahl im Vergleich zur bisherigen Gruppenzahl zu einer überproportionalen Zunahme der gruppeninternen Heterogenität führt. Legt man daher in Abbildung 13 eine gedachte vertikale Linie durch das Dendrogramm, die z.B. dem Heterogenitätswert $h_1$ entspricht, dann werden drei horizontale Linien geschnitten, so dass sich wiederum eine 3-Clusterlösung ergibt.

### 3.6. Clusterdiagnose

Die abschließende Phase der Clusteranalyse beinhaltet die Diagnose der zuvor ermittelten 3-Clusterlösung. Die Clusterdiagnose umfasst zum einen die Beschreibung von Clustern und zum anderen eine Güteprüfung hinsichtlich der Homogenität, der Heterogenität sowie der Stabilität der Clusterlösung.

**(1) Clusterbeschreibung:** Zur inhaltlichen Kennzeichnung von Clustern werden gewöhnlich sowohl Klassifizierungsvariablen als auch zusätzliche, sog. passive Variablen der untersuchten Objekte herangezogen Im vorliegenden Beispiel beinhaltet

der untersuchte Datensatz keine Passualvariablen, so dass die Clusterbeschreibung allein auf Basis der Klassifizierungsvariablen erfolgt. Insofern wird nachfolgend zum einen die Gruppenzugehörigkeit der Fälle und zum anderen das Merkmalsprofil der Klassifizierungsvariablen zur Clusterbeschreibung herangezogen.

❒ **Gruppenzugehörigkeit von Fällen**: Aufschluss über die Gruppenzugehörigkeit bei alternativen Clusterlösungen vermittelt die Zuordnungstabelle (vgl. Tabelle 9). Hiernach setzen sich die Cluster im Fall einer 3-Gruppenlösung wie folgt zusammen:

- ➲ *Cluster 1*: Audi 80 | Citroen GSX | Ford Taunus,
- ➲ *Cluster 2*: BMW 320 | Fiat 131 | Simca | VWPassat,
- ➲ *Cluster 3*: Mercedes 200 | Opel Rekord | Peugeot 244 | Renault 20 | Volvo 244.

Bezogen auf die Gesamtzahl der Fälle betragen somit die Besetzungsanteile für das erste Cluster 25%, für Cluster 2 ca. 33,3% und für Cluster 3 ca. 41,7 %.

**Cluster-Zugehörigkeit**

| Fall | 4 Cluster | 3 Cluster | 2 Cluster |
|---|---|---|---|
| 1:Audi 80 | 1 | 1 | 1 |
| 2:BMW 320 | 2 | 2 | 1 |
| 3:Citroen GSX | 1 | 1 | 1 |
| 4:Fiat 131 | 3 | 2 | 1 |
| 5:Ford Taunus | 1 | 1 | 1 |
| 6:Mercedes 200 | 4 | 3 | 2 |
| 7:Opel Rekord | 4 | 3 | 2 |
| 8:Peugeot 244 | 4 | 3 | 2 |
| 9:Renault 20 | 4 | 3 | 2 |
| 10:Simca | 3 | 2 | 1 |
| 11:VWPassat | 3 | 2 | 1 |
| 12:Volvo 244 | 4 | 3 | 2 |

*Tabelle 9:* Clusterzuordnungen der Beispielobjekte

Sofern die Clusterbildung auf Basis von lediglich zwei metrischen Variablen durchgeführt wurde, kann die Clusterzugehörigkeit der Objekte auch in Form eines Streudiagramms veranschaulicht werden. Hierzu verwenden wir in SPSS das betreffende Grafikmenü und erhalten für unser Beispiel die Abbildung 14, der die folgenden Informationen entnehmbar sind:

- ➲ Durchschnittliche Faktorwerte werden durch Bezugslinien angezgiegt, deren Skalenwert jeweils Null beträgt.
- ➲ Die Faktorwerte der dem Cluster 1 zugehörigen Marken sind durch eine jeweils unterdurchschnittliche Ausprägung bezüglich beider Faktoren gekennzeichnet. Insofern lässt sich diese Gruppe mit dem Begriff „Leistungsarme Klasse“ umschreiben.
- ➲ Im Cluster 2 befinden sich Marken, deren Faktorwerte hinsichtlich der Sportlichkeitsdimension überdurchschnittliche Ausprägungen und bezüglich der Geräumigkeitsdimension lediglich unterdurchschnittliche Ausprägungen

aufweisen. Daher liegt es nahe, diese Gruppe als „Sportlichkeits-Klasse“ zu bezeichnen.

➲ Die dritte Gruppe enthält Marken, deren Faktorwerte bezüglich der Sportlichkeitsdimension als unterdurchschnittlich und hinsichtlich der Geräumigkeitsdimension als überdurchschnittlich einzustufen sind. Demzufolge kann diese Gruppe mit der Bezeichnung „Geräumigkeits-Klasse“ versehen werden.

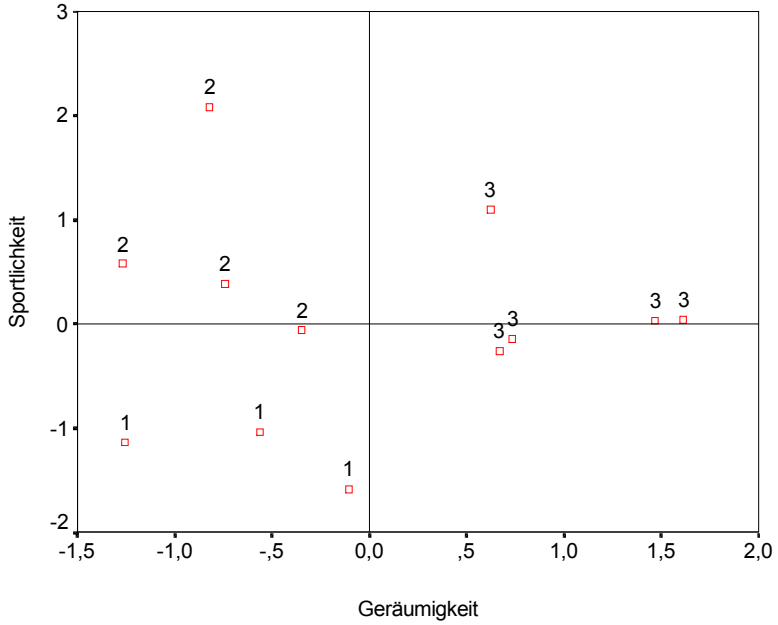



*Abbildung 14: Streudiagramm zur Clusterzugehörigkeit*

❒ **Merkmalsprofil der Cluster**: Bei der Clusterinterpretation mittels Merkmalsprofilen ermittelt man gewöhnlich für jedes Cluster skalenabhängige Lageparameter (z.B. den Mittelwert), anhand derer die typische Merkmalsstruktur der Cluster herausgeschält werden soll. Ein gebräuchliches statistisches Maß stellt bei metrischen Daten der sog. t-Wert dar (vgl. Backhaus et. al. 2000, S. 379 ff.), der sich bestimmt als: $t = \frac{\overline{X}(J,G) - \overline{X}(J)}{S(J)}$ , mit:

$\overline{X}(J,G)$ = Mittelwert der Variable J über die Objekte der Gruppe G

$\overline{X}(J)$ = Gesamtmittelwert der Variable J in der Erhebungsgesamtheit

$S(J)$ = Standardabweichung der Variable J in der Erhebungsgesamtheit

Negative t-Werte verweisen darauf, dass die betreffende Variable im Vergleich zur Erhebungsgesamtheit unterrepräsentiert ist. Demgegenüber zeigt ein positiver t-Wert an, dass die Variable in der Gruppe im Vergleich zur Erhebungsgesamtheit überrepräsentiert ist. Die t-Werte bilden allerdings keinen Bestandteil der SPSS-Ausgabe, so dass eigene Berechnungen mit Hilfe von z.B. Excel durchzuführen sind. Zwecks einer vertiefenden Einsicht in die Clusterstrukturen berechnen wir in unserem Beispiel die t-Werte nicht auf Grundlage der Faktorwerte, sondern auf Basis der Merkmalsausprägungen der Ausgangsvariablen. Hiernach erhalten wir die in Tabelle 10 angeführten t-Werte.

Die t-Werte bestätigen tendenziell die im Zusammenhang mit der Interpreation des des Streudiagramms formulierten Aussagen. So ist beispielsweise für das Cluster 1 („Leistungsarme Klasse“) zu erkennen, daß sämtliche Variablen (mit Ausnahme der „Beschleunigung“) einen negativen t-Wert aufweisen. Das Sportlichkeits-Cluster 2 hingegen ist insbesondere durch positive t-Werte bezüglich der PS-Zahl sowie der Geschwindigkeit gekennzeichnet.

| | FAKTOR 1 | | | | |
|---|---|---|---|---|---|
| | Länge | Breite | Höhe | Gewicht | Verbrauch |
| Mittelwert in der Erhebungsgesamtheit | 4435,25 | 1681,75 | 1397,92 | 1091,25 | 9,79 |
| Standardabweichung in der Erhebungsgesamtheit | 221,03 | 53,78 | 37,63 | 148,42 | 0,94 |
| Mittelwert in Cluster 1 | 4281,00 | 1663,33 | 1358,67 | 955,00 | 8,93 |
| t-Wert für Cluster 1 | -0,70 | -0,34 | -1,04 | -0,92 | -0,91 |
| Mittelwert in Cluster 2 | 4288,50 | 1639,00 | 1377,75 | 1022,50 | 9,30 |
| t-Wert für Cluster 2 | -0,66 | -0,79 | -0,54 | -0,46 | -0,52 |
| Mittelwert in Cluster 3 | 4645,20 | 1727,00 | 1437,60 | 1228,00 | 10,70 |
| t-Wert für Cluster 3 | 0,95 | 0,84 | 1,05 | 0,92 | 0,97 |

| | FAKTOR 2 | | | |
|---|---|---|---|---|
| | PS | Hubraum | Geschwindigkeit | Beschleunigung |
| Mittelwert in der Erhebungsgesamtheit | 79,92 | 1661,00 | 156,92 | 15,31 |
| Standardabweichung in der Erhebungsgesamtheit | 21,06 | 320,32 | 12,11 | 3,05 |
| Mittelwert in Cluster 1 | 55,00 | 1229,33 | 142,33 | 19,53 |
| t-Wert für Cluster 1 | -1,18 | -1,35 | -1,20 | 1,38 |
| Mittelwert in Cluster 2 | 86,75 | 1651,25 | 164,75 | 12,60 |
| t-Wert für Cluster 2 | 0,32 | -0,03 | 0,65 | -0,89 |
| Mittelwert in Cluster 3 | 89,40 | 1927,80 | 159,40 | 14,94 |
| t-Wert für Cluster 3 | 0,45 | 0,83 | 0,20 | -0,12 |

*Tabelle 10:* Variablenspezifische t-Werte für die 3-Cluster-Lösung

**(2) Güteprüfung der Clusterlösung:** Im Rahmen der Güteprüfung ist zunächst die Frage zu beantworten, ob die Clusterlösung dem angestrebten Ziel der internen Clusterhomogenität bzw. externen Clusterheterogenität gerecht wird:

- **Homogenitätsprüfung:** Ein Kriterium zur Überprüfung der Homogenität einer Gruppe stellt der sog. F-Wert dar (vgl. Backhaus et. al. 2000, S. 378 ff.). Dieser ergibt sich aus der Division der Varianz einer Klassifizierungsvariablen in der Erhebungsgesamtheit (V) und der Varianz dieser Variablen in einer Gruppe (Vg) und somit als: $F = \frac{V(J,G)}{V(J)}$. Je kleiner demnach der F-Wert ist, desto geringer ist

die Streuung dieser Variablen in einer Gruppe im Vergleich zur Erhebungsgesamtheit. Ein Cluster kann dann als vollkommen homogen beurteilt werden, wenn die F-Werte für alle Klassifizierungsvariablen kleiner als Eins sind. Wie bereits angesprochen, bilden die beiden faktoriellen Klassifikationsvariablen in unserem Beispiel jeweils standardisierte Werte, so dass in der Erhebungsgesamtheit der Mittelwert gleich 0 ist und die Varianz den Wert 1 aufweist. Das hat zur Folge, dass die Varianz der Klassifikationsvariablen in den Gruppen gleich dem F-Wert ist. Aus diesem Grund empfiehlt es sich, F-Werte für die jeweils „hinter den Faktoren" stehenden Ausgangsvariablen zu berechnen. Das nicht in SPSS ausgewiesene und daher mit Excel ermittelte Ergebnis ist der nachstehenden Tabelle 11 zu entnehmen. Es ist u.a. ersichtlich, dass sich für die Variable „PS" in Cluster 2 ein Wert von größer 1 ergibt. Diese Variable weist demnach im Cluster 2 eine größere Heterogenität auf als in der Erhebungsgesamtheit. Insgesamt lässt sich jedoch der Homogenitätsgrad der 3-Clusterlösung als zufriedenstellend beurteilen.

