METHODEN DER SOZIALFORSCHUNG

HERAUSGEGEBEN VON
ANDREAS DIEKMANN

VS Verlag für Sozialwissenschaften

INHALTSÜBERSICHT

Vorwort 7

I. Einleitung

Andreas Diekmann:
Aktuelle Probleme der empirischen Sozialforschung 8

II. Wissenschaftstheoretische Grundlagen

Volker Gadenne:
Empirische Forschung und normative Wissenschaftstheorie. Was bleibt von der Methodologie des kritischen Rationalismus? 33

Walter Krämer:
Statistik: Vom Geburtshelfer zum Bremser der Erkenntnis in den Sozialwissenschaften? 51

Hans-Hermann Dubben und *Hans-Peter Beck-Bornholdt:*
Die Bedeutung der statistischen Signifikanz 61

Jürgen Rost:
Zählen oder Messen? 75

III. Stichproben, Gewinnung von Daten, Datenqualität

Heike Wirth und *Walter Müller:*
Mikrodaten der amtlichen Statistik – Ihr Potenzial in der empirischen Sozialforschung 93

Siegfried Gabler:
Gewichtungsprobleme in der Datenanalyse 128

Sabine Häder und *Axel Glemser:*
Stichprobenziehung für Telefonumfragen in Deutschland 148

Ulrich Pötter:
Probabilistische Selektionsmodelle 172

Norbert Schwarz und *Bärbel Knäuper:*
Kognitionspsychologie und Umfrageforschung: Altersabhängige Kontexteffekte 203

Mick P. Couper und *Elisabeth Coutts:*
Online-Befragung. Probleme und Chancen verschiedener Arten von Online-Erhebungen 217

IV. Sozialer Kontext und Netzwerke

Christof Wolf:
Egozentrierte Netzwerke. Erhebungsverfahren und Datenqualität 244

Frans N. Stokman und *Manuela Vieth:*
Was verbindet uns wann mit wem? Inhalt und Struktur in der Analyse sozialer Netzwerke 274

Uwe Engel und *Julia Simonson:*
Sozialer Kontext in der Mehrebenenanalyse 303

V. Zeitbezogene Daten

Josef Brüderl und *Stefani Scherer:*
Methoden zur Analyse von Sequenzdaten 330

Götz Rohwer:
Verweildauern und Übergangsraten bei mehreren Folgezuständen 348

Henriette Engelhardt:
Kausalanalysen mit separaten und zusammengefassten Zeitreihen 368

VI. Spezielle Probleme der Datenanalyse

Markus Gangl und *Thomas A. DiPrete:*
Kausalanalyse durch Matchingverfahren 396

Ben Jann:
Diagnostik von Regressionsschätzungen bei kleinen Stichproben 421

Udo Kuckartz:
Computerunterstützte Analyse qualitativer Daten 453

Michael Wagner und *Bernd Weiß:*
Meta-Analyse als Methode der Sozialforschung 479

VII. Modellbildung und Simulation

Klaus G. Troitzsch:
Dynamische Systemmodelle 505

Andreas Flache und *Michael W. Macy:*
„Bottom-up" Modelle sozialer Dynamiken. Agentenbasierte Computermodellierung und methodologischer Individualismus 536

Werner Raub und *Vincent Buskens:*
Spieltheoretische Modellierungen und empirische Anwendungen in der Soziologie 560

Die Autorinnen und Autoren 599

English Summaries 605

VORWORT

Die vorliegende Aufsatzsammlung ist die erste in der langen Reihe von Sonderheften der KZfSS, die sich ausschließlich methodischen Problemen der empirischen Sozialforschung widmet. Die Absicht des Buches besteht aber nicht darin, einen Überblick zu dem großen Spektrum methodischer Varianten zu geben, die sich heute im Werkzeugkasten der empirischen Sozialforschung befinden. Vielmehr sind bevorzugt Beiträge aufgenommen worden, die die methodische Diskussion stimulieren und neue Entwicklungen in der Methodenforschung aufzeigen. Sämtliche Beiträge sind Originalbeiträge, und alle Artikel, mit Ausnahme der Einleitung des Herausgebers, wurden einem Begutachtungsverfahren unterzogen.

Natürlich konnten nicht alle neuen und für die Sozialforschung wichtigen Fragen behandelt werden. So hätten sich Beiträge, z.B. über die neuen explorativen Methoden des „Data-Mining" oder die Nutzung von Daten aus Geoinformationssystemen für die Sozialforschung, gut in den Rahmen des Sonderheftes gefügt. Andere Themen waren geplant, die Bearbeitung konnte in angemessener Frist aber nicht mehr realisiert werden. Dazu zählen die Themen: „Konsequenzen rückläufiger Ausschöpfung in der Surveyforschung", „Gütekriterien qualitativer Forschung", „International vergleichende Surveyprogramme" und „Cross Cultural Studies". Der Band hat aber auch in der vorliegenden Form bereits den geplanten Umfang gesprengt. Dass dies klaglos akzeptiert wurde, sei dem Verlag und der Redaktion gedankt.

Mein Dank gilt allen, die zum Gelingen dieses Sonderheftes beigetragen haben: Autorinnen und Autoren, Gutachterinnen und Gutachtern und Herrn Heine von Alemann für die gute Betreuung des Projekts und die abschließende Redigierung der Texte.

Zürich, September 2005 ANDREAS DIEKMANN

I. Einleitung

AKTUELLE PROBLEME DER EMPIRISCHEN SOZIALFORSCHUNG*

Andreas Diekmann

Zusammenfassung: In diesem einleitenden Beitrag werden Grundfragen der Sozialforschung und Probleme neuer Methoden der Datenerhebung, Datenanalyse und Modellbildung behandelt. Die Diskussion bezieht sich vorrangig auf die in diesem Band versammelten Beiträge. Im Schlussteil werden einige Konsequenzen gezogen, wobei für eine Professionalisierung der empirischen Sozialforschung plädiert wird, die kumulativ orientiert ist, klar definierte Standards setzt, die Qualität von Ergebnissen durch Replikationen sichert und systematische Methodenforschung betreibt.

I. Neue Rahmenbedingungen

Vor nahezu vier Jahrzehnten wurde mit dem „Handbuch der empirischen Sozialforschung" ein umfassendes Kompendium methodischer Techniken vorgelegt, mit dem der Herausgeber René König und seine Mitstreiter, namentlich Erwin K. Scheuch, Maßstäbe gesetzt haben. Schon damals schrieb König (1967) im Vorwort, „dass das Gebiet derart ausgedehnt (ist), dass es niemals von einem Autor allein bewältigt werden kann". Diese Behauptung gilt heute zweifellos umso mehr, nachdem sich sowohl die Datenerhebungsverfahren als auch die statistischen Methoden zur Datenanalyse beständig weiter ausdifferenziert haben.

Einen wesentlichen Anteil an der Ausdifferenzierung der Forschungsmethodik hat die rasante technologische Entwicklung. Technische Innovationen sind die treibende Kraft oder zumindest Voraussetzung methodischer Innovationen der vergangenen Jahrzehnte. Rechen- und Speicherkapazität moderner Computer ermöglichen die rasche Verarbeitung riesiger Datenmengen bis hin zu kompletten Volkszählungen. Dadurch erst wurden Individualdaten aus amtlichen Erhebungen für die Sozialforschung attraktiv. Drei wichtige Aspekte des Forschungsprozesses haben sich infolge der technischen Entwicklung erheblich verändert.

Ein erster Aspekt betrifft die Methoden der Datenerhebung und Stichprobenziehung. Diese Methoden profitieren einerseits von technischen Innovationen. Computerunterstützte Telefoninterviews und Face-to-Face-Interviews mit Laptops, also CATI

* Für Anregungen bedanke ich mich bei Heine von Alemann, Jürgen Friedrichs, Ben Jann und Niels Lepperhoff.

und CAPI, sind heute Standard. Weitere Kommunikationskanäle eröffnet das Internet. Viele Umfrage- und Forschungsinstitute experimentieren inzwischen mit Online-Befragungen. Andererseits bereiten technische Fortschritte auch neue Probleme. In der Mobilfunkgesellschaft, in der manche den Festnetzanschluss durch das Handy substituieren, wird es zunehmend schwieriger, Zufallsstichproben aus der Bevölkerung für telefonische Befragungen zu gewinnen. Eine paradoxe Situation, denn vor wenigen Jahren erst wurde die telefonische Befragung aus der Schmuddelecke der „Quick-and-Dirty-Methoden" herausgeholt und dem Face-to-Face-Interview als ebenbürtig oder gar überlegen an die Seite gestellt. Und nun gräbt die technische Entwicklung, die diese Methode mit steigender Telefondichte befördert hat, dem Telefoninterview bei Bevölkerungsbefragungen schon wieder das Wasser ab. Patentlösungen sind offenbar (noch) nicht in Sicht!

Zweitens ist eine gewaltige Kumulation von Datenregistern zu beobachten. Dazu zählt nicht nur die Infrastruktur sozialwissenschaftlicher Forschung mit einer Vielzahl nationaler und internationaler Surveys. Dazu gehören insbesondere auch Volkszählungen, Ehestandsregister, Geburtenstatistik, Sozialversicherungsdaten und vieles mehr. Derartige Daten werden z.B. in den skandinavischen Ländern unter der Personennummer zusammengeführt und anonymisiert für statistische Analysen zur Verfügung gestellt. Das Max-Planck-Institut für Demografie in Rostock „lebt" von der skandinavischen Statistik. Hier werden nicht mehr Zufallsstichproben aus der Bevölkerung gezogen, sondern ganze Bevölkerungen oder wohl definierte Untergruppen der Bevölkerung unter die Lupe genommen (z.B. sämtliche Zwillinge einer schwedischen Geburtskohorte). Dass damit gleichzeitig schwerwiegende Probleme des Datenschutzes angesprochen werden, liegt auf der Hand. Aber auch hierzulande fallen immer mehr prozessproduzierte Daten an, sei es aus staatlichem Verwaltungshandeln, von Transaktionen im Internet bei eBay und Co. oder in Form digitalisierter Texte der Massenmedien. Mit geeigneten Methoden kann es gelingen, anhand dieses reichhaltigen Materials Zusammenhänge zu entdecken, Hypothesen zu prüfen und die Spuren zu deuten, die über kulturelle und gesellschaftliche Entwicklungen Auskunft geben.

Der dritte Aspekt bezieht sich auf die Methoden der Datenanalyse. Ohne die Quantensprünge in der Computertechnik wären viele der heutigen statistischen Schätzverfahren nicht denkbar. Sie beruhen eigentlich auf alten Prinzipien, wie z.B. die Maximum-Likelihood-Schätzmethode, die in den 1920er Jahren ersonnen wurde. Dass diese Methode heute bei einer Vielzahl statistischer Analyseverfahren in unterschiedlichsten Bereichen zugrunde gelegt wird, ist vor allem dem exponentiellen Wachstum der Rechenleistung zu verdanken. Gleiches gilt für Computersimulationen. Zwar werden Simulationsmodelle, man denke etwa an Meadows „Grenzen des Wachstums" (Meadows et al. 1972), bereits seit Jahrzehnten in den Sozialwissenschaften verwendet, doch bieten moderne Rechner ungleich bessere Möglichkeiten. So hat sich insbesondere im Bereich der Mikrosimulation der methodische Ansatz Agenten-basierter Simulation durchgesetzt. Rudimentäre Vorformen dieser Modelle wie das „Segregationsmodell" von Schelling (1971) gibt es seit langem (dazu *Flache* und *Macy* sowie *Troitzsch* in diesem Band). Heute aber werden dank hoher Rechenkapazität ganze Gesellschaften im Computer nachgebaut, um Verkehrsströme, die Ausbreitung von Epidemien und die Folgen von Katastrophen oder gesellschaftlichen Konflikten zu analysieren.

Ob die wachsende Komplexität statistischer Verfahren und Computermodelle immer ein Segen ist, kann man auch skeptisch betrachten. Noch so raffinierte statistische Methoden helfen wenig, wenn die Qualität der erhobenen Daten zu wünschen übrig lässt, lautet der Einspruch von *Walter Krämer* (in diesem Band). Man kann noch hinzufügen, dass alle technischen Möglichkeiten, Datenberge und methodischen Innovationen wenig nutzen, wenn wir nicht mit informativen Theorien, Hypothesen und klugen Fragestellungen an die empirische Arbeit gehen.

Im Folgenden werden zentrale Probleme empirischer Forschungsmethoden mit Blick auf die Beiträge dieses Bandes diskutiert. Abschnitt II diskutiert einige Grundfragen der Sozialforschung. Abschnitt III bezieht sich auf Probleme der Stichprobenqualität und Erhebungsmethoden, Abschnitt IV auf Methoden der Datenanalyse und Abschnitt V auf Probleme der Modellbildung und Simulation. Im Schlussabschnitt VI wird für eine kumulative Sozialforschung plädiert, die sich stärker als heute vielfach üblich an methodischen Standards orientiert und den Problemen der Datenqualität, der Angemessenheit von Analyseverfahren und der Replizierbarkeit von Ergebnissen mehr Aufmerksamkeit widmet.

II. Grundfragen der Sozialforschung

Wissenschaftstheorie und Sozialforschung haben früher enger kommuniziert als dies heute der Fall ist. Diskussionen über „Forschungsprogramme", „theoretische Begriffe", den „Informationsgehalt von Theorien", Probleme der Falsifikation und Bestätigung haben auch ihren Niederschlag in Kontroversen der empirischen Sozialforschung gefunden. *Volker Gadenne* greift den Faden wieder auf, indem er nach der Aktualität der wissenschaftstheoretischen Lehre von Karl R. Popper fragt. „Was bleibt von der Methodologie des kritischen Rationalismus?", lautet die Fragestellung. Die Diskussion kreist um drei Themen: Das Problem der Falsifikation von Theorien, das Problem der „rationalen Akzeptanz" und die Frage nach der Formulierung normativer methodologischer Regeln. Gadenne räumt zunächst mit der weit verbreiteten Fehlinterpretation des kritischen Rationalismus auf, dass Theorien zwar nicht verifizierbar seien, wohl aber durch einen Basissatz endgültig falsifiziert werden könnten. Da die Akzeptanz von Basissätzen von theoretischen und damit potentiell falschen Vorannahmen abhängt, können auch Falsifikationen immer nur als vorläufig gelten. Der Falsifikationismus ist undogmatischer „Fallibilismus". Sicheres Wissen wird nicht geboten. Andererseits impliziert der Falsifikationismus aber auch kein „anything goes". Denn das „hypothetisch-deduktive Prüfverfahren" verlangt „ernsthafte" (die zu prüfende Hypothese muss aus der zu prüfenden Theorie und nicht aus Hilfsannahmen allein folgen) und „kritische Tests" (z.B. die Konfrontation von zwei sich widersprechenden Theorien im Sinne eines „experimentum crucis").

