

Können Mischverteilungsmodelle das Problem heterogener Daten lösen?

Kühnel, Steffen M.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:

Kühnel, S. M. (1999). Können Mischverteilungsmodelle das Problem heterogener Daten lösen? *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 45, 44-70. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-199459>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Können Mischverteilungsmodelle das Problem heterogener Daten lösen?

von Steffen Kühnel¹

Zusammenfassung

*Datenheterogenität liegt vor, wenn die Untersuchungseinheiten in einer Stichprobe nicht als Realisationen aus einer gemeinsamen Verteilung aufgefaßt werden können. Wird Heterogenität ignoriert, besteht die Gefahr von Fehlschlüssen. Die Berücksichtigung von Heterogenität durch die Spezifikation zusätzlicher exogener Variablen oder durch Gruppenvergleiche setzt voraus, daß Informationen darüber vorliegen, wie sich die Untersuchungseinheiten zu homogenen Subgruppen zusammenfassen lassen. In Mischverteilungsmodellen ist diese Kenntnis dagegen nicht nötig. Zwar wird auch hier vorausgesetzt, daß sich eine Population aus Subpopulationen zusammensetzt. Die Zugehörigkeit der Fälle einer Stichprobe zu diesen als latente Klassen bezeichneten Subpopulationen muß aber nicht bekannt sein. Ein Nachteil dieser größeren Flexibilität von Mischverteilungsmodellen besteht darin, daß es schwierig sein kann, eine eindeutige Lösung zu finden. Trotz dieser Schwierigkeit sind Mischverteilungsmodelle eine fruchtbare Ergänzung des Angebots an statistischen Analysemodellen. Dies wird an einem Anwendungsbeispiel demonstriert, bei dem anhand von Daten der ISSP-Umfrage 1993 der Zusammenhang zwischen der wahrgenommenen Umweltgefährdung durch Autos und der Befürwortung von politischen Maßnahmen zur Verkehrsreduktion untersucht wird. Die Schätzungen von Mischverteilungsmodellen mit dem Programm Mplus von **Muthén** führen hier zu einer Lösung mit zwei latenten Klassen, die sich in der Bewertung der Umweltgefährdung und der Zustimmung zu Maßnahmen zur Verkehrsreduktion deutlich unterscheiden.*

Abstract

Data are heterogeneous if units are not sampled from a unique population with common probability distribution. As a consequence inferences may become invalid. Heterogeneity can be controlled by specification of additional exogenous variables or simultaneous

¹ Dr. **Steffen Kühnel** ist Professor für empirische Sozialforschung am Institut für Politikwissenschaft an der Justus-Liebig-Universität Gießen, Karl-Glöckner-Str. 21 Haus E, 35394 Gießen.

analysis of subgroups. In both approaches, a heterogeneous sample is divided in homogeneous subgroups. Similarly, in latent variable mixture models the total population is divided in subpopulations that are called latent classes. But in contrast to the classical method of group comparisons, it is not necessary to know the membership of a case to a subpopulation. Therefore, latent variable mixture models are more flexible. On the other hand, it becomes more difficult to estimate a mixture model. In this paper, the application of latent variable mixture models and problems of application are demonstrated using data from the ISSP survey 1993. In this example the relation is investigated between perception of environmental risks caused by car traffic and support of policies to reduce car traffic. The estimation of a mixed model using Muthén's program Mplus results in a solution with two latent classes with different views on environmental risks and policies.

Die in der Sozialforschung verwendeten Analysemodelle unterstellen in der Regel, daß die Stichprobe homogene Daten aufweist. Verletzungen dieser Annahme, also heterogene Daten, können zu Fehlschlüssen bei der Dateninterpretation führen. Die Anwendung von Mischverteilungsmodellen verspricht eine Reduzierung dieser Gefahr, geht diese Modellklasse doch explizit auf die Möglichkeit heterogener Datenstrukturen ein. Welche Möglichkeiten, aber auch Probleme die Anwendung von Mischverteilungsmodellen hat, wird in diesem Beitrag diskutiert. Ausgangspunkt ist ein empirisches Beispiel aus der Umweltsoziologie, das im ersten Abschnitt näher vorgestellt wird. Anhand dieses Beispiels werden die Annahme der Datenhomogenität und die möglichen Konsequenzen einer Verletzung dieser Annahme erörtert. Nach einer kurzen Diskussion der bislang üblichen Strategien zur Kontrolle von Datenheterogenität bei der Anwendung von Strukturgleichungsmodellen wird die Logik der Mischverteilungsmodelle vorgestellt. Dabei werden mit dem Programm *Mplus* Mischverteilungsmodelle für das Anwendungsbeispiel berechnet. Es folgt ein Abschnitt über Anwendungsprobleme. Im vorletzten Abschnitt wird die Erweiterung des Modells um Indikatoren für die latenten Klassen vorgestellt. Der Beitrag endet mit einer Diskussion, in der auch Vorschläge für die Vorgehensweise bei der Anwendung von Mischverteilungsmodellen formuliert werden.

1 Inhaltliche Fragestellung: Folgt aus der wahrgenommenen Umweltgefährdung durch Autos die Befürwortung politischer Maßnahmen zur Verkehrsreduktion?

Empirische Untersuchungen haben gezeigt, daß hohes Umweltbewußtsein nicht notwendigerweise in umweltgerechteres Verhalten umgesetzt wird (*Diekmann/Preisendörfer*, 1992; *Franzen*, 1997). Mit einem stärkeren Zusammenhang ist möglicherweise eher zu rechnen, wenn nicht direkt das eigene Verhalten, sondern die Zustimmung zu politischen Maßnahmen angesprochen wird. Ob diese Vermutung zutrifft, soll mit Umfragedaten untersucht werden. 1993 wurde im Rahmen der jährlich in einer Reihe von Staaten durchgeführten Umfrage des „International Social Survey Program“ (ISSP) das Thema Umwelt behandelt.

Tabelle 1: Beurteilung von Autos als Umwelt-Risiko**Umweltgefährdung durch Autos:**

Ganz allgemein betrachtet, ist die durch Autos verursachte Luftverschmutzung Ihrer Ansicht nach ...	Nennungen
1 ... äußerst gefährlich für die Umwelt,	21.8%
2 ... sehr gefährlich für die Umwelt,	42.0%
3 ... etwas gefährlich für die Umwelt,	29.4%
4 ... kaum gefährlich für die Umwelt,	5.5%
5 ... oder überhaupt nicht gefährlich für die Umwelt?	1.4%
	(1001)

Mehr Krankheiten durch Autos

Für wie wahrscheinlich halten Sie es, daß in Deutschland in den nächsten zehn Jahren gesundheitliche Schädigungen bei Stadtbewohnern durch die von Autos verursachte Luftverschmutzung stark zunehmen werden? Ich meine ...

1 ... das wird mit Sicherheit passieren,	25.3%
2 ... das ist sehr wahrscheinlich,	39.5%
3 ... das ist eher wahrscheinlich,	22.3%
4 ... das ist kaum wahrscheinliche,	10.6%
5 ... das wird mit Sicherheit nicht geschehen.	2.2%
	(979)

(Quelle: Westdeutsche Befragte der ISSP-Befragung 1993)

Die Daten dieser ISSP-Studie werden im Zentralarchiv unter der ZA-Studien-Nr. 2450 für Sekundäranalysen bereitgestellt.

In Tabelle 1 ist die Antwortverteilung von zwei Fragen zur Umweltgefährdung durch Autos für die Befragten der westdeutschen Teilstichprobe der ISSP-Umfrage wiedergegeben.² Die Antworten weisen darauf hin, daß die Mehrheit der Befragten der Auffassung ist, daß der Autoverkehr die Umwelt und die Gesundheit von Stadtbewohnern gefährdet. Bedeutet dies auch, daß die Befragten für Maßnahmen zur Reduktion des Umweltverkehrs eintreten? In der Umfrage finden sich auch Fragen, die eine Zustimmung für unterschiedliche verkehrspolitische Maßnahmen zur Reduktion des Autoverkehrs erfassen. Frageformulierungen und Antworten hierzu sind in Tabelle 2 wiedergegeben. Solange es um die Finanzierung von Kampagnen geht, die zu einer freiwilligen Reduktion der Autonutzung aufrufen, ist die Mehrheit dafür. Anders sieht es jedoch aus, wenn die Erhöhung der Benzinsteuern oder die Einführung einer Umweltsteuer für Privatautos gefordert wird. Zwar stimmen jeweils etwa 40% einer solchen Maßnahme eher oder stark zu. Bei der Einführung der Umweltsteuer für private Autos ist aber andererseits die Gruppe derjenigen fast genau so groß, die dies eher oder stark ablehnt. Bei der Erhöhung der Benzinsteuern überwiegt die Zahl der ablehnenden Urteile sogar die Zahl der Zustimmungen.

² Da es denkbar ist, daß sich Personen aus den neuen und den alten Bundesländern bei den Einstellungen zur Umwelt unterscheiden, konzentriert sich die Analyse in diesem Beitrag auf die 1014 Befragten aus den alten Bundesländern.

Tabelle 2: Bewertung verkehrspolitischer Maßnahmen zur Verkehrsreduktion

Zum Schutz der Umwelt sollte die Bundesregierung ...