| | FAKTOR 1 | | | | |
|---|---|---|---|---|---|
| | Länge | Breite | Höhe | Gewicht | Verbrauch |
| Varianz in der Erhebungsgesamtheit | 48854,75 | 2892,20 | 1415,72 | 22027,84 | 0,88 |
| Varianz in Cluster 1 | 19903,00 | 2377,33 | 72,33 | 3325,00 | 0,30 |
| F-Wert für Cluster 1 | 0,41 | 0,82 | 0,05 | 0,15 | 0,34 |
| Varianz in Cluster 2 | 2305,67 | 1080,67 | 160,25 | 10091,67 | 0,15 |
| F-Wert für Cluster 2 | 0,05 | 0,37 | 0,11 | 0,46 | 0,17 |
| Varianz in Cluster 3 | 28190,70 | 1313,00 | 206,30 | 9320,00 | 0,34 |
| F-Wert für Cluster 3 | 0,58 | 0,45 | 0,15 | 0,42 | 0,39 |

| | FAKTOR 2 | | | |
|---|---|---|---|---|
| | PS | Hubraum | Geschwindigkeit | Beschleunigung |
| Varianz in der Erhebungsgesamtheit | 443,36 | 102604,73 | 146,63 | 9,30 |
| Varianz in Cluster 1 | 0,00 | 7436,33 | 21,33 | 3,16 |
| F-Wert für Cluster 1 | 0,00 | 0,07 | 0,15 | 0,34 |
| Varianz in Cluster 2 | 552,25 | 55642,25 | 134,25 | 1,83 |
| F-Wert für Cluster 2 | 1,25 | 0,54 | 0,92 | 0,20 |
| Varianz in Cluster 3 | 180,30 | 7888,20 | 63,30 | 1,74 |
| F-Wert für Cluster 3 | 0,41 | 0,08 | 0,43 | 0,19 |

*Tabelle 11:* Variablenspezifische F-Werte für die 3-Cluster-Lösung

❒ **Heterogenitätsprüfung:** Zur Überprüfung der externen Clusterheterogenität können sowohl Diskriminanzanalysen als auch Varianzanalysen durchgeführt werden. So ist z.B. mittels einer Varianzanalyse zu untersuchen, ob sich die Cluster im Hinblick auf ihre Mittelwerte bezüglich der Klassifizierungsvariablen signifikant voneinander unterscheiden. Im vorgegebenen Fall führen zwei ein-

faktorielle Varianzanalysen zum Ergebnis, dass hoch-signifikante Mittelwert-unterschiede zwischen den Clustern vorliegen (vgl. Tabelle 12).

**ANOVA**

| | | Quadrats umme | df | Mittel der Quadrate | F | Signifikanz |
|---|---|---|---|---|---|---|
| Geräumigkeit | Zwischen den Gruppen | 8,987 | 2 | 4,493 | 20,087 | ,000 |
| | Innerhalb der Gruppen | 2,013 | 9 | ,224 | | |
| | Gesamt | 11,000 | 11 | | | |
| Sportlichkeit | Zwischen den Gruppen | 7,062 | 2 | 3,531 | 8,071 | ,010 |
| | Innerhalb der Gruppen | 3,938 | 9 | ,438 | | |
| | Gesamt | 11,000 | 11 | | | |

*Tabelle 12:* Mittelwertunterschiede zwischen den Markenclustern

- **Stabilitätsprüfung:** Die Stabilitätsprüfung beinhaltet den Vergleich der vorliegenden Clusterlösung mit jenen Clustergebnissen, die sich bei Anwendung alternativer Fusionierungsverfahren ergeben. Für unser Beispiel erbringt z.B. das Averge-Linkage-Verfahren auf Basis des quadrierten Euklidischen Distanzmaßes das in Tabelle 13 angeführte Ergebnis. Hiernach sind im Vergleich zur 3-Clusterlösung des Ward-Verfahrens drei Modelle bzw. 25% der Fälle anderen Gruppen zugeordnet worden: Der Fiat 131 gehört nun nicht mehr der Gruppe 2, sondern dem Cluster 1 an. Darüber hinaus werden nunmehr sowohl der „Simca“ als auch der “Volvo 244“, die nach der Ward-Lösung jeweils der Gruppe 2 angehören, der Gruppe 1 zugewiesen. Ferner ist ersichtlich, dass Cluster 2 nach dem Linkage-Verfahren lediglich aus dem „BMW 320“ besteht. In diesem Zusamammenhang ist augenfällig, dass bei Anwendung des Average-Linkage-Verfahrens das Objekt „BMW 320“ als „Ausreißer“ zu charakterisieren ist, denn dieses bildet auch bei einer 4-Cluster- oder einer 2-Clusterlösung stets das einzige Clusterlement. Vor diesem Hintegrund erachten wir die mit dem Ward-Verfahren erzeugte 3-Clusterlösung als zweckmäßig.

**Cluster-Zugehörigkeit**

| Fall | 4 Cluster | 3 Cluster | 2 Cluster |
|---|---|---|---|
| 1:Audi 80 | 1 | 1 | 1 |
| 2:BMW 320 | 2 | 2 | 2 |
| 3:Citroen GSX | 1 | 1 | 1 |
| 4:Fiat 131 | 3 | 1 | 1 |
| 5:Ford Taunus | 1 | 1 | 1 |
| 6:Mercedes 200 | 4 | 3 | 1 |
| 7:Opel Rekord | 4 | 3 | 1 |
| 8:Peugeot 244 | 4 | 3 | 1 |
| 9:Renault 20 | 4 | 3 | 1 |
| 10:Simca | 3 | 1 | 1 |
| 11:VW Passat | 3 | 1 | 1 |
| 12:Volvo 244 | 4 | 3 | 1 |

*Tabelle 13*: Clusterzugehörigkeit der Pkw-Modelle (Basis: Average-Linkage-Verfahren; quadriertes Euklidisches Distanzmaß)

# 4. Partitionierende Clusterzentrenanalyse mit SPSS

## 4.1. Verfahrensbesonderheiten der Clusterzentrenanalyse

Die in den vorangegangen Abschnitten beschriebene hierarchische Clusteranalyse bietet den Vorteil, dass dem Anwender eine flexible Handhabung im Rahmen der Wahl von Proximitätsmaßen sowie von Fusionierungsverfahren gewährt wird. Darüber hinaus kann jeder einzelne Schritt der Clusterbildung tabellarisch und/oder grafisch dargestellt und darauf aufbauend die Entscheidungsfindung bezüglich einer sachgemäßen Clusterzahl unterstützt werden. Diesen Vorzügen steht jedoch der gravierende Nachteil gegenüber, dass hierarchische Ansätze sehr umfangreiche Berechnungen erfordern, da in jedem Fusionierungsschritt eine neue Distanzmatrix ermittelt werden muss und demzufolge bei großen Stichproben der Rechenaufwand überproportional steigt. Insofern empfiehlt es sich in jenen Situationen, in denen große Stichproben vorliegen und/oder zu Beginn der Analyse bereits Vorkenntnisse über die Zahl der zu bildenden Cluster vorhanden sind, die Clusterzentren-Analyse (K-Means-Analyse; Quick Cluster) einzusetzen. Die Clusterzentrenanalyse stellt ein spezielles partitionierendes Minimal-Distanzverfahren dar (vgl. Abb. 3), dessen Einsatz an drei Voraussetzungen gebunden ist (vgl. SPSS 2003, S. 499 ff):

- **Vorgegebene Clusteranzahl:** Bei der Clusterzentrenanalyse wird eine optimale Zuordnung der Fälle in eine vorgegebene Zahl an Clustern gesucht. Wenn nicht bereits aufgrund von Vorinformationen eine bestimmte Anzahl an Clustern vorgegeben werden kann, ist es vielfach zweckmäßig, zunächst eine hierarchische Clusteranalyse durchzuführen. Diese kann sämtliche untersuchungsrelevanten Objekte einbeziehen oder im Fall von großen Objektmengen auf einer Zufallsstichprobe von Objekten beruhen.

- **Bekannte Clusterzentren:** Eine zweite Voraussetzung betrifft den Aspekt, dass der Anfangswert für das Zentrum jedes einzelnen Clusters bekannt sein muss. Denn bei der Clusterzentrenanalyse wird jedes Objekt jenem Cluster zugeordnet, zu dessen Zentrum seine euklidische Distanz am geringsten ist (Minimal-Distanz-Kriterium). Hierdurch entfällt der rechenintensive paarweise Vergleich des hierarchischen Fusionierungsprocederes und führt damit - ebenso wie die Vorgabe der Clusterzahl - zu einer Verringerung des Rechenaufwandes. Die Ermittlung von Clusterzentren lässt sich in zweifacher Weise vornehmen: Zum einen kann man diese im Zuge einer vorgeschalteten hierarchischen Clusteranalyse bestimmen, indem man die Mittelwerte der Klassifizierungsmerkmale für jedes Cluster berechnet, diese anschließend in einer eigenständigen Datendatei mit einer speziellen Matrixstruktur speichert und schließlich als Dateninput der Clusterzentrenanalyse heranzieht. Zum anderen besteht die Möglichkeit; Anfangswerte für die Clusterzentren von SPSS ermitteln zu lassen. In beiden Fällen bilden die Anfangswerte jedoch nur vorläufige Clusterzentren, die im ersten Schritt nur der Aufteilung von Objekten dienen, sich aber im Analyseprozess verändern werden.

- **Standardisierte Klassifizierungsvariablen:** Da die Clusterzentrennalyse als Distanzmaß grundsätzlich die euklidische Distanz verwendet, ist es zumeist zweckmäßig, die relevanten Klassifizierungsvariablen vorab zu standardisieren. Allerdings bietet SPSS hierzu bei der Clusterzentren-Prozedur im Gegensatz zur

hierarchischen Clusteranalyse keine Möglichkeit. Insofern muss die betreffende Rohdatenmatrix vor einer Durchführung der Clusterzentrenanalyse standardisiert werden.

Im Folgenden soll die Durchführung einer Clusterzentrenanalyse anhand zweier Beispiele demonstriert werden, wobei sich das erste (zweite) Fallbeispiel auf eine Situation bezieht, in der (keine) Vorinformationen über die Clustenzentren vorliegen.

### 4.2. K-Means-Analyse bei bekannten Clusterzentren

Das im vorangegangenen Abschnitt erläuterte Fallbeispiel zur hierarchischen Clusteranalyse hat ergeben, dass sich die betrachteten zwölf Pkw-Modelle auf Basis von Faktorwerten in drei Cluster aufteilen lassen. Mit Hilfe einer Clusterzentrenanalyse soll nun überprüft werden, ob sich die Zuordnung der Pkw-Modelle zu den vorgegebenen Clustern verbessern lässt. Die zur Durchführung der Clusterzentrennalyse benötigten Anfangswerte der drei Clusterzentren wollen wir im allerdings nicht von SPSS ermitteln lassen. Vielmehr sollen diese auf der Grundlage einer hierarchischen Clusteranalyse berechnet werden und als Dateninput der anschließenden Clusterzentrenanalyse dienen. Im Fall von bekannten Clusterzentren vollzieht sich die Clusterzentrenanalyse in vier Phasen:

- ❒ Ermittlung von Clusterzentren,
- ❒ Erstellung der Clusterzentren-Datenmatrix,
- ❒ Festlegung der SPSS-Auswertungsmethodik,
- ❒ Interpretation der Ergebnisse

**I. Ermittlung von Clusterzentren:** Bei der Berechnung der anfänglichen Clusterzentren sind die clusterspezifischen Mittelwerte hinsichtlich der beiden Klassifizierungsvariablen „Geräumigkeit“ und „Sportlichkeit“ zu bestimmen. Hierzu öffnen wir die Datei „Leistungspositionierung“ und führen mittels der SPSS-Befehlsfolge „Analysieren/Mittelwerte vergleichen/Mittelwerte...“ einen Mittelwertvergleich der drei Cluster (unabhängige Variable: „Ward-Method Clu_3“) durch (vgl. Abb. 15).

