3 DETERMINISTISCHE CLUSTERANALYSEVERFAHREN

entnommen aus: Bacher, J., 1996: Clusteranalyse. München, S. 172-198.
Seitenangaben stimmen nicht exakt mit der Buchpublikation überein!

3.1 Einleitende Übersicht

Die Größen bedeuten: V_i = das auf das Intervall [1,7] standardisierte empirische Verschmelzungsniveau des i-ten Schrittes, $E(V_i)$ = das auf das Intervall [1,7] standardisierte erwartete Verschmelzungsniveau des i-ten Schrittes beim Vorliegen einer homogenen, normalverteilten Population. $SA(V_i)$ = Standardabweichung des Verschmelzungsniveaus des i-ten Schrittes beim Vorliegen einer homogenen, normalverteilten Population. Aus dem erwarteten Verschmelzungsniveau und der Standardabweichung läßt sich u.a. mit $E(V_i) \pm 2 \cdot SA(V_i)$ ein 95-Prozent Vertrauensintervall berechnen. Liegen alle empirischen

Verschmelzungsniveaus innerhalb des Vertrauensintervalls, stellen die Ergebnisse ein Artefakt dar, da diese auch erzielt werden können, wenn reine Zufallsdaten untersucht werden. In dem Beispiel liegen alle empirischen Verschmelzungsniveaus innerhalb der Vertrauensintervalle. Die untersuchte Clusterlösung ist somit als Artefakt zu betrachten. Dies stimmt mit den Annahmen des Rechenexperiments überein, bei dem angenommen wurde, daß alle untersuchten Variablen eine homogene Normalverteilung besitzen.

3.1.9 Aufbau des Kapitels

Bei dem Aufbau des Kapitels trat das Problem auf, aus der Vielzahl der Anwendungsmöglichkeiten und der Verfahren bestimmte Verfahren für eine ausführlichere Beschreibung auszuwählen. Wir haben uns dabei an den in Abschnitt 3.1.4 angeführten Grundmodellen und deren Beziehung zu den anderen Verfahren orientiert und das Kapitel wie folgt gegliedert:

Abschnitt 3.2: Gewichtung und Transformation von Variablen und Objekten: Hier wird u.a. behandelt, wie durch Datentransformationen das Problem der Nichtvergleichbarkeit gelöst werden kann.

Abschnitt 3.3: Un- und Ähnlichkeitsmaße: Eine Auswahl von Ähnlichkeits- und Unähnlichkeitsmaßen wird gegeben. Ferner wird der Frage nachgegangen, wie die Signifikanz von Ähnlichkeits- und Unähnlichkeitsmaßen geprüft werden kann.

Abschnitt 3.4: Nächste-Nachbarn-Verfahren und Mittelwertverfahren: Die Logik dieser Verfahren wird am Beispiel des Complete-Linkage dargestellt. Der Single-Linkage, der Complete-Linkage für überlappende Cluster sowie die verallgemeinerten Nächste-Nachbarn-Verfahren werden als Gegentypus (=Single-Linkage) bzw. als Modifikationen (=die anderen Verfahren) des Complete-Linkage behandelt.
Als eine weitere Modifikation des Complete-Linkage werden in diesem Abschnitt auch die Mittelwertverfahren (Weighted-Average-Linkage, Average-Linkage und Within-Average-Linkage) dargestellt.

Abschnitt 3.5: Repräsentantenverfahren

Abschnitt 3.6: Hierarchische Verfahren zur Konstruktion von Clusterzentren. Hier werden das Median-, Zentroid- und Ward-Verfahren dargestellt.

Abschnitt 3.7: Partitionierende Verfahren zur Konstruktion von Clusterzentren. In diesem Abschnitt werden die K-Means-Verfahren behandelt.

3.2 Gewichtung und Transformation von Variablen und Objekten

3.2.1 Vergleichbarkeit von Klassifikationsmerkmalen

Formal müssen - wie bereits im Abschnitt 3.1.7 erwähnt - die Variablen, die in eine deterministische Clusteranalyse einbezogen werden, "vergleichbar" sein. Vergleichbarkeit (Kommensurabilität) von Variablen liegt in folgenden Situationen nicht vor:[1]

1. Die Variablen besitzen unterschiedliches Meßniveau.
2. Die Variablen besitzen gemischtes Meßniveau.
3. Die Variablen sind hierarchisch. Eine Variable kann nur auftreten, wenn in einer anderen Variablen eine bestimmte Ausprägung auftritt. Z.B.: Die Variable "derzeitiger Beruf" wurde nur erfragt, wenn der Befragte erwerbstätig ist.

Unterschiedliche Maßeinheiten: In dem Anwendungsbeispiel des Abschnitts 3.1.7 wurden als Indikatoren für die wirtschaftliche Entwicklung das "Pro-Kopf-Bruttosozialprodukt" und das "jährliche Wirtschaftswachstum in den 80er Jahren" verwendet. Diese beiden Variablen besitzen zwar das gleiche Meßniveau (=quantitativ), sind aber nicht vergleichbar, da das "Pro-Kopf-Bruttosozialprodukt" in einer bestimmten Währungseinheit (=Dollar) gemessen wird, das "jährliche Wirtschaftswachstum in den 80er Jahren" dagegen in Prozenten.

Unterschiedliches Meßniveau: In eine Clusteranalyse sollen die nominalen Variablen "Geschlecht" und "berufliche Tätigkeit", die ordinale Variable "abgeschlossene Schulbildung" und die quantitative Variable "Einkommen" einbezogen werden. In diesem Beispiel liegt Nichtvergleichbarkeit vor, da die Variablen unterschiedliches Meßniveau besitzen.

Hierarchische oder bedingte Variablen: Diese liegen dann vor, wenn das Auftreten einer Variablen von dem Auftreten der Ausprägung(en) einer oder mehrerer anderer Variablen abhängt. Die Variable "derzeitiger Beruf" tritt beispielsweise nur dann auf, wenn die vorausgehenden Variable "Berufstätigkeit" die Ausprägung "derzeit berufstätig" besitzt.

Daneben können *inhaltliche Überlegungen* zu dem Urteil der Nichtvergleichbarkeit führen: Selbst wenn alle in die Analyse einbezogenen Variablen dieselbe Maßeinheit (z.B. Prozente) besitzen, wie z.B. "Industrialisierungsquote in den 80er Jahren", "jährliches Wirtschaftswachstum in den 80er Jahren", kann in Frage gestellt werden, ob beide Klassifikationsmerkmale dieselbe "Maßeinheit" besitzen, ob also eine Differenz von 5 Prozent beim jährlichen Wirtschaftswachstum "dasselbe" bedeutet wie bei der Industrialisierungsquote (siehe dazu auch *Fox* 1982: 132). Weitere inhaltliche Überlegungen beziehen sich darauf, ob den Variablen gemeinsame Dimensionen zugrundeliegen, die durch eine unterschiedliche Anzahl von Indikatoren (=Variablen) repräsentiert sind (Problem der *Über- bzw. Unterrepräsentativität*, siehe dazu Abschnitt 3.1.7).

---

[1] Siehe dazu z.B. *Schlosser* (1976: 60-88), *Sodeur* (1974: 44-59) oder *Vogel (*1975: 50-78).

3.2.2 Lösungsstrategien

Liegt Nichtvergleichbarkeit der Variablen vor, stehen u.a. folgende Strategien zur Verfügung:

1. Die Variablen werden vor der Analyse transformiert bzw. gewichtet.
2. Die Variablen werden bei der Berechnung eines Ähnlichkeits- oder Unähnlichkeitsmaßes gewichtet.
3. Die Variablen werden in der Analyse gewichtet.
4. Es werden getrennte Analysen für jeweils jene Variablengruppen gerechnet, die vergleichbar sind. Beispiel: In einer Untersuchung wurden zwei Fragebatterien verwendet. Mit der ersten Fragebatterie wurden Erziehungsziele durch eine fünfstufige Antwortskala erfaßt. In der zweiten Fragebatterie wurden gemeinsame familiäre Freizeitaktivitäten mit einer dichotomen Antwortskala erfaßt. Um das Problem der Vergleichbarkeit zu umgehen, kann zunächst jede Fragebatterie getrennt untersucht werden.

Allgemein sollte nach Möglichkeit bei der Datenerhebung das *Problem der Nichtvergleichbarkeit vermieden werden*, indem bedeutungsgleiche, der jeweiligen Fragestellung angepaßte Antwortskalen verwendet werden. Liegt dennoch Nichtvergleichbarkeit vor, kann eine der vier genannten Strategien eingesetzt werden, wobei die Strategien 1 bis 3 weitgehend identisch sind. Wir wollen hier die erstgenannte Strategie darstellen.

3.2.3 Theoretische und empirische Standardisierung

Mit der Bezeichnung theoretische und empirische Standardisierung sind folgende Datentransformationen gemeint:

1. Die "eigentliche" *Standardisierung* bzw. *z-Transformation* mit:

(3.2.1) $$z_{gi} = \frac{x_{gi} - \bar{x}_i}{s_i} \quad \text{bzw.} \quad z_{gi} = \frac{x_{gi} - \mu_i}{\sigma_i},$$

   wobei z_{gi} der standardisierte Wert des Objekts g in der standardisierten Variablen Z_i ist. x_{gi} ist der Wert des Objekts g in der nichtstandardisierten Variablen X_i, $\bar{x}_i$ ist der empirische Mittelwert der Variablen X_i, s_i die empirische Standardabweichung, μ_i der theoretische Skalenmittelwert und σ_i die theoretische Standardabweichung.
2. Die *Extremwertnormalisierung* mit

(3.2.2) $$z_{gi} = \frac{x_{gi} - a_i}{b_i - a_i} \quad \text{bzw.} \quad z_{gi} = \frac{x_{gi} - \alpha_i}{\beta_i - \alpha_i},$$

wobei a_i die empirische Untergrenze der Variablen X_i ist. b_i ist die empirische Obergrenze, α_i die theoretische Untergrenze und β_i die theoretische Obergrenze.

In eine Standardisierung können - wie den Berechnungsformeln zu entnehmen ist - theoretische Skalenkennwerte oder empirische Verteilungskennwerte eingehen. Diese Unterscheidung ist *nur für eine objektorientierte Clusteranalyse wichtig*. Im Rahmen einer variablenorientierten Clusteranalyse werden üblicherweise Korrelationskoeffizienten zur Messung der Ähnlichkeit verwendet. Es wird somit implizit eine empirische Standardisierung durchgeführt. Die *nachfolgenden Ausführungen* beziehen sich somit nur auf eine *objektorientierte Clusteranalyse*.