		Kampagnen zur Verkehrsreduktion	Erhöhung der Benzinsteuer	Steuer für private Pkws
		... öffentliche Gelder für Informationskampagnen ausgeben, um die Leute davon zu überzeugen, weniger mit dem Auto zu fahren	... die Steuern für Treibstoffe während der nächsten 10 Jahre in jedem Jahr erhöhen, um die Leute dazu zu bewegen, weniger mit dem Auto zu fahren	... eine besondere Umweltsteuer auf Privatautos erheben und das Geld für die Verbesserung der öffentlichen Verkehrsmittel verwenden
1	stimme stark zu	19.7%	14.1%	15.3%
2	stimme eher zu	37.9%	25.3%	27.2%
3	weder noch	14.8%	15.4%	15.5%
4	stimme eher nicht zu	15.8%	24.3%	18.9%
5	stimme überhaupt nicht zu	11.8%	20.9%	23.0%
		(949)	(963)	(959)

(Quelle: Westdeutsche Befragte der ISSP-Befragung 1993)

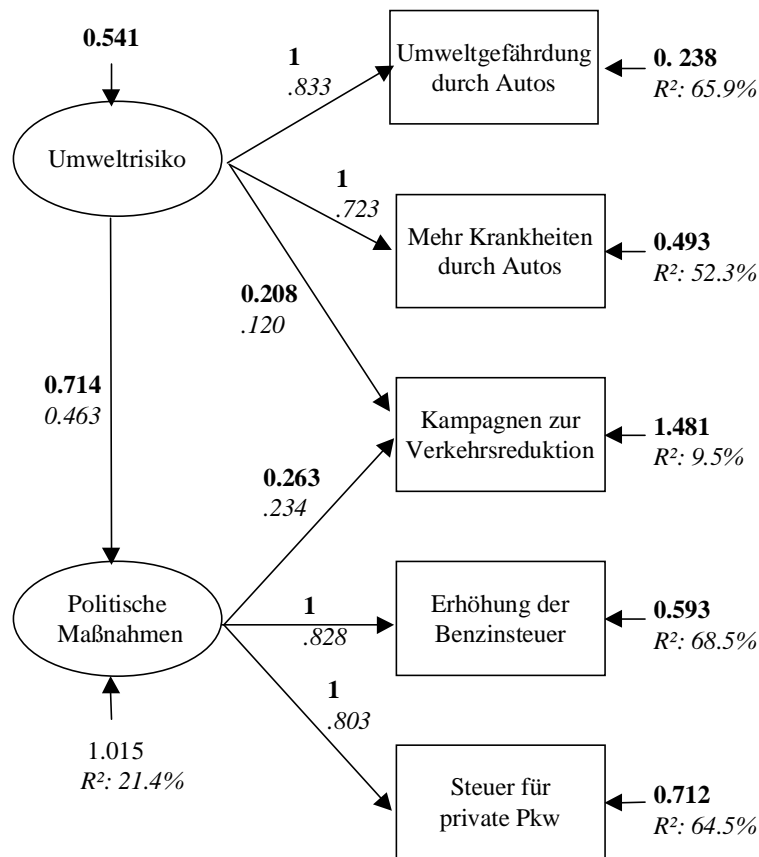
Es liegt nahe, zwischen den in Tabelle 1 und Tabelle 2 vorgestellten Variablen eine positive Beziehung anzunehmen. Danach sollte das Ausmaß der Zustimmung zu den verkehrspolitischen Maßnahmen mit der Höhe der vermuteten Umweltbelastung durch den Autoverkehr ansteigen. Tatsächlich zeigen sich positive Korrelationen zwischen den Variablen. Die in Tabelle 3 wiedergegebenen Werte weisen dabei ein klares Muster auf. Erwartungsgemäß besteht eine hohe Korrelation zwischen den beiden Fragen zur Umweltgefährdung durch Autos. Noch etwas höher ist die Korrelation zwischen den beiden Bewertungen verkehrspolitischer Maßnahmen, die durch Benzin- oder Pkw-Steuer die Autonutzung reduzieren wollen. Die dritte Maßnahme, die über Kampagnen zur Verkehrsreduktion beitragen soll, weist mittlere Korrelationen mit den übrigen Variablen auf, wobei die Korrelationen mit den anderen Maßnahmen höher sind als die mit den Umweltgefährdungen.

Tabelle 3: Korrelationen, Mittelwerte und Standardabweichungen

Korrelationen	Umweltgefährdung durch Autos	Mehr Krankheiten durch Autos	Kampagnen zur Verkehrsreduktion	Erhöhung der Benzinsteuer	Steuer auf private Pkws
Umweltgefährdung	1.000				
Mehr Krankheiten	0.603	1.000			
Kampagnen	0.181	0.182	1.000		
Benzinsteuer	0.325	0.257	0.253	1.000	
Pkw-Steuer	0.321	0.253	0.216	0.664	1.000
Standardabweichung	0.886	1.015	1.279	1.375	1.411
Mittelwert	2.203	2.241	2.618	3.115	3.047

(Quelle: Westdeutsche Befragte der ISSP-Befragung 1993, n=894)

Abbildung 1: Einfluß der Risikobewertung auf die Akzeptanz verkehrspolitischer Maßnahmen



LR-Modellfit: $\chi^2=3.911$, $df=5$, $Prob.=0.5620$ ($n=894$)

(ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*)

Dieses Korrelationsmuster legt es nahe, ein Strukturgleichungsmodell zu formulieren, bei dem die beiden Umweltgefährdungen einen Faktor ‚Umweltrisiko‘ und die drei Maßnahmen einen weiteren Faktor ‚Politische Maßnahmen‘ bilden, der durch den ersten Faktor erklärt wird. Bei der Schätzung eines entsprechenden Modells weist eine nicht befriedigende Modellanpassung und ein hoher Modifikationsindex darauf hin, daß die Bewertung von Kampagnen zur Verkehrsreduktion nicht nur auf den Maßnahmenfaktor, sondern auch auf den Risikofaktor laden. Das schließlich resultierende Strukturgleichungsmodell ist in Abbildung 1 wiedergeben. Mit einem Chi-Quadratwert von nur 3.911 bei fünf Freiheitsgraden ($p=0.562$) weist das Modell eine ausgezeichnete Übereinstimmung zwischen spezifizierter Struktur und empirischen Daten auf.

Nach diesem Modell besteht ein mäßiger Zusammenhang zwischen der Einschätzung des Umweltrisikos und der Zustimmung zu verkehrspolitischen Maßnahmen: Die Unterschiede bei der Wahrnehmung des Umweltrisikos erklären gut 20% der Unterschiede bei den verkehrspolitischen Maßnahmen, was einer Korrelation von 0.46 entspricht. Mit Ausnahme des Indikators ‚Kampagnen‘ weisen alle Indikatoren hohe standardisierte Ladungen zwischen 0.723 und 0.833 auf. Die beiden Faktoren sind daher recht gut durch die jeweiligen Indikatoren erfaßt. Die Doppelladung des Indikators ‚Kampagnen‘ und deren eher geringe Werte legen jedoch den Schluß nahe, daß die Kampagnen zur Verkehrsreduktion weder mit den beiden übrigen Maßnahmen noch mit der Gefährdung der Umwelt durch Autos besonders stark zusammenhängen.

2 Die Bedeutung von Datenhomogenität und Konsequenzen von Datenheterogenität für empirische Analysen

Bei der Analyse der ISSP-Daten mit Hilfe eines linearen Strukturgleichungsmodells wurde unterstellt, daß die Stichprobendaten homogen sind. Datenhomogenität bedeutet, daß alle Fälle der Stichprobe als zufällige und voneinander unabhängige Realisationen aus einer gemeinsamen Population aufgefaßt werden können, deren Verteilung durch ein geeignetes statistisches Modell beschrieben wird. Die in Abbildung 1 geschätzten Beziehungen zwischen den fünf Variablen gelten danach in gleicher Weise für alle Fälle. Dies heißt nun nicht, daß sich die Befragten nicht unterscheiden dürfen. Im Gegenteil, das Modell postuliert ja gerade, daß Personen, die das Umweltrisiko durch Autos relativ höher einschätzen als der Durchschnitt aller Personen, auch eine über dem Durchschnitt liegende Befürwortung verkehrspolitischer Maßnahmen aufzeigen. Unterstellt wird aber, daß der geschätzte Effekt für alle Personen gleich ist, die „wahren“ Regressionskoeffizienten der Regression von ‚Politischen Maßnahmen‘ auf das ‚Umweltrisiko‘ in der Population also für alle Personen gelten. Entsprechendes gilt auch für die Beziehungen zwischen den Indikatoren und den Faktoren. Schließlich wird auch noch unterstellt, daß die Residuen bei allen Fällen gleiche Varianzen aufweisen.

Es ist nicht ausgeschlossen, daß die Homogenitätsannahme falsch ist. So haben wir bei Umfragedaten von Studierenden aus Gießen Hinweise dafür gefunden, daß es zwei Gruppen gibt, die sich nicht nur beim Umweltbewußtsein unterscheiden, sondern die auch unterschiedliche Regressionskoeffizienten bei logistischen Regressionen aufweisen (*Kühnel/Bamberg* 1998; *Bamberg* u.a. 1999). Welche Konsequenzen hätte dies für die Ergebnisse der im ersten Abschnitt vorgestellten Analyse? Die Antwort hängt davon ab, wo und wie die Homogenitätsannahme verletzt ist. So ist aus der klassischen Regressionsanalyse bekannt, daß bei einer Kleinstquadratschätzung ungleiche Residualvarianzen zwar die Schätzung von Standardfehlern verzerren und damit die Gültigkeit inferenzstatistischer Schlüsse in Frage stellen, daß die Regressionskoeffizienten aber trotzdem konsistent und

unverzerrt geschätzt werden. Entsprechendes gilt auch bei der Schätzung von linearen Strukturgleichungsmodellen, die ja aus einer Menge von linearen Regressionsgleichungen bestehen.