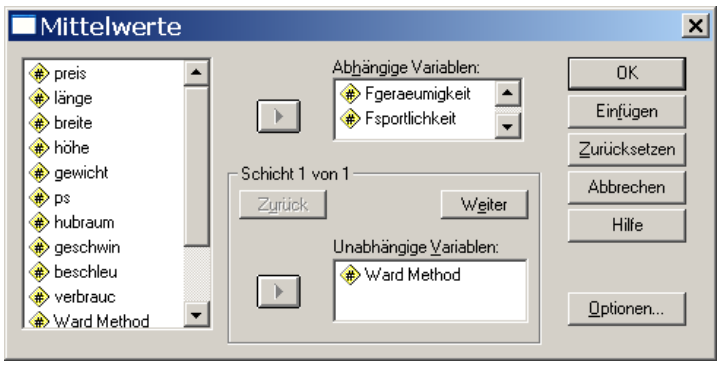



*Abbildung 15*: Dialogbox „Mittelwerte vergleichen/Mittelwerte...“

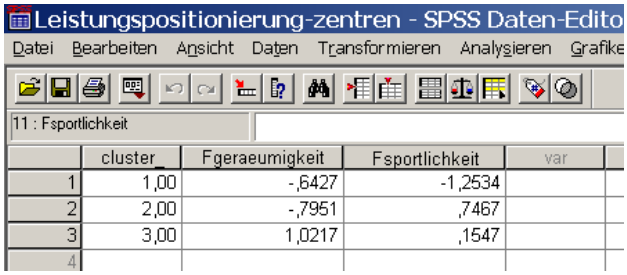
Die daraus resultierenden Mittelwerte bzw. durchschnittlichen Faktorwerte werden in der nachstehenden Tabelle 14 ausgewiesen.

**Bericht**

| Ward Method | | Fgeraeumigkeit | Fsportlichkeit |
|---|---|---|---|
| 1 | Mittelwert | **-,6427** | **-1,2534** |
| | N | 3,0000 | 3,0000 |
| 2 | Mittelwert | **-,7951** | **,7467** |
| | N | 4,0000 | 4,0000 |
| 3 | Mittelwert | **1,0217** | **,1547** |
| | N | 5,0000 | 5,0000 |
| Insgesamt | Mittelwert | ,0000 | ,0000 |
| | N | 12,0000 | 12,0000 |

*Tabelle 14:* Durchschnittliche Faktorwerte der 3-Clusterlösung

**II. Erstellung der Clusterzentren-Datenmatrix:** Damit die SPSS-Prozedur Clusterzentrenanalyse auf die durchschnittlichen Faktorwerte zurückgreifen kann, sind diese als anfängliche Clusterzentren in einer eigenständigen Datendatei einzugeben bzw. zu speichern. Diese erfordert einen ganz bestimmten Aufbau, der für das vorliegende Beispiel in Tabelle 15 dargelegt ist. Diese Datei, die wir unter dem Namen „Leistungspositionierung-zentren" speichern, muss als erste Variable eine Clustervariable enthalten, deren Werte in unserem Beispiel den Bereich 1 bis 3 umfassen. Darüber hinaus muss die Variable unter der Bezeichnung „cluster_" definieren werden. In den darauffolgenden Spalten ist für jede Klassifizierungsvariable die Clustermittelwerte einzutragen.




| | cluster_ | Fgeraeumigkeit | Fsportlichkeit | var |
|---|---|---|---|---|
| 1 | 1,00 | -,6427 | -1,2534 | |
| 2 | 2,00 | -,7951 | ,7467 | |
| 3 | 3,00 | 1,0217 | ,1547 | |
| 4 | | | | |

*Tabelle 15:* Datei „Leistungspositionierung-zentren"

**III. Festlegung der SPSS-Auswertungsmethodik:** Zur Durchführung der Clusterzentrenanalyse gehen wir in den nachfolgenden Schritten vor:

(1) Zunächst öffnen wir die im Rahmen der hierarchischen Clusteranalyse untersuchte **Datei „Leistungspositionierung"**, in der die relevanten Fälle bzw. Pkw-Modelle, faktoriellen Klassifizierungsvariablen sowie die hierarchischen Clusterlösungen enthalten sind (vgl. Abb. 16).

(2) Führen Sie dort die Befehlsfolge „Analysieren/Klassifizieren/Clusterzentrenanalyse.." aus.

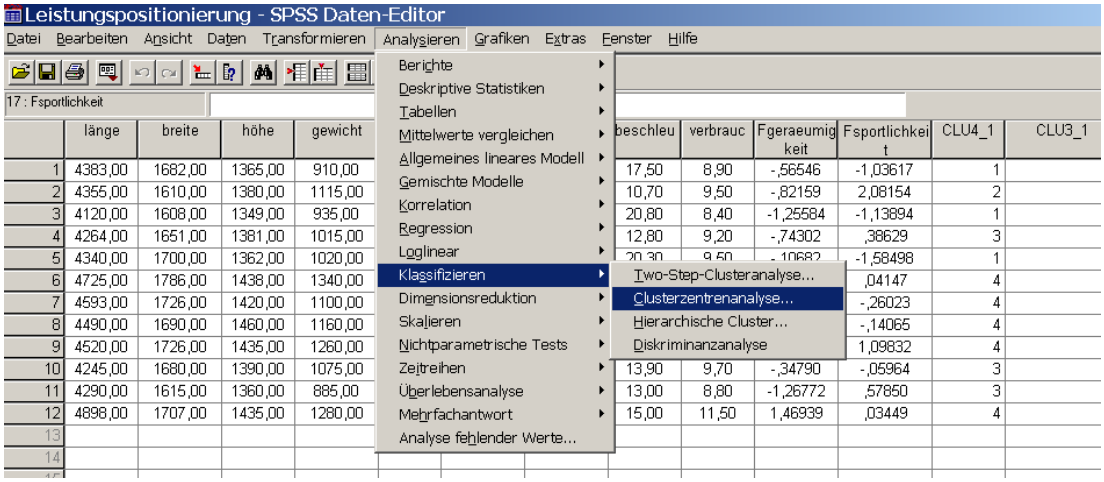



*Abbildung 16*: Datei „Leistungspositionierung“

(3) Hierauf öffnet sich die Dialogbox *„Clusterzentrenanalyse“* (vgl. Abb. 17).

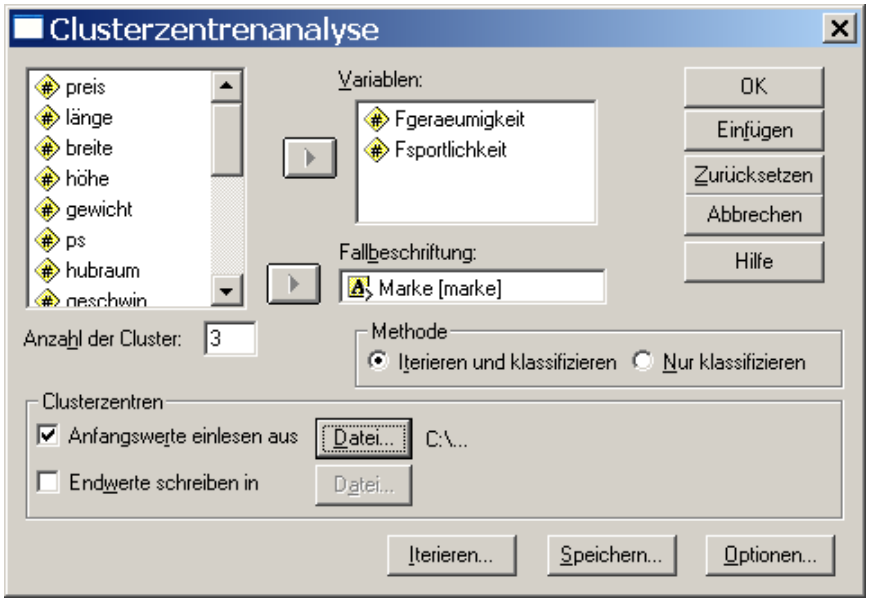



*Abbildung 17: Dialogfeld „Clusterzentrenanalyse“*

- Überführen Sie dort die (standardisierten) Klassifizierungsvariablen „Fgeraeumigkeit“ und „Fsportlichkeit“ in das Feld „Variablen“.
- Zur Fallbeschriftung verwenden wir die Variable „Marke“.
- Geben Sie in das Feld „Anzahl der Cluster“ die Zahl 3 ein.
- Im Feld „Methode“ behalten wir die Voreinstellung „Iterieren und Klassifizieren“ bei, da generell nicht davon ausgegangen werden kann, dass eine optimale Objektzuordnung bereits nsch dem ersten Iterationsschritt gewonnen wird.
- Um SPSS mitzuteilen, dass auf bekannte Clusterzentren zurückgegriffen werden kann, klicken wir im Feld „Clusterzentren“ auf das Kästchen „Anfangswerte einlesen aus“ und dann auf die Schaltfläche „Datei“. Es

öffnet sich die Dialogbox „ *Clusterzentrenanalyse: Aus Datei einlesen"*, in der wir im Feld „Datei" jene Datei öffnen, in der die anfänglichen Clusterzentren enthalten sind bzw. im vorliegenden Beispiel die Datei „Leistungspositionierung-zentren".

⌦ Wir kehren zur Dialogbox „Clusterzentrenanalyse" zurück und können dort die Option „Endwerte schreiben in" aktivieren, mit der SPSS dazu veranlasst wird, die aus der Analyse resultierenden (finalen) Clusterzentren zu speichern. Wir belassen die deaktivierte Voreinstellung und klicken auf die Schaltfläche „Iterieren"

(4) Die Dialogbox „*Clusterzentrennalyse: Iterieren*" wird geöffnet (vgl. Abb. 18). Standardmäßig wird, ausgehend von den anfänglichen Clusterzentren, jeder Fall jeweils jenem Cluster zugeordnet, zu dessen Zentrum er die geringste euklidische Distanz besitzt. Nachdem die Fälle zugeordnet sind, werden die Clusterzentren neu berechnet und die Zuordnung wird wiederholt. Der Prozess wird iterativ solange fortgeführt, bis die eingestellte maximale Zahl an Iterationen, die zwischen 1 und 999 Schritte umfassen kann, erreicht oder das Konvergenzkriteriuum erfüllt ist. Das Konvergenzkriterium bestimmt den Abbruch des Iterationsprozesses und gibt den Anteil der minimalen Distanz zwischen anfänglichen Clusterzentren an. Der Konvergenzwert muss daher größer als 0, darf aber nicht größer als 1 sein. Wenn wir einen Konvergenzwert von z.B. 0,05 festlegen, dann ist der Zuordnungsprozeß dann beendet, wenn eine vollständige Iteration keines der Cluster-Zentren um mehr als fünf Prozent der kleinsten Distanz zwischen zwei Cluster-Zentren der Ausgangslösung verschiebt. Sofern die Option „Gleitende Mittelwerte verwenden" gewählt wird, so wird das Clusterzentrum nach jedem Fall akutalisiert, ansonsten erst nachdem alle Fälle hinzugefügt wurden.

⌦ Wir belassen die Voreinstellungen und bestätigen mit „Weiter".

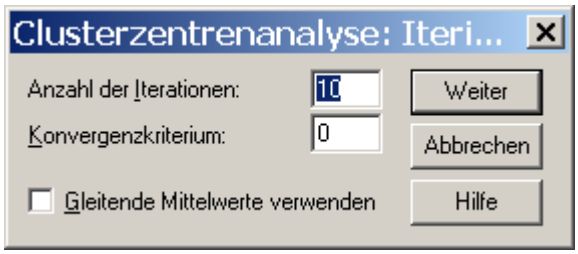



*Abbildung 18:* Dialogbox "Clusterzentrenanalyse: Iterieren'"

(5) Anschließend klicken wir in der Dialogbox „Clusterzentren" auf die Schaltfläche „Optionen", worauf sich das Dialogfenster „*Clusterzentrenanalyse: Optionen"* öffnet (vgl. Abb. 19). Hier wählen wir im Feld „Statistik" die Optionen

⌦ „Anfängliche Clusterzentren",

⌦ „ANOVA-Tabelle",

⌦ „Clusterinformationen für jeden Fall"

⌦ und bestätigen unsere Einstellungen mit „Weiter".

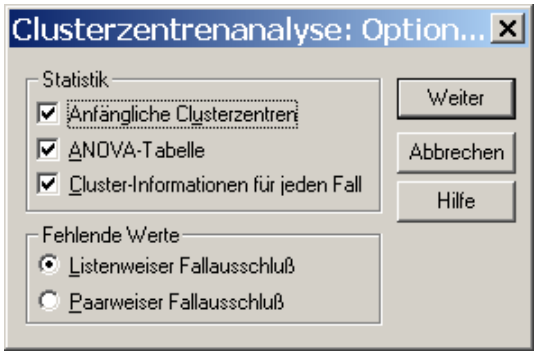



*Abbildung 19:* Dialogfenster „Clusterzentrenanalyse: Optionen“

(6) Im abschließenden Schritt klicken wir in der Dialogbox „Clusterzentren“ auf die Schaltfläche „Speichern“. Wir aktivieren im daraufhin geöffneten Dialogfenster „*Clusterzentrenanalyse: Neue Variablen...*“ die Option „Cluster-Zugehörigkeit“, mit der SPSS der Datendatei eine Variable mit der Bezeichnung „qcl_1“ hinzufügt. Wir beschließen unsere Einstellungen mit „Weiter“ und „OK“.