Die nachfolgenden Aufsätze in diesem Band machen aber auch deutlich, dass weder Aussagen zur Falsifikation, noch zur Bewährung von Theorien einfach zu gewinnen sind. Selektive Stichproben, Verzerrungen durch das Erhebungsinstrument, fehlerhafte statistische Interpretationen, das Problem von Drittvariablen und „unbeobachteter Heterogenität" bei der Prüfung von Kausalhypothesen anhand nicht-experimenteller Da-

ten – die Liste potentieller Fehlerquellen empirischer Projekte ist lang. Jeder einzelne Artikel in diesem Band bemüht sich bei einem einzelnen Aspekt um die Formulierung von Regeln, die dazu verhelfen sollen, die Wahrscheinlichkeit fehlerhafter Aussagen zu verringern. Die Anwendung der im Lichte neuer Forschung stets revisionsbedürftigen Regeln ist dann das, was man als Forschung „lege artis" bezeichnet. Wird aber eine Theorie wiederholt und „lege artis" falsifiziert, ist dies noch lange kein Anlass, die Theorie auszumustern. Theorien werden wohl erst dann ad acta gelegt, wenn sie von anderen Theorien überrundet werden.

Ob die Forschung „lege artis" – oder genauer einzelne Regeln und Verfahren – wirklich immer der „Wahrheitsfindung" dienen, wird auch bisweilen höchst kontrovers betrachtet. Diskussionsstoff hierzu liefern die anregenden Artikel von *Hans-Herrmann Dubben* und *Hans-Peter Beck-Bornholdt* zur Signifikanzstatistik sowie der Beitrag von *Walter Krämer* über den Nutzen der Statistik. Probleme statistischen Schließens stellen eine in den Sozialwissenschaften besonders typische Quelle der Unsicherheit bei der Falsifikation von Hypothesen dar, von der in Lehrbüchern der Wissenschaftstheorie eher selten die Rede ist.

Nicht nur wegen theoretischer und damit potentiell falscher Vorannahmen, sondern auch wegen der Logik statistischen Schließens ist eine Falsifikation in den Sozialwissenschaften immer vorläufig. Wenn eine Theorie einen „Effekt" prognostiziert (z.B. einen Zusammenhang zwischen zwei Variablen), eine Prüfung aber zu Gunsten der Nullhypothese spricht, so wird die Theorie mit der Wahrscheinlichkeit β irrtümlich falsifiziert. Denn β ist ja per Definitionem die Wahrscheinlichkeit, dass eine korrekte Alternativhypothese fälschlicherweise abgelehnt wird.

Dubben und Beck-Bornholdt gehen aber viel grundsätzlicher mit dem Missbrauch der Signifikanzstatistik ins Gericht. Die Fehler bei der Planung und der Interpretation von Studien, die sie an illustrativen Beispielen und mit Blick auf die klinische Forschung aufzeigen, findet man nicht minder in der sozialwissenschaftlichen Forschung. Es ist eigentlich nicht so sehr die Neyman-Pearson-Signifikanzstatistik per se, die im Zentrum der Kritik steht, sondern vielmehr die zum Ritual erstarrte Anwendung und die fehlerhafte Interpretation von Testergebnissen, insbesondere von „p-Werten". Statistisch informierte Forscherinnen und Forscher wissen natürlich, dass immer unter der Nullhypothese geprüft wird und p-Werte überhaupt nichts darüber aussagen, mit welcher Wahrscheinlichkeit die tatsächlich interessierende (Alternativ-)Hypothese korrekt ist. Nicht wenige Anwender der Biostatistik wie auch der Statistik in den Sozialwissenschaften haben offenbar nur eine sehr ungenaue Vorstellung von der Logik der Tests und neigen regelmäßig dazu, die Aussagekraft der Prüfverfahren zu überziehen (vgl. auch Krämer in diesem Band). Deshalb sollte auch die Methodenausbildung wesentlich verbessert werden, um weit verbreitete Fehlinterpretationen nicht nur bei Studierenden, sondern auch bei „gestandenen" Forscherinnen und Forschern zu vermeiden. Welche „Power" oder „Macht" ein Test hat, wovon diese abhängt und wie man die Macht eines Tests in einer konkreten Situation berechnen kann, müsste in der Statistik-Ausbildung viel stärker betont werden. Erst wenn man sieht, dass bei vielen Studien, wie Dubben und Beck-Bornholdt betonen, β-Fehler von 0,75 oder 75 Prozent nicht ungewöhnlich sind, wird man die Resultate von Signifikanztests etwas misstrauischer beäugen.

Die Autoren machen aber auch einen konkreten, praktischen Vorschlag. Sie plädieren für ein Untersuchungsdesign mit einem Signifikanzniveau von $\alpha = 0{,}001$ und einer Power ($= 1 - \beta$) von 0,90. Gegen diesen Vorschlag lassen sich allerdings auch Einwände formulieren. Die Power hängt vom Typ des Tests, von α, von der Größe des als bedeutsam angenommenen Unterschieds oder Zusammenhangs und von der Fallzahl ab. Bei Studien über Nebenwirkungen von Medikamenten kann eine Differenz von einem Prozentpunkt zweifellos bedeutsam sein, während sie bei einer bestimmten sozialwissenschaftlichen Hypothese als wenig relevant gelten wird. Um solche kleinen Unterschiede bei den Vorgaben der Autoren ($\alpha = 0{,}001$ und $\beta = 0{,}10$) nachweisen zu können, benötigt man immens hohe Fallzahlen. Sollte man nicht dann besser die Festlegung von α und β vom Ziel der Studie und den Konsequenzen eines Irrtums abhängig machen? Welche Kosten entstehen, wenn ich einen Effekt behaupte, der nicht existiert oder wenn ich einen als bedeutsam angenommenen Effekt übersehe? Die Berücksichtigung solcher Kriterien bei der Untersuchungsplanung, d.h. der Wahl des Signifikanzniveaus und des β-Fehlers bzw. der Fallzahlen, wird in der Entscheidungstheorie seit langem gefordert. Wie auch immer man also den Vorschlag von Dubben und Beck-Bornholdt bewerten wird, er wird sicher Stoff für Diskussionen liefern. Vor allem aber ist ihr Beitrag ein heilsames Gegengift gegen die bedenkenlose Anwendung von Signifikanztests im Alltag der Forschung.

Methoden der Datenanalyse wie die gewöhnliche multiple Regression sind unbestritten eine große Hilfe bei der Untersuchung mutmaßlicher Effekte verschiedener unabhängiger Variablen auf eine abhängige Variable. Aber welchen Mehrwert erzeugen demgegenüber die „Zwei-Stufen-Schätzmethode“ oder weitere Verfeinerungen der Regressionstechniken, fragt Walter Krämer und liefert die Antwort gleich mit. Der Nutzen komplexer mathematisch-statistischer Verfahren, so die These, ist äußerst begrenzt. Allzu oft wird mit ökonometrischen Kanonen auf Spatzen geschossen. Stattdessen sollten Statistiker sich wieder stärker auf die Verbesserung der Datenerhebung konzentrieren. Zahlreiche illustrierende Beispiele machen die These plausibel.

Man kann die zweite These akzeptieren, ohne der ersten in vollem Umfang folgen zu müssen. Zugegeben, manche Verfahren, die sehr restriktive Annahmen erfordern, können mehr Schaden als Nutzen stiften. Aber was ist z.B. mit neuen statistischen Methoden wie die Mehrebenenanalyse, neuen Methoden der Kausalanalyse nicht-experimenteller Daten oder Techniken zur Analyse von Sequenzdaten, die auch in diesem Band behandelt werden? Eine wichtige Botschaft des stimulierenden Essays ist sicher, dass statistische Methoden nicht nur zum Reparaturbetrieb für Defizite und Fehler verkümmern sollten, die bereits bei der Erhebung der Daten entstanden sind. Deshalb muss die Statistik der Messung, der Definition der Begriffe und der Qualität der Datengewinnung wieder mehr Aufmerksamkeit widmen.

Die statistische und ökonometrische Fachliteratur offeriert heute – abhängig vom Typ der Daten und der jeweiligen Modellierung der Datengenerierung – eine Vielzahl von Analysetechniken, die auch für Spezialisten kaum noch überschaubar ist. Dass weniger Gewicht auf Probleme der Messung gelegt wird, mag auch damit zusammenhängen, dass Ökonometriker und Biometriker „harte“ Daten auf Verhältnis- oder Absolutskalenniveau wie Preise, Mengen oder physikalische Größen zugrunde legen. In Soziologie, Psychologie und Politikwissenschaften spielt dagegen die Messung der Intensität

von Einstellungen und anderen nicht direkt beobachtbaren Merkmalen eine ungleich größere Rolle. Um eine Einstellung wie z.B. das Umweltbewusstsein per Befragung zu messen, wird in der Regel eine Batterie von Fragen vorgelegt und aus den Antworten ein Summenscore konstruiert. Mit Konsistenz- und Faktorenanalysen der Korrelationsmuster können sodann Maßzahlen der Reliabilität und Validität ermittelt werden. Diese Art der Einstellungsmessung wird in einem grundlegenden Artikel von *Jürgen Rost* kritisch unter die Lupe genommen. Rost greift dabei auf das Rasch-Modell zurück, das seit langem in der Psychometrie bekannt ist, in die Praxis der Sozialforschung aber kaum Eingang gefunden hat. Ein Vorzug des Modells besteht darin, dass die Summenbildung von Items einer Prüfung unterzogen wird. Weiterhin kann es der Fall sein, dass die Einstellungen der befragten Personen nicht uniform skalierbar sind. Mit so genannten „mixed Rasch-Modellen" können Untergruppen von Befragten identifiziert werden, die einzelnen Items die gleiche Bedeutung zumessen, während diese zwischen den Gruppen variiert. Rost demonstriert die Itemanalyse ausführlich anhand einer Fragebatterie zu normabweichenden Verhaltensweisen aus dem Allbus 2002. Für eine in manchen Routinen erstarrte Sozialforschung ist es sicher heilsam, sich wieder mit grundlegenden Fragen der Messung auseinanderzusetzen.

III. Stichproben und Datenerhebung

In der Sozialforschung werden erhebliche Anstrengungen unternommen, um die Qualität von Stichproben zu sichern. Man muss aber vorausschicken, dass Stichprobenqualität nicht bei allen Fragestellungen von vorrangiger Bedeutung ist. „Repräsentativität" zu erlangen ist keineswegs immer das höchste Ziel der Sozialforschung. Prüft man z.B. eine Kausalhypothese in einem Experiment mit Zufallsaufteilung auf Kontroll- und Versuchsgruppen, dann wird man andere Anforderungen an die Stichprobenziehung stellen als bei allgemeinen Bevölkerungsumfragen. Genügen im ersteren Fall durchaus willkürliche Stichproben, so bemisst sich die Qualität der Stichprobe bei Bevölkerungsumfragen an dem Ideal der Zufallsstichprobe.

Die Orientierung am Ideal der Zufallsstichprobe ist wichtig, wenn deskriptive Schätzwerte der Stichprobe auf Populationen verallgemeinert werden sollen. Dies ist bei deskriptiven Studien fast immer der Fall, denn man interessiert sich ja nicht für die Wahlabsichten oder die Einkommensverteilung in der Gruppe zufällig ausgewählter befragter Personen, sondern für die Werte in der Population. Explorative und hypothesenprüfende Studien sind dagegen nicht unbedingt auf Zufallsstichproben angewiesen. In allen Fällen geht es aber darum, an Hand des empirischen Materials Aussagen zu formulieren, die über Einzelfälle hinausgehen. Ob Exploration, Hypothesenprüfung oder Deskription: Es geht letztlich um die Gewinnung oder Prüfung generalisierender Aussagen.

Die Bedeutung von Zufallsstichproben zur Schätzung von Merkmalen der Population hat einen guten Grund. Denn nur für Zufallsstichproben existiert eine statistische Theorie, die Aussagen über die Qualität von Schätzungen und Fehlermargen erlaubt.

Zufallsstichprobe heißt, dass die Inklusionswahrscheinlichkeit für jedes Element der Population größer als Null und bekannt sein muss. Bei telefonischen Umfragen z.B.

werden per Zufall Telefonnummern von Privathaushalten ausgewählt. Sodann wird die zu befragende Person im Haushalt mit einem Zufallsverfahren ermittelt. Personen in größeren Haushalten haben somit eine geringere Inklusionswahrscheinlichkeit als Personen in Single-Haushalten. Mit dem Stichprobendesign werden die Inklusionswahrscheinlichkeiten ex-ante festgelegt. Fast alle Bevölkerungsumfragen wie die Allgemeine Bevölkerungsumfrage (Allbus), der International Social Survey (ISS), der European Social Survey (ESS) usw. basieren auf einem mehrstufigen Design, z.T. mit zusätzlicher Schichtung. *Siegfried Gabler* spricht in seinem Beitrag ein merkwürdiges Paradox an. In der Praxis der Sozialforschung wird die Design-Gewichtung weitgehend ignoriert. Vertrauensintervalle z.B. werden routinemäßig unter der Annahme einfacher Zufallsstichproben berechnet und damit in der Regel unterschätzt. Warum eine ganze Branche die Fehlermargen zu niedrig ansetzt, sollte schon einmal hinterfragt werden. Dagegen sind Nachgewichtungen (Redressment) zur Anpassung an bekannte Randverteilungen äußerst populär. Im Unterschied zu Designgewichten beruht die Nachgewichtung aber nicht auf einer statistischen Theorie. Ob eine Nachgewichtung systematische Verzerrungen auch bei anderen als den Gewichtungsvariablen verringert, bedarf schon einer genaueren Begründung, die über simple Ad-hoc-Annahmen hinausgehen sollte.

Stichprobenprobleme plagen in je spezifischer Weise alle Befragungsmethoden. Nicht der Typ des Kommunikationskanals – Face-to-Face-Interview, telefonisch, schriftlich, online – bereitet Kopfzerbrechen, sondern die mit einem bestimmten Kommunikationskanal verbundenen Schwierigkeiten der Stichprobenziehung und Teilnahmebereitschaft. Neben Non-Response stellen sich bei der Ziehung von Stichproben für eine telefonische Befragung zwei weitere Probleme: Erstens die Berücksichtigung von Anschlüssen ohne Eintrag im Telefonverzeichnis und zweitens Haushalte mit Mobiltelefon, die nicht mehr über das Festnetz erreichbar sind. Über den Stand der Forschung zur Gewinnung von Stichproben für telefonische Interviews orientieren *Sabine Häder* und *Axel Glemser* in diesem Band.