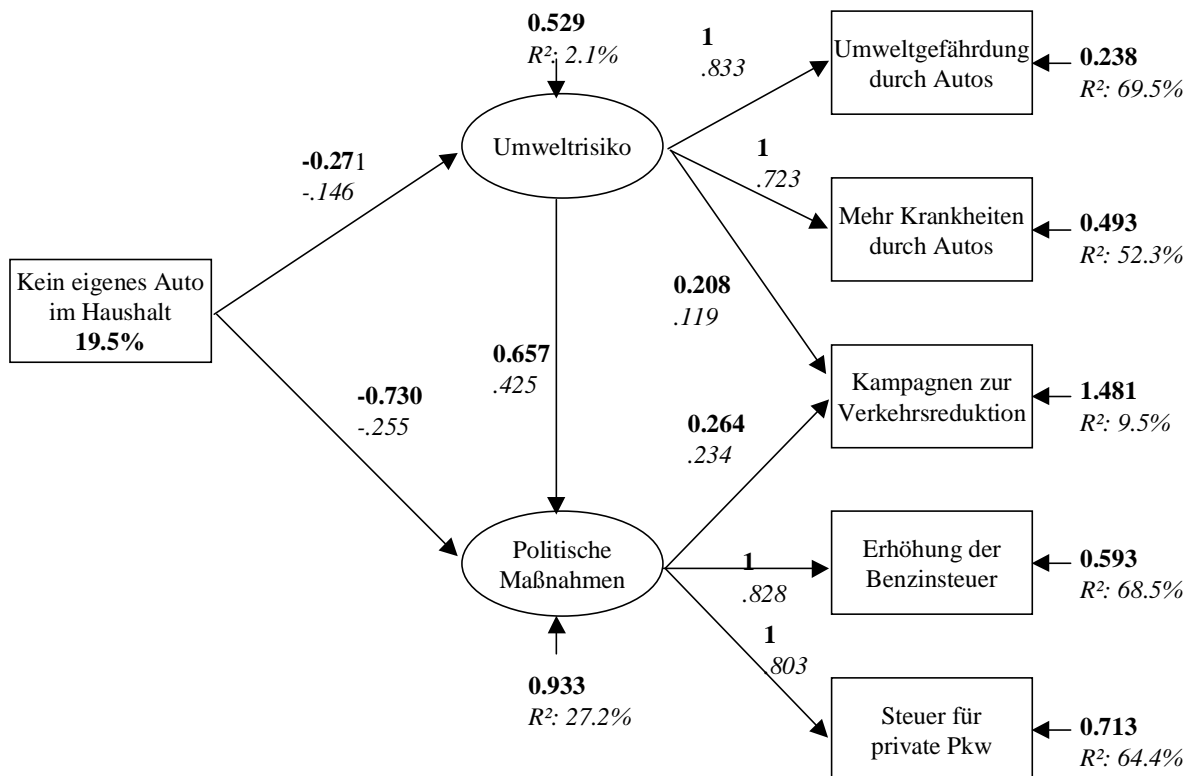
Problematischer ist es, wenn die Regressionskoeffizienten über die Fälle variieren. Die bei der Schätzung berechneten Werte sind dann mittlere Werte, die die tatsächlichen Gegebenheiten möglicherweise sehr unzutreffend beschreiben. Beim Einfluß des Umweltrisikos auf die politischen Maßnahmen wäre z.B. der Fall denkbar, daß es in einer Teilgruppe von Befragten eine sehr starke Beziehung gäbe, während in einer anderen Teilgruppe nur eine geringe oder gar keine Beziehung bestünde. Falls eine solche Heterogenität der Population nicht entdeckt würde und etwa auf der Basis der Ergebnisse aus Abbildung 1 versucht würde, die Einschätzung der Umweltgefährdung durch Autos zu beeinflussen und damit indirekt die Zustimmung zu politischen Maßnahmen zu erhöhen, die den Autoverkehr reduzieren sollen, dann würde eine solche Politik zumindest in der Gruppe mit dem geringen Zusammenhang wenig Erfolg haben. Die sehr negativen Reaktionen auf den Vorschlag von „Bündnis 90 / die Grünen“ vor der letzten Bundestagswahl, den Benzinpreis schrittweise auf 5 DM zu erhöhen, zeigen, welche Konsequenzen eine Fehleinschätzung der öffentlichen Meinung haben kann.³

3 Klassische Lösungsansätze bei heterogenen Daten

Das Ziel jeder Datenanalyse ist es, ein zutreffendes Bild über einen Ausschnitt der Wirklichkeit zu liefern. Bei heterogenen Daten wird entsprechend versucht, die Heterogenität durch ein geeignetes Analysemodell zu berücksichtigen. Bei der Anwendung linearer Strukturgleichungsmodelle kamen in der Vergangenheit vor allem zwei Strategien zum Zuge (vgl. *Muthén* 1989). Eine Strategie basiert darauf, daß eine Verletzung der Homogenitätsannahme durch nichtberücksichtigte Variablen ausgelöst werden kann. Bei der Analyse des Zusammenhangs zwischen ‚Umweltrisiko‘ und ‚politischen Maßnahmen‘ liegt es nahe, daß der Besitz eines Autos eine Größe ist, deren Ignorierung unberücksichtigte Heterogenität im Datensatz hervorruft und dadurch die geschätzten Beziehungen zwischen den beiden Faktoren aus Abbildung 1 verzerrt. Erst die Berücksichtigung von Autobesitz als zusätzliche erklärende Variable kann dieses Problem möglicherweise lösen.

³ Zu diesem Thema siehe auch den Beitrag von *Markus Klein* in diesem Heft.

Abbildung 2: Die Berücksichtigung von Autobesitz als zusätzliche erklärende Variable



LR-Modellfit: $\chi^2=4.012$, $df=8$, $Prob.=0.8560$ ($n=894$)

(ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*)

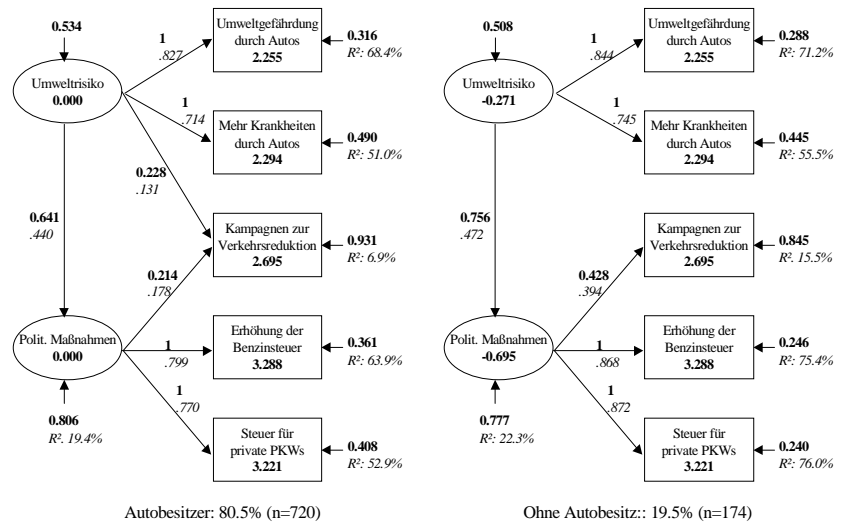
In der ISSP-Umfrage 1993 wurde der Besitz von Autos im Haushalt der Befragten erfaßt. Abbildung 2 zeigt die Ergebnisse der Schätzung eines Strukturgleichungsmodells, bei dem Autobesitz sowohl das ‚Umweltrisiko‘ als auch die ‚politischen Maßnahmen‘ beeinflusst. Die Verfügung über ein Auto wird dabei durch eine 0/1-kodierte dichotome Variable gemessen, die den Wert eins aufweist, wenn es kein Auto im Haushalt gibt. Dies trifft für 19.5% der 894 Fälle zu. Die Ergebnisse weisen darauf hin, daß die Verfügung über ein Auto im Haushalt tatsächlich sowohl die Einschätzung der Umweltgefährdung durch Autos als auch die Bewertung politischer Maßnahmen zur Verkehrsreduktion beeinflusst. Da die Effekte negativ sind und niedrigere Werte für die Wahrnehmung eines größeren Umwelt-

risikos durch Autos (vgl. Tabelle 1) und eine größere Zustimmung zu verkehrspolitischen Maßnahmen (vgl. Tabelle 2) stehen, erhöht das Fehlen eines Autos sowohl die Wahrnehmung des Umweltrisikos wie auch die Zustimmung zu Maßnahmen zur Verkehrsreduktion. Umgekehrt halten Autobesitzer die Umwelt für weniger gefährdet und sind eher gegen Maßnahmen zur Verkehrsreduktion. Der Vergleich mit Abbildung 1 weist zudem darauf hin, daß die Nichtberücksichtigung des Autobesitzes die geschätzte Beziehung zwischen ‚Umweltrisiko‘ und ‚politischen Maßnahmen‘ überschätzt. Das unstandardisierte wie das standardisierte Regressionsgewicht sind in Abbildung 2 geringer als die entsprechende Werte in Abbildung 1.

Das Beispiel zeigt, daß die Berücksichtigung zusätzlicher erklärender Variablen verzerrte Schätzungen korrigieren könne. Diese Vorgehensweise ist jedoch unzureichend, wenn unberücksichtigte Variablen nicht-additive Effekte aufweisen. Würden sich etwa die Regressionsgewichte der ‚politischen Maßnahmen‘ auf das ‚Umweltrisiko‘ bei Autobesitzern und Personen ohne Auto unterscheiden, dann könnte dies durch die Einbeziehung von Autobesitz als zusätzliche erklärende Variable nur unzureichend aufgefangen werden. Stattdessen böte sich hier die zweite Strategie der Kontrolle von Datenheterogenität an, nämlich für beide Gruppen ein eigenes Modell zu spezifizieren und die Regressionskoeffizienten in einem simultanen Gruppenvergleich zu schätzen. Bei einem solchen Gruppenvergleich ist es möglich, unterschiedliche Beziehungen zwischen den Gruppen zu modellieren.