**IV. Interpretation der Ergebnisse:** Die tabellarische SPSS-Ergebnisausgabe umfasst

- anfängliche Clusterzentren,
- das Iterationsprotokoll,
- finale Clusterzentren,
- Distanzen zwischen den finalen Clusterzentren,
- den varainzanalytischen Mittelwertvergleich der Cluster,
- die Clusterzugehörigkeiten der Fälle.

In der SPSS-Ausgabe wird zunächst eine Tabelle mit den **anfänglichen Clusterzentren** ausgewiesen (vgl. Tabelle 16). Die darin enthaltenen Werte entsprechen den von uns zuvor ermittelten und als Dateninput vorgegebenen Gruppenzentroiden.

**Anfängliche Clusterzentren**

| | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Fgeraeumigkeit | -,64270 | -,79510 | 1,02170 |
| Fsportlichkeit | -1,25340 | ,74670 | ,15470 |

Aus Unterbefehl FILE eingeben

*Tabelle 16:* Anfängliche Clusterzentren

Aus dem in Tabelle 17 angeführten **Iterationsprotokoll** ist ersichtlich, dass der Zuordnungsprozess nach der zweiten Iteration abgebrochen wurde. Demzufolge sind von den maximal 10 Iterationen lediglich zwei Prozeßstufen benötigt worden, um eine optimale Zuordnungslösung bzw. die angestrebte Minimaldistanz-Partition zu erzeugen. Für jeden Schritt (Iteration) werden die quantitativen Verschiebungen bzw. Änderungen der Clusterzentren – bzw. Mittelwerte angezeigt. Diese sind allerdings

schon nach der ersten Iteration überaus gering und erfüllen bereits nach der zweiten Iteration das vorgegebene Konvergenzkriterium von 0.

**Iterationsprotokoll[a]**

| Iteration | Änderung in Clusterzentren | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | ,00004 | ,00005 | ,00004 |
| 2 | ,000 | ,000 | ,000 |

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist ,000. Die aktuelle Iteration lautet 2. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 1,911.

Tabelle 17: Iterationsprotokoll

Die aus dem Iterationsprozeß resultierenden Zentrenänderungen finden in den **finalen Clusterzentren** ihren Niederschlag. Für das vorliegende Beispiel lässt die Tabelle 18 erkennen, dass sich die finalen Clusterzentren nur marginal von den anfänglichen Zentrenwerten unterscheiden.

**Clusterzentren der endgültigen Lösung**

| | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Fgeraeumigkeit | -,64271 | -,79506 | 1,02167 |
| Fsportlichkeit | -1,25336 | ,74667 | ,15468 |

*Tabelle 18*: Finale Clusterzentren

Mit Hilfe der in Tabelle 19 enthaltenen **Distanzen zwischen den Clusterzentren der finalen Clusterlösung** ist eine erste Beurteilung der Gruppeneinteilung möglich. Sofern das mit einer Clusteranalyse verfolgte Ziel einer extern möglichst heterogenen Clusterstuktur erreicht wurde, müssen die Distanzunterschiede zwischen den finalen Clusterzentren möglichst groß sein. Nach Tabelle 19 ist die euklidische Distanz zwischen den Zentren der Gruppen 1 und 3 (Distanzwert: 2,180) jeweils größer als die Distanz zwischen den Gruppen 1 und 2 ( Distanzwert: 2,006) sowie den Gruppen 2 und 3 (Distanzwert: 1,911).

**Distanz zwischen Clusterzentren der endgültigen Lösung**

| Cluster | 1 | 2 | 3 |
|---|---|---|---|
| 1 | | 2,006 | 2,180 |
| 2 | 2,006 | | 1,911 |
| 3 | 2,180 | 1,911 | |

*Tabelle 19:* Euklidische Distanzen zwischen finalen Clusterzentren

Einen Aufschluss darüber, ob die betreffenden Gruppenunterschiede signifikant sind, vermittelt die in Tabelle 20 angeführte **Varianztabelle**. Der betreffende F-Test geht von der Grundüberlegung aus, dass die Werte der Klassifizierungs-variablen bei intern

homogenen Clustern nur geringfügig vom jeweiligen Clustermittelwert abweichen und bei extern heterogenen Clustern recht stark um den Mittelwert der betreffenden Variablen für die Gesamtheit der Fälle streuen. Daher weist die Varianztabelle für jede der beiden Klassifizierungsvariablen die Quadratsumme der Cluster (Spalte „Fehler: Mittel der Quadrate“) sowie die Quadratsumme zwischen den Clustern (Spalte „Cluster: Mittel der Quadrate“ ) aus. Der als Testgröße dienende F-Wert (Spalte „F“) ergibt sich als Quotient aus der Quadratsumme zwischen den Clustern und der Quadratsumme innerhalb der Cluster. Ein hoher F-Wert und ein geringer Signifikanzwert (Spalte „Sig.“) zeigen somit an, dass die Werte der betreffenden Variablen innerhalb der Gruppen homogener sind als die Merkmalswerte in unterschiedlichen Clustern. Insofern verweisen die Befunde der beiden variablenpezifischen einfaktoriellen Varianzanalysen darauf, dass signifikante Mittelwertunterschiede zwischen den Gruppen vorliegen.

**ANOVA**

| | Cluster | | Fehler | | | |
|---|---|---|---|---|---|---|
| | Mittel der Quadrate | df | Mittel der Quadrate | df | F | Sig. |
| Fgeraeumigkeit | 4,493 | 2 | ,224 | 9 | 20,087 | ,000 |
| Fsportlichkeit | 3,531 | 2 | ,438 | 9 | 8,071 | ,010 |

Die F-Tests sollten nur für beschreibende Zwecke verwendet werden, da die Cluster so gewählt wurden, daß die Differenzen zwischen Fällen in unterschiedlichen Clustern maximiert werden. Dabei werden die beobachteten Signifikanzniveaus nicht korrigiert und können daher nicht als Tests für die Hypothese der Gleichheit der Clustermittelwerte interpretiert werden.

*Tabelle 20:* Varianztabelle

Zur Beurteilung der Güte der Clusterlösung ist der F-Test jedoch in nur geringem Maße geeignet; denn die Clusterbildung erfolgte gezielt unter der Bedingung, dass sich möglichst große Distanzen zwischen den Gruppen ergeben. Im konkreten Anwendungsfall erscheint es daher geboten, die Güteprüfung um eine diskriminanz-analytische Untersuchung zu ergänzen.

Die endgültige **Gruppenzugehörigkeit der Objekte** sowie die gruppenspezifische Besetzungzahl werden in Tabelle 21 sowie Tabelle 22 (diese enthält zusätzlich die Distanzen der Objekte zu den relevanten Clusterzentren) angezeigt. Aus dem Vergleich mit den Clusterergebnissen der hierarchischen Analyse (vgl. z.B. Tabelle 8) wird ersichtlich, dass sich im Zuge der Clusterzentrenanalyse keine Objekt-umgruppierungen ergeben haben. Aus diesem Grund darf die 3-Clusterlösung als recht stabil erachtet werden.

**Anzahl der Fälle in jedem Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 3,000 |
| | 2 | 4,000 |
| | 3 | 5,000 |
| Gültig | | 12,000 |

*Tabelle 21.* Anzahl der Objekte in den Clustern

**Cluster-Zugehörigkeit**

| Fallnummer | Marke | Cluster | Distanz |
|---|---|---|---|
| 1 | Audi 80 | 1 | ,231 |
| 2 | BMW 320 | 2 | 1,335 |
| 3 | Citroen GSX | 1 | ,624 |
| 4 | Fiat 131 | 2 | ,364 |
| 5 | Ford Taunus | 1 | ,630 |
| 6 | Mercedes 200 | 3 | ,603 |
| 7 | Opel Rekord | 3 | ,544 |
| 8 | Peugeot 244 | 3 | ,413 |
| 9 | Renault 20 | 3 | 1,025 |
| 10 | Simca | 2 | ,922 |
| 11 | VW Passat | 2 | ,502 |
| 12 | Volvo 244 | 3 | ,464 |

*Tabelle 22:* Finale Gruppenzugehörigkeit der Fälle

### 4.3. K-Means-Analyse bei unbekannten Clusterzentren

Im Folgenden soll ein Beispiel beahndelt werden, bei welchem die Clusterzentren nicht vorgegeben, sondern von SPSS ermittelt werden. Dabei sei davon ausgegangen, dass bei einer Stichprobe von 30 Nachfragern eine Reihe ausgewählter Personenmerkmale erhoben wurde, deren Werte in der Tabelle 23 angeführt sind.

| Person | ALTER | GESCHLECHT (1= m; 2 = w) | EINKOMMEN (TSD. €/Jahr) | FAMSTAND (1 = Single; 2 = verh.) | MAWAHL (1= A; 2 = B) | KAUFRATE (ME/Monat) | PREISAKZ (max. € /ME) |
|---|---|---|---|---|---|---|---|
| 1 | 37 | 2 | 28 | 1 | 1 | 3 | 16 |
| 2 | 25 | 2 | 23 | 1 | 1 | 4 | 15 |
| 3 | 20 | 1 | 26 | 2 | 2 | 1 | 14 |
| 4 | 40 | 2 | 25 | 1 | 1 | 1 | 15 |
| 5 | 27 | 2 | 41 | 1 | 2 | 8 | 18 |
| 6 | 30 | 2 | 37 | 1 | 1 | 3 | 15 |
| 7 | 45 | 2 | 33 | 1 | 1 | 5 | 17 |
| 8 | 34 | 1 | 34 | 1 | 2 | 4 | 16 |
| 9 | 38 | 2 | 32 | 1 | 1 | 4 | 16 |
| 10 | 35 | 2 | 23 | 1 | 1 | 1 | 14 |
| 11 | 35 | 2 | 27 | 2 | 2 | 2 | 15 |
| 12 | 43 | 1 | 42 | 2 | 2 | 7 | 18 |
| 13 | 32 | 1 | 30 | 1 | 1 | 5 | 17 |
| 14 | 42 | 2 | 30 | 1 | 1 | 2 | 14 |
| 15 | 22 | 2 | 30 | 2 | 2 | 6 | 16 |
| 16 | 32 | 2 | 39 | 2 | 2 | 6 | 17 |
| 17 | 27 | 2 | 33 | 2 | 1 | 5 | 16 |
| 18 | 40 | 1 | 35 | 1 | 1 | 7 | 17 |
| 19 | 30 | 2 | 35 | 1 | 2 | 7 | 18 |
| 20 | 35 | 1 | 44 | 1 | 2 | 8 | 19 |
| 21 | 22 | 2 | 34 | 2 | 2 | 7 | 18 |
| 22 | 41 | 1 | 37 | 1 | 2 | 5 | 16 |
| 23 | 31 | 2 | 23 | 2 | 1 | 1 | 14 |
| 24 | 27 | 1 | 26 | 1 | 2 | 3 | 15 |
| 25 | 25 | 2 | 29 | 2 | 2 | 2 | 14 |
| 26 | 45 | 1 | 39 | 1 | 1 | 5 | 16 |
| 27 | 34 | 2 | 36 | 1 | 1 | 5 | 16 |
| 28 | 37 | 1 | 33 | 1 | 2 | 6 | 17 |
| 29 | 38 | 1 | 40 | 1 | 1 | 8 | 19 |
| 30 | 20 | 2 | 22 | 2 | 1 | 2 | 14 |

*Tabelle 23:* (Fiktive) Stichprobe von 30 Nachfragern

Mit Hilfe einer Clusterzentranalyse soll nun untersucht werden, ob sich die 30 Nachfrager hinsichtlich ausgewählter Personenmerkmale in Teilgruppen bzw. Marktsegmente aufteilen lassen. Als relevante Segmentierungskriterien sollen dabei die die metrischen Variablen „Alter“, „Einkommen“ und „Preisbereitschaft“ dienen.

Im Fall von unbekannten Clusterzentren vollzieht sich die Clusterzentrenanalyse in drei Phasen:

- ❒ Erstellung der Datenmatrix,
- ❒ Festlegung der SPSS-Auswertungsmethodik,
- ❒ Interpretation der Ergebnisse

**I. Erstellung der Datenmatrix:** Die Erstellung der Datenmatrix umfasst im vorliegenden Beispiel zwei Schritte. Zunächst sind die Daten der Tabelle 23 in den Dateneditor von SPSS einzugeben. Hieran anschließend sind die drei Segmentierungsvariablen zu **standardisieren**, da diese in unterschiedlichen Dimensionen (z.B. Alter in Jahren, Einkommen in Tsd. €) gemessen wurden. Hierzu gehen wir in der folgenden Weise vor:

(1) Wir führen die Befehlsfolge Wählen Sie „Analysieren/Deskriptive Statistiken/ Deskriptive Statistiken...“ durch, worauf sich die Dialogbox *„Deskriptive Statistiken“* öffnet (vgl. Abb. 20).

(2) Dort überführen wir die drei Segmentierungsvariablen in das Feld „Variable(n)“ und klicken anschließend auf die Option „Standardisierte Werte als Variable speichern“.