Einige Stichprobenprobleme wie Non-Response und unvollständige Angaben können aber noch auf einem anderen Wege nach Abschluss der Erhebung angegangen werden: Durch die Modellierung der Antwortwahrscheinlichkeiten unter Nutzung relevanter Stichprobeninformationen. Das Ergebnis kann allerdings stark mit der verwendeten Modellklasse variieren. Immerhin zeigt die Modellierung genauer auf, unter welchen Annahmen ein bestimmtes Schätzverfahren, z.B. die Schätzung des Einkommensmittelwerts unter Weglassung fehlender Angaben, zulässig ist. Über „probabilistische Selektionsmodelle" informiert *Ulrich Pötter* am Beispiel der Schätzung von Einkommensmittelwerten und Konfidenzintervallen für die Allbus-1996-Daten.

Bei Online-Befragungen steht überhaupt das Problem der Stichprobenziehung im Vordergrund. Solange der Internetzugang noch nicht den Verbreitungsgrad des Telefons aufweist, dürfte die ausschließliche Nutzung dieses Kommunikationskanals für Surveys der Wohnbevölkerung noch erhebliche Probleme aufwerfen. Ob Access-Panels für Online-Bevölkerungsbefragungen einen Ausweg bieten, wird in der vorliegenden Literatur kritisch betrachtet (Faas 2003 und Schoen 2004). Anders verhält es sich mit dem Einsatz von Online-Umfragen bei speziellen Populationen oder „Mixed-Mode-Erhebungen", bei denen ein Teil der Personen in der Stichprobe Online antwortet, während die anderen Personen z.B. gebeten werden, einen per Briefpost zugesandten Frage-

bogen auszufüllen. Diese und weitere Stichprobenprobleme von Online-Erhebungen thematisieren in diesem Band *Michael Couper* und *Elizabeth Coutts*.

Von großem Einfluss auf die Datenqualität ist ferner die Qualität des Erhebungsinstruments, bei Interviews also der Fragebogen, bei Online-Interviews, die eine Variante der schriftlichen Befragung darstellen, auch die Präsentation der Fragen. Über die „Lehre von der Frage" und die Konstruktion von Fragebögen, über Kontexteffekte, Interviewereffekte usw. liegt eine umfangreiche Literatur mit Ergebnissen aus hunderten von Methodenexperimenten vor. Woran es fehlt, ist eine Systematisierung der Resultate aus zahlreichen Einzelexperimenten. Die Kognitionspsychologie liefert hierbei eine übergreifende Perspektive zur Integration von Forschungsergebnissen und zur Entwicklung von Hypothesen über Befragtenreaktionen. Die Lehren aus den experimentellen Ergebnissen können dann zur Verbesserung von Befragungsinstrumenten beitragen. Norbert Schwarz, Seymour Sudman und andere verfolgen dieses Forschungsprogramm seit gut zwei Jahrzehnten (Überblick in Sudman, Bradburn und Schwarz 1995). In dem vorliegenden Band präsentieren *Norbert Schwarz* und *Bärbel Knäuper* exemplarisch einige Experimente über Fragereiheneffekte, bei denen altersspezifische Antwortmuster u.a. durch die Funktionsfähigkeit des Arbeitsgedächtnisses erklärbar sind.

Gegen die empirische Umfrageforschung wurde bisweilen das Argument vorgebracht, dass sie nur „atomisierte Individuen" erfasse. Dieses Argument steht schon deshalb auf schwachen Füßen, da mit Umfragen vielfältige soziale Bezüge der befragten Personen erhoben werden können wie z.B. der Familien- und Haushaltskontext, regionale Bezüge, berufliche Einbindungen u.a.m. Darüber hinaus wurde die Umfrageforschung durch die Netzwerkanalyse angeregt, soziale Beziehungsnetze von Befragten in allgemeinen Bevölkerungsumfragen zu berücksichtigen. Die Debatte zur Bedeutung von Sozialkapital hat die Erhebung von Beziehungsnetzen in Umfragen zusätzlich stimuliert. Mit so genannten Namensgeneratoren werden zunächst Bezugspersonen – Freunde, Bekannte, Verwandte – erhoben und sodann die zwischen den genannten Personen bestehenden Beziehungen und ihre Attribute. Die Größe und Art der resultierenden egozentrierten Netzwerke hängt allerdings stark von der Befragungstechnik und insbesondere auch von dem jeweils verwendeten Namensgenerator ab. Experimente mit unterschiedlichen Varianten geben darüber Aufschluss, welche Technik sich für bestimmte Zielsetzungen besonders eignet. *Christof Wolf* setzt sich in diesem Band mit den alternativen Verfahren auseinander und befasst sich mit der Reliabilität und Validität der gewonnenen Daten.

Mit der wachsenden Zahl großer Erhebungen wächst gleichzeitig die Bedeutung der Sekundäranalyse. Die Sozialstrukturanalyse und komparative Forschung hat in den vergangenen zwei Jahrzehnten erheblich von den großen nationalen und internationalen Projekten, beispielsweise vom „Sozioökonomischen Panel" profitiert. Seit einigen Jahren ist eine gewichtige, neue Datenquelle hinzugekommen, die beständig ausgebaut wird. Gemeint sind anonymisierte „Public Use Files" oder „faktisch anonymisierte" Erhebungen der amtlichen Statistik, die der wissenschaftlichen Forschung als Individualdaten („Mikrodaten") zur Verfügung gestellt werden. Eine engere Kooperation zwischen statistischen Ämtern und wissenschaftlichen Nutzern wurde möglich, nachdem das Konzept „faktischer Anonymität von Mikrodaten" mittels empirischer Studien geklärt und präzisiert werden konnte (Müller et al. 1991). Weitere wichtige Anstöße gin-

gen von der „Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik“ (KVI 2001) aus. Daten der amtlichen Statistik sind prozessproduziert wie z.B. die Beschäftigtenstatistik oder werden durch Befragung wie beim Mikrozensus oder der „Einkommens- und Verbrauchsstichprobe“ gewonnen. Bei der Mikrozensus-Erhebung muss man sich über statistischen Aufwand zur Korrektur von Non-Response nicht den Kopf zerbrechen. Die Teilnahme ist gesetzlich vorgeschrieben, die Verweigerungsquote entsprechend gering. Weitere Vorteile sind der große Stichprobenumfang und damit die Möglichkeit, detailliert Teilgruppen der Bevölkerung zu untersuchen, die regelmäßige Wiederholung der Erhebungen und die Verknüpfung mit „Kontextmerkmalen“ (Wirth und Müller in diesem Band). So können beispielsweise Angaben aus der Beschäftigtenstatistik mit Merkmalen des Betriebes verknüpft werden. *Heike Wirth* und *Walter Müller* skizzieren das Potenzial amtlicher Mikrodaten ebenso wie die mit Sekundäranalysen verbundenen Nachteile. Anhand von Mikrozensus, Beschäftigtenstatistik und der Einkommens- und Verbrauchsstichprobe werden die Möglichkeiten der Nutzung auch für Hypothesentests oder zur Untersuchung der differenzierten Wirkung von Arbeitsmarktpolitik aufgezeigt.

Neue Forschungschancen bieten zudem amtliche Daten mit räumlichem Bezug, der über Landeskoordinaten hergestellt wird. So genannte Geoinformationssysteme (GIS-Daten) bilden für die Sozialwissenschaften ein großes Potenzial, das erstaunlicherweise bislang eher selten genutzt wird. Geo-referenzierte Daten sind mehr oder minder kleinräumig zugeordnete Merkmale, z.B. der an einem Wohnblock gemessene Straßenlärm, die Arbeitslosenquote in einer Gemeinde oder die Anzahl von Kindergartenplätzen in der näheren Umgebung. Werden diese Merkmale mit Interviewdaten verknüpft, könnten neue Einsichten über den Einfluss des sozialen Kontextes auf die Einstellungen und Handlungen der Befragten gewonnen werden.

Man kann davon ausgehen, dass die Nutzung von Public-Use-Files der amtlichen Statistik erst am Anfang steht. Dies hat wohl auch damit zu tun, dass nur ein kleiner Teil der Soziologieabsolventen und Absolventinnen bislang über das Know-how verfügt, um diese Forschungschancen überhaupt nutzen zu können.

IV. Probleme der Datenanalyse

Eine grundlegende Annahme der Soziologie ist die Kontextabhängigkeit menschlichen Handelns. Man kann die Annahme, dass das, was Leute denken, für richtig halten und tun, maßgeblich von der sozialen Umgebung beeinflusst wird, fast als ein Axiom der Soziologie und Sozialforschung ansehen. Max Webers Auffassung vom sozialen Handeln, Georg Simmels „Kreuzung sozialer Kreise“, Emile Durkheims „soziale Tatsachen“, Gabriel Tardes „Gesetz der Nachahmung“ weisen doch bei aller Verschiedenheit der Auffassungen auf eines hin, nämlich dass Einstellungen und Verhaltensweisen von Personen durch den „sozialen Kontext“ geprägt werden und auf diesen zurückwirken. Auch hier besteht häufig eine Diskrepanz zwischen Theorie und empirischer Forschung, die die Kontextabhängigkeit sozialen Handelns oftmals ignoriert.

Weiter oben wurde bereits bemerkt, dass Kontextmerkmalen in der Umfrageforschung eine wichtige Rolle zukommt. So gibt z.B. die Erhebung egozentrierter Netz-

werke Aufschlüsse über den sozialen Beziehungsraum von Personen. Weiterhin können Individualdaten aus Befragungen auch „externe" Kontextmerkmale zugespielt werden. So wäre es z.B. möglich, Befragungsdaten mit geo-referenzierten Daten zu verknüpfen, um auf diese Weise Einflüsse der sozialen, kulturellen und physischen Umgebung auf Einstellungen, Verhalten und Befindlichkeiten von Personen zu studieren. Welche Auswirkungen hat der am Wohnort gemessene Lärm auf die Gesundheit? Welchen Einfluss hat die Ausstattung mit Kindergartenplätzen oder die Existenz von Ganztagsschulen in Wohnortnähe auf die Erwerbstätigkeit von Müttern? Durch die Verknüpfung von Befragungsdaten mit raumbezogenen Registerdaten werden Untersuchungen zu derartigen Fragestellungen möglich. Bemerkenswert ist auch, dass in der empirisch-soziologischen Forschung raumbezogene Daten sträflich vernachlässigt werden. Eine weitere Quelle von Kontextmerkmalen sind amtliche Mikrodaten. Wirth und Müller (in diesem Band) machen auf diesen Vorzug einiger großer, amtlicher Erhebungen aufmerksam. Mit den Registerdaten der Beschäftigtenstatistik liegen z.B. Informationen über Erwerbstätige und über den Betrieb vor, in dem eine Person tätig ist oder war.

Um nun Zusammenhänge zwischen Kontextmerkmalen (z.B. Wohnort in einer Gemeinde mit wenig Infrastruktur zur Kinderbetreuung), Individualmerkmalen (Bildung, berufliche Qualifikation usw.) und Handlungsweisen (z.B. die wöchentliche Arbeitszeit für eine bezahlte Erwerbstätigkeit) zu untersuchen, sind einfache Regressionsanalysen nicht ausreichend, da die Annahmen über den Zufallsfehler verletzt sind. Beispielsweise könnte der Einfluss der Bildung auf die Erwerbstätigkeit von Müttern je nach dem Angebot von Betreuungsmöglichkeiten für Kinder, also mit dem Kontext, variieren. Die Methoden der Mehrebenenanalyse erlauben unter bestimmten Bedingungen die korrekte Schätzung derartiger Interaktionen von Kontext- und Individualmerkmalen auf das Verhalten. *Uwe Engel* und *Julia Simonson* skizzieren in diesem Band die Traditionslinien der Kontextanalyse in Anknüpfung an die Typologie von Lazarsfeld und Menzel und erläutern die Techniken der Mehrebenenanalyse am Beispiel von Untersuchungen zur Statusinkonsistenz.

Eine andere Methode, mit sozialem Kontext umzugehen, ist die Netzwerkanalyse. Seit langem existiert ein großes Arsenal formaler Techniken zur Bestimmung struktureller Eigenschaften von Netzwerken wie Zentralität, Dichte, die Anzahl von Cliquen usw. Besonderes Augenmerk gilt seit den Arbeiten von Granovetter (1973) und Burt (1992) den Verbindungen zwischen unterschiedlichen Verdichtungszonen in Netzwerken. Inhaber von „Maklerpositionen" mit Beziehungen zu verschiedenen Kreisen verfügen über mehr Informationen, die z.B. für den beruflichen Aufstieg in Organisationen von Nutzen sind. In dem Beitrag von *Frans Stokman* und *Manuela Vieth* für diesen Band wird allerdings die Einseitigkeit dieser Perspektive kritisiert. Stokman und Vieth argumentieren, dass die gleiche Struktur je nach Bedeutung der Beziehung ganz unterschiedliche Wirkungen entfalten kann. Burts „Maklerpositionen" sind privilegiert, wenn es um den Informationsfluss geht. Zur Produktion gemeinsamer Güter sind hingegen dichte Netzwerke von Vorteil. Die Wirkung der Struktur darf also nicht losgelöst vom Inhalt gesehen werden. Ähnliches gilt für Anwendungen der Netzwerkanalyse in der Tauschtheorie und bei der kollektiven Entscheidungsfindung. Ziel des Beitrags

ist es, verschiedene Netzwerkansätze zu integrieren und die Netzwerkanalyse handlungstheoretisch zu fundieren.

Auch in einem ganz anderen Zusammenhang spielt der Kontext eine Rolle, nämlich bei der Inhaltsanalyse von Texten. Gegen Computer unterstützte quantitative Inhaltsanalysen wurde oftmals eingewendet, dass sie z.B. mit bloßen Frequenzanalysen den Kontext vernachlässigen. Entsprechend wird das Erfordernis kontextbezogener Analysen besonders in der qualitativen Inhaltsanalyse betont. Dass heute die qualitative Inhaltsanalyse von Computer unterstützten Techniken profitieren kann, beschreibt *Udo Kuckartz* in seinem Beitrag für diesen Band.