In Abbildung 3 sind die Ergebnisse eines simultanen Gruppenvergleichs von Autobesitzern und Personen ohne Auto wiedergegeben. Die Pfaddiagramme weisen bei beiden Gruppen eine deutlich unterschiedliche Beziehungsstruktur auf. Während das Modell für die Autobesitzer dem Ausgangsmodell aus Abbildung 1 recht ähnlich sieht, zeigt sich für die Gruppe ohne Autobesitz die ursprünglich unterstellte Struktur, nach der die zwei Fragen zur Umweltgefährdung den ersten und die Bewertung der drei politischen Maßnahmen den zweiten Faktor bilden. Unterschiedliche Werte zwischen den beiden Gruppen weisen auch die Beziehung zwischen ‚Umweltrisiko‘ und ‚politischen Maßnahmen‘, die Ladung der ‚Kampagnen zur Verkehrsreduktion‘ auf die ‚politischen Maßnahmen‘ und die Residualvarianzen und Reliabilitäten (erklärte Varianzen) der Indikatoren auf. Differenzen gibt es schließlich auch zwischen den Mittelwerten in den beiden Gruppen.⁴ In der Abbildung sind in den Symbolen für die Variablen unter den Variablenbezeichnungen die Interzepte (Regressionskonstanten) eingetragen. Aus diesen lassen sich die Mittelwerte berechnen.

4 Die Mittelwerte bzw. Interzepte latenter Variablen können nicht eindeutig bestimmt werden. Es ist allerdings möglich, Differenzen zwischen Mittelwerten und Interzepten zu schätzen. Aus diesem Grund sind in Abbildung 3 die Interzepte der beiden Faktoren in der Gruppe der Autofahrer auf null fixiert. Die geschätzten Werte in der zweiten Gruppe messen dann die Differenzen zu den Werten der ersten Gruppe.



LR-Modellfit: $\chi^2=19.856$, $df=14$, $Prob.=0.1347$

(ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*, Interzepte unter Variablenbeschreibung)

Abbildung 3: Unterschiedliche Beziehungen bei Autobesitzern und Personen ohne Auto

Da sich bei den ISSP-Daten nahezu alle Modellgrößen zwischen der Gruppe der Autobesitzer und der Nichtbesitzer unterscheiden, ist in diesem Beispiel die zweite Strategie des simultanen Gruppenvergleichs (Abbildung 3) der ersten Strategie der Spezifikation einer zusätzlichen erklärenden Variable (Abbildung 2) vorzuziehen. Bei der Schätzung der Beziehung zwischen ‚Umweltrisiko‘ und ‚politischen Maßnahmen‘ scheint die erste Strategie sogar zu einem schlechteren Ergebnis zu führen als die Ignorierung der durch Autobesitz hervorgerufenen Datenheterogenität. Zumindest das standardisierte Regressionsgewicht unterschätzt in Abbildung 2 mit einem Wert von 0.425 die Werte in den beiden Gruppen (0.440 für Autobesitzer und 0.472 für Personen ohne Auto). In Abbildung 1 liegt der geschätzte Wert mit 0.463 dagegen zwischen den beiden Gruppenwerten.

4 Die Anwendung von Mischverteilungsmodellen

Die beiden üblichen Strategien bei Datenheterogenität setzen voraus, daß die Ursache der mangelnden Homogenität bekannt ist. Beim Gruppenvergleich ist dies offensichtlich, ist doch die Gruppenzugehörigkeit das Kriterium, das die heterogenen Fälle voneinander unterscheidet. Innerhalb der Gruppen wird weiterhin Homogenität angenommen. Auch bei der Spezifikation einer zusätzlichen erklärenden Variable bilden die Ausprägungen dieser Variablen Teilgruppen, für die Homogenität unterstellt wird. Beide Lösungsstrategien sind daher ungeeignet, wenn unbekannt ist, wie die einzelnen Fälle einer Stichprobe unterschiedlichen Subgruppen zuzuordnen sind.⁵

Einen Ausweg bieten Mischverteilungsmodelle an. Die Idee dieser Modelle entspricht im wesentlichen dem simultanen Gruppenvergleich. Im Unterschied zum üblichen Gruppenvergleich ist es aber nicht notwendig, die Stichprobe explizit in Subgruppen zu unterscheiden. Stattdessen wird nur angenommen, daß sich die Gesamtpopulation aus unterschiedlichen Subpopulationen zusammensetzt und für diese Subpopulationen jeweils unterschiedliche Parameter gelten. Entsprechend setzt sich auch die Gesamtstichprobe aus unterschiedlichen Teilstichproben zusammen, wobei jedoch die Gruppenzugehörigkeit der einzelnen Fälle unbekannt ist. Da die Gruppenzugehörigkeit zu den Subpopulationen bzw. Teilstichproben unbeobachtbar ist, werden die Gruppen in Mischverteilungsmodellen als *latente Klassen* bezeichnet.⁶ Die Bezeichnung „Mischverteilungsmodelle“ (engl.: *latent variable mixture models*) kommt daher, daß sich die Verteilung in der Gesamtpopulation als eine Mischung aus verschiedenen Verteilungen in den latenten Klassen zusammensetzt. Wenn $F(Y)$ die (multivariate) Verteilungsfunktion einer Gesamtpopulation bezeichnet,

5 Dies gilt auch für eine dritte Strategie, bei der Heterogenität durch die Spezifikation von Mehrebenenmodellen kontrolliert werden soll (zu Mehrebenenmodellen mit latenten Variablen vgl. **Muthén** 1989).

6 Tatsächlich können die hier vorgestellten Mischverteilungsmodelle als eine Kombination aus latenter Klassenanalyse (LCA) und linearen Strukturgleichungsmodellen verstanden werden. Eine Einführung in die LCA findet sich in dem Lehrbuch von **Andrefß** u.a. (1997).

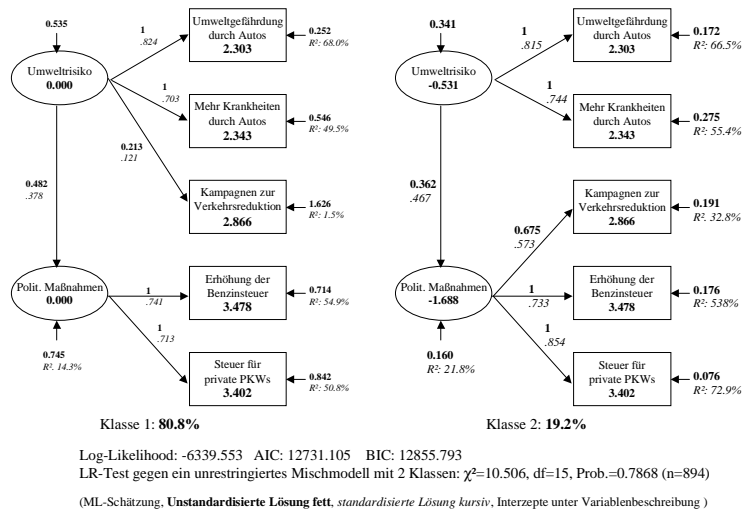
$F_c(Y)$ die Verteilungsfunktion in einer latenten Klasse c , und π_c die Größe (den Anteil) der latenten Klasse c an der Gesamtpopulation, dann ergibt sich $F(Y)$ als gewichtetes Produkt aller Verteilungsfunktionen in den latenten Klassen:

$$F(Y) = \prod_c \pi_c \cdot F_c(Y) \quad (1)$$

Erst seit kurzer Zeit stehen für die Schätzung von Mischverteilungsmodellen leistungsfähige Schätzalgorithmen zur Verfügung. Die nachfolgenden Mischverteilungsmodelle sind - wie bereits die Schätzungen der in den Abbildungen 1 bis 3 wiedergegebenen Strukturgleichungsmodelle - mit dem Programm *Mplus* Version 1.03 von **Bengt Muthén** (*Muthén/Muthén* 1998) berechnet worden. Bei der Schätzung eines Mischverteilungsmodells wird zunächst die Anzahl der latenten Klassen festgelegt. Wie beim simultanen Gruppenvergleich wird dann für jede Klasse ein gruppenspezifisches lineares Strukturgleichungsmodell spezifiziert. Dabei ist es möglich, verschiedene Modellparameter (Mittelwerte, Varianzen und Kovarianzen exogener Variablen sowie Regressionskoeffizienten) innerhalb einer latenten Klasse oder auch über latente Klassen hinweg gleichzusetzen. Das Programm schätzt die unbekanntes Modellparameter mittels ML-Schätzung, wobei für die Standardfehler auch robuste, d.h. verteilungsunabhängige, Schätzer zur Verfügung stehen.

Die Ergebnisse der Schätzung eines Mischverteilungsmodells mit zwei Klassen für die ISSP-Daten sind in Abbildung 4 wiedergegeben. In diesem Modell unterscheiden sich die beiden latenten Klassen noch deutlicher als beim simultanen Gruppenvergleich. In der ersten und mit 80.8% größeren Klasse läßt die Befürwortung von Kampagnen zur Verkehrsreduktion nur auf dem ersten Faktor, in der zweiten Klasse dagegen nur auf dem zweiten Faktor. Inhaltlich bedeutet dies, daß in den beiden Klassen offenbar unterschiedliche Sichtweisen bestehen. In der zweiten Klasse können alle drei Maßnahmen zu einem Faktor zusammengefaßt werden, wobei die ‚Kampagnen zur Verkehrsreduktion‘ allerdings einen etwas unzuverlässigeren Indikator bilden. In der ersten Klasse erfassen dagegen nur die beiden Maßnahmen einen gemeinsamen Faktor, die über eine finanzielle Belastung der Autofahrer durch Benzinpreiserhöhungen oder Steuer für private Autos weniger motorisierten Individualverkehr erzielen wollen. Die Ladung der ‚Kampagnen zur Verkehrsreduktion‘ auf den Faktor ‚Umweltrisiko‘ in der ersten Klasse sollte allerdings nicht dahingehend interpretiert werden, daß in dieser Klasse die Zustimmung zu Kampagnen ein Indikator der wahrgenommenen Umweltgefährdung ist. Schon die mit nur 1.5% äußerst geringe erklärte Varianz spricht dagegen. Statt dessen ist davon auszugehen, daß in der ersten Klasse eine klare Unterscheidung getroffen wird zwischen verkehrspolitischen Maßnahmen, die Autofahrer direkt belasten, und solchen, die dies nicht tun. Die ‚Kampagnen zur Verkehrsreduktion‘ bilden einen Indikator für die zweite Maßnahmengruppe. Da es hierzu aber keinen weiteren Indikator gibt, ist es nicht möglich, explizit einen eigenen Faktor zu spezifizieren.