Zur Verdeutlichung des Unterschieds von theoretischen und empirischen Kennwerten wollen wir folgendes Beispiel betrachten. In eine Clusteranalyse soll u.a. das Item "Eine berufstätige Mutter kann ein genauso herzliches Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist" mit den Ausprägungen 1 (=stimme voll zu), 2 (=stimme zu), 3 (=stimme eher nicht zu) und 4 (=stimme überhaupt nicht zu) einbezogen werden. Die Befragten verteilen sich auf die Antwortkategorien wie folgt: 1 = 35%, 2 = 45%, 3 = 20% und 4 = 0%. Hinsichtlich dieser Variablen können folgende Skalenkennwerte berechnet werden:

1. Theoretische Unter- und Obergrenze: Diese sind 1 und 4.
2. Empirische Unter- und Obergrenze: Diese sind 1 und 3, da empirisch die Ausprägung 4 nicht auftritt.
3. Theoretischer Skalenmittelwert: Dieser ist gleich (1+2+3+4)/4 = 2.5.
4. Empirischer Mittelwert: Dieser ist gleich $1 \cdot 0.35 + 2 \cdot 0.45 + 3 \cdot 0.20 + 4 \cdot 0.0 = 1.850$.
5. Theoretische Skalenstandardabweichung: Diese ist gleich Wurzel aus $((1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2)/4$ =Wurzel aus 1.25 =1.12.
6. Empirische Standardabweichung: Diese ist gleich der Wurzel aus $0.35 \cdot (1-1.850)^2 + 0.45 \cdot (2-1.850)^2 + .. = 0.73$.

Die Abbildung 3.2.1 verdeutlicht nochmals den Unterschied zwischen theoretischen und empirischen Skalenkennwerten. Empirische Skalenwerte werden somit auf der Grundlage der Verteilung der untersuchten Objekte berechnet. *In die Berechnung der theoretischen Skalenwerte geht dagegen die Verteilung der untersuchten Objekte (=Personen) nicht ein.* Sie sind daher populationsunabhängig. Die theoretischen Skalenkennwerte werden aus der Bedeutung der vorgegebenen Antwortskalen abgeleitet, wobei zur Berechnung des theoretischen Skalenmittelwerts und der theoretischen Skalenstandardabweichung zusätzlich eine Gleichverteilung der Objekte auf den Skalen (Variablen) angenommen wird.[1] Die theoretischen Skalenkennwerte

[1] Auch die Annahme einer Normalverteilung ist möglich (*Schlosser* 1976: 64). Dadurch ensteht aber ein größerer Berechnungsaufwand.

können in Abhängigkeit vom Skalentyp nach einer der in der Tabelle 3.2.1 wiedergegebenen Formeln berechnet werden.

Abbildung 3.2.1: Theoretische und empirische Skalenkennwerte

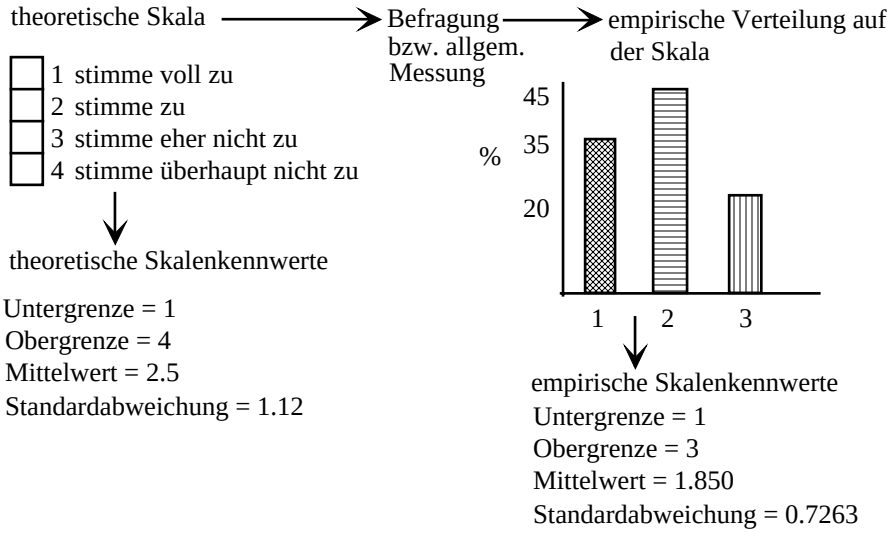


Tabelle 3.2.1: Berechnungsformeln für theoretische Skalenkennwerte

theoretische Skalen-kennwerte	Symbol	Skalentypen		
		Skalentyp I	Skalentyp II	Skalentyp III
theoretische Untergrenze	α_i	aus der Skala unmittelbar ablesbar		
theoretische Obergrenze	β_i	aus der Skala unmittelbar ablesbar		
theoretischer Skalenmittelwert	μ_i	$(\beta_i - \alpha_i)/2$	$(\beta_i - \alpha_i)/2$	$\sum X_{ij}/m_i$
theoretische Skalenvarianz	σ_i^2	$(\beta_i - \alpha_i)^2/1$	$\frac{[(m_i+1)\cdot(m_i}{12}$	$\sum(X_{ij} - \mu_i)^2$
theoretische Skalenstandardab-weichung	σ_i	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$

m_i = Zahl der Ausprägungen der untersuchten Variablen, X_{ij} = Wert der j-ten Ausprägung der Variablen X_i.

Die drei Skalentypen der Tabelle 3.2.1 sind:

1. *Skalentyp I:* Variablen mit einer kontinuierlichen Skala in dem Wertebereich zwischen α_i und β_i, wie z.B. Stimmenanteile von Parteien (Wertebereich von 0 bis 100 Prozent).

2. *Skalentyp II:* Variablen mit einer diskreten Skala mit ganzzahligen äquidistanten Ausprägungen: $X_{1i} = \alpha_i$, $X_{2i} = \alpha_i + 1$, ... wie z.B. eine Einstellungsfrage mit den Ausprägungen 1 (=stimme voll zu), 2 (=stimme zu), 3 (=stimme eher nicht zu), 4 (=stimme überhaupt nicht zu).
3. *Skalentyp III:* Variablen mit einer diskreten Skala mit nicht äquidistanten Ausprägungen, wie z.B. eine Aktivitätsfrage (z.B. nach der Häufigkeit einer bestimmten Freizeitaktivität oder nach dem Kontakt mit einer bestimmten Person) mit den Ausprägungen täglich (=7.0), mehrmals wöchentlich (=3.5), wöchentlich (=1.0) und seltener (=0.20).

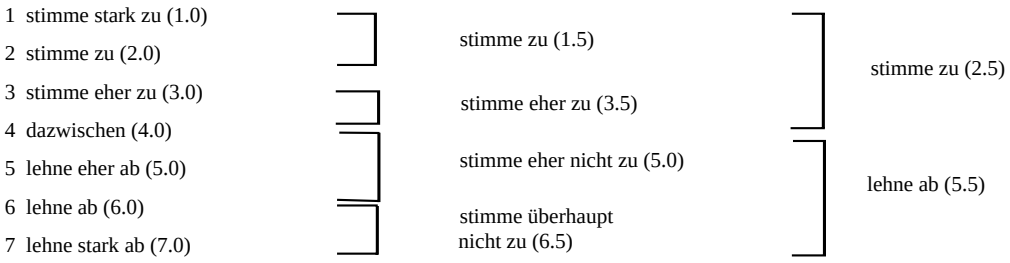
Da der Skalentyp II in der sozialwissenschaftlichen Praxis der Regelfall ist, enthält nachfolgende Tabelle die theoretischen Skalenkennwerte für eine unterschiedliche Zahl von Ausprägungen. Dabei wurde von einer Kodierung ausgegangen, die mit 1 beginnt.

Tabelle 3.2.2: Theoretische Skalenkennwerte für den Skalentyp II in Abhängigkeit von der Zahl der Ausprägungen

Zahl der Ausprägungen	theoretische Skalenkennwerte			
	Untergrenze	Obergrenze	Skalenmittelwert	Skalenstandardabweichung
	α_i	β_i	μ_i	σ_i
2	1	2	1.5	0.50
3	1	3	2.0	0.82
4	1	4	2.5	1.12
5	1	5	3.0	1.41
6	1	6	3.5	1.71
7	1	7	4.0	2.00
8	1	8	4.5	2.29
9	1	9	5.0	2.59
10	1	10	5.5	2.87

Wird beispielsweise eine fünfstufige Skala verwendet, kann durch die Transformation $z_{gi} = (x_{gi} - 3.0)/1.41$ eine theoretische z-Transformation durchgeführt werden, indem die entsprechenden Tabellenwerte verwendet werden. Die fünf Ausprägungen erhalten dadurch folgende standardisierte Skalenwerte: 1 = -1.42, 2 = -0.71, 3 = 0, 4 = 0.71, 5 = 1.42.

Ziel der theoretischen Standardisierung ist, vergleichbare Variablen zu erhalten, wobei die standardisierten Skalenwerte die ursprüngliche Bedeutung der Ausprägungen abbilden sollen. Mitunter kann dies auch durch eine einfache Umkodierung erreicht werden. So ist beispielsweise für nachfolgende Antwortskalen folgende Umkodierung denkbar, die zu einer Vergleichbarkeit führt:



Zahlenwerte in Klammern = theoretische Skalenwerte entsprechend der Bedeutung der Ausprägungen

Eine theoretische Extremwertnormalisierung würde hier zu dem unerwünschten Effekt führen, daß die Antwortkategorien "stimme zu" und "stimme stark zu" denselben Skalenwert von 0.0 erhalten würden (siehe Tabelle 3.2.3).