Immer noch kann die Frage nach der Kausalität von Zusammenhängen als eines der zentralen Probleme der empirischen Sozialforschung angesehen werden. Genauer gesagt lautet die Frage: Wie können kausale Zusammenhänge anhand nicht-experimenteller Daten identifiziert werden? Dieses Problem ist nicht nur von akademischem Interesse. Man möchte wissen, ob Arbeitsmarktmaßnahmen wirklich die Jobchancen verbessern, ob und unter welchen Bedingungen kleinere Schulklassen den Lernerfolg erhöhen, in welchem Ausmaß Tempolimits die Unfallzahlen reduzieren. Die Aufdeckung, Prüfung und Erklärung kausaler Zusammenhänge zwischen sozialen Merkmalen ist das Kerngeschäft einer analytisch-empirischen Sozialwissenschaft (dazu auch *Markus Gangl* und *Thomas DiPrete* in diesem Band). Paul Lazarsfeld hat mit seinen Untersuchungen über Korrelationen und Typen kausaler und nicht-kausaler Zusammenhänge eine Analysetechnik in der Soziologie begründet, die später mit immer komplexeren statistischen Modellen fortgesetzt wurde. In den 1960er und 1970er Jahren wurden diese Ideen mit der Pfad- und Kausalanalyse systematisiert (z.B. Blalock 1985). Im Wesentlichen handelte es sich dabei um multiple Regressionstechniken mit z-standardisierten Skalen. Aber auch multivariate Regressionsschätzungen führen oftmals in die Irre, wenn Stichprobenselektionsfehler und „unbeobachtete Heterogenität" die Schätzwerte verzerren oder sogar zu einer Verkehrung des Vorzeichens führen. Ein Beispiel mag dies verdeutlichen. So zeigte sich in demografischen Studien, dass Ehepartner, die vor der Eheschließung in einem gemeinsamen Haushalt zusammenlebten („Kohabitation"), ein signifikant höheres Scheidungsrisiko aufwiesen als Ehepartner, die erst nach der Heirat zusammenzogen. Erhöht also die „Probeehe" das Scheidungsrisiko? Schon früh wurde der Verdacht geäußert, dass der positive Schätzwert des Kohabitationseffekts durch Selbstselektion verzerrt worden sei. Personen mit einem höheren Scheidungsrisiko hatten auch eine größere Neigung zur Lebensform der Kohabitation. Auch multivariate Schätzungen ergaben zunächst einen positiven Effekt der Kohabitation auf das Risiko einer Ehescheidung. Tatsächlich zeigen aber verschiedene Tests sowie die Verwendung von Schätzmodellen, die Selektionsverzerrungen berücksichtigen, dass der positive „Scheineffekt" verschwindet oder das Vorzeichen des Koeffizienten sich sogar umkehrt (vgl. Brüderl et al. 1997).

Modelle zur Korrektur von Selektionsverzerrungen stellen eine Möglichkeit dar, kausalen Beziehungen auf die Spur zu kommen. Sind die Annahmen nicht erfüllt, können sie aber auch die Krankheit verschlimmern, statt sie zu heilen (vgl. auch *Walter Krämer* in diesem Band). Fehler durch Überdosierung von Ökonometrie sind bisweilen bei dem von Heckman entwickelten Verfahren zur Korrektur von Selektionsfehlern beobachtet worden (vgl. *Ulrich Pötter* in diesem Band). Ein alternativer Weg zur Auf-

deckung kausaler Zusammenhänge wird dagegen mit dem relativ neuen Verfahren der Kausalanalyse durch „Matching" mittels Propensity-Scores beschritten (*Markus Gangl* und *Thomas DiPrete* in diesem Band). Beim „Matching" wird einer Person in der Kontrollgruppe ein „Zwillingspartner", d.h. eine Person mit gleichen Merkmalen bezüglich einer Reihe von Eigenschaften in der Versuchsgruppe zugeordnet. Dieses Verfahren wird seit langem praktiziert. Eine wesentliche Weiterentwicklung besteht jedoch darin, dass die Zuordnung bei den neuen Matching-Methoden mittels statistisch geschätzten Maßen der Ähnlichkeit der Personen erfolgt. Ein solches Maß ist der Propensity-Score, d.h. die bedingte Wahrscheinlichkeit, dass eine Person mit einer bestimmten Merkmalskombination Mitglied der „Versuchsgruppe" ist oder – allgemein gesprochen – die kausal relevante Eigenschaft aufweist. Ein Vorteil von Matching-Methoden gegenüber herkömmlichen Regressionsverfahren besteht darin, dass erstere keine parametrischen Modellannahmen unterstellen, so dass Matching-Methoden relativ robuste Schätzungen kausaler Effekte liefern. Gangl und DiPrete demonstrieren das Verfahren anhand der Auswirkungen eines Arbeitsplatzverlusts auf die nachfolgende Berufskarriere.

In der Statistik hat sich in den letzten Jahren eine lebhafte Diskussion über Konzepte und Definitionen von „Kausalität" entwickelt (zu einem Überblick Engelhardt 2004). Ein prominentes Modell ist das Rubin-Modell der Kausalität, das auf dem Konzept „kontrafaktischer Kausalität" basiert (vgl. Gangl und DiPrete). Dies ist auch die Philosophie, die den Matching-Methoden zugrunde liegt.

Ökonometrische Methoden zur Prüfung von Kausalhypothesen sind Werkzeuge, um Fehlschlüsse bei der Interpretation nicht-experimenteller Daten vermeiden zu helfen. Die Statistik und Ökonometrie ist eine Art nachträglicher Reparaturbetrieb, wenn eigentlich angemessene experimentelle Designs nicht realisierbar sind. Aus der Not wird aber nicht immer eine Tugend. Auch in Soziologie, Ökonomie und Politikwissenschaften gibt es viele grundlegende theoretische Fragestellungen – z.B. zur Entstehung von Normen, Kooperation, Vorurteilen, Guppenverhalten, Sanktionsneigung usw. – die direkt mit experimentellen Methoden, im Labor oder mit Feldexperimenten im sozialen Raum untersucht werden können. Auch paradox, dass die Soziologie, die Begriffe wie Normen und Sanktionen geprägt hat oder jedenfalls zu ihren Grundbegriffen zählt, die experimentelle Erforschung zentraler Hypothesen auf diesem Gebiet weitgehend der experimentellen Ökonomie überlässt. Es gibt aber auch Ausnahmen, wie das folgende Zitat belegt: „Ist Soziologie eine nomographische Wissenschaft, und erhebt sie den Anspruch, Regeln aufzustellen, die in exakter Weise die funktionalen und kausalen Zusammenhänge auf dem Gebiete der sozialen Erscheinungen beschreiben, so muss früher oder später der Zeitpunkt kommen, an dem ihre Probleme experimentell untersucht werden." Das Zitat stammt aus der Feder von Pitirim A. Sorokin. Sein Artikel über eine experimentelle Untersuchung von „Arbeitsleistung und Entlohnung" ist in dieser Zeitschrift im Jahre 1928 erschienen. Manchmal lohnt sich die Besinnung auf die Anfänge.

Ein weiteres Thema der Datenanalyse ist der Umgang mit zeitbezogenen Daten wie Zeitreihen-, Panel-, Ereignis- und Sequenzdaten. Auch Zeitreihendaten eröffnen Möglichkeiten zur Untersuchung kausaler Zusammenhänge. Zur Prüfung von Kausalhypothesen benötigt man aber offenkundig Zeitreihendaten von mindestens zwei Merkma-

len. Da Zeitreihen oftmals einen Trend aufweisen, findet man häufig „Scheinkorrelationen“ zwischen solchen Daten. Zur Lösung dieses Problems gibt es in der Statistik eine ganze Reihe neuer Vorschläge. So kann man Zeitreihen um den Trend bereinigen, indem zunächst nicht-stationäre in stationäre Zeitreihen transformiert werden. Weiterhin wurden Kointegrationsmodelle und entsprechende Tests vorgeschlagen, um zu prüfen, ob langfristige strukturelle Zusammenhänge nachweisbar sind. *Henriette Engelhardt* gibt einen Überblick zu den neuen Entwicklungen und demonstriert sie an Untersuchungen möglicher Zusammenhänge zwischen der Erwerbstätigkeit von Frauen und der Geburtenrate in Deutschland, Italien und Schweden.

Zeitintervalle zwischen Ereignissen wie Heirat und Scheidung, Geburt des ersten und Geburt des zweiten Kindes, Eintritt in eine Firma bis zum Firmenwechsel usw. werden in der sozialwissenschaftlichen Statistik als Ereignisdaten bezeichnet. Ein Ereignisdatum ist ein Informationstripel von Ausgangszustand, Zielzustand und Zeitintervall bis zum Wechsel in den Zielzustand, wobei in der Regel noch Informationen über Kovariate der Beobachtungseinheit hinzukommen. Existieren mehrere „Zielzustände“ – eine Ehe kann z.B. durch eine Scheidung, durch den Tod des Ehemannes oder durch den Tod der Ehefrau beendet werden – spricht man in der Terminologie medizinischer Statistik von „konkurrierenden Risiken“. In der Ereignisanalyse schätzt man nun u.a. Überlebensfunktionen für spezifische Risiken („Pseudo-Survivorfunktionen“, *Götz Rohwer* in diesem Band), wobei die nicht-interessierenden Risiken als „zensiert“ betrachtet werden. Bei der Schätzung der Pseudo-Survivorfunktion für Scheidung geht demnach eine durch Verwitwung aufgelöste Ehe als zensiertes Zeitintervall in die Schätzfunktion ein. Dies führt nach Rohwers Argumentation, die er an konkreten Beispielen mit Ehescheidungsdaten illustriert, zu Interpretationsproblemen. Es fragt sich allerdings, ob Pseudo-Survivorfunktionen nicht doch eine sinnvolle, kontrafaktische Bedeutung verliehen werden könnte. Bei dieser Lesart lautet die Interpretation, dass Pseudo-Survivorfunktionen darüber Auskunft geben, welche Anteile im Ausgangszustand nach einem Zeitintervall verbleiben würden, wenn außer dem interessierenden Risiko alle anderen Risiken hypothetisch ausgeschaltet werden. Rohwer plädiert demgegenüber für einen anderen Weg. Anstelle der Schätzung von Pseudo-Survivorfunktionen schlägt er die Berechnung von Anteilsfunktionen vor. Diese Funktionen informieren über die Anteile der Personen, die nach t Zeiteinheiten in den jeweils betrachteten Zielzustand übergewechselt sind.

Ereignissequenzen sind Abfolgen von Episoden, wobei – anders als bei Ereignisdaten – die Dauer der Episoden ignoriert wird. Berufsverläufe, Familienzyklen, Migrationsgeschichten und andere Arten von Verläufen bilden den Stoff der Sequenzanalyse. Bekannter geworden ist die Methodik in der Soziologie durch eine Studie von Abbott und Hrycak (1990) über die Lebensverläufe deutscher Musiker. Eine Berufskarriere z.B. kann aus verschiedenen Phasen bestehen: Ausbildung, erster Beruf, Arbeitslosigkeit, Weiterbildung, zweiter Beruf usw. Muster ähnlicher Sequenzen können dann zu Clustern gruppiert werden. Weiterhin kann man fragen, welche Bedingungen dazu führen, dass Personen oder allgemein Untersuchungseinheiten mit hoher Wahrscheinlichkeit bestimmten Clustern zugeordnet werden können. Was aber heißt Ähnlichkeit von Sequenzen? Ist die Sequenz BAAB der Sequenz BBBB ähnlicher als der Sequenz ABBA? Zur Bestimmung der Ähnlichkeit werden Techniken aus der biologischen

DNS-Analyse herangezogen. Der Grad der Ähnlichkeit ist grob gesprochen die Anzahl der Schritte, die erforderlich sind, um eine Ausgangssequenz in eine Zielsequenz zu transformieren. Sequenzanalysen sind relativ neu und wurden bislang in den Sozialwissenschaften noch relativ selten angewandt. *Josef Brüderl* und *Stefani Scherer* informieren über die Logik des Verfahrens u.a. am Beispiel der Pluralisierung von Lebensformen.

Sofern einmal die Daten aufbereitet wurden, sind in der Regel statistische Schätzungen auch komplexer Modelle mit den üblichen Softwarepaketen ohne größeren Aufwand durchführbar. Wie steht es aber um die Qualität und Robustheit der Schätzungen? Insbesondere bei kleinen Fallzahlen und nicht-experimentellen Daten stellt sich die Frage, inwieweit man den Schätzergebnissen vertrauen kann. Dabei sind kleine Fallzahlen allein noch nicht das Problem. Hinzu kommt, dass es sich um nicht-experimentelle Daten handelt und ein Interesse an der Prüfung von Kausalhypothesen besteht. Der übliche Weg besteht in der Schätzung multivariater Modelle unter Berücksichtigung entsprechender „Kontrollvariablen". In den Sozialwissenschaften ist das Problem der Schätzung multivariater Modelle anhand kleiner Stichproben insbesondere bei Makroanalysen virulent. Hypothesen über kausale Zusammenhänge zwischen z.B. Sozialkapital, Einkommensungleichheit, Bildungsinvestitionen und wirtschaftlichem Wachstum werden zumeist anhand von Stichproben einzelner Staatengruppen wie OECD-Staaten, EU-Staaten etc. untersucht. Oftmals wird auch die Stichprobe von Makroanalysen durch die teilnehmenden Länder an internationalen Surveys wie dem „European Social Survey (ESS)", dem „International Social Survey Program (ISSP)", dem „World Value Survey (WVS)" oder dem „Fertility and Family Survey (FFS)" begrenzt. Die Stichproben umfassen maximal 200 Länder, sehr häufig aber nur 20 bis 30 Staaten. Eine Vielzahl von Publikationen mit weit reichenden Schlussfolgerungen bis hin zu Empfehlungen für politische Strategien basiert auf solchen Makroanalysen. Bei derart kleinen Stichproben können sich aber die Schätzwerte als recht instabil erweisen. Signifikanztests allein liefern keine genügenden Anhaltspunkte, da bei kleinen Fallzahlen und nicht-experimentellen Daten einzelne Extremwerte, Messfehler oder fehlspezifizierte Modelle hoch signifikante Werte für nicht existierende Effekte zur Konsequenz haben können. Um Fehlschlüsse zu vermeiden, ist eine genaue Diagnostik der geschätzten Modelle anzuraten. Die dafür erforderlichen Methoden werden von Ben Jann in diesem Band vorgestellt und auf Beispiele von Studien aus der Forschungspraxis angewandt. Welches sind die einflussreichen Datenpunkte und welche Auswirkung haben diese auf die Schätzwerte einzelner Parameter? Lassen sich Spezifikationsfehler anhand von partiellen Residuenplots erkennen? Die Beantwortung dieser und weiterer Fragen sowie ein Überblick zu den einzelnen Maßzahlen ist überaus hilfreich bei der Durchführung von Tests, die die Regressionsdiagnostik zur Verfügung stellt.