(3) Abschließend bestätigen wir unsere Einstellungen mit „OK“.

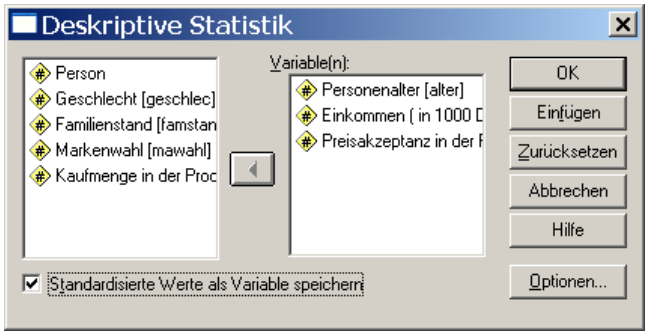



*Abbildung 20*: Dialogbox „Deskriptive Statistiken“

(4) SPSS fügt nun der Datendatei drei standardisierte Variablen „Zalter“, „Zeinkomme“ und „zpreisakz“ an (vgl. Abb. 21).

(5) Wir speichern die Datendatei unter der Bezeichnung „konsum-standardisiert“.

| | Person | alter | geschlec | einkomme | famstand | mawahl | kaufrate | preisakz | Zalter | Zeinkomme | Zpreisakz |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 37 | 2 | 28 | 1 | 1 | 3 | 16 | ,54870 | -,67985 | -,04358 |
| 2 | 2 | 25 | 2 | 23 | 1 | 1 | 4 | 15 | -1,0838 | -1,48920 | -,69727 |
| 3 | 3 | 20 | 1 | 26 | 2 | 2 | 1 | 14 | -1,7640 | -1,00359 | -1,35096 |
| 4 | 4 | 40 | 2 | 25 | 1 | 1 | 1 | 15 | ,95682 | -1,16546 | -,69727 |
| 5 | 5 | 27 | 2 | 41 | 1 | 2 | 8 | 18 | -,81171 | 1,42445 | 1,26380 |
| 6 | 6 | 30 | 2 | 37 | 1 | 1 | 3 | 15 | -,40359 | ,77697 | -,69727 |
| 7 | 7 | 45 | 2 | 33 | 1 | 1 | 5 | 17 | 1,63702 | ,12950 | ,61011 |
| 8 | 8 | 34 | 1 | 34 | 1 | 2 | 4 | 16 | ,14058 | ,29136 | -,04358 |
| 9 | 9 | 38 | 2 | 32 | 1 | 1 | 4 | 16 | ,68474 | -,03237 | -,04358 |
| 10 | 10 | 35 | 2 | 23 | 1 | 1 | 1 | 14 | ,27662 | -1,48920 | -1,35096 |
| 11 | 11 | 35 | 2 | 27 | 2 | 2 | 2 | 15 | ,27662 | -,84172 | -,69727 |
| 12 | 12 | 43 | 1 | 42 | 2 | 2 | 7 | 18 | 1,36494 | 1,58632 | 1,26380 |
| 13 | 13 | 32 | 1 | 30 | 1 | 1 | 5 | 17 | -,13151 | -,35611 | ,61011 |
| 14 | 14 | 42 | 2 | 30 | 1 | 1 | 2 | 14 | 1,22890 | -,35611 | -1,35096 |
| 15 | 15 | 22 | 2 | 30 | 2 | 2 | 6 | 16 | -1,4919 | -,35611 | -,04358 |
| 16 | 16 | 32 | 2 | 39 | 2 | 2 | 6 | 17 | -,13151 | 1,10071 | ,61011 |
| 17 | 17 | 27 | 2 | 33 | 2 | 1 | 5 | 16 | -,81171 | ,12950 | -,04358 |
| 18 | 18 | 40 | 1 | 35 | 1 | 1 | 7 | 17 | ,95682 | ,45323 | ,61011 |
| 19 | 19 | 30 | 2 | 35 | 1 | 2 | 7 | 18 | -,40359 | ,45323 | 1,26380 |
| 20 | 20 | 35 | 1 | 44 | 1 | 2 | 8 | 19 | ,27662 | 1,91006 | 1,91749 |
| 21 | 21 | 22 | 2 | 34 | 2 | 2 | 7 | 18 | -1,4919 | ,29136 | 1,26380 |
| 22 | 22 | 41 | 1 | 37 | 1 | 2 | 5 | 16 | 1,09286 | ,77697 | -,04358 |
| 23 | 23 | 31 | 2 | 23 | 2 | 1 | 1 | 14 | -,26755 | -1,48920 | -1,35096 |
| 24 | 24 | 27 | 1 | 26 | 1 | 2 | 3 | 15 | -,81171 | -1,00359 | -,69727 |
| 25 | 25 | 25 | 2 | 29 | 2 | 2 | 2 | 14 | -1,0838 | -,51798 | -1,35096 |
| 26 | 26 | 45 | 1 | 39 | 1 | 1 | 5 | 16 | 1,63702 | 1,10071 | -,04358 |
| 27 | 27 | 34 | 2 | 36 | 1 | 1 | 5 | 16 | ,14058 | ,61510 | -,04358 |
| 28 | 28 | 37 | 1 | 33 | 1 | 2 | 6 | 17 | ,54870 | ,12950 | ,61011 |
| 29 | 29 | 38 | 1 | 40 | 1 | 1 | 8 | 19 | ,68474 | 1,26258 | 1,91749 |
| 30 | 30 | 20 | 2 | 22 | 2 | 1 | 2 | 14 | -1,7640 | -1,65107 | -1,35096 |

*Abbildung 21:* SPSS- Datei „konsum-standardisiert“

**II. Festlegung der SPSS-Auswertungsmethodik:** Zur Durchführung der Clusterzentrenanalyse gehen wir in den folgenden Schritten vor:

(1) Wir öffnen die Dialogbox „Clusterzentrenanalyse“ (vgl. Abb.22) und wählen dort die nachstehenden Einstellungen:

- ⌦ Die drei standardisierten Segmentierungsvariablen werden in das Feld „Variablen“ und die Variable „Person“ als Fallbeschriftung übertragen.
- ⌦ Aufgrund von sachlogischen Überlegungen gehen wir davon aus, dass sich die Stichprobe von 30 Nachfragern in drei Cluster aufteilen lässt. Daher tragen wir unter „Anzahl der Cluster“ die Zahl 3 ein.
- ⌦ Im Feld „Methode“ belassen wir die Voreinstellung „Iterieren und klassifizieren“.
- ⌦ Im Gegensatz zur Analyse bei bekannten Clusterzentren nehmen wir im Feld „Clusterzentren“ keine Eintragungen vor, da wir im vorliegenden Beispiel über keine Vorinformationen hinsichtlich der Clusterzentren verfügen.

⌦ Demgegenüber nehmen wir hinsichtlich der Schaltflächen „Iterieren“, „Speichern“ und „Optionen“ dieselben Einstellungen wie im vorangegangen Beispiel der Clusterzentrenanalyse bei bekannten Zentren vor.

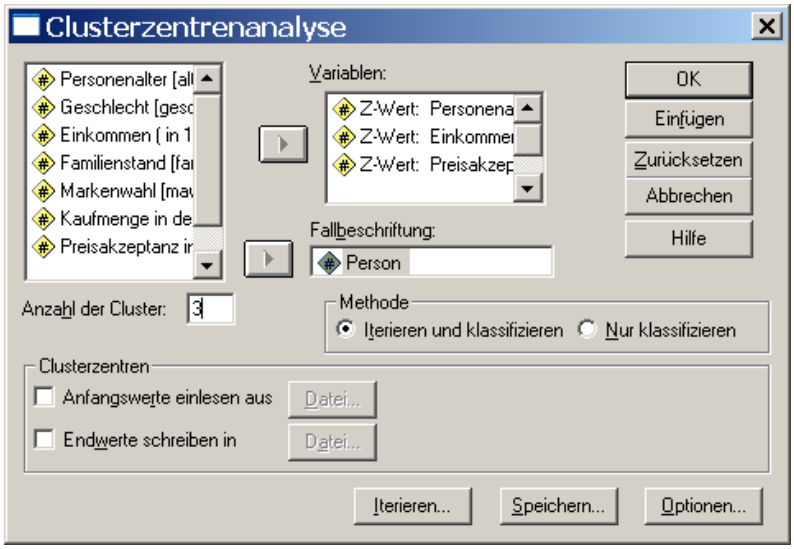



*Abbildung 22*: Dialogbox „Clusterzentrenanalyse“

**III. Interpretation der Ergebnisse:** Die tabellarische SPSS-Ergebnisausgabe umfasst

- ❑ anfängliche Clusterzentren,
- ❑ das Iterationsprotokoll,
- ❑ finale Clusterzentren,
- ❑ Distanzen zwischen den finalen Clusterzentren,
- ❑ den varainzanalytischen Mittelwertvergleich der Cluster,
- ❑ die Clusterzugehörigkeiten der Fälle.

In der SPSS-Ausgabe wird zunächst eine Tabelle mit den **anfänglichen Clusterzentren** ausgewiesen (vgl.. Tabelle 24). Zur Bestimmung der Startwerte für die Clusterzentren geht SPSS folgendermassen vor: Wenn insgesamt k Cluster (hier: 3) gebildet werden sollen, dann werden die ersten k Fälle der Datendatei als vorläufige Clusterzentren verwendet. Im darauffolgenden Schritt werden die übrigen Fälle daraufhin untersucht, ob sie möglicherweise bessere Clusterzentren darstellen als die ersten k Fälle. Hierbei wird ein als provisorisches Clusterzentrum ausgewählter Fall dann durch einen anderen Fall ersetzt, wenn seine kleinste euklidische Distanz zu einem Clusterzentrum größer ist als die Distanz zwischen den beiden nächsten Gruppen (vgl. ausführlich Bortz 1993, S. 535 ff.). Jenes Clusterzentrum, welches dem betreffenden Fall näher ist, wird dann ersetzt.

**Anfängliche Clusterzentren**

| | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Z-Wert: Personenalter | 1,22890 | ,27662 | -1,76400 |
| Z-Wert: Einkommen ( in 1000 DM/Jahr) | -,35611 | 1,91006 | -1,65107 |
| Z-Wert: Preisakzeptanz in der Produktart (max. Preisbereitschaft in €/Stück) | -1,35096 | 1,91749 | -1,35096 |

*Tabelle 24:* Anfängliche Clusterzentren

Gemäß des **Iterationsprotokolls** der Tabelle 25 wurde das Konvergenzkriterium nach drei Iterationsschritten erreicht. So hat sich beispielsweise das Zentrum von Cluster 1 nach dem ersten Schritt durch eine Neuzuteilung bzw. Umgruppierung von Nachfragern um 1,307 und im zweiten Schritt um 0,203 verschoben. Das in der zweiten Iteration gewonnene Ergebnis lässt sich im Hinblick auf das Minimal-Distanzkriterium nicht mehr verbessern, so dass der Prozeß abgeschlossen wird.

**Iterationsprotokoll[a]**

| | Änderung in Clusterzentren | | |
|---|---|---|---|
| Iteration | 1 | 2 | 3 |
| 1 | 1,307 | 1,095 | 1,080 |
| 2 | ,203 | ,181 | ,178 |
| 3 | ,000 | ,000 | ,000 |

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist ,000. Die aktuelle Iteration lautet 3. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 3,261.

*Tabelle 25*: Iterationsprotokoll

Die Tabelle der **endgültigen Clusterzentren** zeigt die Zentrenwerte nach Abschluss des Iterationsprozesses für die vorgegebene 3-Clusterlösung an (vgl. Tabelle 26). Ein Vergleich mit den anfänglichen Clusterzentren macht deutlich, dass sich die variablenspezifischen Zentren bei allen drei Clustern beträchtlich verändert haben.

**Clusterzentren der endgültigen Lösung**

| | Cluster | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Z-Wert: Personenalter | ,66530 | -,07320 | -,97798 |
| Z-Wert: Einkommen ( in 1000 DM/Jahr) | ,06012 | 1,14696 | -,98560 |
| Z-Wert: Preisakzeptanz in der Produktart (max. Preisbereitschaft in €/Stück) | -,09027 | 1,35718 | -,91516 |

*Tabelle 26:* Finale Clusterzentren

Die in Tabelle 27 ausgewiesenen euklidischen **Distanzen zwischen den Clusterzentren der finalen Clusterlösung** lassen erkennen**,** dass die Gruppen 2 und 3 (Distanzwert: 3,245) die größte Unähnlichkeit besitzen, während sich die Gruppen 1 und 2 ( Distanzwert: 1,955) am ähnlichsten sind.