Tabelle 3.2.3: Konsequenzen der theoretischen Extremwertnormalisierung und z-Transformation

Variable 1		Variable 2		Variable 3	
urspr. Skala	transf. Skala	urspr. Skala	transf. Skala	urspr. Skala	transf. Skala
theoretische Extremwertnormalisierung					
1 stimme voll zu	0.00	1 stimme zu	0.00	1 stimme stark zu	0.00
2 stimme zu	0.33	2 lehne ab	1.00	2 stimme zu	0.17
3 stimme eher nicht zu	0.67			3 stimme eher zu	0.33
4 stimme überhaupt nicht zu	1.00			4 dazwischen	0.50
				5 lehne eher ab	0.67
				6 lehne ab	0.83
				7 lehne stark ab	1.00
theoretische z-Transformation					
1 stimme voll zu	-1.34	1 stimme zu	-1.00	1 stimme stark zu	-1.50
2 stimme zu	-0.45	2 lehne ab	1.00	2 stimme zu	-1.00
3 stimme eher nicht zu	0.45			3 stimme eher zu	-0.50
4 stimme überhaupt nicht zu	1.34			4 dazwischen	0.00
				5 lehne eher ab	0.50
				6 lehne ab	1.00
				7 lehne stark ab	1.50

Das Beispiel macht somit deutlich, daß *eine bestimmte theoretische Transformation nicht automatisch durchgeführt werden soll*. Vielmehr sind in jedem konkreten Anwendungsbeispiel die Konsequenzen einer Skalierung zu prüfen.

Voraussetzung für eine theoretische Transformation ist, daß die Skalenkennwerte der untersuchten Variablen definiert sind und die Variablen quantitatives Meßniveau be-

sitzen oder als "quantitativ" betrachtet werden können. Nicht alle Variablen erfüllen diese Voraussetzungen, wie man sich leicht anhand von Beispielen verdeutlichen kann:

1. Die Variable "Wirtschaftswachstum" ist zwar eine kontinuierliche Skala (Skalentyp I), die Unter- und Obergrenzen sich aber nicht definiert.
2. Das Einkommen - durch eine offene Frage erfragt - ist eine kontinuierliche Skala (Skalentyp I), die Obergrenze ist nicht definiert.
3. Die Zahl der Kinder in einem Haushalt ist eine diskrete, ganzzahlige Variable (Skalentyp II). Die Obergrenze dieser Variablen ist nicht exakt definiert.
4. Die Variable "berufliche Tätigkeit" erfüllt nicht diese Voraussetzung, da sie nominalskaliert ist.

Sind die Voraussetzungen nicht erfüllt, ist eine theoretische Transformation von Variablen nicht möglich. Liegen quantitative Variablen vor, kann eine empirische Transformation, z.B. in Form einer empirischen z-Transformation, durchgeführt werden. Diese hat *folgende Konsequenzen*:

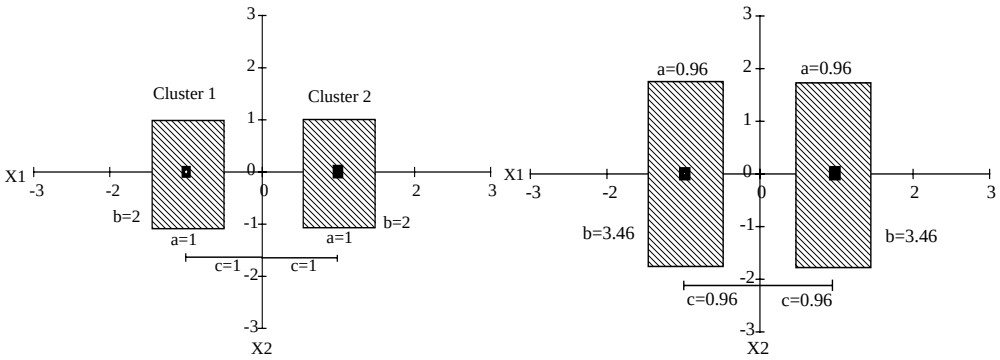
1. *Vergleiche von Objekten sind nur mehr innerhalb einer Variablen möglich*. Eine Aussage der Art "Objekt A hat in der Variablen V4 einen größeren Wert als Objekt B" ist zulässig, da der Vergleich innerhalb der Variablen V4 stattfindet. Eine Aussage der Art "Objekt A hat in der Variablen V4 einen größeren Wert als in der Variablen V5" ist dagegen nicht zulässig.
   Auf der anderen Seite sind empirisch standardisierte Werte innerhalb einer Variablen leicht zu interpretieren. Ein hoher positiver bzw. negativer Wert (z.B. von -2 bzw. +2 oder von +3 bzw. -3) bedeutet eine (relativ) hohe Abweichung vom Gesamtmittelwert.
2. *Unterschiedliche empirische Standardabweichungen werden beseitigt*. Die empirische Varianz kann man sich als Summe der Varianz zwischen den Clustern und der Varianz innerhalb der Cluster vorstellen. Die Varianz innerhalb der Cluster kann wie bei der Varianzanalyse als Fehlervarianz interpretiert werden. Eine hohe empirische Varianz kann zwei Ursachen haben:

a) Hohe Varianz zwischen den Clustern. Die Variable trennt die Cluster sehr gut. In diesem Fall ist eine Standardisierung unerwünscht, da die Variable in der Analyse ein kleineres Gewicht erhalten würde.

b) Hohe Fehlervarianz[1]. Die Variable trennt die Cluster nicht. Die Unterschiede sind rein zufällig. In diesem Fall führt eine Standardisierung zu einem positiven Effekt. Zufällige Unterschiede werden beseitigt bzw. reduziert.

Beide Effekte können wir uns anhand eines Beispiels veranschaulichen: Es sollen zwei Variablen X_1 und X_2 vorliegen. Variable X_1 trennt die Objekte in zwei Cluster, Variable X_2 ist dagegen eine "irrelevante" Variable: Ihre Varianz ist ausschließlich durch Zufallsfehler verursacht. Die Ausgangskonstellation ist in der nachfolgenden Abbildung dargestellt:

Abbildung 3.2.2: Negativer Effekt der empirischen Standardisierung auf eine vorhandene Clusterstruktur (=höhere Bewertung der zufallsbehafteten Variablen X_2)



(a) nichtstandardisierte Konfiguration

(b) standardisierte Konfiguration (die Cluster werden in der irrelevanten Variablen X2 gestreckt)

a=Breite der Rechtecke
b=Höhe der Rechtecke
c=Abweichung des Clusterzentrums vom Nullpunkt
■ =Clusterzentren

Die Objekte sollen sich in den beiden Rechtecken gleichverteilen. Die Varianzen in den beiden Variablen setzen sich dann allgemein wie folgt zusammen:

Variable	Varianz in den Clustern	Varianz zwischen den Clustern	Gesamtvarianz
X_1	$a^2/12 = 1/12$	$c^2=1$	13/12 =1.083
X_2	$b^2/12 = 4/12$	0	4/12 = 0.333
Gesamt	5/12 = 0.416	1	1.416
in %	29.6%	71.4%	100%

[1] Diese zweite Fehlerursache wurde von *Schlosser* (1976: 79) in seiner Kritik an der empirischen Standardisierung übersehen. Schlosser nimmt an, daß kleine Varianzen ein hohes Fehlerausmaß bedeuten und große Varianzen ein Hinweis auf Clusterunterschiede sind.

In dem Beispiel besitzt die Variable X_1 eine höhere Varianz, die auf Unterschiede zwischen den Clustern zurückzuführen ist. Der Anteil der Varianz in den Clustern ist insgesamt gleich 29.6 Prozent, jener zwischen den Clustern gleich 71.4 Prozent. Wird eine empirische Standardisierung durchgeführt, ergibt sich folgendes Bild:

Variable	Varianz in den Clustern	Varianz zwischen den Clustern	Gesamtvarianz
X_1	(1/12)/1.085=0.08	1/1.085=0.92	1.085/1.085=1.00
X_2	(4/12)/0.335=1.00	0/0.335=0.00	(4/12)/0.335=1.00
Gesamt	1.08	0.92	2.00
in %	54.0%	46.0%	100%

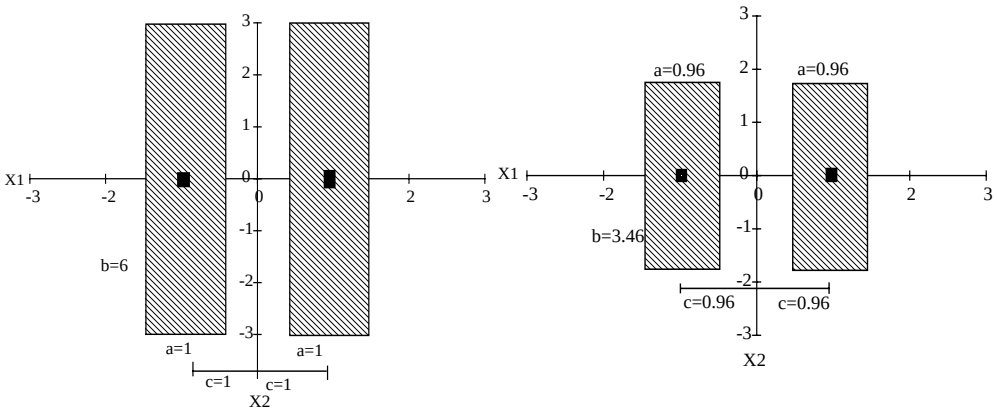
Durch die Standardisierung erhöht sich der Varianzanteil in den Clustern (=Fehlervarianz) auf 54.0 Prozent. Graphisch dargestellt ergibt sich nach der Transformation die in der Abbildung 3.2.2b dargestellte Konfiguration. Eine Standardisierung führt hier zu dem nachteiligen Effekt, daß die Cluster in der Variablen X_1 näher aneinander rücken und die Clusterunterschiede in der relevanten Variablen X_1 reduziert werden, während in der irrelevanten Variablen X_2 eine Streckung stattfindet.

Der *umgekehrte positive Effekt einer empirischen Standardisierung* läßt sich dadurch aufzeigen, daß für die ursprüngliche Datenkonstellation die Höhe der beiden Rechtecke gleich 6 ist. In diesem Fall ergeben sich folgende Varianzanteile:

Variable	Varianz in den Clustern	Varianz zwischen den Clustern	Gesamtvarianz
	vor der Standardisierung		
X_1	$a^2/12 = 1/12$	$c^2=1$	13/12 = 1.083
X_2	$b^2/12 = 36/12$	0	36/12 = 3.000
Gesamt	37/12 = 3.083	1	4.083
in %	75.5%	24.5%	100%
	nach der Standardisierung		
X_1	(1/12)/1.083=0.08	1/1.083=0.92	1.00
X_2	(36/12)/3.000=1.00	0	1.00
Gesamt	1.08	0.92	2.00
in %	54.0%	46.0%	100%

Die Standardisierung führt hier zu dem positiven Effekt, daß der Einfluß der irrelevanten Variablen, deren Varianz nur aus Fehlern besteht, reduziert wird. Graphisch dargestellt ergibt sich das in der Abbildung 3.2.3 dargestellte Bild.