V. Simulation und Modellbildung

Interessante Entdeckungen ergeben sich oftmals als nicht-intendierte Nebenfolgen von Forschungsvorhaben, die ursprünglich anderen Zielen gewidmet waren. Häufig findet man neue Zusammenhänge erst im Verlauf des Forschungsprozesses. Bestimmte Muster oder Regelmäßigkeiten fallen bei Sichtung des Materials, der Lektüre von Protokollen

und Interviews, bei Beobachtung und Feldarbeit, bei der Inspektion von Grafiken oder der Analyse von Tabellen und Korrelationsmatrizen auf und werden in einem weiteren Schritt als Hypothesen formuliert. Induktive Techniken wie „Data Mining" mit neuronalen Netzen systematisieren diese Strategie der Hypothesengenerierung. Ob die so formulierten Hypothesen Geltung beanspruchen können, kann allerdings erst durch die Prüfung der Hypothesen anhand neuen empirischen Materials entschieden werden.

Ein anderer Weg der Hypothesengenerierung führt über formale Modelle und Computersimulationen. Der Kern eines analytischen oder Simulationsmodells besteht aus einer Menge formalisierter Annahmen. Aus diesen Annahmen werden Hypothesen, und dabei mitunter auch überraschende, neue Zusammenhänge auf deduktivem Wege abgeleitet. Ein oft zitiertes und bereits weiter oben erwähntes Beispiel ist das Segregationsmodell von Thomas C. Schelling (1971). Angenommen werden relativ milde Präferenzen zur Bevorzugung von Nachbarn der eigenen Gruppe bei gleichzeitig weitgehender Toleranz von Fremden. Als Makroeffekt folgt aber die vollständige räumliche Segregation der anfangs zufällig verteilten Mitglieder der beiden Gruppen. Zudem lässt sich mit Modellen experimentieren. Man kann an verschiedenen „Schrauben drehen" und Parameter verändern. Was passiert im Segregationsmodell, wenn die Anfangskonstellation einem bestimmten Muster folgt, die Gruppen ungleich groß sind, das Toleranzniveau verringert wird oder weitere soziale Gruppen auf den Plan treten? Annahmen können also systematisch variiert und die Konsequenzen ausgelotet werden.[1] Formale Modelle und Computersimulationen sind hilfreiche Werkzeuge zur Konstruktion und Formalisierung von Theorien und der Gewinnung prüfbarer Hypothesen, die aus der formalisierten Theorie folgen.

Die Ableitung von Hypothesen erfolgt bei analytischen Modellen, z.B. in der Spieltheorie oder Netzwerkanalyse, mit mathematischen Methoden. Der Vorteil ist, dass eine allgemeine Lösung genau über die Bedingungen Auskunft gibt, unter denen ein bestimmter Effekt zu erwarten ist. Leider ist es bei komplexen Annahmen häufig nicht mehr möglich, eine allgemeine analytische Lösung zu bestimmen. Die Computersimulation ist in diesem Sinne ein Notbehelf, der erst dann zum Einsatz kommen sollte, wenn analytische Methoden versagen. Andererseits machen es die Fortschritte der Computertechnik möglich, fast beliebig komplexe Modelle zu konstruieren und zu analysieren. Das kann bei bestimmten Anwendungen, z.B. der Prognose von Folgen konkreter Maßnahmen der Verkehrsplanung, ein Vorteil sein. Für die sozialwissenschaftliche Theoriebildung hat diese Strategie aber auch Schattenseiten. Bei manchen Problemen gibt ein Globus eine bessere Orientierung als ein Stadtplan. Auch bei Computersimulationen kann eine sparsame Modellierung von Vorteil sein, wenn es darum geht, allgemeine theoretische Erkenntnisse zu erlangen.

Diese und weitere Probleme von Methoden der Computersimulation werden in den Beiträgen von *Klaus G. Troitzsch* sowie *Andreas Flache* und *Michael Macy* angesprochen. Troitzsch gibt anhand von Beispielen aus der Forschung einen Überblick zu verschiedenen Typen von Simulationsmodellen dynamischer Systeme. Ausgehend von Ma-

1 Wer das ausprobieren möchte, sei auf eine wesentlich erweiterte und grafisch sehr schön gestaltete Computersimulation des Schelling-Modells verwiesen, die eine studentische Arbeitsgruppe im Rahmen einer soziologischen Semesterarbeit an der ETH-Zürich erstellt hat. Das Simulationsmodell findet man unter der folgenden Web-Adresse: http://www.socio.ethz.ch/news/index

krosimulationsmodellen in der Tradition der „System Dynamics" hat sich heute das Interesse in Richtung auf Mikrosimulation, insbesondere agentenbasierte Simulationsmodelle, verlagert. Die Makromodelle sind im Kern große Systeme von Differenzengleichungen. Die Werte der Zustandsvariablen werden Zeitschritt für Zeitschritt vom Computer errechnet. Als Lösung des Gleichungssystems erhält man die Trajektorien, die Entwicklungspfade der Zustandsvariablen im Zeitablauf. Wenn man jetzt die Anfangswerte oder bestimmte Parameter verändert, kann man verschiedene Szenarien „durchspielen", wie man sie etwa in den „Grenzen des Wachstums" von Meadows et al. (1972) findet. Diese Modelle kennen aber keine handelnden Akteure. Mit der Entwicklung der Methoden agentenbasierter Simulation hat sich die Situation grundlegend verändert. Ausgangspunkt der Modelle sind zielorientiert handelnde Akteure, die mit anderen Akteuren interagieren. Dadurch können soziale Strukturen, Systeme von interagierenden Akteuren und Formen der Selbstorganisation entstehen, wie sie von Flache und Macy an Beispielen beschrieben werden. Solche Mikrosimulationsmodelle lösen gleichzeitig das Mikro-Makro-Problem. Aus den Aktionen der interagierenden Agenten lassen sich jeweils die Systemzustände oder die Zustände höherer Hierarchieebenen (der soziale Kontext) errechnen, der wiederum auf die Akteure zurückwirkt. Analytische Modelle scheitern rasch an der Komplexität von Mikro-Makro- (Aggregation) und Makro-Mikro-Relationen (Kontexthypothesen), während Simulationsmodelle die Komplexität mühelos verarbeiten können.

Troitzsch spricht in seinem Beitrag von der prognostischen und explanatorischen Funktion von Simulationsmodellen. Man kann hinzufügen, dass diese auch eine heuristische Aufgabe bei der Generierung von Hypothesen erfüllen. Das in dem Beitrag erwähnte Modell von Troitzsch zur Simulation von Wählerpräferenzen bietet ein gutes Beispiel. Zunächst ist die Verteilung unimodal. Durch eine Drift zu Orten größerer Dichte plus eine Zufallsüberlagerung, können unter bestimmten Bedingungen bimodale Verteilungen entstehen. Mit dem Modell kann eventuell herausgearbeitet werden, welche Bedingungen zu einem Prozess der Polarisierung führen. Diese Hypothesen können sodann mit den Methoden der empirischen Sozialforschung an realen Daten überprüft werden.

Wie die Agentensimulation stellt auch die Spieltheorie Werkzeuge zur Verfügung, mit denen Modelle für spezifische Konstellationen der Interaktion zielorientierter Akteure konstruiert werden können. Auf diese Weise können soziologische Grundprobleme wie das Problem sozialer Ordnung, der Umgang mit Vertrauen, sozialer Tausch, verschiedene Formen sozialer Einbettung, die Entstehung von Konventionen, die Befolgung sozialer Normen und vieles mehr präzise dargestellt werden. Im Unterschied zu Simulationsmodellen ermöglichen spieltheoretische Modelle die Ableitung von Hypothesen auf analytischem Wege. Die Standard-Spieltheorie geht dabei von der Annahme strikter Rationalität aus. Unter dieser Annahme kann aus einem Modell die teilspielperfekte Nash-Gleichgewichtsstrategie abgeleitet werden (Raub und Buskens in diesem Band). Gibt es mehrere Lösungen, können weitere Einschränkungen („Verfeinerungen" des Nash-Gleichgewichts) vorgenommen werden. Die Lösung des Spiels ist dann erstens eine Hypothese über das Verhalten der Akteure und zweitens informiert sie über den Makro-Effekt, die kollektive Verteilung der Ergebnisse. Das Konzept des Nash-Gleichgewichts ist in der deskriptiven Spieltheorie somit eine Verhaltenshypothese und

zugleich eine Lösung des Mikro-Makro-Problems. Dieses sehr elegante Konzept hat nur einen Haken. Die strengen Rationalitätserfordernisse sind, wie zahlreiche Experimente zeigen, oftmals nicht erfüllt. Die Agentensimulation ist eine Methode, um spieltheoretische Modelle durch Annahmen begrenzter Rationalität zu erweitern (vgl. Flache und Macy). Eine andere Alternative ist der Ansatz der verhaltensorientierten Spieltheorie („Behavioural Game Theory", Camerer 2003). Bei diesem Ansatz werden die Verhaltensannahmen durch Motive wie Fairness und „Ungleichheitsaversion" ergänzt, die zu einfachen Verhaltensannahmen der Spieltheorie werden sozusagen durch Psychologie angereichert. Beide Alternativen fordern aber auch ihren Preis. Deshalb kann es sich zunächst einmal empfehlen, relativ einfache Modelle unter Standardannahmen zu konstruieren. Erst wenn diese Methoden versagen, d.h. die Verhaltensprognosen systematisch fehlerhaft sind, sollte man zu komplexeren Modellen übergehen.

Werner Raub und *Vincent Buskens* schlagen in ihrem Beitrag für diesen Band die Brücke zwischen spieltheoretischen Modellen und empirischer Forschung. Nach einer Einführung in die „Bausteine" der Spieltheorie rücken sie das Vertrauensproblem in den Mittelpunkt, das für wirtschaftliche Transaktionen grundlegend ist. Sie demonstrieren dabei das Potential der Standard-Spieltheorie für die Gewinnung von Hypothesen über kooperatives oder opportunistisches Verhalten unter Bedingungen sozialer Einbettung. Die aus dem Modell abgeleiteten Hypothesen über Lern-, Kontroll- und Netzwerkeffekte werden anhand einer Befragungsstudie von Angestellten über Einkäufe von Informationstechnologie, einer Vignettenstudie zum Gebrauchtwagenkauf und einer experimentellen Studie über den Effekt von Verpflichtungen überprüft. Der Beitrag schließt mit einem Überblick zu soziologischen Anwendungen der Spieltheorie. Die von ihnen vorgestellten Studien sind Musterbeispiele für eine „theoriegeleitete" empirische Forschung.

VI. Professionalisierung der Sozialforschung

Der Ausbau sozialwissenschaftlicher Infrastruktur, die Nutzung neuer Techniken und systematische Methodenforschung, von der auch die Beiträge in diesem Band zeugen, haben die empirische Sozialforschung auf einen hohen Stand gebracht. Empirische Sozialforschung ist nicht mehr Angelegenheit einer bestimmten „Schule" oder einzelner Institute, sondern alltägliche Praxis und Routine in den Sozialwissenschaften. Die Praxis weist aber auch manche Defizite auf und ist oftmals von einer „best practice" weit entfernt (vgl. auch Kaase 1999). So wäre es wichtig, dass Standards der Datenqualität klar definiert werden. Was wäre z.B. von einem Autohersteller zu halten, der den Benzinverbrauch seiner Autos nach seiner eigenen (optimistischen) Messmethode angibt, statt einer vergleichbaren Norm zu folgen? Genau so verhält es sich aber z.B. mit den Angaben von Befragungsinstituten zur Ausschöpfung einer Stichprobe.

1. Standards zur Verbesserung von Datenqualität

Kommen wir noch einmal auf die Probleme von Stichproben und Non-Response zurück, von denen weiter oben die Rede war. An diesem Beispiel lässt sich aufzeigen, wie schwierig es heute mangels vergleichbarer Standards ist, die Qualität einer Befragung zu beurteilen. Über die Risiken und Nebenwirkungen einer Kopfschmerztablette informiert bekanntlich ein umfangreicher Beipackzettel. Die Inhaltsstoffe eines Joghurtbechers werden fein säuberlich auf der Verpackung notiert, und für jede Schraube oder Eigenschaft eines Geräts existiert eine Industrienorm. Nur der Auftraggeber oder Nutzer einer Umfrage weiß häufig nicht, welche Qualitätsmerkmale das Produkt erfüllt.

Stichprobenumfang und Ausschöpfungsquote sind übliche Kennziffern. Aber diese Angaben können auch in die Irre führen. Werden systematisch verzerrte Stichproben gezogen, so führt die Erhöhung des Stichprobenumfangs dazu, dass sozusagen der falsche Wert mit größerer Genauigkeit getroffen wird. Zwar sollte eine hohe Ausschöpfungsquote systematische Fehler vermindern helfen. Aber was besagt eine von Umfrageinstituten angegebene Ausschöpfung von z.B. 50, 60 oder 70 Prozent, die man gerne erreichen möchte oder die sogar in Verträge hineingeschrieben wird? Am Beispiel der Ausschöpfungsquote ist erkennbar, wie Informationen über Qualitätsmerkmale von Umfragen verzerrt werden, weil die Berechnung keinen exakten Standards folgt.