Abbildung 4: Risikobewertung und Akzeptanz verkehrspolitischer Maßnahmen im Mischverteilungsmodell mit 2 Klassen



Neben der Zuordnung der beobachteten Variablen zu den Faktoren unterscheiden sich auch die geschätzten Koeffizienten zwischen den Klassen. So ist zwar das unstandardisierte Regressionsgewicht der Regression des Maßnahmenfaktors auf den Risikofaktor in der zweiten Klasse geringer als in der ersten Klasse. Aufgrund der geringeren Streuung in der zweiten Klasse ist der Anteil der erklärten Varianz mit 21.8% aber deutlich höher als in der ersten Klasse, wo nur 14.3% der Varianz erklärt wird. Am prägnantesten sind die Mittelwertunterschiede, die in Tabelle 4 wiedergegeben sind. Da alle beobachteten Variablen auf Antwortskalen mit fünf Antwortkategorien basieren, gilt diese Skalierung auch für die beiden Faktoren. Verglichen mit der ersten Klasse wird die Umweltgefährdung durch Autos in der zweiten Klasse um mehr als eine halbe Einheit größer eingeschätzt. Bei den ‚politischen Maßnahmen‘ beträgt die Differenz sogar annähernd zwei Einheiten. Während in der ersten Klasse nur die Kampagnen im Durchschnitt eine leichte Zustimmung erhalten und bei den beiden anderen Maßnahmen eher mit Ablehnung zu rechnen ist, weisen die Mittelwerte in der zweiten Klasse bei allen drei Maßnahmen auf eine deutliche Zustimmung hin.

Tabelle 4: Mittelwerte in den latenten Klassen

	Klasse 1	Klasse 2
<i>Faktoren</i>		
Umweltrisiko	0.000	-0.531
Politische Maßnahmen	0.000	-1.880
<i>Indikatoren</i>		
Umweltgefährdung durch Autos	2.303	1.772
Mehr Krankheiten durch Autos	2.343	1.812
Kampagnen zur Verkehrsreduktion	2.866	1.597
Erhöhung der Benzinsteuern	3.478	1.598
Steuer für private Pkw	3.402	1.522

Das Mischverteilungsmodell weist somit auf eine deutliche Heterogenität in der ISSP-Stichprobe hin, die durch die Existenz von zwei sehr unterschiedlichen Subpopulationen hervorgerufen wird. In der mit gut 19% deutlich kleineren Subpopulation wird die Umweltgefährdung durch Autos als recht groß angesehen. Maßnahmen zur Verkehrsreduktion finden in dieser Gruppe eine deutliche Zustimmung. Die Zustimmung hängt mit der Wahrnehmung der Umweltgefährdung zusammen, wobei der Zusammenhang bei Maßnahmen, die die individuellen Kosten des Individualverkehrs erhöhen, stärker ist als bei Kampagnen, die für eine freiwillige Reduzierung der Autonutzung werben. In der mit fast 81% deutlich größeren Gruppe gibt es dagegen eine klare Ablehnung von verkehrspolitischen Maßnahmen, die die individuellen Kosten der Autonutzung erhöhen. Lediglich für die Kampagnen gibt es eine knappe Mehrheit. Die Zustimmung hierzu hat in dieser Subpopu-

lation aber kaum etwas mit der Einschätzung der Umweltgefährdung durch Autos zu tun. Verglichen mit der kleineren Subpopulation wird in dieser größeren Gruppe die Gefahr der Umweltgefährdung für deutlich weniger groß gehalten. Dieses Ergebnis repliziert die Befunde aus unseren Untersuchungen bei Gießener Studierenden, wo mittels Cluster-Analysen ebenfalls zwei Gruppen identifiziert worden sind (*Kühnel/Bamberg* 1998). Im Unterschied zu den Studierenden scheint bei der Gesamtpopulation die Gruppe mit einer ökologischen Orientierung aber deutlich kleiner zu sein.

Verglichen mit dem simultanen Gruppenvergleich (Abbildung 3) wird beim Mischverteilungsmodell die Unterschiedlichkeit der beiden Gruppen noch stärker akzentuiert. Die Ähnlichkeit zwischen den Modellen legt die Vermutung nahe, daß in der ersten Klasse des Mischverteilungsmodells eher Autobesitzer zu erwarten sind und in der zweiten Klasse eher Personen ohne Auto im Haushalt. Da das für die Modellschätzung verwendete Programm *Mplus* es ermöglicht, für jeden Fall die geschätzten Wahrscheinlichkeiten auszugeben, mit der ein Fall den latenten Klassen angehört, kann dieser Vermutung nachgegangen werden. Jeder Fall wurde dabei der latenten Klasse zugeordnet, die die höchste Zuordnungswahrscheinlichkeit aufweist. Von den der ersten Klasse zugeordneten 702 Fällen haben tatsächlich 86.3% ein Auto in der Familie. Entgegen der Erwartung haben aber auch 57.7% der 182 Fälle, die der zweiten Klasse zugeordnet werden, ein Auto in der Familie. Obwohl die Prozentsatzdifferenz von immerhin 28.6 Prozentpunkten auf einen deutlichen Zusammenhang zwischen Klassenzugehörigkeit und Autobesitz schließen läßt, ist doch festzustellen, daß es auch Autobesitzer gibt, die der zweiten Klasse angehören und bei denen daher mit einer Zustimmung zu verkehrspolitischen Maßnahmen zu rechnen ist, deren Realisierung ihre eigenen Kosten erhöhen würde. Auf der anderen Seite gibt es aber auch Personen, die kein Auto haben und trotzdem der ersten Klasse angehören. Aufgrund der Größenordnung der Klassen überwiegt ganz eindeutig der Personenkreis, der gegen Maßnahmen ist, die das Autofahren verteuern. Der Versuch, auf indirektem Weg über Zustimmung zu politischen Maßnahmen, die umweltschädlicheres Verhalten verteuern, ein umweltgerechteres Verhalten zu erreichen, scheint kein leichtes Unterfangen zu sein.

5 Anwendungsprobleme von Mischverteilungsmodellen

Gegenüber dem simultanen Gruppenvergleich weist die Anwendung von Mischverteilungsmodellen zwei für die Sozialforschung interessante Vorteile auf. Anstelle einer von vornherein festgelegten, deterministischen Zuordnung der Untersuchungseinheiten zu Gruppen erscheint es in vielen Situationen realistischer, daß die Zuordnung nicht starr und deterministisch ist. So ist es in dem Anwendungsbeispiel durchaus plausibel, daß nicht alle Autobesitzer automatisch in die Gruppe derjenigen fallen, die eine weniger kritische Sicht der Umweltgefährdung durch Autos aufweisen und sich gegen verkehrspolitische Maßnahmen wenden, die sie direkt belasten würden. Umgekehrt ist es genauso plausibel, daß

nicht alle Fußgänger, Radfahrer und Nutzer von Bus und Bahn Autos als äußerst umweltschädlich betrachten und das Autofahren möglichst verteuern wollen. Die Konzeption der latenten Klassen in Mischverteilungsmodellen ermöglicht es, diese flexiblere und vermutlich realitätsgerechtere Sicht bei der statistischen Modellierung zu berücksichtigen.

Ein zweiter Vorteil der Mischverteilungsmodelle besteht darin, daß die Notwendigkeit entfällt, explizit eine Gruppierungsvariable zu erheben. In die Berechnung des Mischverteilungsmodells aus Abbildung 4 sind allein die Antworten auf die fünf in den Tabellen 1 und 2 vorgestellten Fragen eingegangen. Obwohl der Autobesitz bei der Schätzung nicht berücksichtigt wurde, ergab sich doch eine Lösung, die einen Zusammenhang mit dem Autobesitz aufweist. Mit Mischverteilungsmodellen ist es somit möglich, auch verborgene Datenstrukturen aufzudecken. Dies ist nicht nur inhaltlich interessant, sondern auch aus statistischen Gründen von Bedeutung, da verborgene Datenstrukturen die Homogenitätsannahme und damit die Ergebnisse konventioneller Datenanalysen in Frage stellen können.

Auf der anderen Seite gibt es jedoch auch Anwendungsprobleme, die nicht verschwiegen werden sollen. Diese ergeben sich aus dem problematischen Identifikationsstatus von Mischverteilungsmodellen. So wie sich eine Summe auf verschiedene Weise aus einzelnen Summanden zusammensetzen kann, so ist es auch möglich, daß in einer Population dieselbe Wahrscheinlichkeitsverteilung auf unterschiedliche Weise entsprechend Gleichung (1) aus den Verteilungen von Subpopulationen zusammengesetzt werden kann. Bei der Anwendung von Mischverteilungsmodellen kommt es daher darauf an, die "richtige" Aufteilung in Subpopulationen zu finden.

Auf der Anwendungsebene ergeben sich daraus drei Probleme:

- Es ist sehr leicht möglich, daß der Schätzalgorithmus überhaupt keine Lösung findet.
- Es ist möglich, mehrere Lösungen für einen Datensatz zu finden.
- Es gibt kein eindeutiges Kriterium für die Übereinstimmung von Modell und Daten.