Abbildung 3.2.3: Positiver Effekt der empirischen Standardisierung auf eine vorhandene Clusterstruktur (=geringere Bewertung der zufallsbehafteten Variablen X_2)



(a) nichtstandardisierte Konfiguration

(b) standardisierte Konfiguration (die Cluster werden in der irrelevanten Variablen X_2 zusammengestaucht)

Um den Einfluß der beiden dargestellten Effekte auf die Clusteranalyse zu überprüfen, wurden Simulationsrechnungen durchgeführt, wobei angenommen wurde, daß jedes Cluster der Abbildung 3.2.2a und 3.2.3a aus 50 Objekten besteht. Als Clusteranalyseverfahren wurde das Ward-Verfahren ausgewählt. Mit Ausnahme der Konfiguration 3.2.3a (irrelevante Variable mit hoher Streuung) wird die Clusterstruktur reproduziert. Die Konfiguration 3.2.3a ist durch eine irrelevante Variable mit einer hohen Fehlerstreuung gekennzeichnet. Die Standardisierung führt hier zu dem positiven Effekt, daß die vorgegebene Clusterstruktur reproduziert wird. Umgekehrt führt die Standardisierung der Konfiguration 3.2.2a nicht zu einer Zerstörung der Clusterstruktur. Die mit der Standardisierung verbundene Aufwertung der irrelevanten Variablen ist zu gering. Allgemein ist festzuhalten, daß der Effekt einer empirischen Standardisierung nicht überschätzt werden darf, dies zeigen auch die Simulationsstudien von *Milligan* (1980). In der Forschungspraxis kann man sich an folgenden Regeln orientieren:

1. Eine empirische Standardisierung muß immer dann durchgeführt werden, wenn eine theoretische Standardisierung nicht möglich ist (siehe dazu das Beispiel der Entwicklungsländerdaten).
2. Ist eine theoretische Standardisierung möglich, wird man diese durchführen, wenn größere Varianzen in einer oder mehreren Variablen ein Hinweis auf Unterschiede zwischen den Clustern (= Varianz zwischen den Clustern) sind.
3. Sind größere Varianzen dagegen auf eine höhere Fehlerstreuung (= Varianz in den Clustern) zurückzuführen, wird man sich für eine empirische Standardisierung entscheiden.
4. Zufällige Meßfehler sind eine Ursache für eine Fehlerstreuung.

4a. Messen also alle Variablen gleich gut eine Zieldimension, wird man sich für eine theoretische Standardisierung entscheiden.

4b. Wird mit mittleren Gesamtpunktwerten gerechnet, ist die Gefahr von zufälligen Meßfehlern geringer, da diese durch die Mittelwertbildung reduziert werden. Man wird sich daher für eine theoretische Standardisierung entscheiden. Besitzen die Variablen, die in die Berechnung des mittleren Gesamtpunktwertes eingehen, unterschiedliche Skaleneinheiten, wird man diese theoretisch standardisieren.

5. Ist eine eindeutige Entscheidung für eine theoretische oder empirische Standardisierung nicht möglich, wird man eine Analyse mit beiden Varianten rechnen.

6. Als ein weiteres Entscheidungskriterium können inferenzstatistische Überlegungen dienen. So z.B. setzen statistische Signifikanztests für die euklidische Distanz oder die quadrierte euklidische Distanz empirisch standardisierte, unabhängige und normalverteilte Variablen voraus (siehe Abschnitt 3.3.4). Sollen derartige Tests durchgeführt werden, wird man sich für eine empirische Standardisierung entscheiden. Dieses Kriterium sollte aber keineswegs den Vorrang vor den oben genannten Kriterien und vor inhaltlichen Überlegungen haben.

Von der Frage, ob die in einer objektorientierten Analyse einbezogenen, nichtvergleichbaren Variablen theoretisch oder empirisch zu standardisieren sind, ist die Interpretation einer Clusterlösung zu unterscheiden. Zur Interpretation wird man i.d.R. sowohl empirisch standardisierte Werte als auch die Rohwerte und die theoretisch standardisierten Werte verwenden, um die Vorteile der einzelnen Skalen bei der Interpretation zu nutzen. Diese sind:

1. Die empirisch standardisierten Variablen geben Anhaltspunkte, wie stark ein Cluster in einer Variablen vom Gesamtmittelwert abweicht.
2. Die nichtstandardisierten Rohwerte ermöglichen eine Interpretation in der ursprünglichen Skala und Vergleiche zwischen jenen Variablen, die in derselben Einheit gemessen sind.
3. Die theoretisch standardisierten Variablen schließlich ermöglichen auch Vergleiche zwischen jenen Variablen, die in unterschiedlichen Skalen gemessen wurden.

Zu Punkt 1 sei abschließend angemerkt, daß zur Beantwortung der Fragestellungen, wie stark ein Cluster in einer Variablen vom Gesamtmittelwert der Variablen abweicht, anstelle von standardisierten Variablenwerten auch folgende Teststatistik verwendet werden kann:

$$(3.2.3) \qquad \bar{z}_{ki} = \frac{\bar{x}_{ki} - \bar{x}_{i}}{\bar{s}_{ki}},$$

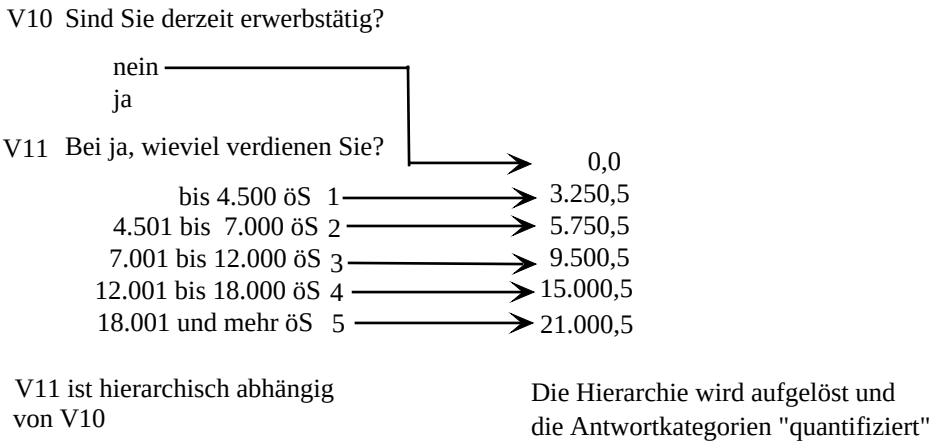
wobei $\bar{x}_{ki}$ der Mittelwert bzw. Anteilswert des Clusters k in der Variablen i ist, $\bar{s}_{ki}$ ist die Standardabweichung des Mittelwerts bzw. des Anteilswerts des Clusters k in der Variablen i und $\bar{x}_i$ schließlich ist der Gesamtmittelwert. Dieser z-Wert unterscheidet sich von dem aus den empirisch standardisierten Werten berechneten z-Wert

$$z_{ki} = \frac{\bar{x}_{ki} - \bar{x}_i}{s_i} \qquad (3.2.4)$$

dadurch, daß im Nenner nicht die Gesamtstandardabweichung s_i steht, sondern die Standardabweichung des Clusterzentrums. In die Berechnung von (3.2.3) geht neben der Clustergröße nur die tatsächliche Fehlerstreuung des Clusters ein, während die in (3.2.4) verwendete Gesamtstandardabweichung neben den Fehlerstreuungen auch die Varianz zwischen den Clustern enthält.

3.2.4 Hierarchische Variablen

In dem vorausgehenden Abschnitt wurde bereits darauf hingewiesen, daß Vergleichbarkeit von Variablen auch durch Umkodierungen erreicht werden kann. Durch Umkodierungen können die meisten hierarchischen Variablen in nicht hierarchische Variablen aufgelöst werden. Betrachten wir dazu folgendes Beispiel: In einer Untersuchung wurden die Befragten zunächst danach gefragt, ob sie erwerbstätig sind und wieviel sie gegebenenfalls verdienen. Die Variable "Einkommen" hängt also "hierarchisch" von der Variablen "Erwerbstätigkeit" ab. Aus diesen beiden Variablen läßt sich eine neue Variable erzeugen, indem Personen, die derzeit nicht erwerbstätig sind, ein Einkommen von 0 zugewiesen wird. Die anderen Einkommenskategorien können durch Bildung der Intervallmitten (z.B. 2 = (4.501+7.000)/2 = 5.750,5) quantifiziert werden. Nachfolgendes Schema verdeutlicht das Vorgehen.



3.2.5 Gemischte Variablen

Das Problem der Nichtvergleichbarkeit von gemischten Variablen kann durch folgendes Vorgehen gelöst werden:

1. Die nominalen Variablen werden in Dummies aufgelöst. Die Dummies können in der weiteren Analyse wie quantitative Variablen behandelt werden.
2. Dichotome und ordinale Variablen können wie quantitative Variablen behandelt werden (siehe dazu Abschnitt 3.3.1 und 3.3.3).
4. Die Variablen (Dummies und ordinale und quantitative Variablen) werden empirisch oder theoretisch standardisiert.
5. Für die Dummies der nominalen Variablen ist eine weitere Gewichtung erforderlich: Sie müssen mit 0.5 multipliziert werden, da ansonsten die Unterschiede in den nominalen Variablen ein doppeltes Gewicht erhalten würden. Von diesem Sachverhalt kann man sich leicht überzeugen. Betrachten wir dazu die Variable "Geschlecht". Da diese Variable dichotom ist, kann sie als quantitative Variable behandelt werden. Tut man dieses, ist die maximale Differenz für Personen mit unterschiedlichem Geschlecht gleich 1. Wird das Geschlecht dagegen als nominale Variable betrachtet und in zwei Dummies aufgelöst, nimmt die maximale Differenz einen Wert von 2 an. Durch eine Gewichtung mit 0.5 der Dummies kann dieser Effekt beseitigt werden.
6. Die Unähnlichkeit zwischen den Objekten wird durch die City-Block-Metrik oder ein anderes Distanzmaß (z.B. durch quadrierte euklidische Distanzen) gemessen.
7. Die Clusteranalyse wird durchgeführt.