Die Ausschöpfungsquote ist der Quotient aus der Anzahl auswertbarer Interviews und der um neutrale Ausfälle bereinigten Bruttostichprobe. Doch was ist die Bruttostichprobe und was sind neutrale Ausfälle? Hier gibt es erhebliche gestalterische Möglichkeiten, die von Befragungsinstituten weidlich genutzt werden. Alles was die Bruttostichprobe im Nenner reduziert, erhöht rechnerisch die Ausschöpfung. Ein Beispiel für eine Schönrechnung durch eine äußerst großzügige Definition „neutraler" Ausfälle um immerhin zehn Prozentpunkte wird in der Arbeit von Diekmann (2002) zur methodischen Qualität sozialwissenschaftlicher Forschung erwähnt. Besonders instruktiv sind die Angaben zur „Allgemeinen Bevölkerungsumfrage (Allbus)" von Koch (2002), der uns einen „Blick in die Black Box" der regelmäßigen Allbus-Erhebung erlaubt. Der Allbus basiert auf persönlichen Interviews einer Zufallsstichprobe von ca. 3000 Befragten. 1994, 1996 und 2000 wurde die Stichprobe aus Einwohnermeldedaten gewonnen („Registerstichprobe"). In den anderen Jahren wurden die Stichproben mal per „Adress Random", mal per „Random Route" oder „Standard Random" gezogen. Bei den letzten beiden Verfahren bestimmen die Interviewer gemäß vorgegebenen, aber nicht kontrollierbaren Begehungsregeln, wer in die Bruttostichprobe gelangt. Bei Random Route und Standard Random liegen die Auswahl der Adressen und die Durchführung der Interviews in einer Hand. Dass dabei nicht erlaubte und nicht kontrollierbare Substitutionen erfolgen und entsprechend befragungsbereite Personen mit höherer Wahrscheinlichkeit in der „Bruttostichprobe" enthalten sind, ist offenkundig.[2] Die preislich güns-

2 Beim Familiensurvey 1988 des Deutschen Jugendinstituts wurde ein instruktives Methodenexperiment unternommen. Ein Teil der Stichprobe wurde aus Gemeinderegistern, der andere Teil per Random Route gezogen. Bei der Registerstichprobe mit vorgegebenen Adressen wurden beim ersten Kontaktversuch 25 Prozent der Interviews realisiert, beim Random-Route-Verfahren waren es dagegen 44 Prozent (Alt 1991). Diese Differenz bestätigt die Vermutung, dass die Interviewer bei Random Route Freiheitsgrade bei der Auswahl der Adressen haben und leichter zu befragende Personen bevorzugt in die Stichprobe aufnehmen.

tigen Random-Route-Stichproben ergeben regelmäßig höhere Ausschöpfungsquoten (bis zu 70 Prozent) als die korrekten und teuren Registerstichproben. Bei der Registerstichprobe des Allbus 2000 lag die Ausschöpfungsquote nur noch bei 49,1 Prozent, während der Preis eines Interviews die Rekordhöhe von 250 DM erreichte (Koch 2002). Die Malaise ist aber nicht nur die Täuschung durch übertriebene Ausschöpfungsquoten. Problematisch sind besonders die falschen Anreize, die durch diese Praxis gesetzt werden, wie die Denkschrift der Deutschen Forschungsgemeinschaft zu den „Qualitätskriterien der Umfrageforschung" bemerkt. „Je ‚härter' die Verfahrensregeln für die Bestimmung und Dokumentation des Stichprobenbruttos sind – was unter Qualitätsgesichtspunkten positiv zu bewerten wäre –, umso niedriger ist die errechnete Stichprobenausschöpfung. Da diese als wesentliches Qualitätsmaß gilt, wird hohe Qualitätsorientierung geradezu bestraft" (Kaase 1999: 103f.). Die Registerstichprobe des Allbus 2000 mit einer korrekt berechneten Ausschöpfungsquote von unter 50 Prozent hat möglicherweise eine höhere Datenqualität als eine Random Route Befragung mit erschwindelten 70 Prozent Ausschöpfung.

Dass es auch anders geht, zeigen die genauen Regeln und Definitionen der „American Association of Public Opinion Research" (AAPOR). Nach den Standards der AAPOR (2000) wird im Detail klassifiziert, aus welchem Grund ein Interview nicht stattgefunden hat. Dabei werden selbst exotische Kategorien für zeitweilige Ausfallgründe wir z.B. Trunkenheit oder gar ein Schweigegelübde berücksichtigt. Falls Adressen substituiert werden, gilt der Grundsatz: „Any use of substitutions must be reported" (AAPOR 2000: 10). Nach der detaillierten Kategorisierung wird sodann nach genauen Regeln die Ausschöpfungsquote berechnet. Der Regelfall ist dabei die Registerstichprobe. Wenn nicht vorab eine Liste von Adressen vorliegt (Bruttostichprobe), ist es vermutlich ohnehin wenig sinnvoll, eine Ausschöpfungsquote zu berechnen. Wenn solche Standards, wie sie z.B. die AAPOR vorschlägt, ignoriert werden, sind Ausschöpfungsquoten sinnlose Zahlen, denen man genau so wenig wie einer gefälschten Bilanz trauen sollte.[3]

Die Ausschöpfungsquote ist nur ein Beispiel dafür, dass eine professionelle Sozialforschung Standards benötigt, die verhindern, dass als Qualitätsmerkmale vorgesehene Kennziffern beliebig manipulierbar sind. Standards sind keine Dogmen und offen für Diskussion und Revision. Sie sollen Klarheit schaffen, Vergleiche ermöglichen und generell die Orientierung über Qualitätsmerkmale erleichtern statt durch nebulöse und scheinpräzise Zahlen zu verwirren.

Natürlich lösen Standards noch nicht das Problem schwindender Response-Quoten. Aber auch hier gibt es Möglichkeiten. Bei ehrlicher Berechnung, die wie gesagt eigentlich nur bei vorgegebenen Adressen von Zielpersonen z.B. aus Gemeinderegistern möglich ist, darf man heute bei deutschen Face-to-Face-Bevölkerungsumfragen schon froh sein, wenn die Non-Response-Quote 50 Prozent nicht überschreitet. Wie Mohler et al. (2003) aufzeigen, sind geringe Ausschöpfungsquoten kein Naturgesetz. Der „Produktionsprozess" der Institute, so ihre Vermutung, ist aufgrund mangelnder Nachfrage gar nicht darauf angelegt, Qualitätsstichproben mit hoher Ausschöpfung zu produzieren.

3 „In reporting response rates and all other outcome rates calculated according to the rules and formulas indicated above, researchers must precisely define which rates are being used. For example, a statement that ‚the response rate is X' is unacceptable" (AAPOR 2000: 40).

Dass es bei Optimierung der Feldarbeit auch anders geht, demonstrieren ausländische Vorbilder. Man kann also Qualitätsstichproben erzeugen, wenn man möchte.

In Deutschland gibt es in der wissenschaftlichen empirischen Sozialforschung eine klare Arbeitsteilung, die bei Face-to-Face-Interviews nie durchbrochen wurde. Die akademische Sozialforschung plant Projekte, analysiert Daten und betreibt Methodenforschung, aber die eigentliche Erhebung, die Feldarbeit, wird bei größeren Umfragen von kommerziellen Meinungsforschungsinstituten besorgt. Manche Angehörige der akademischen Sozialforschung plädieren daher für die Etablierung eines nicht-kommerziellen Erhebungsinstituts. Ein solches Institut könnte bei wissenschaftlichen Umfragen Maßstäbe setzen und den Wettbewerb für qualitativ hochwertige Umfragen beleben.

2. Qualitätssicherung von Forschung durch Replikationen

Dank einer ausgebauten Infrastruktur zur Datenarchivierung werden immer mehr Erhebungen zu kollektiven Gütern, auf die interessierte Wissenschaftlerinnen und Wissenschaftler gegen geringe Entgelte Zugriff haben. Nochmals erweitert werden die Datenbestände durch die Kooperation mit der amtlichen Statistik (Wirth und Müller in diesem Band). Damit werden nicht nur Sekundäranalysen zu neuen Themen und Fragestellungen ermöglicht, die die Primärforscher nicht selbst untersucht haben, sondern es besteht auch eine bessere Chance zur Kritik und Kontrolle publizierter Forschungsergebnisse. Systematische Replikationen haben gezeigt, dass ein erheblicher Anteil von Untersuchungen, die in einer Fachzeitschrift erschienen sind, nicht vollständig reproduzierbar waren und oftmals fehlerhafte statistische Analysen enthalten (Dewald et al. 1986; McCullough et al. 2005). In manchen Fällen sind die Fehler nicht gravierend. Häufig kommt es aber auch vor, dass die in einer Studie behaupteten Befunde durch korrekte statistische Reanalysen nicht bestätigt werden können. Bei einer weiteren Kategorie von Arbeiten wiederum existieren unterschiedliche Ansichten über die Angemessenheit alternativer Auswertungsmethoden. Replikationen ermöglichen eine Prüfung der Robustheit von Methoden durch Sekundärforschung und eröffnen die Chance zu einer Diskussion über methodische Zugriffe und Resultate. Fehler und fahrlässige Irrtümer kommen auch aus Nachlässigkeit unter Publikationsdruck zu Stande. Würden Replikationen häufiger als heute üblich durchgeführt, hätte dies zur Konsequenz, dass die Sorgfalt und Qualität der Arbeiten zunehmen würde. Eine professionelle Sozialforschung benötigt eine „Replikationskultur" und sollte diese, wo immer möglich, durch institutionelle Regeln und Normen unterstützen. Aus öffentlichen Mitteln geförderte Projekte sollten verpflichtet werden, die Daten, eventuell nach einer kurzen Sperrfrist, für wissenschaftliche Auswertungen zugänglich zu machen. Gleichzeitig müssen die Förderungsinstitutionen Mittel für eine sorgfältige Dokumentation und Archivierung der Daten bewilligen. Wissenschaftliche Fachzeitschriften müssen verlangen, dass bei Veröffentlichungen die Datenquellen verfügbar sind.[4] Heute bietet sich zudem die Möglichkeit an, dass Primärforscher ihren Datensatz in dokumentierter Form im Inter-

4 Immer mehr Fachzeitschriften fordern von ihren Autoren und Autorinnen, dass die Rohdaten bei Veröffentlichung in dokumentierter Form für Sekundäranalysen zur Verfügung gestellt werden. In Deutschland hat die Zeitschrift für Soziologie kürzlich diese Praxis eingeführt.

net auf ihrer Homepage oder dem Server der Fachzeitschrift zugänglich machen. Neben der Erleichterung des Zugangs müssen auch Anreize für die Durchführung von Replikationen existieren. In der Ausbildung z.B. sollte vermehrt dazu angeregt werden Replikationsstudien zu unternehmen.

Natürlich sollte nicht nur versucht werden, Ergebnisse „intern" zu replizieren. Schon wegen der höheren Wahrscheinlichkeit zur Publikation signifikanter Ergebnisse werden in Fachzeitschriften sehr häufig Artefakte berichtet (dazu auch Dubben und Beck-Bornholdt in diesem Band). Deshalb ist es wichtig, Untersuchungen durchzuführen, die darüber Aufschluss geben, ob die berichteten Befunde an neuen Daten reproduziert werden können. Aus der Wissenschaftstheorie haben wir gelernt, dass Hypothesen wiederholt möglichst strengen Tests unterzogen werden sollen (dazu Gadenne in diesem Band). Wenn in einer Fachzeitschrift ein neuer „Effekt" diskutiert wird, darf diese Nachricht zunächst einmal mit Skepsis aufgenommen werden. Erst wenn sich die Hypothese über einen Effekt wiederholt bewährt hat und keine widersprechenden Ergebnisse vorliegen, wird das Vertrauen in das Resultat wachsen.

3. Die Bedeutung von Metaanalysen

In einigen Bereichen der Sozialforschung, beispielsweise auf Teilgebieten der Demographie, wird heute so verfahren, dass von einer kumulativen Sozialforschung gesprochen werden kann. Wo dies der Fall ist und mehrere Replikationen vorliegen, sind Survey-Artikel hilfreich, um über den Stand der Forschung zu orientieren. Jede Theorierichtung und jedes Forschungsgebiet verschafft sich mit einer oder mehreren Fachzeitschriften Resonanz. Dagegen gibt es nur wenige Publikationsorgane wie z.B. „Annual Review of Sociology", „Annual Review of Psychology" oder das „Journal of Economic Literature", die mit Überblicks-Artikeln die diversifizierten Einzelergebnisse wieder zusammentragen.

Einen Schritt weiter geht die Metaanalyse. Zwar wurden vereinzelt Metaanalysen bereits vor mehreren Jahrzehnten durchgeführt. In der Soziologie ist dieses Verfahren aber noch relativ selten anzutreffen. Während Survey-Artikel einen qualitativen Überblick vermitteln, versucht die Metaanalyse die Ergebnisse der vorliegenden Replikationsstudien mit statistischen Methoden quantitativ zusammenzufassen. *Michael Wagner* und *Bernd Weiß* (in diesem Band) informieren über Grundlagen und Anwendungen dieser Methode. In der Regel wird davon ausgegangen, dass nicht die Rohdaten, sondern nur die Stichprobenschätzwerte der einzelnen Studien wie Mittelwerte, Korrelations- oder Regressionskoeffizienten, Standardabweichungen usw. vorliegen. Man verfügt also über die Informationen, die üblicherweise in einem Fachartikel berichtet werden. Wenn nun z.B. zwanzig Studien Regressionsschätzwerte über den Zusammenhang zwischen der Größe von Schulklassen und der Lernleistung der Schüler ermittelt haben, dann können die Ergebnisse mit einem „Meta-Schätzwert" zusammengefasst werden. Dies ist natürlich nicht einfach der arithmetische Mittelwert, sondern ein gewichtetes Mittel, in das die Fehlervarianzen der einzelnen Schätzungen eingehen. Wagner und Weiß diskutieren auch Stärken und Schwächen der Methode. Ein Problem ist, dass es methodisch und von der Datenqualität her gesehen gute und schlechte Studien

gibt. Die Metaanalyse vermischte damit „Spreu und Weizen" bei der Berechnung eines einzelnen Gesamtwerts (vgl. auch die Kritik von Brüderl 2004 und die Antwort von Wagner und Weiß 2004). Die Autoren diskutieren als Problemlösung ein Gewichtungsverfahren, das methodisch schwächeren Studien ein geringeres Gewicht beimisst als methodisch ausgewiesenen Arbeiten. Bei der Einstufung der methodischen Qualität kommt aber auch ein subjektives Element hinzu, das durch Expertenrating allenfalls abgeschwächt werden kann. Die Bewertung der methodischen Qualität könnte ferner durch eine Kriterienliste von Gütemerkmalen „objektiviert", d.h. einer inter-subjektiven Prüfung zugänglich gemacht werden. Ein weiteres Problem ist der oben erwähnte „publication bias": Signifikante Ergebnisse werden mit höherer Wahrscheinlichkeit publiziert, nicht-signifikante bleiben in der Schublade des Forschers. Die Metaanalyse würde dann nur mehrere Artefakte zu einem einzelnen Artefakt zusammenfassen. Diese Kritik richtet sich allerdings nicht gegen die Metaanalyse an sich, sondern greift ein Problem von Signifikanztests und Publikationsentscheidungen auf. In der Metaanalyse werden zudem Verfahren vorgeschlagen, um einen „publication bias" zu erkennen. Ein einfaches grafisches Verfahren, das die Autoren demonstrieren, basiert auf der Idee, dass die Effektstärken bei größeren Fallzahlen weniger stark streuen sollten als bei kleinen Stichproben. Die grafische Darstellung von Fallzahlen und Effektstärken im Koordinatensystem sollte einem Trichter ähneln, wobei Abweichungen von diesem Muster auf einen „publication bias" hindeuten.