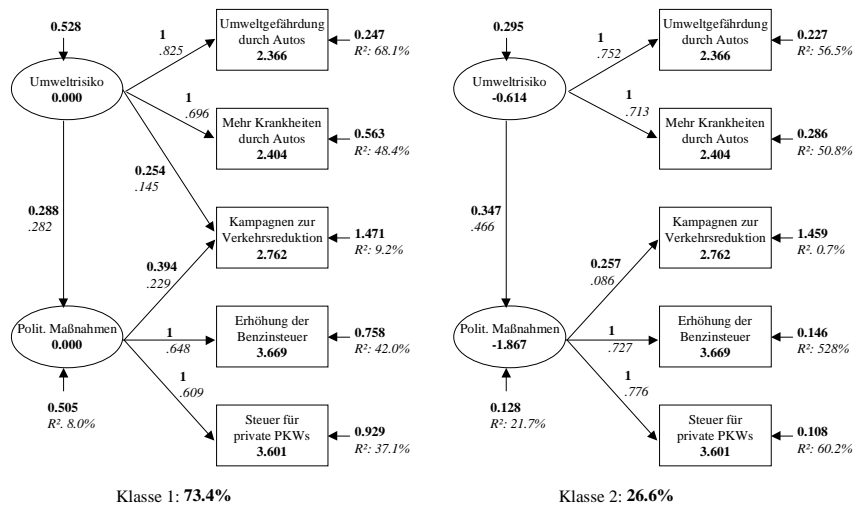
Nach Gleichung (1) ist die Wahrscheinlichkeitsverteilung in einer Population das gewichtete Produkt der Wahrscheinlichkeitsverteilungen in den Subpopulationen. Bei der Schätzung eines Mischverteilungsmodells werden die unbekanntes Modellparameter so bestimmt, daß die maximale Wahrscheinlichkeitsdichte besteht, nach der die beobachteten Daten durch das spezifizierte Modell generiert worden sein könnten. Diese als Maximum-Likelihood-Methode (ML-Schätzung) bezeichnete Vorgehensweise ist bei Mischverteilungsmodellen mathematisch sehr aufwendig zu berechnen und basiert grundsätzlich auf einem iterativen Vorgehen, bei dem eine suboptimale Ausgangslösung schrittweise immer näher an das Optimum, die maximale Wahrscheinlichkeitsdichte, herangeführt wird. Falls nun eine Ausgangslösung sehr weit von einer optimalen Lösung entfernt ist, kann es sehr

leicht vorkommen, daß die schrittweise Verbesserung nicht gelingt, d.h. gar kein Optimum gefunden wird.

Das Auffinden einer Lösung setzt daher voraus, daß eine Ausgangslösung vorgegeben wird, die der endgültigen Lösung schon recht nahe ist. Dies schränkt ganz offensichtlich die oben angedeutete Möglichkeit ein, mit Mischverteilungsmodellen verborgene Strukturen aufzudecken. Da sich durch Zufall kaum eine gute Ausgangslösung finden läßt, bedarf es schon gewisser Vorkenntnisse oder Erwartungen, um eine dann doch nicht mehr ganz verborgene Struktur aufzudecken.

Wenn es gelungen ist, eine Lösung zu finden, bleibt unklar, ob es sich wirklich um die "richtige" Lösung handelt. In Abhängigkeit von der Ausgangslösung ist es nämlich auch möglich, verschiedene Optima zu finden. Statistiker sprechen in diesem Zusammenhang von der Existenz lokaler Maxima. Bei Mischverteilungsmodellen besteht nun eine besonders große Chance, auf ein solches lokales Maximum zu stoßen. Dies gilt auch für die oben verwendeten ISSP-Daten. Als Beispiel zeigt Abbildung 5 eine alternative Lösung, die sich bei anderen Ausgangswerten für den Schätzalgorithmus einstellt.

Der Unterschied zu dem in Abbildung 4 vorgestellten Modell besteht in erster Linie darin, daß die größere Klasse nur 73.4% statt 80.8% der Fälle aufweist und, daß die ‚Kampagnen zur Verkehrsreduktion‘ anders laden. Vor dem Hintergrund der Ergebnisse des Gruppenvergleichs aus Abbildung 3 ist auch dieses zweite Modell inhaltlich nicht unplausibel. Es stellt sich daher die Frage, welches Modell vorzuziehen ist.



Log-Likelihood: -6349.637 AIC: 12753.273 BIC: 12882.757
 LR-Test gegen ein unrestringiertes Mischmodell mit 2 Klassen: $\chi^2=15.336$, $df=14$, $Prob.=0.1203$ (n=894)
 (ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*, Interzepte unter Variablenbeschreibung)

Abbildung 5: Ein alternatives Mischverteilungsmodell mit 2 Klassen

Hier stellt sich das dritte Problem, daß es bei Mischverteilungsmodellen kein eindeutiges Kriterium für die Übereinstimmung von Modell und Daten gibt. Bei der Anwendung der üblichen Strukturgleichungsmodelle wird dagegen standardmäßig ein Likelihood-Ratio-Test berechnet, der die Nullhypothese prüft, daß die durch das geschätzte Modell implizierten Mittelwerte, Varianzen und Kovarianzen nicht signifikant von den beobachteten Mittelwerten, Varianzen und Kovarianzen abweichen. Bei Mischverteilungsmodellen entfällt dieser LR-Goodness-of-fit-Test, da die Zugehörigkeit der Fälle zu den latenten Klassen ja unbekannt ist und somit auch keine klassenspezifischen empirischen Mittelwerte, Varianzen und Kovarianzen zu beobachten sind. Für die in Abbildung 4 und 5 wiedergegebenen Modelle wurde als Alternative ein Likelihood-Ratio-Test des wiedergegebenen Modells gegen ein unrestringiertes Mischverteilungsmodell gleicher Klassenzahl durchgeführt. In diesem Baseline-Modell können die Mittelwerte, Varianzen und Kovarianzen in den latenten Klassen beliebige Werte annehmen. Die zweifache Differenz der Log-Likelihood-Funktion eines solchen Modells zu dem inhaltlich interessierenden Modell kann als Teststatistik für die Modellanpassung verwendet werden. Die Statistik prüft, ob das inhaltlich interessierende Modell signifikant schlechter mit den Daten vereinbar ist als das unrestringierte Modell. Die Teststatistik ist bei Gültigkeit der Nullhypothese, daß kein signifikanter Unterschied besteht, chiquadratverteilt. Die Zahl der Freiheitsgrade ergibt sich aus der Differenz der in den beiden Modellen geschätzten freien Parameter. Nach diesem Test weisen beide Modelle eine gute Modellanpassung auf, wobei der Fit des ersten Modells (Abbildung 4) sogar ganz ausgezeichnet ist.

Für diesen Test der Modellanpassung muß zusätzlich zu dem eigentlich interessierenden Modell ein Baseline-Modell geschätzt werden. Bei dessen Schätzung stellt sich aber wiederum das Problem des Auffindens einer optimalen Lösung. Ich bin bei meinen Analysen so vorgegangen, daß ich die geschätzten Klassengrößen und die modellimplizierten Mittelwerte, Varianzen und Kovarianzen des in Abbildung 4 bzw. 5 wiedergegebenen Modells als Ausgangswerte für die Schätzung des Baseline-Modells verwendet habe. Diese Vorgehensweise erleichtert das Auffinden einer unrestringierten Lösung. Auf der anderen Seite wird dadurch die Chance verringert, eine ganz andere Lösung zu finden, die vielleicht deutlich besser ist. Die Vorgehensweise ergibt zudem unterschiedliche Baseline-Modelle für die beiden Mischverteilungsmodelle aus den Abbildungen 4 und 5.⁷ Im Unterschied zum LR-Modellfitt klassischer Strukturgleichungsmodelle ist der LR-Test

7 Nach der reinen Lehre der Statistik ist eigentlich für alle Mischverteilungsmodelle mit gleicher Variablen- und Klassenzahl das Baseline-Modell heranzuziehen, das den höchsten Log-Likelihood-Wert aufweist. Im Beispiel wäre dies das unrestringierte Modell, das aus der Ausgangslösung aus Abbildung 4 gefunden wurde. Danach wäre die Anpassung des Modells aus Abbildung 5 unbefriedigend ($\chi^2=30.674, df=15, p=.0097$). Wenn für die ISSP-Daten nur das Modell aus Abbildung 5 geschätzt worden wäre, wäre jedoch auch nur das „passende“ Baseline-Modell geschätzt.

gegen ein unrestringiertes Modell daher mit größerer Unsicherheit über die Gültigkeit des Testergebnisses behaftet.

Zur Beurteilung der Gültigkeit eines geschätzten Modells sind in den Abbildungen 4 und 5 weitere Hinweise aufgeführt. Angegeben ist die Log-Likelihood-Funktion, die bei der Schätzung der Koeffizienten maximiert wird. Je größer diese Funktion, desto größer ist die Wahrscheinlichkeitsdichte, daß die empirischen Daten durch das Modell generiert sein könnten. Leider lassen sich die Funktionswerte nur bei identischer Modellstruktur direkt vergleichen. Eine Lösung bieten die Informationskriterien AIC und BIC, die die Anzahl der zu schätzenden Modellparameter mit dem Wert der Likelihood-Funktion in Beziehung setzen.⁸ Für beide Koeffizienten gilt: Je kleiner der Wert von AIC und BIC, desto besser erscheint das Modell mit den empirischen Daten vereinbar. Folgt man diesen Kriterien, scheint das Modell aus Abbildung 4 dem Modell aus Abbildung 5 überlegen zu sein. Nach formal-statistischen Kriterien ist somit das Modell aus Abbildung 4 die Lösung, die die ISSP-Daten angemessen beschreibt. Auch inhaltliche Gründe sprechen meiner Ansicht nach für diese Lösung. Grundsätzlich bleibt jedoch die Unsicherheit über die Gültigkeit der Lösung bestehen.