Wir wollen das Vorgehen anhand eines Beispiels darstellen. Dazu sollen die Wertedaten von *Denz* (1989) verwendet werden (siehe Abschnitt 1.5). Ziel der Analyse ist eine Clusterung der befragten Schüler und Schülerinnen (n=221) aufgrund folgender sozialstruktureller Variablen:

- besuchter Schultyp (=nominalskalierte Variable) der befragten Jugendlichen mit den Ausprägungen "BHS" (=1), "AHS" (=2) und "BS" (=3)
- Geschlecht der Befragten (=dichotome Variable) mit den Ausprägungen "männlich" (=1) und "weiblich" (=2)
- Erwerbstätigkeit der Mutter (=dichotome Variable) mit den Ausprägungen "ja" (=1) und "nein" (=2)
- Berufsprestige des Vaters (=ordinale Variable) mit den Ausprägungen von 1 (=geringes Berufsprestige) bis 7 (=hohes Berufsprestige)

- abgeschlossene Schulbildung der Eltern (ordinale Variable) mit den Ausprägungen "keine abgeschlossene Schulausbildung" (=1), "Pflichtschule ohne Lehre" (=2) usw.

Wir wollen zunächst die Verfahrensschritte für zwei befragte Schüler darstellen (siehe Tabelle 3.2.4).

Tabelle 3.2.4: Datenmatrix für zwei befragte Jugendliche in den untersuchten sozialstrukturellen Variablen

	Ausgangsdatenmatrix							
	Schtyp	Geschl	BerufV	ErwM	SchV	SchM		
g	3	1	4	1	1	1		
g*	1	1	5	1	4	8		
Datenmatrix nach Dummy-Auflösung der nominalen Variablen								
	BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM	
g	0	0	1	1	4	1	1	1	
g*	1	0	0	1	5	1	4	8	
theoretische Skalenkennwerte									
μ_i	0.50	0.50	0.50	1.50	4.00	1.50	4.50	4.50	
σ_i	0.50	0.50	0.50	0.50	2.00	0.50	2.29	2.29	
standardisierte Datenmatrix (Dummies mit 0.5 multipliziert)									
	BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM	
g	-0.50	-0.50	0.50	-1.00	0.00	-1.00	-1.53	-1.53	
g*	0.50	-0.50	-0.50	-1.00	0.50	-1.00	-0.22	1.53	
Absolute Differenzen in den einzelnen Variablen									Σ
	1	0	1	0	0.50	0	1.31	3.06	6.87

Die Abkürzungen bedeuten: Schtyp=besuchter Schultyp (1=BHS, 2=AHS, 3=BS), Geschl=Geschlecht; BerufV=Berufsprestige des Vaters (1=gering bis 7=hoch), ErwM=Erwerbstätigkeit der Mutter (1=ja, 2=nein), SchV=Schulbildung des Vaters (1=gering bis 8=hoch), SchM=Schulbildung der Mutter (1=gering bis 8=hoch)

Die Tabelle 3.2.4 enthält folgende Informationen:

Ausgangsdatenmatrix: Die Person g ist männlich (Geschl=1) und besucht derzeit eine Berufsschule (Schtyp=3). Der Vater übt eine Tätigkeit mit einem mittleren Berufsprestige (BerufV=4) aus und hat eine geringe Schulbildung (SchV=1) abgeschlossen. Die Mutter ist erwerbstätig (ErwM=1) und hat ebenfalls eine geringe Schulbildung (SchM=1). Die Person g* ist ebenfalls männlich, besucht aber eine berufsbildende höhere Schule (Schtyp=1). Der Vater hat eine mittlere Schulbildung (SchV=4) abgeschlossen und übt derzeit einen Beruf mit einem mittleren Berufsprestige (BerufV=5) aus. Die Mutter hat eine hohe Schulbildung (SchM=8) und ist ebenfalls erwerbstätig (ErwM=1).

Dummy-Auflösung: Der derzeit von den Jugendlichen besuchte Schultyp ist eine nominalskalierte Variable mit drei Ausprägungen. Diese Variable wird daher für die Analyse in drei Dummies aufgelöst. Die Person g erhält in der Dummy-Variablen "BS" den Wert 1, da sie derzeit eine Berufsschule besucht. Die Person g* erhält dagegen in der Dummy-Variablen "BHS" den Wert 1.

Theoretische Standardisierung: Die theoretischen Skalenkennwerte können der Tabelle 3.2.2 entnommen werden und sind in der Tabelle 3.2.4 wiedergegeben: Die dichotomen Variablen "Geschlecht" und "Erwerbstätigkeit der Mutter" haben zwei Ausprägungen. Ihr theoretischer Skalenmittelwert ist daher gleich 1.5 und ihre theoretische Standardabweichung gleich 0.5 (siehe Tabelle 3.2.4). Das Berufsprestige des Vaters ist eine ordinale Variable mit Werten von 1 bis 7. Für eine siebenstufige Skala ist entsprechend Tabelle 3.2.2 der theoretische Skalenmittelwert gleich 4.0 und die theoretische Standardabweichung gleich 2.0. Die "Variablen Schulbildung des Vaters" und "Schulbildung der Mutter" besitzen acht Ausprägungen. Die theoretische Skalenmittelwerte sind daher gleich 4.5 und die entsprechenden Standardabweichungen gleich 2.29. Die Dummy-Variablen sind ebenfalls dichotome Variablen. Ihre theoretische Standardabweichung ist daher gleich 0.5. Da sie nicht mit 1 und 2 sondern mit 0 und 1 kodiert sind, ist der theoretische Skalenmittelwert gleich 0.5.

Standardisierung: Unter Verwendung der theoretischen Skalenkennwerte werden die Ausprägungen der Personen entsprechend der Gleichung (3.2.1) standardisiert. Für die Person g ergibt sich in der Dummy-Variablen BHS eine Wert von -1.0 (=(0-0.5)/0.5). Dieser wird aus den genannten Gründen für die Analyse mit 0.5 multipliziert.

Berechnung der City-Block-Metrik: In den einzelnen Variablen werden die absoluten Abweichungen für die Personen g und g* berechnet und summiert. Dieser Vorgang ergibt die City-Block-Metrik, die einen Wert von 6.87 hat.

Wenn wir das hier dargestellte Vorgehen für alle Befragten anwenden und die berechnete Distanzmatrix mit dem Weighted-Average-Linkage untersuchen, ergibt sich das Verschmelzungsschema der Tabelle 3.2.5.

In dem Verschmelzungsniveau wurden nur die letzten 30 Schritte eingetragen. Eine deutliche Zunahme im Verschmelzungsniveau tritt an mehreren Stellen auf: beim Übergang von 23 zu 22 Clustern, beim Übergang von 15 zu 14 Clustern usw. In einem ersten Interpretationszugang haben wir uns für die 8-Clusterlösung entschieden, da auch beim Übergang von 8 zu 7 Clustern ein deutlicher Zuwachs im Verschmelzungsniveau vorliegt. Die Ergebnisse sind in der Tabelle 3.2.6 dargestellt.

In dem unteren Teil der Tabelle 3.2.6 wurden die nach Gleichung (3.2.3) berechneten z-Werte eingetragen, wobei für Werte größer 3 ein "+" und für Werte kleiner -3 ein "-" eingetragen wurde. Ein "+" bedeutet somit, daß der entsprechende Clustermittelwert

deutlich über dem Gesamtmittelwert liegt, ein "-", daß er deutlich unter dem Gesamtmittelwert liegt.

Tabelle 3.2.5: Verschmelzungsschema der Clusteranalyse der sozialstrukturellen Merkmale für den Weighted-Average-Linkage unter Verwendung der City-Block-Metrik (die letzten 30 Schritte)

Clusterzahl	Verschm. Niveau	Zunahme	Clusterzahl	Verschm. Niveau	Zunahme
30	2.300	0.014	15	3.223	0.005
29	2.344	0.044	14	3.628	**0.405**
28	2.397	0.053	13	3.633	0.005
27	2.454	0.056	12	3.675	0.043
26	2.559	0.105	11	3.937	**0.262**
25	2.567	0.008	10	4.002	0.064
24	2.576	0.009	9	4.174	0.173
23	2.592	0.016	8	4.319	0.145
22	2.836	**0.244**	7	4.673	**0.353**
21	2.891	0.055	6	4.717	0.045
20	2.955	0.064	5	4.943	**0.226**
19	3.089	0.134	4	5.337	**0.394**
18	3.118	0.029	3	5.446	0.109
17	3.143	0.025	2	5.849	**0.403**
16	3.218	0.075	1	6.675	**0.825**

Fettgedruckte Werte = deutliche Zunahme des Verschmelzungsniveaus.

Tabelle 3.2.6: Clusterzentren der 8-Clusterlösung der Tabelle 3.2.5

		männl. Jugdl.	erwerb. Mütter	Beruf-Vater	SchV	SchM	BHS	AHS	BS
	n=	Mittelwerte bzw. Anteilswerte							
C1	34	1.00	0.97	4.41	5.74	5.76	0.47	0.53	0.00
C2	63	1.00	0.14	3.32	3.90	3.35	0.40	0.21	0.39
C3	6	0.33	1.00	2.67	2.33	2.33	1.00	0.00	0.00
C4	2	1.00	1.00	3.50	7.50	7.50	0.00	0.00	1.00
C5	67	0.00	0.48	3.09	3.91	3.21	0.37	0.21	0.42
C6	21	0.00	0.67	4.80	6.20	5.52	1.00	0.00	0.00
C7	10	0.00	1.00	4.90	6.60	5.80	0.00	1.00	0.00
C8	13	0.46	0.00	5.55	7.42	4.92	0.00	1.00	0.00
		z-Werte (+=größer 3, -=kleiner-3)							
C1	34	+	+	+	+	+		+	-
C2	63	+	-		-	-			
C3	6		+	-	-	-	+	-	-
C4	2	+	+		+	+	-	-	+
C5	67	-		-	-	-			
C6	21	-	+	+	+	+	+	-	-
C7	10	-	+	+	+	+	-	+	-
C8	13		-	+	+		-	+	-

Die gefundene Clusterlösung läßt sich wie folgt interpretieren: Es liegen zwei sehr große Cluster (C2 und C5) vor, die durch eine mittlere Bildungs- und Berufsherkunft gekennzeichnet sind. Der Mittelwert im Berufsprestige des Vaters beträgt 3.32 bzw. 3.09 auf der siebenstufigen Prestigeskala. Die Mittelwerte in der abgeschlossenen Schulbildung der Eltern variieren zwischen 3.21 und 3.91. Bezogen auf die achtstufige Skala (1=geringe Schulbildung, 8=hohe Schulbildung) liegt somit ebenfalls eine mittlere bis geringe Schulbildung vor. Diesen beiden Clustern gehört der Großteil der in die Analyse einbezogenen Befragten (130 von 216 = 60 Prozent) an. Die beiden Cluster unterscheiden sich nur darin, daß das Cluster C2 von männlichen Jugendlichen gebildet wird, das Cluster C5 dagegen von weiblichen. Eine BHS und BS wird zu jeweils ungefähr 40 Prozent besucht, eine AHS zu 20 Prozent. 60 Prozent der Befragten kommen somit aus der Mittelschicht und besuchen zu jeweils 40 Prozent eine BHS oder BS und zu 20 Prozent eine AHS. Diese Verteilung auf die untersuchten Schultypen entspricht jener der Gesamtpopulation. Es liegen daher keine z-Werte größer +3 oder kleiner -3 hinsichtlich des Schultyps vor.