**Distanz zwischen Clusterzentren der endgültigen Lösung**

| Cluster | 1 | 2 | 3 |
|---|---|---|---|
| 1 | | 1,955 | 2,115 |
| 2 | 1,955 | | 3,245 |
| 3 | 2,115 | 3,245 | |

*Tabelle 27*: Distanzen der finalen Clusterzentren

Darüber hinaus kann gemäß der in Tabelle 28 angeführten **Varianztabelle** davon ausgegangen werden, dass sich die Gruppenmittelwerte bezüglich aller drei Segmentierungsvariablen signifikant voneinander unterscheiden.

**ANOVA**

| | Cluster | | Fehler | | | |
|---|---|---|---|---|---|---|
| | Mittel der Quadrate | df | Mittel der Quadrate | df | F | Sig. |
| Z-Wert: Personenalter | 7,421 | 2 | ,524 | 27 | 14,153 | ,000 |
| Z-Wert: Einkommen ( in 1000 DM/Jahr) | 9,001 | 2 | ,407 | 27 | 22,097 | ,000 |
| Z-Wert: Preisakzeptanz in der Produktart (max. Preisbereitschaft in €/Stück) | 10,273 | 2 | ,313 | 27 | 32,806 | ,000 |

Die F-Tests sollten nur für beschreibende Zwecke verwendet werden, da die Cluster so gewählt wurden, daß die Differenzen zwischen Fällen in unterschiedlichen Clustern maximiert werden. Dabei werden die beobachteten Signifikanzniveaus nicht korrigiert und können daher nicht als Tests für die Hypothese der Gleichheit der Clustermittelwerte interpretiert werden.

*Tabelle 28*: Varianztabelle

Zur abschließenden **Beschreibung der Clusterprofile** stehen im Rahmen der Clusterzentrenanalyse drei Informationsbereiche zur Verfügung:

- ❒ clusterspezifsche Fallzahlen,
- ❒ die Clusterzugehörigkeit von Nachfragern sowie
- ❒ eine inhaltliche Kennzeichnung mittels finaler Clusterzentren.

Tabelle 29 vermittelt einen Eindruck über die **clusterspezifischen Fallzahlen**. So gehören 14 von 30 Nachfragern bzw. 46,7 % der Stichprobe dem Cluster 1 an, das somit als das besetzungsstärkste Segment zu beurteilen ist.

**Anzahl der Fälle in jedem Cluster**

| | | |
|---|---|---|
| Cluster | 1 | 14,000 |
| | 2 | 7,000 |
| | 3 | 9,000 |
| Gültig | | 30,000 |

*Tabelle 29:* Clusterspezifische Fallanzahl

Darüber hinaus kann die **Gruppenzugehörigkeit der Nachfrager** betrachet werden. Nach Tabelle 30, in welcher zusätzlich die Distanzen der Objekte zu den relevanten Clusterzentren angezeigt werden, gehört z.B. Nachfrager 1 dem Cluster 1 an, zu dessen Clustermittelwert er eine Distanz von 0,751 besitzt..

**Cluster-Zugehörigkeit**

| Fallnummer | Person | Cluster | Distanz |
|---|---|---|---|
| 1 | 1 | 1 | ,751 |
| 2 | 2 | 3 | ,559 |
| 3 | 3 | 3 | ,899 |
| 4 | 4 | 1 | 1,398 |
| 5 | 5 | 2 | ,794 |
| 6 | 6 | 1 | 1,423 |
| 7 | 7 | 1 | 1,200 |
| 8 | 8 | 1 | ,575 |
| 9 | 9 | 1 | ,105 |
| 10 | 10 | 3 | 1,420 |
| 11 | 11 | 1 | 1,154 |
| 12 | 12 | 2 | 1,507 |
| 13 | 13 | 1 | 1,140 |
| 14 | 14 | 1 | 1,442 |
| 15 | 15 | 3 | 1,192 |
| 16 | 16 | 2 | ,751 |
| 17 | 17 | 3 | 1,425 |
| 18 | 18 | 1 | ,854 |
| 19 | 19 | 2 | ,774 |
| 20 | 20 | 2 | 1,009 |
| 21 | 21 | 2 | 1,659 |
| 22 | 22 | 1 | ,836 |
| 23 | 23 | 3 | ,974 |
| 24 | 24 | 3 | ,275 |
| 25 | 25 | 3 | ,648 |
| 26 | 26 | 1 | 1,425 |
| 27 | 27 | 1 | ,765 |
| 28 | 28 | 1 | ,713 |
| 29 | 29 | 2 | ,950 |
| 30 | 30 | 3 | 1,118 |

*Tabelle 30:* Clusterzugehörigkeit der Nachfrager

Bei der **inhaltlichen Kennzeichnung** der Clusterprofile mit Hilfe von finalen Clusterzentren ist darauf zu achten, dass es sich dabei im vorliegenden Beispiel um Mittelwerte standardisierter Segmentierungsvariablen handelt. Aufgrund der Skalierung der Ausgangsvariablen signalisieren positive (negative) Zentrenwerte jeweils eine hohe bzw. überdurchschnittliche (geringe bzw. unterdurchschnittliche ) Ausprägung der betreffenden Personenmerkmals. Hiernach lassen sich die Cluster auf Basis der finalen Clusterzentren der Tabelle 26 wie folgt interpretieren:

- In *Cluster 1* befinden sich Nachfrager, die ein überdurchschnittliches Personenalter, ein durchschnittlliches Einkommen sowie eine durchschnittliche Preisbereitschaft. aufweisen.
- *Cluster 2* setzt sich aus Nachfragern zusammen, die ein mittleres Personenalter, ein überdurchschnittliches Einkommen sowie eine überdurchschnittliche Preisbereitschaft besitzen.

➲ Dem *Cluster 3* gehören Nachfrager an, deren Personenalter, Einkommen sowie Preisbereitschaft jeweils unterdurchschnittlich ausgeprägt ist.

Im praktischen Anwendungsfall wäre eine ergänzende bzw. detailliertere Clusterbeschreibung mit Hilfe jener Zusatzvariablen (sog. Passualvariablen) vorzunehmen, die nicht als Segmentierungsvariablen herangezogen wurden. Da die Variable „Clusterzugehörigkeit" nominales Datenniveau besitzt, ist die Untersuchung von Segmentunterschieden hinsichtlich nominaler Zusatzvariablen (z.B. Markenwahl) mittels Kontingenzanalysen durchzuführen, während Gruppenvergleiche auf Basis metrischer Zusatzvariablen (z.B. Kaufmenge) im Zuge von Varianzanalysen vorzunehmen sind. Einen alternativen Ansatz bietet die SPSS-Prozedur Two-Step-Clusteranalyse, die Anwendern seit der Einführung der SPSS Version 12.0 zur Verfügung steht. Mit dieser können sowohl gemischt-skalierte Klassifizierungsvariablen ausgewertet als auch simultan Zusatzvariablen zur Clusterbeschreibung analysiert werden.

# 5. Fallbeispiele aus der Marketingpraxis

## 5.1. Serviceanalyse im Automobilhandel

Im Rahmen einer vom Verfasser durchgeführten Serviceanalyse für einen Vertragshändler einer bundesdeutschen Premium-Automobilmarke, wurde auf Basis einer Stichprobe von 397 Automobilkunden u.a. die Servicezufriedenheit der Kunden analysiert (Meßinstrument: standardisierter Fragebogen; Ratingskalierung: 1 = sehr zufrieden,..3 = weder zufrieden noch unzufrieden....,5 = sehr unzufrieden). Neben der Durchführung von Standardauswertungen, wie z.B. der Ermittlung von Zufriedenheitsprofilen oder der Gegenüberstellung von Zufriedenheitsausprägungen sowie den Beurteilungsgewichten relevanter Servicemerkmale in der sog. Kundenzufriedenheitsmatrix (vgl. Müller 1996), bestand ein wesentliches Analyseziel darin zu untersuchen, ob sich die betreffenden Kunden hinsichtlich ihres Zufriedenheitsgrades voneinander unterscheiden und in sog. Zufriedenheitsgruppen unterteilen lassen. Als Gruppierungsvariablen dienten hierbei die Einzelzufriedenheiten bezüglich von 14 Merkmalen des händlerseitigen Serviceangebots sowie die (globale) Gesamtzufriedenheit der Kunden. Mittels einer Clusterzentrenanalyse, die auf der zufriedenheitstheoretisch begründeten Vorgabe von drei Clustern (vgl. hierzu Müller 1997) beruhte, wurden u.a. die nachstehenden Befunde gewonnen:

❒ Dem **Iterationsprotokoll** zufolge, erreichte der iterative Prozeß der distanzminierenden Fallzuordnung nach 16 Schritten das Konvergenzkriterium.

**Iterationsprotokoll[a]**

| Iteration | Änderung in Clusterzentren | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 4,441 | 2,986 | 4,303 |
| 2 | ,796 | ,196 | ,538 |
| 3 | ,867 | ,176 | ,517 |
| 4 | ,413 | ,210 | ,355 |
| 5 | ,378 | ,176 | ,284 |
| 6 | ,302 | ,305 | ,331 |
| 7 | ,421 | ,339 | ,294 |
| 8 | ,260 | ,239 | ,177 |
| 9 | ,214 | ,103 | ,096 |
| 10 | ,212 | ,036 | ,072 |
| 11 | ,117 | ,038 | ,046 |
| 12 | ,114 | ,000 | ,035 |
| 13 | ,107 | ,018 | ,040 |
| 14 | ,000 | ,052 | ,029 |
| 15 | ,000 | ,022 | ,012 |
| 16 | ,000 | ,000 | ,000 |

a. Konvergenz wurde aufgrund geringer oder keiner Änderungen der Clusterzentren erreicht. Die maximale Änderung der absoluten Koordinaten für jedes Zentrum ist ,000. Die aktuelle Iteration lautet 16. Der Mindestabstand zwischen den anfänglichen Zentren beträgt 8,602.

*Tabelle 31*: Iterationsprotokoll

❒ Die in Tabelle 32 angeführten euklidischen **Distanzen zwischen den Clusterzentren der finalen Clusterlösung** machen u.a. deutlich**,** dass die Gruppen 1 und 3 (Distanzwert: 6,417) die größte Unähnlichkeit besitzen.

**Distanz zwischen Clusterzentren der endgültigen Lösung**

| Cluster | 1 | 2 | 3 |
|---|---|---|---|
| 1 | | 6,417 | 3,435 |
| 2 | 6,417 | | 3,039 |
| 3 | 3,435 | 3,039 | |

*Tabelle 32:* Distanzen zwischen finalen Clusterzentren

❒ Die Tabelle der **endgültigen Clusterzentren** zeigt die Zentrenwerte nach dem Abschluss des Iterationsprozesses bzw. die finalen Clusterzentren für die vorgegebene 3-Clusterlösung an (vgl. Tabelle 33), die sich im Vergleich zu den anfänglichen Clusterzentren erheblich verschoben haben.

**Clusterzentren der endgültigen Lösung**

| | Cluster[a] | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| atmosphäre im warteraum | 2,39 | 1,41 | 2,07 |
| vereinbarte servterm werden eingehalten | 2,19 | 1,07 | 1,67 |
| positives Preis-Leistungsverhältnis | 3,72 | 1,85 | 2,72 |
| vereinbarte Leistungsumfänge einhalten | 2,76 | 1,24 | 1,93 |
| kurze Wartezeit auf Insp. /Rep.termine | 2,46 | 1,29 | 1,92 |
| beratungsqualität des personals | 3,21 | 1,22 | 2,11 |
| Qualität der Reparaturarbeiten | 3,03 | 1,23 | 2,01 |
| Qualität der Inspektionsarbeiten | 2,76 | 1,23 | 1,99 |
| freundl. u. hilfsb. Auftreten d. Personals | 2,52 | 1,11 | 1,83 |
| Rechng. u. arbeiten werden erklärt | 2,96 | 1,40 | 2,19 |
| verhalten bei kundenbeschwerden | 3,66 | 1,52 | 2,35 |
| keine mängel bei auslieferung | 3,33 | 1,24 | 2,08 |
| unbuerok. verhalten b. kleinarbeiten | 3,21 | 1,17 | 2,17 |
| Fahrzeugabholung verlaeuft reibungslos | 2,07 | 1,05 | 1,79 |
| **Gesamtzufriedenheit** | **3,09** | **1,19** | **2,05** |

a. hoch-signifikante Gruppenunterschiede bezüglicher aller Merkmale

*Tabelle 33*: Finale Clusterzentren

❒ Die **Clusterprofile** können anhand der finalen Clusterzentren, die sich signifikant voneinander unterscheiden, wie folgt beschrieben werden:

➲ In *Cluster 1* befinden sich Automobilkunden, deren Servicezufriedenheit lediglich durchschnittlich (Ratingwert der Gesamtzufriedenheit = 3) ausgeprägt ist und diese demzufolge als die Gruppe der „Indifferenten“ gekennzeichnet werden kann. Dieser Gruppe gehören 67 Kunden bzw. 16,7 % der Stichprobe an.