6 Mischverteilungsmodelle mit Indikatoren für die latenten Klassen

In den in den Abbildungen 4 und 5 vorgestellten Mischverteilungsmodellen wurden die latenten Klassen aus den Werten der fünf Variablen identifiziert, die auch die Basis für die spezifizierten Strukturgleichungsmodelle bilden. Die Unsicherheit über die Gültigkeit der Modelle läßt sich möglicherweise dadurch reduzieren, daß zusätzliche Daten herangezogen werden, die Hinweise auf die Existenz der latenten Klassen bilden. Das Programm *Mplus* bietet hierfür zwei Möglichkeiten.⁹ Zum einen können exogene Variablen als Prädiktoren der latenten Klassen spezifiziert werden. Zum anderen können dichotome Variablen als Indikatoren der latenten Klassen eingesetzt werden. Die zweite Möglichkeit soll für die ISSP-Daten genutzt werden. Für das Modell aus Abbildung 4 wurde bereits gezeigt, daß die latenten Variablen mit dem Autobesitz zusammenhängen. In der ISSP-Umfrage 1993 ist auch danach gefragt worden, ob und wie oft die befragte Person - unabhängig vom Besitz eines Autos - normalerweise in der Woche Auto fährt. Aus den Antworten auf diese Frage ist eine 0/1-kodierte Variable gebildet worden, die die Autonutzung erfaßt. Das Modell aus Abbildung 4 kann als Ausgangslösung für ein Mischverteilungsmodell verwendet

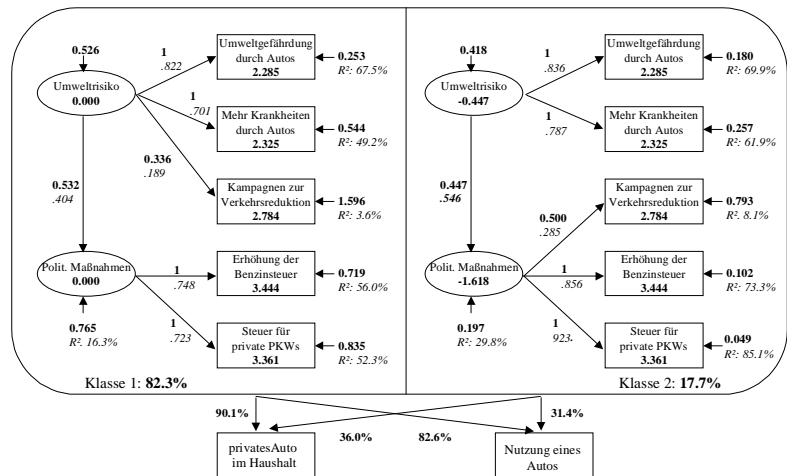
8 Der Wert von AIC ergibt sich aus der zweifachen Summe der negativen Log-Likelihood-Funktion und der Zahl der geschätzten Modellparameter. Beim BIC wird die Summe der zweifachen negativen Log-Likelihood-Funktion und das Produkt aus der logarithmierten Fallzahl und der Zahl der Modellparameter gebildet.

9 Eine dritte Möglichkeit besteht, wenn für einige Fälle die Zugehörigkeit zu einer Klasse bekannt ist. Diese Information kann in *Mplus* bei der Schätzung eines Mischverteilungsmodells genutzt werden.

werden, bei dem Autobesitz und Nutzung eines Autos Indikatoren für die latenten Klassen bilden.

Das Ergebnis der Schätzung ist in Abbildung 6 wiedergegeben. Die zentralen Ergebnisse stimmen mit denen aus Abbildung 4 überein. Im Detail sind allerdings Unterschiede zu erkennen. So ist der Unterschied in der Klassengröße noch etwas angestiegen. Die kleinere Klasse ist von 19.2% auf 17.7% gesunken. Die Beziehung zwischen dem ‚Umweltrisiko‘ und den ‚politischen Maßnahmen‘ ist dadurch stärker geworden. In der kleineren Klasse werden nun knapp 30% der Varianz erklärt, in der größeren Klasse gut 16%. Bei der Verwendung von Indikatoren für die latenten Klassen werden die ‚Kampagnen zur Verkehrsreduktion‘ in der ersten Klasse nun etwas besser erklärt. Umgekehrt wird die Variation dieser Variablen aber in der zweiten Klasse deutlich schlechter erklärt. Inhaltlich läßt sich dieses Ergebnis so deuten, daß sich die ‚politischen Maßnahmen‘ auch in der zweiten Klasse eigentlich aus zwei untergeordneten Faktoren zusammensetzen. Ein Faktor bezieht sich auf Maßnahmen, welche die Kosten des Autofahrens erhöhen. Ein zweiter Faktor wird nur durch die ‚Kampagnen zur Verkehrsreduktion‘ gemessen. Während die beiden Faktoren in der zweiten Klasse einen gemeinsamen Faktor (2. Ordnung) bilden, sind sie in der ersten Gruppe nur über die Bewertung des Umweltrisikos miteinander verbunden. Bei den Mittelwerten gibt es nur geringe Unterschiede zwischen den Modellen aus Abbildung 4 und Abbildung 6. Der Vergleich der Werte in den Tabellen 4 und 5 zeigt, daß die Mittelwertunterschiede geringer ausfallen, wenn Autobesitz und Autonutzung als Indikatoren der latenten Klassen spezifiziert werden.

Die Beziehung zwischen den latenten Klassen und deren Indikatoren ist inhaltlich interessant. In der größeren Klasse, deren Angehörige eine geringere Umweltgefährdung durch Autos wahrnehmen und die eher gegen Maßnahmen sind, die das Autofahren verteuern, besteht eine Wahrscheinlichkeit von 90%, daß es ein Auto im Haushalt gibt, und eine Wahrscheinlichkeit von fast 83%, daß regelmäßig ein Auto genutzt wird. In der kleineren Klasse, deren Mitglieder ein größeres Umweltrisiko durch Autos sehen und die auch für Maßnahmen eintreten, die das Autofahren verteuern, besteht nur eine Chance von 36%, daß ein Auto besessen wird und eine Chance von gut 31%, daß ein Auto regelmäßig genutzt wird. Alles in allem erhöht die Zunahme von Indikatoren für die latenten Klassen die Plausibilität der Lösung. Der starke, aber nicht perfekte Zusammenhang mit Autobesitz und Nutzung rechtfertigt die Anwendung eines Mischverteilungsmodells für die hier untersuchte Fragestellung.



Log-Likelihood: -7228.357 AIC: 14516.714 BIC: 14660.585
 LR-Test gegen ein unrestringiertes Mischmodell mit 2 Klassen: $\chi^2=14.000$, $df=15$, $Prob.=0.5255$ ($n=894$)
 (ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*, Interzpte unter Variablenbeschreibung)

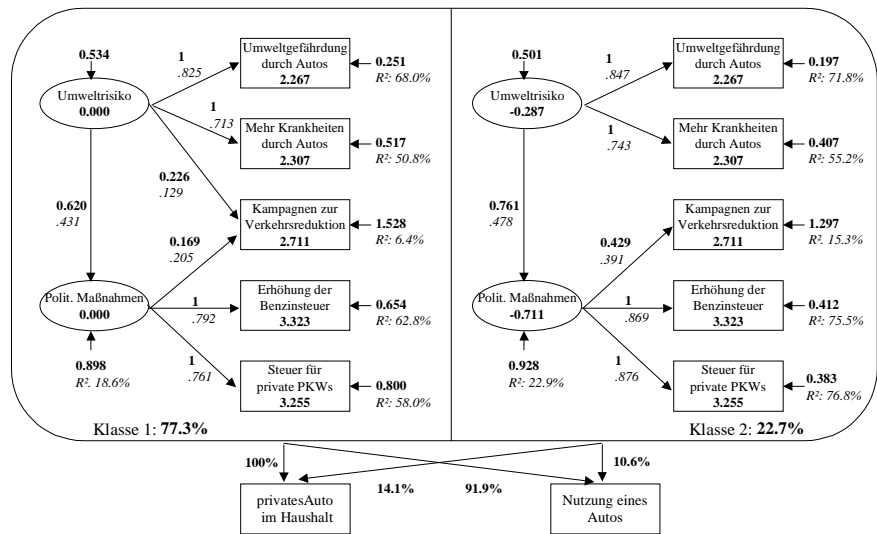
Abbildung 6: Mischverteilungsmodell mit Indikatoren für die latenten Klassen

Tabelle 5: Mittelwerte in den latenten Klassen des Modells aus Abbildung 6

	Klasse 1	Klasse 2
<i>Faktoren</i>		
Umweltrisiko	0.000	-0.447
Politische Maßnahmen	0.000	-1.818
<i>Indikatoren</i>		
Umweltgefährdung durch Autos	2.285	1.838
Mehr Krankheiten durch Autos	2.325	1.878
Kampagnen zur Verkehrsreduktion	2.784	1.875
Erhöhung der Benzinsteuer	3.444	1.626
Steuer für private Pkw	3.361	1.543

Obwohl die Hinzuziehung von Indikatoren die Ergebnisse als sehr plausibel erscheinen läßt, bleibt jedoch die Ungewißheit über die Gültigkeit der Analyse bestehen. Wie schon bei den Mischverteilungsmodellen ohne Indikatoren für die latenten Klassen gibt es wiederum mehrere Lösungen. In Abbildung 7 ist eine alternative Lösung wiedergegeben, die sich einstellt, wenn die Ergebnisse aus Abbildung 5 als Ausgangslösung für ein Mischverteilungsmodell mit Indikatoren für die latenten Klassen herangezogen werden.

Die Ähnlichkeit zwischen diesem Modell und dem Modell aus Abbildung 5 ist etwas geringer als die Ähnlichkeit zwischen den Modellen aus Abbildung 4 und Abbildung 6. Der Größenzuwachs der ersten Klasse ist etwas größer und die Koeffizienten der Beziehung zwischen ‚Umweltrisiko‘ und ‚politischen Maßnahmen‘ unterscheiden sich deutlicher. Dies gilt auch für die Interzepte bzw. die daraus berechneten Mittelwerte. Eine sehr deutliche Ähnlichkeit besteht dagegen mit dem Ergebnis des simultanen Gruppenvergleichs (Abbildung 3). Tatsächlich haben in dem Modell aus Abbildung 7 alle Angehörigen der größeren Klasse ein Auto im Haushalt. In der kleineren Klasse beträgt der Anteil der Autobesitzer nur 14%.