Neben diesen beiden Clustern gibt es noch ein weiteres relativ großes Cluster C1 (n=34) von männlichen Jugendlichen, die eine Schulbildung mit Maturaabschluß (=BHS oder AHS) besuchen. Die Jugendlichen kommen aus einer mittleren bis höheren Bildungs- und Berufsschicht, ihre Mütter sind erwerbstätig. Auf Seiten der weiblichen Jugendlichen wird dieser Typus in zwei Cluster aufgespalten: in ein Cluster der BHS-Schülerinnen (=C6), wo etwa zwei Drittel der Mütter erwerbstätig sind, und in ein Cluster der AHS-Schülerinnen (=C7), wo alle Mütter erwerbstätig sind. Die verbleibenden drei Cluster sind nur mehr schwach besetzt: Das Cluster C3 ist durch BHS-Schüler und Schülerinnen aus der unteren Bildungs- und Berufsschicht gekennzeichnet. Das Cluster C4 wird von 2 Jugendlichen gebildet, die sich trotz einer sehr hohen Bildungsherkunft für eine Berufsschule entschieden haben. Das Cluster C8 schließlich wird von AHS-Schülern/innen gebildet, deren Väter aus einer sehr hohen Bildungsschicht kommen, die Mütter dagegen nur aus einer mittleren.

Zusammenfassend zeigen die Ergebnisse: In mittleren sozialen Schichten ist die Erwerbsbeteiligung der Mütter gering. Einer Berufsbildung - in Form einer BHS oder BS - wird ein größeres Gewicht beigemessen als dem Besuch einer AHS. In höheren sozialen Schichten liegt i.d.R. eine hohe Erwerbsbeteiligung der Mütter vor. Eine Ausnahme bilden Eltern, die nicht bildungshomogam geheiratet haben. Hier besitzen die Mütter eine geringere Schulbildung und sind i.d.R. nicht erwerbstätig. In diesen höheren sozialen Schichten wird bis auf wenige Ausnahmen ausschließlich eine höhere Schule besucht.

Das Vorgehen zur Behandlung von gemischten Variablen ist also denkbar einfach. Daneben wurde noch eine Reihe anderer Strategien entwickelt (siehe dazu auch *Gordon*

1981, *Opitz* 1980: 57-64 u.a.), von denen hier nur einige aus Gründen der Vollständigkeit angeführt seien:

1. Reduktion des Meßniveaus: Diese Strategie kann mitunter mit einem erheblichen Informationsverlust verbunden sein.
2. Gewichtung der Distanzen (siehe dazu Abschnitt 3.3.6).
3. Verwendung der probabilistischen Clusteranalyse: Hier tritt das Problem der Nichtvergleichbarkeit von gemischten Variablen nicht auf, da nicht mit Distanzen oder Korrelationen gerechnet wird, sondern mit Wahrscheinlichkeiten, die unabhängig vom Meßniveau auf das Intervall [0,1] normiert sind (siehe dazu Abschnitt 4.5).

3.2.6 Standardisierung von Objekten

Die bisherigen Ausführungen bezogen sich ausschließlich auf die Variablen, die in die Analyse einbezogen werden sollen. Neben einer empirischen oder theoretischen Standardisierung von Variablen kann auch eine (empirische) *Mittelwertzentrierung oder Standardisierung von Objekten* durchgeführt werden.[1] Dazu wird für jedes Objekt (=Zeile einer Datenmatrix) der Mittelwert und die Standardabweichung in den zu analysierenden Variablen berechnet mit:

(3.2.5) $$\bar{x}_g = \sum_i x_{gi} \Big/ m \ .$$

(3.2.6) $$s_g = \left[\sum_i \left(x_{gi} - \bar{x}_g \right) \Big/ m \right]^{1/2}.$$

wobei über alle Variablen m summiert wird. Der Mittelwert $\bar{x}_g$ eines Objekts g wird in der clusteranalytischen Literatur als Profilhöhe bezeichnet, die Standardabweichung s_g als Profilstreuung oder Profilstandardabweichung. Diese Namensgebung rührt daher, daß der Antwortvektor eines Objekts g als Profil dargestellt werden kann. Die nachfolgende Abbildung gibt ein Beispiel für ein Profil aus der Wertestudie von *Denz* (1989). Der Befragte hat eine starke Präferenz für Ziele der demokratischen Mitbestimmung (Variablen b, d und g). Diese werden für sehr wichtig gehalten. Als wichtig gelten noch humanistische Werte (Variablen h und k), Umweltschutz (=l), wirtschaftliche Werte (Variablen c und i) sowie Aufrechterhaltung der Ordnung (=a). Stark abgelehnt wird eine starke Landesverteidigung (=f) sowie die Bekämpfung der Inflation (=e) und von Verbrechen (=j).

Für den Befragten ergibt sich eine Profilhöhe von

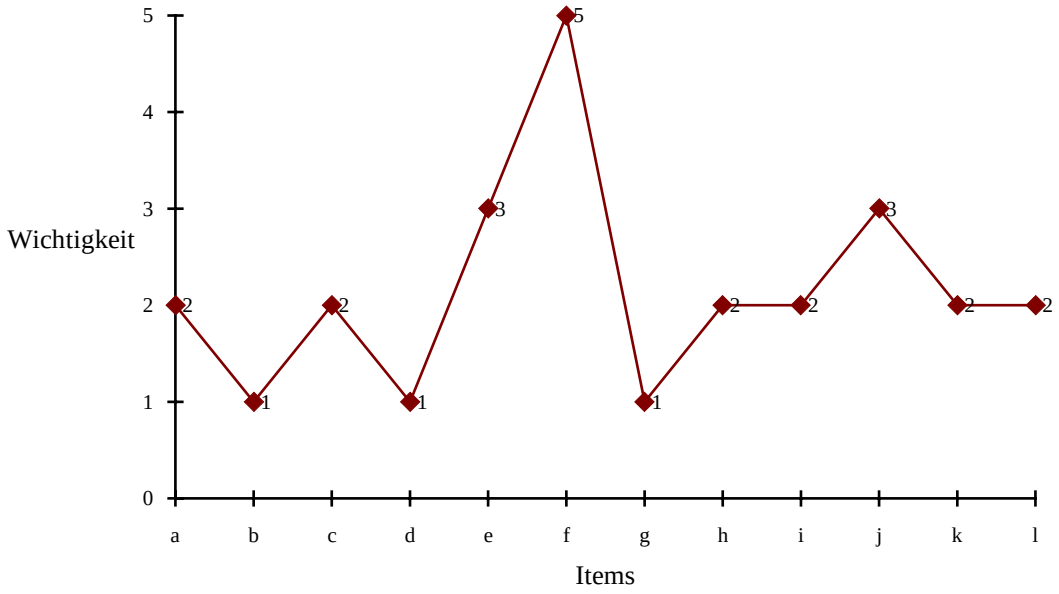
$$\bar{x}_g = \tfrac{1}{12} \cdot (2+1+2+1+\ldots+2) = 2.17$$

[1] Eine theoretische Standardisierung ist hier i.d.R. nicht sinnvoll.

und eine Profilstandardabweichung von

$$s_g = \left[\tfrac{1}{12} \cdot \left((2-2.17)^2 + \ldots + (2-2.17)^2\right)\right]^{1/2} = 1.07 .$$

Abbildung 3.2.4: Ein Beispiel für ein Antwortprofil eines Befragten g



Abkürzungen:
a=Aufrechterhaltung der Ordnung, b=mehr Mitsprache bei wichtigen Regierungsentscheidungen, c=Kampf gegen steigende Preise, d=Schutz der freien Meinungsäußerung, e=Erhaltung des wirtschaftlichen Wachstums, f=eine starke Landesverteidigung, g=verstärktes Mitspracherecht am Arbeitsplatz, h=menschengerechte Städte, i=stabile Wirtschaft, j=Kampf gegen das Verbrechen, h=eine humane und weniger unpersönliche Gesellschaft, l=Erhaltung der Natur.
1=sehr wichtig, 2=wichtig, 3=weniger wichtig, 4=eher unwichtig, 5=vollkommen unwichtig.

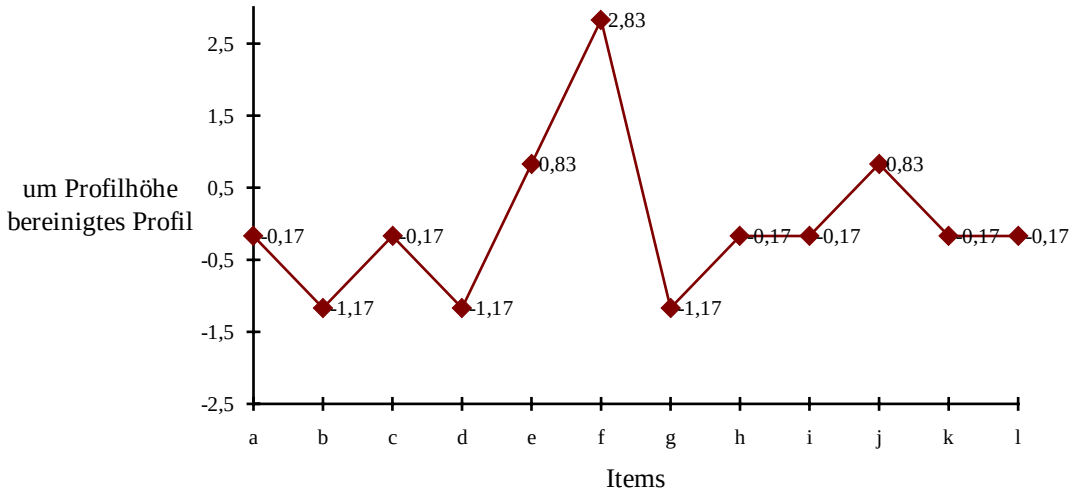
Mit diesen Kennwerten können folgende Transformationen durchgeführt werden:

1. Mittelwertzentrierung: Diese ist definiert als:

(3.2.7) $$x'_{gi} = x_{gi} - \overline{x}_g .$$

Aus dem Profil jeder Person g bzw. allgemein des Objekts g wird die Profilhöhe herausgenommen. Diese Transformation führt dazu, daß alle Personen dieselbe Profilhöhe besitzen. Inhaltlich bedeutet sie, daß nicht mehr die absolute Bewertung jedes Items untersucht wird, sondern die Präferenzen jeder Person. Für das in der Abbildung 3.2.4 dargestellte Profil ergibt sich folgendes mittelwertzentrierte Profil:

Abbildung 3.2.4a: Mittelwertzentriertes Profil der Person g der Abbildung 3.2.4



2. Standardisierung: Diese ist definiert mit:

$$x''_{gi} = \frac{x_{gi} - \overline{x}_g}{s_g} = \frac{x'_{gi}}{s_g}. \qquad (3.2.8)$$

Aus dem Profil jeder Person g wird die Profilhöhe und die Profilstreuung herausgenommen. Alle Profile erhalten dieselbe Profilhöhe von 0 und die Profilstandardabweichung von 1. Ist die Profilstreuung gleich Null, sind die standardisierten Profilwerte gleich Null. In unserem Beispiel entspricht das standardisierte Profil weitgehend jenem des mittelwertzentrierten Profils, da die Standardabweichung nahe bei 1 liegt. Das Profil wird in der Vertikalen nur geringfügig zusammengestaucht. Aus dem Wert 2.83 wird der Wert 2.64 (=2.83/1.07) und aus dem Wert -1.17 der Wert -1.09 (=-1.17/1.07).

Die Standardisierung der Profile führt allgemein dazu, daß nur mehr relative Präferenzen abgebildet werden. In unserem Beispiel würden sich u.a. für die in der Tabelle 3.2.7 dargestellten Rohprofile dieselben standardisierten Profile ergeben, wenn wir zur besseren Veranschaulichung nur vier Items herausgreifen.

Tabelle 3.2.7: Rohprofile, die dasselbe standardisierte Profil besitzen

Person	Rohprofil				Profil-höhe	Profil-stand.	stand. Profil			
	a)	b)	c)	d)	$\overline{x}_g$	s_g	a)	b)	c)	d)
1	1	1	2	2	1.5	1.5	-1	1	1	-1
2	1	3	3	1	2.0	1.0	-1	1	1	-1
3	1	4	4	1	2.5	1.5	-1	1	1	-1
4	1	3	5	1	3.0	2.0	-1	1	1	-1
5	2	3	3	2	2.5	0.5	-1	1	1	-1
6	4	5	5	4	4.5	0.5	-1	1	1	-1

Alle in der Tabelle angeführten Personen besitzen dasselbe standardisierte Profil. Es werden nur mehr relative Präferenzen abgebildet. Die (absolute) Wichtigkeit der einzelnen Items spielt keine Rolle mehr. So besitzen die Person 1 und 6 dasselbe standardisierte Profil, obwohl sich ihre Antwortmuster (=Rohprofile) deutlich unterscheiden. Auch die Stärke der Präferenz spielt keine Rolle mehr. So z.B. hat Person 4 eine starke Präferenz für die Items a) und d), da die Items b) und c) deutlich abgelehnt werden, die Person 1 dagegen nur eine sehr schwache Präferenz.

Zusammenfassend hat eine Mittelwertzentrierung oder Standardisierung der Objekte somit folgende Effekte:

1. Die Mittelwertzentrierung führt dazu, daß nur mehr Präferenzen interpretierbar sind. Aussagen über die Einstellung (Bewertung) einzelner Items sind nicht mehr möglich.
2. Die Standardisierung führt darüber hinaus dazu, daß auch die Stärke der Präferenz eliminiert wird.

Ob in einer Analyse einer der beiden Effekte erwünscht ist, hängt von der inhaltlichen Fragestellung ab. I.d.R. sind *beide Effekte nicht erwünscht*, so daß keine Standardisierung der Objekte durchgeführt wird. *Bezüglich der Auswahl eines Un- oder Ähnlichkeitsmaßes ist dann darauf zu achten, daß dabei nicht implizit eine Standardisierung der Objekte durchgeführt wird.* Dies ist beispielsweise bei der Verwendung von Korrelationskoeffizienten der Fall.

3.2.7 Exkurs: Die Problematik einer automatischen Orthogonalisierung

In der Literatur (siehe z.B. *Green und Tull* 1982: 414*; Lüdtke* 1989; *Kaufman* 1985) werden die Variablen vor der Clusteranalyse oft orthogonalisiert. Dies wird damit begründet, daß das verwendete Verfahren unabhängige Variablen bzw. einen orthogonalen Merkmalsraum[1] voraussetzt. Eine automatische Orthogonalisierung ist aber nicht unproblematisch, da empirische Korrelationen ein Hinweis auf eine vorhandene Clusterstruktur sein können. Von diesem Sachverhalt kann man sich leicht überzeugen, wenn man sich folgendes Modell vorstellt:

1. Es liegen drei Cluster 1, 2 und 3 vor mit den Anteilswerten p_k $(k = 1,2,3;\ \sum p_k = 1)$.
2. Jedes Cluster k besitzt in den in die Clusteranalyse einbezogenen Variablen X_i die Clustermittelwerte $\bar{x}_{ik}$.

[1] Die Annahme eines orthogonalen Merkmalsraumes ist nur für bestimmte statistische Testverfahren erforderlich.

3. Die Variablen X_i besitzen einen Erwartungswert von 0 und eine Varianz von 1.
4. Die Ausprägungen der Objekte g eines Clusters k kommen durch folgendes Modell zustande: $x_{ig(k)} = \bar{x}_{ik} + e_{ig(k)}$, wobei $e_{ig(k)}$ ein zufälliger normalverteilter Fehlerterm mit Erwartungswert 0 und Varianzen σ^2_{ik} ist.
5. Die zufälligen Fehlerterme sind voneinander unabhängig.

Unter diesen Annahmen ist die Korrelation zwischen zwei Variablen gleich der Kovarianz, da entsprechend der Modellannahme 3 die Variablen eine Varianz von 1 besitzen:
(3.2.9)

$$KORR(X_i, X_{i^*}) = \frac{KOV(X_i, X_{i^*})}{\left[\underbrace{VAR(X_i)}_{=1.0} \cdot \underbrace{VAR(X_{i^*})}_{=1.0}\right]^{1/2}} = KOV(X_i, X_{i^*}).$$

Für eine Person g aus dem Cluster k ergibt sich wegen der Modellannahmen 2 und 4 für den Kovarianzterm folgende Darstellung:

(3.2.10)
$$\left(x_{ig(k)} - \underbrace{\bar{x}_i}_{0}\right) \cdot \left(x_{i^*g(k)} - \underbrace{\bar{x}_{i^*}}_{0}\right) = \left(\bar{x}_{ik} + e_{ig(k)}\right) \cdot \left(\bar{x}_{i^*k} + e_{i^*g(k)}\right).$$

Die Kovarianz zwischen den Variablen i und i* ist somit:
(3.2.11)

$$\begin{aligned} KOV(X_i, X_{i^*}) &= E\left(\left(\bar{x}_{ik} + e_{ig(k)}\right) \cdot \left(\bar{x}_{i^*k} + e_{i^*g(k)}\right)\right) \\ &= \sum_k p_k \cdot \left(\bar{x}_{ik} \cdot \bar{x}_{i^*k^*} + \bar{x}_{ik} \cdot E(e_{i^*g(k)}) + \bar{x}_{i^*k} \cdot E(e_{ig(k)}) + E(e_{ig(k)} \cdot e_{i^*g(k)})\right). \\ &= \sum p_k \cdot \bar{x}_{ik} \cdot \bar{x}_{i^*k} \end{aligned}$$

da entsprechend den Modellannahmen die Fehlerterme unabhängig voneinander sind und einen Erwartungswert von 0 aufweisen. *Unter den getroffenen Modellannahmen hängt die Korrelation zwischen zwei Variablen nur von den Clustermittelwerten ab.* Eine positive oder negative Korrelation ist ein Hinweis auf das Vorliegen einer Clusterstruktur auf der manifesten Ebene. Wird die Korrelation beseitigt, kann die zugrundeliegende Clusterstruktur zerstört werden.

Umgekehrt kann e*ine Korrelation von Variablen aber auch dadurch zustande kommen, daß die Variablen eine gemeinsame Zieldimension messen*. Wir können uns beispielsweise folgendes *Clustermodell auf einer latenten Ebene* vorstellen:

1. Die in die Clusteranalyse einbezogenen Variablen messen eine gemeinsame Zieldimension Y. Es gilt folgendes Meßmodell: $X_i = b_i \cdot Y + u_i$, wobei die Meßfehler u_i normalverteilte Zufallsvariablen mit Erwartungswerten 0 und Varianzen $\sigma^2_{u_i}$ sind.
2. Auf der gemeinsamen Dimension gibt es drei Cluster mit den Anteilswerten p_k (k = 1, 2, 3; $\sum p_k = 1$) und den Clusterzentren $\bar{y}_k$.
3. Die Objekte g eines Clusters k streuen zufällig um ihr Clusterzentrum. Es soll gelten: $y_{g(k)} = \bar{y}_k + e_{g(k)}$, wobei $e_{g(k)}$ normalverteilte Zufallsgrößen mit Erwartungswerten von 0 und Varianzen von σ^2_k sind.
4. Die latente Dimension Y soll einen Erwartungswert von 0 und eine Varianz von 1 besitzen.
5. Die zufälligen Meßfehler u_i und die Zufallsgrößen $e_{g(k)}$ sind paarweise unabhängig.

Aus diesen Modellannahmen ergibt sich zunächst folgende Darstellung für die Varianzen der latenten Variablen Y:

$$(3.2.12) \qquad VAR(Y) = 1 = \sum_k p_k \cdot \left(\bar{y}_k - \underbrace{\bar{y}}_{0} \right)^2 + \sum_k p_k \cdot \sigma^2_k = \sigma^2_{zw} + \sigma^2_{in},$$

wobei $\sigma^2_{zw} = \sum p_k \cdot \bar{y}^2_k$ die Varianz zwischen den Clustern und $\sigma^2_{in} = \sum p_k \cdot \sigma^2_k$ die Varianz innerhalb der Cluster ist.