➲ *Cluster 2* setzt sich aus Nachfragern zusammen, die mit dem Händlerservice sehr zufrieden sind (Ratingwert der Gesamtzufriedenheit = 1,19). Insofern liegt es nahe, diese Gruppe, in der sich 118 Kunden bzw. 29,7 % aller Kunden befinden, als die „Begeisterten“ zu interpretieren.

➲ Dem *Cluster 3* gehören Kunden an, die bezüglich der angebotenen Händlerservices zufrieden sind (Ratingwertwert der Geamtzufriedenheit = 2,05). Daher kann diese Gruppe, der 213 Kunden bzw. 53,6 % der Stichprobe zugehörig sind, als die „Zufriedenen“ beschrieben werden.

❒ Im Zuge **weiterführender Analysen** wurde ferner der Frage nachgegangen, ob sich die Gruppen bezüglich ausgewählter Passualvariablen voneinander unterscheiden. So lässt z.B. eine Kreuztabellierung der beiden Variablen „Fahrzeugklasse“ und „Gruppenzugehörigkeit der Fälle“ erkennen, dass von den insgesamt 62 Besitzern eines Oberklassen-Pkw's ca. 21 % ( = 13/62 * 100) der Gruppe der „Indifferenten“ angehören, während dieser Gruppe lediglich 13% der Besitzer eines Kompakt-Pkw's zugehörig sind. Hieraus kann die Hypothese formuliert werden, dass die Besitzer von Oberklassen-Pkw's einen tendenziell geringeren Zufriedenheitsgrad aufweisen als die Besitzer anderer Fahrzeugklassen. Diese Aussage steht im Einklang der theroetisch begründbaren Einsicht, dass die Besitzer von Oberklassen-Pkw's über höhere Servicerwartungen verfügen als die Besitzer von Mittelklasse- oder Kompaktfahrzeugen.

Anzahl

| | | Cluster-Nr. des Falls | | | **Gesamt** |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | |
| Fahrzeugklasse | Oberklasse | 13 | 18 | 31 | 62 |
| | Mittelklasse | 41 | 68 | 132 | 241 |
| | Kompaktklasse | 12 | 32 | 50 | 94 |
| **Gesamt** | | 66 | 118 | 213 | 397 |

*Tabelle 34*: Kreuztabelle

## 5.2. Strategische Wettbewerbergruppen im Großhandel

Im Rahmen einer umfassenden Analyse zu den Zukunftsperspektiven des regionalen Großhandels im Kammerbezirk Dortmund war dem Verfasser u.a. die Aufgabe gestellt, auf der Grundlage der Konzeption der strategischen Erfolgsfaktorenforschung strategische Wettbewerbergruppen zu identifizieren, voneinander abzugrenzen und die relevanten Erfolgsfaktoren der Marktbearbeitung herauszuschälen (vgl. Müller 2004 a). Die Erfassung und Abgrenzung von Wettbewerbergruppen auf Basis einer Zufallsstichprobe von 120 Großhandelsbetrieben (Meßinstrument: schriftliche Befragung) wurde mittels einer hierarchischen Clusteranalyse (Ward-Verfahren, quadriertes euklidisches Distanzmaß) vorgenommen, bei welcher insgesamt zehn ökonomische, psychographische und soziale Unternehmensziele von Großhandelsbetrieben als Gruppierungskriterien dienten. Der Analyseprozess erbrachte die nachstehenden zentralen Ergebnisse:

❒ Im Großhandel agieren zwei Wettbewerbergruppen, die sich als **Wettbewerbsführer** und **Wettbewerbsfolger** kennzeichnen lassen ( vgl. Abb. 23). Die Gruppe der Wettbewerbsführer, der 57 % der Großhandelsbetriebe angehören, verfolgt eine Zielkonzeption, in der alle zehn strategischen Zielbereiche einen jeweils hohen Stellenwert einnehmen. Hierbei werden die Unternehmensaktivitäten zukünftig primär auf die Erreichung der vier Kernziele „Unternehmensrentabilität“,

„betriebliche Kostensituation“ „Gewinnwachstum“ und „Kundenzufriedenheit“ ausgerichtet. In (signifikantem) Unterschied hierzu ist das Zielsystem der Wettbewerbsfolger dadurch gekennzeichnet, dass dieses einerseits die drei Kernziele der „Unternehmensrentabilität“, Gewinnwachstum“ und „Kundenzufriedenheit“ umfasst und andererseits den übrigen Zielbereichen einen lediglich mittleren oder geringen Stellenwert beimisst. Augenfällig ist ferner, dass Wettbewerbsfolger die „Sicherung von Arbeitsplätzen“ als unbedeutsam erachten.

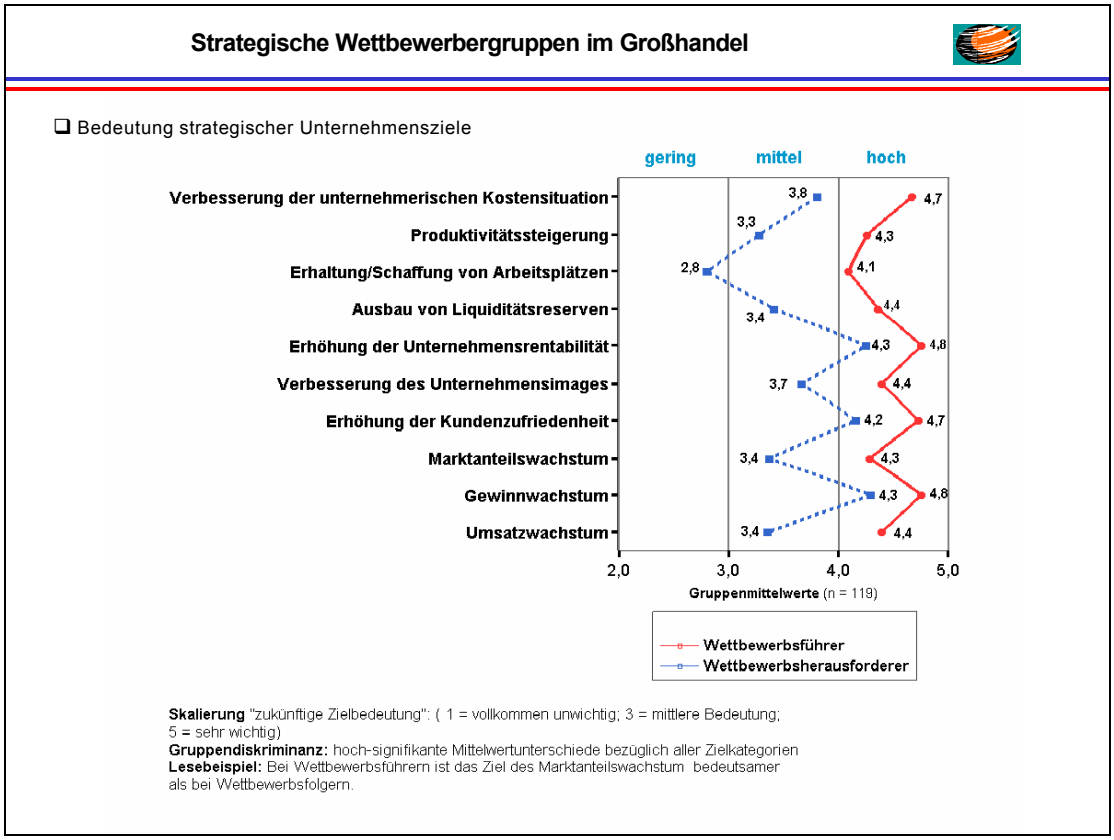



*Abbildung 23*: Unternehmensziele strategischer Wettbewerbergruppen im Großhandel

❒ Die Umsetzung strategischer Unternehmensziele erfolgt durch Marketing-, Beschaffungs-, Kooperations – und Managmentstrategien (vgl. Müller 2004 a). Im Bereich der **Marketingstrategien** forcieren Wettbwerbsführer die strategische Erschließung neuer Kundengruppen sowie die Bindung des Kundenstamms (vgl. Abb. 24). Nicht zuletzt vor dem Hintergrund des angestrebten Gewinnwachstums und der beabsichtigten Erhöhung der Kundenzufriedenheit, erscheint eine derartige strategische Schwerpunktbildung als sachlich angemessen. Im Rahmen marktfeldstrategischer Überlegungen nimmt die Durchdringung des nationalen Absatzgebietes einen zentralen Stellenwert ein; ein strategischer Ansatz, der angesichts der bislang praktizierten Absatzfokussierung auf NRW als folgerichtig gewertet werden muß..Der Aufbau bzw. die Festigung von Wettbewerbsvorteilen ist im regionalen Großhandel vorrangig auf das Konzept der Qualitätsführerschaft ( im Verbund mit einem überdurchschnittlich hohen Preisniveau) ausgerichtet. Unternehmen, welche diese wettbewerbsstrategische Variante praktizieren, bemühen sich darum, einerseits den gewachsenen Qualitätsanforderungen ihre

Abnehmergruppen umfassend gerecht zu werden und andererseits dem aggressiven Preiswettbewerb zu entgehen. Demgegenüber lassen Wettbewerbsfolger eine Fokussierung der Marketingstrategien vermissen. Vielmehr ist die strategische Marketingbearbeitung dieser Wettbewerbergruppe durch eine größtenteils durchschnittliche Einsatzintensität von zudem nicht auf das Zielsystem abgestimmten Marktstrategien gekenzeichnet.

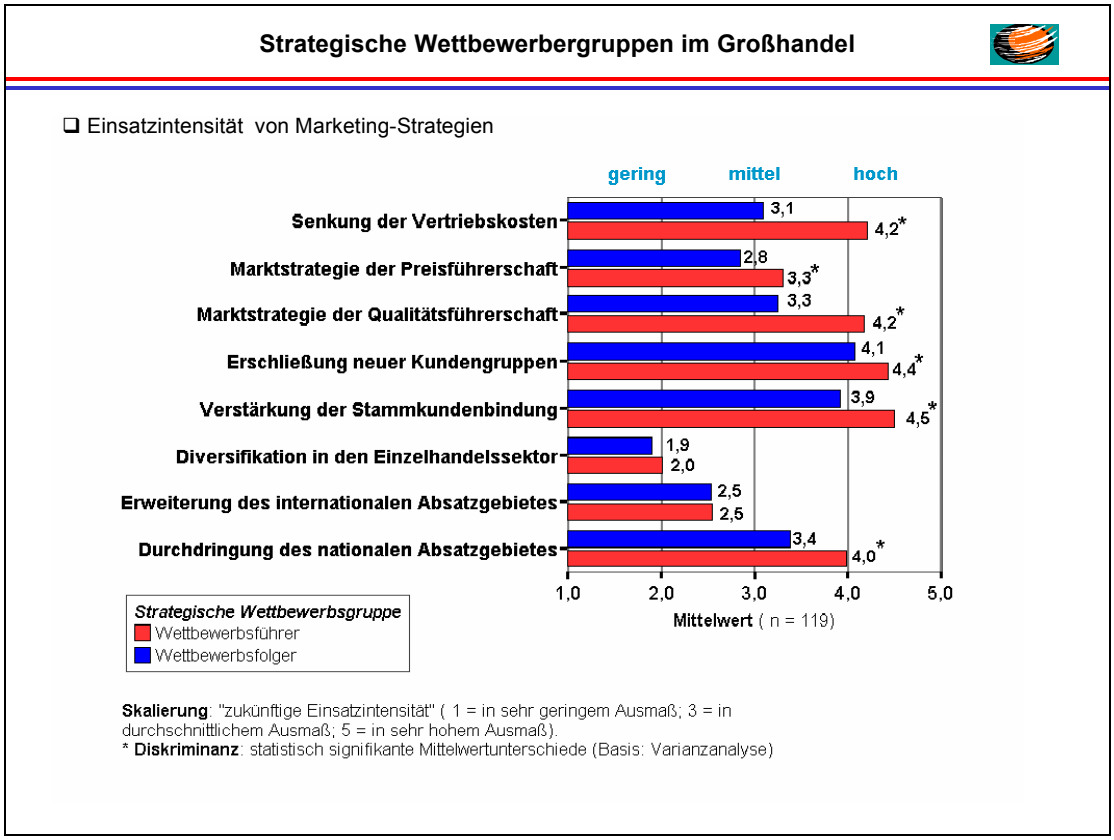



*Abbildung 24* Marketingstrategien strategischer Wettbewerbergruppen im Großhandel

❒ Die Bedeutsamkeit leistungsfähiger **Management-Strategien** wird in der betrieblichen Praxis vielfach übersehen. So werden beispielsweise aus einem mangelndem Unternehmenserfolg häufig falsche Schlüsse gezogen und demzufolge die an sich zielgerechten Unternehmensaktivitäten fälschlicherweise verändert oder vollkommen eingestellt. Demgegenüber wäre es im Sinne eines „ structure follows strategy“ vielfach zweckmäßiger, die betrieblichen Managementkonzepte einer kritischen Prüfung zu unterziehen und den geplanten Marktaktivitäten anzupassen, denn diese befinden sich in zahlreichen Fällen auf einem veralteten und wenig effizienten Leistungsniveau. Management-Strategien beinhalten den Aufbau der erforderlichen betrieblichen Ressourcenstruktur, mit deren Hilfe die marktgerichteten strategischen und operativen Unternehmensaktivitäten geplant, umgesetzt und kontrolliert werden können. Personalstrategien bilden den zukünftig bedeutsamsten Managementbereich des Großhandels. Dies gilt insbesondere für die Gruppe der Wettbewerbsführer, welche der Flexibilisierung des Personaleinsatzes, der Verbesserung von

Mitarbeiterqualifikationen sowie der Personalkostensenkung einen gewichtigen Stellenwert beimessen. Von gleichfalls überdurchschnittlicher Bedeutsamkeit ist die Implementierung leistungsfähiger Managementsysteme. Wettbewerbsführer werden zukünftig verstärkt den Einsatz neuer Informations-, Controlling und Logistiksysteme forcieren. Darüber hinaus erfolgt im Rahmen des Investitionsmanagements zukünftig eine Fokussierung auf Erweiterungs- sowie auf Rationalisierungsinvestitionen, während Ersatzinvestitionen in vergleichsweise unterdurchschnittlichem Maße getätigt werden. Für die Gruppe der Wettbewerbsfolger ist hingegen – analog zum Bereich der Marktstrategien - keine managementstrategische Schwerpunktsetzung zu erkennen.