Log-Likelihood: -7193.580 AIC: 14447.159 BIC: 14591.030
 LR-Test gegen ein unrestringiertes Mischmodell mit 2 Klassen: $\chi^2=21.522$, $df=14$, $Prob.=0.0890$ ($n=894$)

(ML-Schätzung, **Unstandardisierte Lösung fett**, *standardisierte Lösung kursiv*, Interzepte unter Variablenbeschreibung)

Abbildung 7: Alternatives Mischverteilungsmodell mit Indikatoren für die latenten Klassen

Wie bereits bei den Modellen ohne Indikatoren für die latenten Klassen stellt sich wiederum die Frage, ob das Modell aus Abbildung 6 oder das aus Abbildung 7 vorzuziehen ist. Aufgrund der Einfachheit der Struktur und damit der Klarheit der Interpretation, der größeren Ähnlichkeit zwischen den Modellen mit und ohne Indikatoren für die latenten Klassen und der besseren Übereinstimmung mit einem passenden unrestringierten Baseline-Modell bevorzuge ich das Modell aus Abbildung 6. Anders als bei den Modellen aus Abbildung 4 und 5, wo auch die formalen Kriterien AIC und BIC für das Modell aus Abbildung 4 sprechen, weist nun aber das Modell aus Abbildung 7 nach diesen Kriterien einen besseren Fit auf als das Modell aus Abbildung 6. Die beiden Informationskriterien führen also in den Modellen mit und ohne Indikatoren für die latenten Klassen zu gegensätzlichen Ergebnissen. Letztlich bleibt also auch bei der Berücksichtigung von Indikatoren für die latenten Klassen die Frage der Gültigkeit der Ergebnisse offen. Dies bedeutet aber nicht, daß auf die Anwendung von Mischverteilungsmodellen besser verzichtet werden sollte. Tatsächlich stellt sich bei allen Datenanalysen die Frage der Gültigkeit. Bei vielen Analysemodellen ist dies allerdings nicht so offensichtlich wie bei den Mischverteilungsmodellen.

7 Diskussion

Im Titel dieses Beitrages wird die Frage aufgeworfen, ob Mischverteilungsmodelle das Problem heterogener Daten lösen können. Ich denke mit diesem Beitrag gezeigt zu haben, daß Mischverteilungsmodelle in der Tat eine interessante Erweiterung der statistischen Modellangebote darstellen und insbesondere bei heterogenen Daten hilfreich sein können. Allerdings wäre es nicht sinnvoll, stets Mischverteilungsmodelle anzuwenden, um so einer möglichen Datenheterogenität zu begegnen. Mischverteilungsmodelle können nur dann Datenheterogenität kontrollieren, wenn diese durch unbekannte Subpopulationen hervorgerufen wird und diese Subpopulationen durch die latenten Klassen eines Mischverteilungsmodells identifiziert werden. Mischverteilungsmodelle sind somit vor allem in Situationen angemessen, bei denen von vornherein zu erwarten ist, daß sich eine Stichprobe aus verschiedenen Teilpopulationen zusammensetzt und die Zuordnung der Fälle zu den Teilpopulationen unsicher oder unbekannt ist. In diesem Sinne verstehe ich Mischverteilungsmodelle als eine Ergänzung und Erweiterung des simultanen Gruppenvergleichs.

Unrealistisch erscheint mir dagegen die Vorstellung, mit Mischverteilungsmodellen gänzlich unbekannte Strukturen aus den empirischen Daten zu entdecken. Zwar können durch Mischverteilungsmodelle Subgruppen ohne zusätzliche Informationen identifiziert werden, doch setzt dies bereits Erwartungen über die Eigenschaften dieser Gruppen voraus. Ohne klare Vorerwartungen dürfte die Chance, die einer Stichprobe zugrundeliegende Struktur aufzufinden, praktisch null sein. Wenn ohne Vorerwartungen ein Modell mit latenten Klassen identifiziert würde, wäre dieses Modell auch nicht notwendigerweise das „richtige“

Modell. Aufgrund lokaler Maxima können in der Regel bei einem Datensatz verschiedene Modelle resultieren.

Angemessener erscheinen mir Anwendungen, in denen mit Hilfe von Mischverteilungsmodellen Erwartungen über Zahl und Eigenschaften von Subpopulationen geprüft werden. Findet sich keine Lösung, die den Erwartungen entspricht, kann dies als eine Widerlegung der Erwartungen interpretiert werden. In diesem Sinne sind Mischverteilungsmodelle eher konfirmatorisch als explorativ zu nutzen. So habe ich bei der vorliegenden Analyse der ISSP-Daten mit Mischverteilungsmodellen davon profitiert, daß ich aufgrund der Ergebnisse anderer Studien damit rechnen konnte, daß es zwei Gruppen gibt, die sich in ihren Umweltorientierungen deutlich unterscheiden.

Anwendungsprobleme von Mischverteilungsmodellen ergeben sich vor allem dadurch, daß es oft schwierig ist, eine eindeutige Lösung zu finden. Vor der Schätzung eines Mischverteilungsmodells sollten daher zunächst vorhandene Vorerwartungen expliziert werden. Dabei ist zunächst zu klären, wie viele latenten Klassen spezifiziert werden sollen. Nach meinen Erfahrungen sollte zunächst von einer möglichst geringen Zahl von latenten Klassen ausgegangen werden. Je mehr latente Klassen unterstellt werden, desto schwieriger wird es, eine Lösung zu finden. Für jede der unterstellten Klassen sollte dann überlegt werden, welche Parameterwerte in dieser Klasse zu erwarten sind und wie sich die Werte von denen in anderen Klassen unterscheiden. Am Anfang sollte von einem möglichst restriktiven Modell ausgegangen werden. Zu den Vorüberlegungen gehört auch, sich über die relative Größe der latenten Klassen Gedanken zu machen.

Die Vorüberlegungen sind notwendig, um die Parameter für ein Ausgangsmodell zu spezifizieren, auf dem der Schätzalgorithmus der Mischverteilungsmodelle aufbaut. Falls es in einem Datensatz Variablen gibt, deren Ausprägungen möglicherweise mit den latenten Klassen korrespondieren, ist es sinnvoll, vor der Anwendung eines Mischverteilungsmodells einen Gruppenvergleich durchzuführen. Die Ergebnisse eines solchen Gruppenvergleichs können als Ausgangswerte für ein Mischverteilungsmodell verwendet werden. So habe ich auch bei meinen Analysen die Ergebnisse des simultanen Gruppenvergleichs (Abbildung 3) als erste Startlösung verwendet. Wenn in einem ersten Schritt ein Gruppenvergleich durchgeführt wurde, sollte auch bei der Spezifikation von Mischverteilungsmodellen die Gruppierungsvariable für Indikatoren der latenten Klassen genutzt werden.

Relativ leicht kann es vorkommen, daß der Schätzalgorithmus deswegen zu keiner Lösung kommt, weil die geschätzte Größe einer latenten Klasse null wird. In solchen Situationen kann es hilfreich sein, zunächst die Größen der latenten Klassen zu fixieren. In einer späteren Analyse kann diese Beschränkung dann wieder aufgegeben werden. Nachdem eine Lösung gefunden ist, sollten aus den Modellparametern die modellimplizierten Mittelwerte,

Varianzen und Kovarianzen berechnet werden. Diese können dann als Ausgangslösung für ein unrestringiertes Baseline-Modell herangezogen werden.

Zur Absicherung der Ergebnisse sollte geprüft werden, ob sich bei gleicher Zahl latenter Klassen nicht bessere Lösungen finden lassen. Hinweise kann das unrestringierte Baseline-Modell liefern. So kann versucht werden, durch Schätzung von Strukturgleichungsmodellen auf der Basis der für eine latente Klasse geschätzten Mittelwerte, Varianzen und Kovarianzen eine andere Ausgangslösung zu finden. Eine andere Möglichkeit besteht darin, die geschätzten Werte des Baseline-Modells im Sinne einer größeren Unterschiedlichkeit zwischen den Gruppen geringfügig zu verändern und als neue Ausgangswerte zu verwenden.

Diese auf meinen Erfahrungen beruhenden Hinweise für die Anwendung von Mischverteilungsmodellen verdeutlichen, daß die statistische Datenanalyse mit Mischverteilungsmodellen recht aufwendig ist. Es ist auch nicht garantiert, daß es überhaupt eine sinnvolle Lösung gibt. In vielen Fällen dürfte sich der Aufwand aber lohnen. Alles in allem halte ich die Möglichkeiten von Mischverteilungsmodellen daher für eine fruchtbare Ergänzung zu den klassischen Analysemodellen der empirischen Sozialforschung.

Literatur:

Andreß, Hans -Jürgen , Hagnaars, Jacques A. u. Kühnel, Steffen, 1997: Analyse von Tabellen und kategorialen Daten. Berlin u.a.: Springer.

Bamberg, Sebastian, Kühnel, Steffen M. u. Schmidt, Peter, 1999: The impact of general attitude on decisions. A framing approach. *Rationality and Society*, Jg. 11, S. 5-25.

Diekmann, Andreas u. Preisendörfer, Peter, 1992: Persönliches Umweltverhalten. Diskrepanzen zwischen Anspruch und Wirklichkeit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 44: 226-251.

Franzen, Axel, 1997: Umweltbewußtsein und Verkehrsverhalten. Empirische Analysen zur Verkehrsmittelwahl und der Akzeptanz umweltpolitischer Maßnahmen. Chur u. Zürich: Ruegger.

Kühnel, Steffen u. Bamberg, Sebastian, 1998: Überzeugungssysteme in einem zweistufigen Modell rationaler Handlungen. Das Beispiel umweltgerechten Verkehrsverhaltens. *Zeitschrift für Soziologie*, Jg. 27, S. 256-270.

Muthén, Bengt O., 1989: Latent variable modeling in heterogeneous populations. *Psychometrika*. Jg. 54, S. 557-585.

Muthén, Linda K. u. Muthén, Bengt O., 1998: *Mplus*. User's Guide. Los-Angeles: StatModel.