Für die Varianzen der manifesten Variablen i ergibt sich folgende Darstellung:

$$(3.2.13) \qquad VAR(X_i) = b_i^2 \cdot \left(\sigma^2_{zw} + \sigma^2_{in}\right) + \sigma^2_{u_i} = b_i^2 + \sigma^2_{u_i}.$$

Die Varianz in den empirisch beobachteten Variablen setzt sich aus der Meßfehlervarianz und der Faktorladung b_i zusammen. Für die Kovarianz und die Korrelation von zwei Variablen ergibt sich schließlich folgende Darstellung:

$$(3.2.14) \qquad KOV(X_i, X_{i^*}) = b_i \cdot b_{i^*} \cdot \left(\sigma^2_{zw} + \sigma^2_{in}\right) = b_i \cdot b_{i^*}.$$

$$(3.2.15) \qquad KORR(X_i, X_{i^*}) = \frac{b_i \cdot b_{i^*}}{\sqrt{\left(b_i^2 + \sigma^2_{u_i}\right) \cdot \left(b_{i^*}^2 + \sigma^2_{u_{i^*}}\right)}}.$$

Unter *den getroffenen Modellannahmen hängt die Korrelation nur von den Meßmodellparametern (Faktorladungen und Meßfehlervarianzen), nicht aber von der Streuung innerhalb und zwischen den Clustern ab*.

Eine empirische Korrelation von 0.75 (=3/4) kann somit beispielsweise durch folgende Modelle zustandekommen:

Modell I (Clusteranalysemodell auf der manifesten Ebene ohne Meßmodell): Es liegen drei Cluster mit folgenden Kennwerten vor:

$p_1 = 1/6; \quad \bar{x}_{i1} = -3/2; \quad \bar{x}_{i*1} = -3/2;$ $\qquad p_2 = 4/6; \quad \bar{x}_{i2} = 0; \quad \bar{x}_{i*2} = 0;$

$p_3 = 1/6; \quad \bar{x}_{i3} = 3/2; \quad \bar{x}_{i*3} = 3/2;$

Unter diesen Annahmen ergibt sich eine Korrelation von:

$$KORR(X_i, X_{i*}) = \tfrac{1}{6} \cdot \left(-\tfrac{3}{2}\right)^2 + \tfrac{4}{6} \cdot (0)^2 + \tfrac{1}{6} \cdot \left(\tfrac{3}{2}\right)^2 = 3/4 = 0.75.$$

Modell II (Clusteranalysemodell auf der latenten Ebene mit Meßmodell): Die beiden Faktorladungen sind gleich 1 und die Meßfehlervarianzen gleich 1/3.

$$KORR(X_i, X_{i*}) = \frac{(b_i = 1) \cdot (b_{i*} = 1)}{\sqrt{\left((b_i^2 = 1) + (\sigma_{u_i}^2 = 1/3)\right) \cdot \left((b_{i*}^2 = 1) + (\sigma_{u_{i*}}^2 = 1/3)\right)}} = \frac{1}{4/3} = 0.75$$

.

Die Ausführungen zeigen somit, daß eine automatische Beseitigung von empirischen Korrelationen, indem die Variablen z.B. orthogonalisiert werden, problematisch ist. Eine *inhaltlich und empirisch begründete Entscheidung ist erforderlich, ob empirische Korrelationen ein Hinweis auf ein Clustermodell auf einer latenten Ebene oder auf ein Clustermodell auf einer manifesten Ebene sind*. Empirisch begründet werden kann eine Entscheidung durch eine Faktorenanalyse. In der weiteren Analyse werden dann aber nicht die unrotierten Faktoren (=Hauptkomponenten) verwendet, sondern die inhaltlich interpretierten (rotierten) Faktoren.

Hinzu kommt, daß ein automatisches Vorgehen der oben dargestellten Art - im Unterschied zu einer dimensionalen Analyse durch eine Faktorenanalyse[1] - kaum zur Elimination von Fehlerquellen beiträgt. *Kaufman* (1985) untersuchte in Simulationsstudien, ob durch eine Orthogonalisierung Meßfehler beseitigt werden können. Er vergleicht folgende Vorgehensweisen:

1. Verwendung der empirisch standardisierten Werte
2. Verwendung der bedeutsamen ungewichteten Hauptkomponenten[2] (Hauptkomponenten mit Eigenwerten größer 1)
3. Verwendung der bedeutsamen gewichteten Hauptkomponenten
4. Verwendung aller ungewichteten Hauptkomponenten
5. Verwendung aller gewichteten Hauptkomponenten

---

[1] Mit Faktorenanalyse ist hier gemeint, daß die berechneten Dimensionen inhaltlich interpretiert werden, z.B. indem die bedeutsamen Faktoren (Hauptkomponenten bzw. Faktoren mit Eigenwerten größer 1) rotiert werden (siehe Abschnitt 2.4.3.2).

[2] Dies sind die unrotierten Faktorladungen, wenn keine Kommunalitätenschätzung durchgeführt wird.

Die Ergebnisse zeigen, daß bei der Verwendung aller ungewichteten Hauptkomponenten wesentlich schlechtere Ergebnisse erzielt werden als mit den anderen vier Methoden. Der Anteil der Fehlklassifikationen betrug für die untersuchten Modellkonstellationen (im Durchschnitt 5 Prozent Meßfehler und durchschnittlich 5 Prozent fehlende Werten) 45.87 Prozent, wenn alle ungewichteten Hauptkomponenten verwendet werden. Bei den anderen Methoden variiert die Fehlerquote zwischen 2.00 Prozent (alle gewichteten Hauptkomponenten) und 5.65 Prozent (bedeutsame ungewichtete Hauptkomponenten). Bei einer Analyse mit standardisierten Werten werden im Vergleich zur Verwendung von gewichteten Hauptkomponenten (Methode 3 und 5) nur geringfügig schlechtere Ergebnisse erzielt. Der Fehlklassifikationsanteil betrug 2.65 Prozent. Clusteranalyseprogramme, die automatisch eine Orthogonalisierung durchführen sind daher mit äußerster Vorsicht anzuwenden.

Von diesem Sachverhalt werden wir uns nachfolgend überzeugen, zunächst soll aber geklärt werden, auf was sich die Forderung der Unabhängigkeit von Variablenkann man sich leicht überzeugen, wenn man sich folgendes Modell vorstellt:

Theoretische Unabhängigkeit der Variablen: Mit theoretischer Unabhängigkeit ist gemeint, daß zwei oder mehrere Variablen unabhängig voneinander erhoben oder berechnet wurden. Theoretische Unabhängigkeit ist immer dann gegeben, wenn zwei oder mehrere Fragen unabhängig voneinander beantwortet werden können, wie dies beispielsweise bei Fragebatterien der Fall ist. Werden dagegen Befragte ersucht, eine Menge von Items in eine Rangreihe zu bringen, liegt theoretische Abhängigkeit vor, da der Rangplatz, der einem Item zugewiesen wird, nicht einem anderen Item zugewiesen wird. Die Abbildung 2.2.1 verdeutlicht diesen Unterschied anhand einer Kurzform der Materialismus-Postmaterialismusskala von Inglehart. Werden die Items in Form einer Fragebatterie präsentiert, bei der jedes Items isoliert mit "sehr wichtig" bis "vollkommen unwichtig" bewertet werden kann, liegt theoretische Unabhängigkeit vor. Wird der Befragte dagegen aufgefordert, die vier Items in eine Rangreihe zu bringen, liegt theoretische Abhängigkeit vor. Wird das Item A beispielsweise an erster Stelle gereiht, können die anderen Items nicht mehr an erster Stelle genannt werden.

Abbildung 3.2.5: Beispiele für theoretisch abhängige und unabhängige Variablen

Erhebungsdesign, das zu theoretisch unabhängigen Variablen führt

Nachfolgend sind vier Ziele angeführt, denen heute in der Politik unterschiedliches Gewicht beigemessen wird. Wie wichtig sind nachfolgende Ziele für Sie?

	sehr wichtig	wichtig	wenig wichtig	eher unwichtig	vollk. unwichtig	Variable
Verstärktes Mitsprache bei wichtigen Regierungsentscheidungen	1	2	3	4	5	MitReg
Eine humane und weniger unpersönliche Gesellschaft	1	2	3	4	5	humGesell
Erhaltung des wirtschaftliches Wachstum	1	2	3	4	5	wirtWachst
Aufrechterhaltung der Ruhe und Ordnung im Staat	1	2	3	4	5	OrdnSich

Erhebungsdesign, das zu theoretisch abhängigen Variablen führt.

Nachfolgend sind vier Ziele angeführt, denen heute in der Politik unterschiedliches Gewicht beigemessen wird. Reihen Sie diese bitte die Ziele nach der Wichtigkeit, die diese für Sie persönlich haben! Rangreihe

		Variable
A	Verstärktes Mitsprache bei wichtigen Regierungsentscheidungen	RangMitSpr
B	Eine humane und weniger unpersönliche Gesellschaft	RangHumGesell
C	Erhaltung des wirtschaftliches Wachstum	RangWirtWachst
D	Aufrechterhaltung der Ruhe und Ordnung im Staat	RangSich

am wichtigsten:.....
am Zweitwichtigsten.....
am Drittwichtigsten.....
am Viertwichtigsten.....

Die theoretische Un- bzw. Abhängigkeit hängt somit von dem Erhebungsdesign bzw. der Berechnungsmethode von Variablen ab. Der Anwender muß entscheiden, ob er aus der theoretischen Abhängigkeit eine Nichtvergleichbarkeit ableitet. Formal kann dies damit begründet werden, daß die Signifikanzprüfung der Teststatistiken, wie sie im Rahmen einer Clusteranalyse verwendet werden, theoretische Unabhängigkeit der Variablen voraussetzt. Ob man sich für eine Beseitigung der theoretischen Unabhängigkeit entscheidet, hängt von dem Stellenwert von Signifikanztests im Vergleich zu anderen inhaltlichen Überlegungen und Interpretationsfragen ab. Kommt der Signifikanzprüfung ein zentraler Stellenwert zu, wird man die theoretische Abhängigkeit beseitigen.

Empirische Korreliertheit von Variablen: Die theoretische Unabhängigkeit wird in der Literatur oft mit der empirischen Korreliertheit von Variable