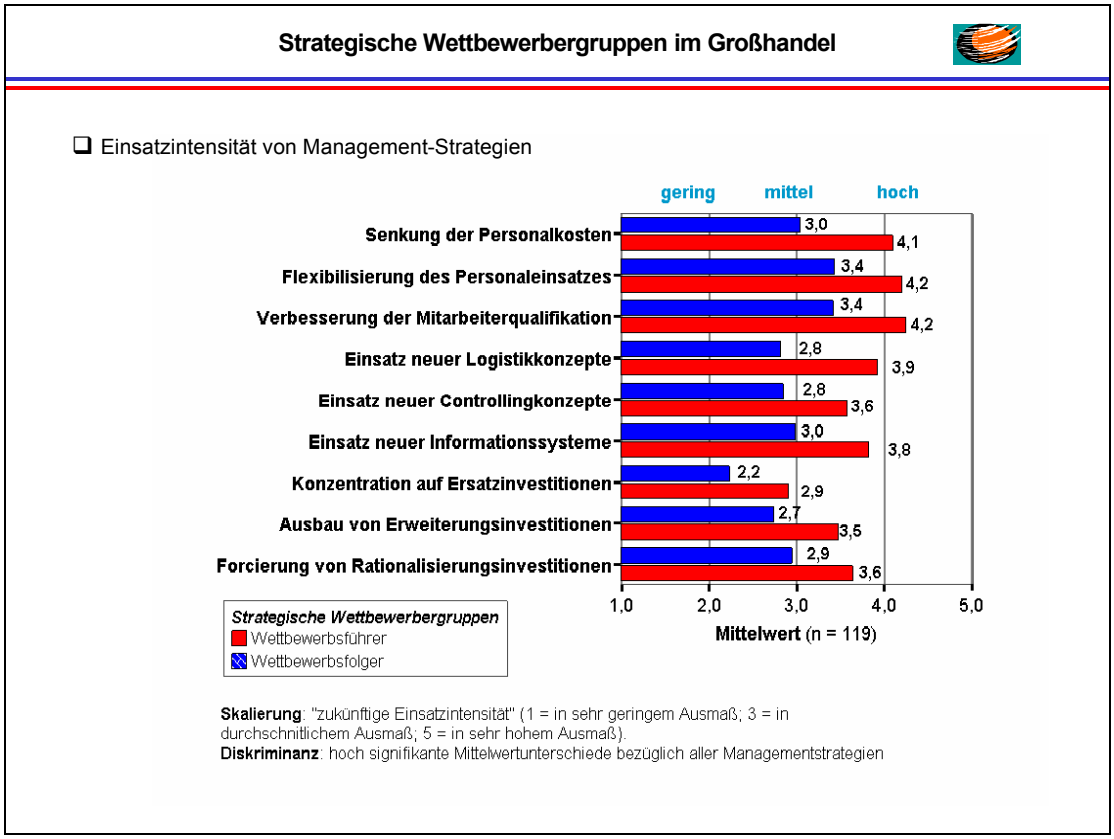



*Abbildung 25:* Managementstrategien strategischer Wettbewerbergruppen im Großhandel

❒ Eine abschließende Beschreibung der Wettbewerbergruppen mit Hilfe von ausgewählten **Merkmalen der Unternehmensstruktur** macht deutlich, dass der Gruppe der Wettbewerbsführer tendenziell mehr größere Großhandelsbetriebe, d.h. Unternehmen angehören, deren Mitarbeiterzahl, Kundenzahl und Umsatzhöhe vergleichsweise hoch ist (vgl Tabelle 35).

| | | | Strategische Wettbewerbergruppen | | Gesamt |
|---|---|---|---|---|---|
| | | | Wettbewerbs-führer | Wettbewerbs-folger | |
| **Branchentyp** | Produktionsgüterhandel | | 55,8% | 44,2% | 100,0% |
| | Konsumgüterhandel | | 58,1% | 41,9% | 100,0% |
| **Kundenzahl** | < 250 Kunden | *a | 42,1% | 57,9% | 100,0% |
| | 251 - 1000 Kunden | | 56,4% | 43,6% | 100,0% |
| | > 1000 Kunden | * | 71,8% | 28,2% | 100,0% |
| **Umsatzklasse** | 6 - 10 Mio. DM | | 54,9% | 45,1% | 100,0% |
| | 11 - 50 Mio. DM | | 56,7% | 43,3% | 100,0% |
| | > 50 Mio. DM | * | 70,0% | 30,0% | 100,0% |
| **Beschäftigtenklasse** | < 10 Beschäftigte | | 45,5% | 54,5% | 100,0% |
| | 10 - 50 Beschäftigte | * | 59,2% | 40,8% | 100,0% |
| | > 50 Beschäftigte | * | 70,8% | 29,2% | 100,0% |

a. * = statistisch signifikante Merkmalsunterschiede (Basis: Verteilungstest)

*Tabelle 35:* Unternehmensmerkmale strategischer Wettbewerbergruppen im Großhandel

## Literaturverzeichnis


Aaker, D., Kumar, V., Day, G. (2001): Marketing Research, 7th Edition, New York, Chichester u.a..

Bacher, J. (1996): Clusteranalyse, 2. Auflage, München.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2003): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, 10. Auflage, Berlin, Heidelberg, New York.

Berekhoven, L., Eckert, W., Ellinger, P. (2004): Marktforschung, 10. Auflage, Wiesbaden 2004.

Böhler, H. (2004): Marktforschung, 3. Auflage, Stuttgart, Berlin u.a.

Bortz, J. (1993): Statistik für Sozialwissenschaftler, 4. Auflage, Berlin, Heidelberg u.a..

Brosius, F. (2002): SPSS 11, Landsberg a. Lech.

Büschgen, J., Thaden, Ch. (1999): Clusteranalyse, in: Herrmann, A., Homburg, Ch. (Hrsg.): Marktforschung, Wiesbaden, S. 337-380.

Churchill, G., Iacobucci, D. ( 2005): Marketing Research, 9th Edition, Mason.

Diehl, J., Kohr, H. (1999): Deskriptive Statistik, 12. Auflage, Frankfurt/M..

Eckey, H.-F., Kosfeld, R., Rengers, M. (2002): Multivariate Statistik. Grundlagen – Methoden – Beispiele, Wiesbaden.

Freter, H., Obermaier, O. (2000): Marktsegmentierung, in: Herrmann, A., Homburg, Ch. (Hrsg.): Marktforschung, Wiesbaden, S.739-763.

Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1998): Multivariate data analysis, Fifth Edition, New Jersey.

Hammann, P., Erichson, B. (2000): Marktforschung, 4. Auflage, Stuttgart.

Hartung, J., Elpelt, B. (1995): Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik, 5. Auflage, München, Wien.

Homburg, Ch. (1998): Quantitative Betriebswirtschaftslehre, 2. Auflage, Wiesbaden.

Hüttner, M (1997): Grundzüge der Marktforschung, 5. Auflage, München, Wien.

Hüttner, M., Schwarting, U. (1999): Exploratorische Faktorenanalyse, in: Herrmann, A., Homburg, Ch. (Hrsg.): Marktforschung, Wiesbaden, S. 383-412.

Kachigan, S. K. (1991): Multivariate Statistical Analysis, 2nd Edition, New York.

Kinnear, Th., Taylor, J. (1996): Marketing Research. An Applied Approach, 5th Edition, NewYork,St. Louis u.a..

Litz, P. (2000): Multivariate Statistische Methoden, München, Wien.

Marinell, G. (1998): Multivariate Verfahren, 5. Auflage, München, Wien.

Malhotra, N. (1999): Marketing Research. An Applied Orientation, 3rd Edition, Upper Saddle River.

Müller, W. (1996): Angewandte Kundenzufriedenheitsforschung, in: Marktforschung & Management, Heft 4, S. 149-159.

Müller, W. (1997): Erfolgsfaktoren im Dienstleistungsmanagement des Automobilhandels, in: Jahrbuch der Absatz- und Verbrauchs-forschung, Heft 1, 1997, S. 41-65.

Müller, W. (1998): Service-Qualität im Firmenkundengeschäft von Banken, , in: Bank und Markt, Heft 1, 1998, S. 45-51.

Müller, W. (1999): Gerechtigkeitsheoretische Modelle zur Kundenzufriedenheit, in: Jahrbuch der Absatz- und Verbrauchsforschung, Heft 3, 1999, S. 239-266.

Müller, W. (2004 a): Marktorientierte Unternehmensführung im Großhandel – eine empirische Bestandsaufnahme, in: Baumgarth, C. (Hrsg.): Marktorientierte Unternehmensführung, Festschrift zum 60. Geburtstag von Univ.-Prof. Dr. Hermann Freter, Peter Lang Verlag, Frankfurt/M., S. 345-371.

Müller, W. (2004 b): Multivariate Statistik im Quantitativen Marketing - Teil III: Faktorenanalyse, Band 8 des Instituts für Angewandtes Markt-Management, Dortmund.

Rinne, H. (2000): Statistische Analyse multivariater Daten, München, Wien.

Rudolf, M., Müller, J. (2004): Multivariate Verfahren, Göttingen, Bern.

Sudman, S., Blair, E. (1998): Marketing Research. A Problem Solving Approach, Boston, Burr Ridge u.a..

Tabachnick, B., Fidell, L. (2001): Using Multivariate Statistics, 4th Edition, Boston, London u.a..

Trommsdorf, V., Bookhagen, A., Hess, C. (2000): Produktpositionierung, in: Herrmann, A., Homburg, Ch. (Hrsg.): Marktforschung, Wiesbaden, S.767-787.

Voß, W. (2004) (Hrsg.): Taschenbuch der Statistik, 2. Auflage, München, Wien.

## Dokumentation der Forschungsreihe

Die Forschungspapiere des iamm erscheinen in unregelmäßigen Abständen. Bisher sind erschienen:

1) Satisfaction-based Measurement of Service Quality. Forschungspapier, Band 1 des Instituts für Angewandtes Markt-Management, Dortmund 2000.
2) Customer Satisfaction at the German Travel Agency Marktet. An empirical Analysis, Forschungspapier, Band 1 des Instituts für Angewandtes Markt-Management, Dortmund 2001.
3) Marktorientierte Unternehmensführung im mittelständischen Großhandel – eine empirische Bestandsaufnahme, Band 3 des Instituts für Angewandtes Markt-Management, Dortmund 2002.
4) Gerechtigkeitstheoretische Modelle zur Kundenzufriedenheit, Band 4 des Instituts für Angewandtes Markt-Management, Dortmund 2003.
5) Grundlagen der quantitativen Marketinganalyse, Band 5 des Instituts für Angewandtes Markt-Management, Dortmund 2004.
6) Multivariate Statistik im Quantitativen Marketing - Teil I: Regressionsanalyse, Band 6 des Instituts für Angewandtes Markt-Management, Dortmund 2004.
7) Multivariate Statistik im Quantitativen Marketing - Teil II: Varianzanalyse, Band 7 des Instituts für Angewandtes Markt-Management, Dortmund 2004.
8) Multivariate Statistik im Quantitativen Marketing - Teil III: Faktorenanalyse, Band 8 des Instituts für Angewandtes Markt-Management, Dortmund 2004.
9) Multivariate Statistik im Quantitativen Marketing - Teil IV: Clsuteranalyse, Band 9 des Instituts für Angewandtes Markt-Management, Dortmund 2004.