Shang et al. (2005) setzten kürzlich dieses Verfahren bei einer Metaanalyse der Wirksamkeit homöopathischer Heilverfahren ein. Studien von geringerer methodischer Qualität und Studien, die auf kleineren Fallzahlen basierten, sprachen eher für Heilerfolge, die über Placebo-Effekte hinausgehen, als größere Studien. Die Erklärung dafür lautet, dass bei kleineren Studien oftmals das Ergebnis nur dann publiziert wird, wenn es den Erwartungen entspricht und signifikant ist. Bei größeren Studien ist dagegen der Publikationsbias geringer. Größere Studien zeichnen sich damit nicht nur durch den Vorteil kleinerer Standardfehler aus, sondern auch dadurch, dass die Befunde aus solchen Studien seltener in Schubladen verschwinden. Diese Zusammenhänge wären ohne die diagnostischen Methoden der Metaanalyse kaum erkennbar. Die in der Sozialforschung noch recht selten angewandten Diagnostiken sind ein weiteres Argument für die Durchführung von Metaanalysen.

Wagner und Weiß erwähnen auch eine wichtige Voraussetzung für Metaanalysen, die wiederum mit dem Problem der Standardisierung bzw. einer Minimalberichterstattung über Forschungsergebnisse in Fachzeitschriften zu tun hat. Für Metaanalysen werden nämlich die Standardschätzfehler (oder t-Werte) der Koeffizienten benötigt. Sozialwissenschaftliche Zeitschriften haben häufig keine professionellen Standards für die Präsentation von Ergebnissen in Tabellenform. Werden nun – wie häufig noch üblich – nur die Schätzwerte mit „Sternchen" für die Signifikanz angegeben, dann kann die entsprechende Studie in der Metaanalyse nicht angemessen berücksichtigt werden (vgl. Wagner und Weiß in diesem Band). Es empfiehlt sich schon aus diesem Grunde, dass Fachzeitschriften Regeln für eine Minimalberichterstattung empirischer Ergebnisse aufstellen.

4. Methodenforschung

Methodische Standards, Archivierung, Dokumentation und Bereitstellung von Daten für die Sekundärforschung, Replikationen und Metaanalysen, innovative Erhebungsverfahren und Auswertungstechniken bilden das Fundament einer kumulativen und professionellen Sozialforschung. Methodenforschung verhilft dazu, bestehende Standards kritisch zu überdenken und neue Standards zu setzen. Über einige Neuerungen der Methodenforschung wird in den Beiträgen dieses Sonderhefts berichtet. Die Beiträge diskutieren Lösungen für methodische Probleme, werfen dabei aber auch zahlreiche neue Fragen auf. Insbesondere die eingangs skizzierten technischen Entwicklungen verändern massiv die Möglichkeiten sozialwissenschaftlicher Datengewinnung. Forschungsbedarf besteht bei der ganzen Bandbreite des Prozesses der Datenerhebung mit den neuen Medien, vor allem aber bei der Stichprobenproblematik und der Untersuchung von Instrumenteneffekten. Wenn allgemeine Bevölkerungsumfragen auch in Zukunft via Telefoninterview möglich sein sollen, bedarf es dringend einer Lösung für die Konstruktion von Zufallsstichproben, die auch Haushalte berücksichtigen, welche nur über Mobiltelefone erreichbar sind. Zahlreiche Fragen stellen sich bei Online-Erhebungen. Ist die Technik auch geeignet, um valide Antworten für „heikle" Fragen zu erhalten, z.B. bei Viktimisierungssurveys oder bei Erhebungen zu selbst berichteter Delinquenz? Wie und auf welche Weise sind Anonymisierungstechniken für heikle Fragen einsetzbar wie z.B. die „Randomized-Response-Technik"? Führen bewährte Skalen bei Online-Erhebungen zu den gleichen Antwortmustern wie etwa bei Face-to-Face-Interviews?

Diesen und zahlreichen weiteren Forschungsfragen kann mit Methodensplit-Experimenten nachgegangen werden. Kaum irgendwo ist die klassische Methode des Experiments mit randomisierten Versuchsgruppen so einfach anwendbar wie in der Umfrage-Methodenforschung. Einst wurde noch von der „Kunstlehre" der Befragung gesprochen. Und tatsächlich erfordert es eine solide Kenntnis des sozialen Umfelds und semantisches Fingerspitzengefühl, um „gute" Fragen zu konstruieren. Ob die Fragen aber tatsächlich brauchbar sind, d.h. zuverlässige und valide Antworten hervorbringen, kann mit objektiven Methoden überprüft werden. Jede Frageformulierung ist im Prinzip eine Hypothese, die genau wie andere Hypothesen mittels der Methoden empirischer Sozialforschung einer Prüfung unterzogen werden kann. Empirische Sozialforschung ist aber nicht nur Surveyforschung. Zwar nimmt die „Survey-Methodologie"[5] einen zentralen Platz ein. Es sollte aber nicht vergessen werden, dass das Spektrum der Methoden viel weiter reicht und experimentelle Designs sowie Feldexperimente ebenso einschließt wie die Analyse von Textmaterial. Für alle Bereiche gilt, dass Methodenforschung dazu beiträgt, die Qualität von Daten und die Aussagekraft empirischer Analysen zu verbessern.

5 So der Name eines neu bei der DFG beantragten Schwerpunktprogramms, siehe Engel et al. (2005).

Literatur

AAPOR. The American Association for Public Opinion Research, 2000: Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Lenexa, Kansas.

Abbott, Andrew, und *Alexandra Hrycak*, 1990: Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians Careers. American Journal of Sociology 96: 144–185.

Alt, Christian, 1991: Stichprobe und Repräsentativität. S. 497–531 in: *Hans Bertram* (Hg.), Die Familie in Westdeutschland. Stabilität und Wandel familialer Lebensformen, DJI: Familien-Survey 1. Opladen: Leske + Budrich.

Blalock, Hubert M. (Hg.), 1985: Causal Models in the Social Sciences, 2. Aufl. Chicago: Aldine.

Brüderl, Josef, 2004: Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsforschung oder „statistischer Fruchtsalat"? Zeitschrift für Soziologie 33: 84–86.

Brüderl, Josef, Andreas Diekmann und *Henriette Engelhardt*, 1997: Erhöht die Probeehe das Scheidungsrisiko? Eine empirische Untersuchung mit dem Familiensurvey. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 205–222.

Burt, Ronald S., 1992: Structural Holes. The Social Structure of Competition. Cambridge, MA: Harvard University Press.

Camerer, Colin, 2003: Behavioral Game Theory. Experiments on Strategic Interaction. Princeton: Princeton University Press.

Dewald, William G., Jerry G. Thursby, und *Richard G. Anderson*, 1986: Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. American Economic Review 76: 587–603.

Diekmann, Andreas, 2002: Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Institut für Technikfolgen-Abschätzung der Österreichischen Akademie der Wissenschaften: Wien.

Engel, Uwe, Rainer Schnell und *Gert G. Wagner*, 2005: Survey Methodologie und Statistik. Diskussionsgrundlage für das DFG-Rundgespräch am 24. Juni 2005. Manuskript: Universität Bremen.

Engelhardt, Henriette, 2004: Causal Analysis in Population Studies. Theoretical Considerations and Empirical Applications to Fertility. Unveröff. Habilitationsschrift. Universität Bern.

Faas, Thorsten, 2003: Offline-rekrutierte Access-Panels: Königsweg der Online-Forschung? ZUMA-Nachrichten 27/53: 58–76.

Granovetter, Mark S., 1973: The Strength of Weak Ties. American Journal of Sociology 78: 1360–1380.

Kaase, Max (Hg.), 1999: Qualitätskriterien der Umfrageforschung. Memorandum der Deutschen Forschungsgemeinschaft. Berlin: Akademie Verlag.

Koch, Achim, 2002: 20 Jahre Feldarbeit im Allbus. Ein Blick in die Blackbox. ZUMA-Nachrichten 26/51: 9–37.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) (Hg.), 2001: Wege zu einer besseren informationellen Infrastruktur. Baden-Baden: Nomos.

König, René (Hg.), 1967: Handbuch der Empirischen Sozialforschung, Bd I und II. Stuttgart: Enke.

McCullough, B. D., Kerry Anne McGeary und *Teresa D. Harrison*, 2005: Lessons from the JMCB Archive. Journal of Money, Credit, and Banking 37.

Meadows, Dennis, Donella Meadows, Erich Jahn und *Peter Milling*, 1972: Die Grenzen des Wachstums. Bericht des Club of Rome zur Lage der Menschheit. Stuttgart: Deutsche Verlagsanstalt.

Mohler, Peter Ph., Achim Koch und *Siegfried Gabler*, 2003: Alles Zufall oder? Ein Diskussionsbeitrag zur Qualität von face to face-Umfragen. ZUMA-Nachrichten 27/53: 10–15.

Müller, Walter, Uwe Blien, Peter Knoche und *Heike Wirth*, 1991: Die faktische Anonymität von Mikrodaten. Band 19 der Schriftenreihe Forum der Bundesstatistik. Stuttgart: Metzler-Poeschel.

Schelling, Thomas C., 1971: Dynamic Models of Segregation. Journal of Mathematical Sociology 1: 143–186.

Schoen, Harald, 2004: Online-Umfragen – schnell, billig, aber auch valide? Ein Vergleich zweier Internetbefragungen mit persönlichen Interviews zur Bundestagswahl 2002, ZA-Information 54: 27–52.

Shang, Aijing, Karin Huwiler-Müntener, Linda Nartey, Peter Jüni, Stephan Doerig, Jonathan A. C. Sterne, Daniel Pewsner und *Matthias Egger*, 2005: Are the Clinical Effects of Homoeopathy Placebo Effects? Comparative Study of Placebo-Controlled Trials of Homoeopathy and Allopathy. Lancet 366: 726–732.

Sorokin, Pitirim A., 1928: Arbeitsleistung und Entlohnung (Experimentelle Untersuchungen bei Kindern im Alter von 3 – 4 und von 13 – 14 Jahren). Kölner Vierteljahreshefte für Soziologie 7: 186–198.

Sudman, Seymour, Norman M. Bradburn und *Norbert Schwarz*, 1995: Thinking About Answers. The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Wagner, Michael, und *Bernd Weiß*, 2004: Die Bilanz der Scheidungsforschung bleibt eine Bilanz. Eine Antwort auf Josef Brüderl. Zeitschrift für Soziologie 33: 87–89.

II. Wissenschaftstheoretische Grundlagen

EMPIRISCHE FORSCHUNG UND NORMATIVE WISSENSCHAFTSTHEORIE

Was bleibt von der Methodologie des kritischen Rationalismus?

Volker Gadenne

Zusammenfassung: Ist der kritische Rationalismus für die Methodologie der empirischen Sozialforschung noch von Bedeutung? Benötigt die empirische Wissenschaft überhaupt eine normative wissenschaftstheoretische Fundierung? Zur Beantwortung dieser Fragen werden Poppers Regeln der Falsifikation sowie seine Ablehnung der Induktion und des Rechtfertigungsdenkens analysiert. Es soll gezeigt werden, dass einige (obwohl nicht alle) Falsifikationsregeln als nützliche heuristische Prinzipien angesehen werden können. Was die Rechtfertigungsproblematik angeht, lässt sich zeigen, dass die empirische Wissenschaft nicht auf ein methodologisches Prinzip verzichten kann, das es erlaubt, Hypothesen vorläufig als wahr zu akzeptieren. Abschließend werden noch einige Funktionen von methodologischen Regeln und normativen Prinzipien für die empirische Wissenschaft aufgezeigt und diskutiert.

I. Einleitung

Die Entwicklung der modernen Wissenschaftstheorie wurde entscheidend geprägt durch Ideen Karl Poppers, vor allem durch seine Methodologie, die er 1934 in seiner *Logik der Forschung* darlegte und später zum kritischen Rationalismus (kurz KR) weiter entwickelte. In den 1960er Jahren war der KR Gegenstand jener Debatten, an denen unter anderem Thomas Kuhn, Imre Lakatos und Paul Feyerabend mitwirkten. Um diese Zeit bestand auch in den Sozialwissenschaften ein großes Interesse an der Wissenschaftstheorie. 1961 hielt Popper in Tübingen den Eröffnungsvortrag zu einer Tagung, in der es um die Logik der Sozialwissenschaften ging; Theodor W. Adorno hielt das Korreferat. Die Debatte wurde anschließend vor allem durch Hans Albert und Jürgen Habermas in der *Kölner Zeitschrift für Soziologie und Sozialpsychologie* fortgesetzt und ist als *Positivismusstreit* (Adorno et al. 1969) bekannt geworden – eine irreführende Bezeichnung, zu der es dadurch kam, dass der KR von der Gegenseite, den Vertretern der kritischen Theorie, als „positivistisch" eingestuft wurde; tatsächlich war Popper aber seit den 1930er Jahren der wichtigste Kritiker des Positivismus (vgl. zum Positivismusstreit aus späterer Sicht Schmid 1993).

Ab den 1960er Jahren wurden gegen den KR jene Einwände vorgebracht, die schließlich viele davon überzeugten, dass er sich als unhaltbar erwiesen hätte. Zu den bekanntesten Kritikern gehörten Thomas Kuhn und Paul Feyerabend. Manche sehen deren Kritik als so schwerwiegend an, dass sie zum Ergebnis kommen, normative Wissenschaftstheorie hätte sich insgesamt als eher schädlich oder zumindest nicht begründbar erwiesen und sollte ganz durch deskriptive Wissenschaftsforschung ersetzt werden.

Wie ist der KR aus heutiger Sicht zu bewerten? Kann er zu den aktuellen wissenschaftstheoretischen Fragen noch etwas beitragen? Ich denke, dass man dies anhand der folgenden drei Fragenkomplexe diskutieren kann:

– Gibt es ein rationales Verfahren zur empirischen Prüfung von Theorien? Welche Rolle kommt im Rahmen eines solchen Verfahrens Poppers *Falsifikationsidee* zu?
– Benötigt eine Methodologie Regeln zur *rationalen Akzeptanz* oder *Präferenz* von Hypothesen? Erfordert eine Akzeptanz von Hypothesen aufgrund von Beobachtungsergebnissen ein *Induktionsprinzip*? Kann der KR ohne Induktion auskommen?
– Lässt sich *normative Wissenschaftstheorie* überhaupt noch als sinnvoll und gerechtfertigt ansehen? Welche Funktion könnte sie für die Wissenschaften haben?

Es ist klar, dass in einem Artikel nicht alle wissenschaftstheoretischen Fragen behandelt werden können, die für die Sozialwissenschaften von Interesse sind und mit dem KR in Zusammenhang stehen. Zu den hier nicht behandelten Themen gehören z.B. die Werturteilsproblematik, Poppers Situationslogik sowie seine Drei-Welten-Theorie (vgl. zu den beiden letzten Punkten etwa die Debatte in *Angewandte Sozialforschung* 1999/2000: Heft 3/4). Meine Gründe für die Auswahl der drei genannten Problemstellungen bestehen vor allem darin, dass einerseits Poppers bedeutendste Beiträge meines Erachtens auf eben diesen Gebieten liegen, und dass es sich andererseits um Probleme handelt, die sich nicht nur aus der Perspektive des KR stellen, sondern die von allen wissenschaftstheoretischen Positionen aus als grundlegend gelten können.

II. Theorieprüfung und Falsifikation

Erinnern wir uns an die zentralen Punkte von Poppers Lehre, wie er sie in *Logik der Forschung* (im Folgenden kurz *Logik*) darlegte. Er begann mit der Frage, wodurch sich *empirische Wissenschaft* auszeichnet und von den Formalwissenschaften einerseits, von Metaphysik und Pseudowissenschaft andererseits unterscheidet. Die *Verifizierbarkeit* von Theorien mit Hilfe der induktiven Methode kann es nicht sein. Popper verwirft die Induktion mit den Argumenten, die schon David Hume vorbrachte. Statt des induktiven Vorgehens schlug er ein *deduktives Prüfverfahren* vor: Aus der zu prüfenden Theorie T zusammen mit bestimmten Randbedingungen wird eine Prüfaussage P abgeleitet. Wird P als Ergebnis von Beobachtungen akzeptiert, gilt T als bewährt. Wird P verworfen, so gilt T als widerlegt, als *falsifiziert*. Die *Falsifizierbarkeit* ist das gesuchte *Abgrenzungskriterium*: Empirische Wissenschaft formuliert ihre Theorien dergestalt, dass diese mit Beobachtungsresultaten in Widerspruch geraten können, und sie versucht *aktiv*, diese Theorien zu falsifizieren, indem sie *strenge Tests* durchführt. Falsifikationsentscheidungen sind der Motor des Erkenntnisfortschritts, denn sie weisen jeweils

den Weg zur Formulierung besserer Theorien, die mehr erklären können als ihre Vorgänger.

Die Bewährung einer Theorie bedeutet keine Wahrheitsgarantie. Eine solche gibt es nicht. Popper plädiert für einen *Fallibilismus:* Alle menschlichen Erkenntnisversuche gelten als *grundsätzlich fehlbar* und daher *revidierbar.* Beobachtungsaussagen sind davon nicht ausgenommen. Gewissheiten gibt es nicht.

Dass Beobachtungsaussagen trotz ihrer prinzipiellen Fehlbarkeit zum Test von allgemeinen Hypothesen herangezogen werden, hat den Grund, dass man sich über Beobachtungsaussagen leichter einig wird. Wissenschaftler können sich leichter darüber einig werden, dass eine Person bei einer bestimmten Frage eines Fragebogens die Antwort „ja" angekreuzt hat, als über eine allgemeine Hypothese, die etwa besagt, dass Gewalt in Filmszenen bei den Zuschauern die Bereitschaft zu aggressivem Verhalten fördert. Es ist sinnvoll, gemäß der methodologischen Regel zu verfahren, die schwieriger zu beurteilenden Aussagen anhand der leichter zu beurteilenden zu prüfen.

Beobachtungsaussagen gelten als akzeptabel, so lange keine Hinweise vorliegen, dass sie auf einer Täuschung beruhen. Wenn es aber solche Hinweise gibt, dann dürfen auch Beobachtungsaussagen in Zweifel gezogen und anhand neuer Beobachtungen überprüft werden. Poppers Methodologie nimmt an, dass es auch ohne eine absolut gesicherte Beobachtungsbasis möglich ist, aus Beobachtungen zu lernen bzw. sie effektiv zur Korrektur von Hypothesen heranzuziehen.

Zu der Frage, wann genau eine Theorie als falsifiziert gelten soll, äußerte sich Popper in *Logik* nicht ganz klar. Er sprach davon, dass Falsifikationsentscheidungen nicht auf einzelnen Beobachtungen beruhen sollten, sondern eher auf einem empirischen *„Effekt",* wobei er an replizierbare empirische Befunde dachte. Doch legte er sich nicht auf eine bestimmte Zahl von erforderlichen Replikationen fest. Und an einer Textstelle, auf die seine Kritiker später ausgiebig Bezug nahmen, sagte er: „Wir betrachten also im allgemeinen eine (methodisch entsprechend gesicherte) intersubjektiv nachprüfbare Falsifikation als endgültig ..." (1935/1994: 214). Es ist nicht ganz klar, was „methodisch gesichert" heißen soll. Mit Bezug auf die Sozialwissenschaften könnte man diese Forderung so interpretieren, dass der falsifizierende empirische Befund auf einem Untersuchungsplan beruhen muss, der eine dem Problem angemessene Kontrolle von Störvariablen ermöglicht. Eine weitergehende Forderung (die Popper allerdings nicht erhob) wäre, dass der falsifizierende Befund mindestens einmal repliziert werden konnte. Soviel zu der Methodologie in *Logik.*

Zum richtigen Verständnis des KR muss man nun wissen, dass es in ihm eine Entwicklung gibt, die unter anderem auch eine gemäßigtere Auffassung der Falsifikation mit sich brachte. In den 1940er Jahren begann Popper, sozialphilosophische Theorien zu behandeln, und in diesem Zusammenhang entwickelte er die Auffassung, dass man nicht nur empirische, sondern auch *philosophische* (z.B. erkenntnistheoretische, metaphysische, sozialphilosophische) Aussagen rational diskutieren und bewerten kann (vgl. dazu Wettersten 1992). Dies eröffnete ihm die Möglichkeit, den erkenntnistheoretischen *Realismus* und die *Korrespondenztheorie der Wahrheit* zu verteidigen, gegenüber denen er in *Logik* noch eine Neutralität bewahrt hatte. Damit vollzog er den Übergang vom frühen Falsifikationismus, der noch durch die vom logischen Positivismus beeinflusste Distanz zur Metaphysik geprägt war, zum *kritischen Rationalismus.* Letzterer um-

fasst nach Albert (1982: 9ff.) den *Fallibilismus*, den *Realismus* und den *methodischen Rationalismus*. Der Fallibilismus wurde schon erläutert. *Realismus* ist die Position, dass es eine objektive Welt gibt, die unabhängig von unserem Wahrnehmen und Denken existiert und dass wir dazu in der Lage sind, zumindest über Teile dieser Realität ein (fehlbares) Wissen zu erlangen. Der *methodische Rationalismus* ist die Ausdehnung der Idee der kritischen Prüfung, des kritischen Argumentierens auf alle Problemlösungen, sei es in der Wissenschaft, der Moral, dem Recht oder der Politik. Er verlangt, dass man die jeweils vorgeschlagenen Problemlösungen der Kritik unterzieht und vergleicht, welche dabei am besten abschneiden.

Im Rahmen der geschilderten Entwicklung änderte Popper auch seine Auffassung über die Möglichkeit und Strenge von Falsifikationsentscheidungen. Im Postskript zu *Logik*, das Anfang der 1950er Jahre entstand, betonte er, dass es endgültige Falsifikationen *nicht* gäbe (1983: XXIIff.; damit widersprach er wörtlich seiner früheren Aussage, er wollte dies jedoch nur als „Klarstellung" verstanden wissen. Popper tendierte dazu, Änderungen seiner Auffassung herunter zu spielen. Dies hat es begünstigt, dass seine Lehre von vielen auch heute noch irrtümlich mit dem gleichgesetzt wird, was in *Logik* steht). Dies ist auch die Auffassung, die von den bekannten Schülern Poppers bzw. heutigen Vertretern des KR, wie etwa Joseph Agassi, Hans Albert, Gunnar Andersson oder Alan Musgrave vertreten wird. Falsifikationsentscheidungen müssen ebenfalls als revidierbar angesehen werden. Alles andere wäre dogmatischer Falsifikationismus und würde dem Fallibilismus zuwider laufen.

Betrachten wir nun die Kritik am KR. Ein Teil der Einwände nimmt Bezug auf die Wissenschaftsgeschichte (Kuhn 1967; Feyerabend 1976: Kap. 5). Es wird darauf verwiesen, dass bedeutende Wissenschaftler an ihren Theorien festgehalten haben, obwohl diese eindeutig mit anerkannten Fakten in Widerspruch standen. Aus späterer Sicht stellte sich dieses Ignorieren der Fakten als vorteilhaft für den Erkenntnisfortschritt heraus. Nach Poppers Lehre hätten diese Wissenschaftler ihre Theorien jedoch angesichts der widersprechenden Fakten längst aufgeben müssen.

Der größte Teil der Einwände knüpft weiterhin an eine Einsicht an, die schon Pierre Duhem formulierte und die später als Duhem-Quine-These bekannt geworden ist (Duhem 1908/1978: 243ff.; Gadenne 1998). Duhem zeigte, dass bei der Prüfung wissenschaftlicher Theorien zur deduktiven Ableitung einer Prüfaussage P stets eine größere Menge an Aussagen als Prämissen benötigt wird, die neben den Annahmen der Theorie T bestimmte *Hilfsannahmen* H umfassen. Sollte nun die Untersuchung ein negatives Ergebnis erbringen, so widerspricht dies nicht T allein, geschweige denn einer einzelnen Aussage von T, sondern dem ganzen Aussagensystem T & H. Welche Aussage soll man dann als falsifiziert betrachten und verwerfen?

Die Hilfsannahmen können z.B. zum Gegenstand haben, dass ein verwendeter Fragebogen eine bestimmte Variable der Theorie zu erfassen vermag. Bei sozialwissenschaftlichen Fragestellungen muss auch in Betracht gezogen werden, dass die zu prüfenden Aussagen oft *ceteris-paribus-Hypothesen* sind. In diesem Fall gehört zu den vorausgesetzten Hilfsannahmen jeweils eine ceteris-paribus-Annahme, die behauptet, dass in der gegebenen Untersuchungssituation außer den kontrollierten Variablen keine weiteren eine Rolle spielen (Gadenne 1984, 1998). Weiterhin wird die zu prüfende Theorie eventuell statistische Aussagen enthalten oder auf dem Umweg über statistische

Hypothesen geprüft werden, und dann kommen statistische Entscheidungsregeln zur Anwendung. Die damit verbundenen Entscheidungsrisiken bilden, zusätzlich zu den eventuell falschen Hilfsannahmen, eine weitere Quelle möglicher Irrtümer.

Bevor wir entscheiden, ob hieraus ein Einwand gegen den kritischen Rationalismus resultiert, müssen wir uns klar machen, dass es angesichts des Duhemschen Problems durchaus eine systematische Vorgehensweise gibt, die sich der hypothetisch-deduktiven Prüfmethode bedient. Sie garantiert zwar keine kurzfristige Lösung, verspricht aber dennoch einen Fortschritt auf längere Sicht: Man ändert versuchsweise eine bestimmte Annahme von T oder eine bestimmte Hilfsannahme und untersucht, ob nun zutreffende Prüfvorhersagen resultieren. Wir ändern z.B. die in Zweifel gezogene Theorie T_1 und machen daraus T_2. Wir kombinieren beide mit denselben Hilfsannahmen derart, dass aus T_1 & H die Prüfvorhersage P_1 folgt, während aus T_2 & H die von P_1 abweichende Vorhersage P_2 ableitbar ist. Wenn in diesem Fall P_1 verworfen und P_2 akzeptiert wird, so haben wir einen weiteren Hinweis darauf gewonnen, dass T_1 falsch ist. Auf diese Weise kann es einen Fortschritt auch ohne ein Kriterium für eine endgültige Falsifikation geben.

Ohne hier näher auf Lakatos' (1974) *Methodologie wissenschaftlicher Forschungsprogramme* eingehen zu können, sei angemerkt, dass auch sie kein Kriterium für zwingende, nicht revidierbare Falsifikationsentscheidungen anzubieten hat. Im Rahmen dieser Methodologie verschiebt sich das Problem nur auf die Ebene der Forschungsprogramme: Wie viel Kredit soll einem Forschungsprogramm gegeben werden, bis es als *degeneriert* erklärt und aufgegeben werden soll? Dafür gibt es in Lakatos' Methodologie keine genaue Angabe (Feyerabend 1976: 256ff.). Er lässt es zu, dass sich Programme, die als degeneriert erscheinen, wieder erholen.

Mit dem Fallibilismus steht Lakatos' Methodologie insofern nicht in Einklang, als sie vorsieht, die Kernannahmen des Programms (sie entsprechen den Grundgesetzen einer Theorie) zeitweilig für tabu gegen Kritik zu erklären. Doch warum sollte man dies tun? Es ist denkbar, dass ein Wissenschaftler frühzeitig, d.h. so lange das Forschungsprogramm noch lange nicht als degeneriert gilt, eine geniale Idee zur Änderung eines grundlegenden Gesetzes hat. Es ist nicht einzusehen, warum eine solche Änderung abgelehnt werden sollte. Viel sinnvoller erscheint es, sie zuzulassen und den prognostischen Erfolg dieses Vorgehens mit dem Erfolg jener Wissenschaftler zu vergleichen, die es vorziehen, die Grundannahmen zu belassen und Hilfsannahmen zu ändern.

Was folgt nun daraus für die Methodologie des KR? Der in *Logik* enthaltene Falsifikationismus wird, soweit er „endgültige" Falsifikationsentscheidungen vorsieht, von den Einwänden getroffen. Die spätere kritisch-rationale Methodologie sagt dagegen explizit, dass Falsifikationsentscheidungen nicht zwingend oder endgültig sein können, sie wird also durch die Einwände nicht berührt. Doch ist sie dafür mit einem anderen Problem konfrontiert (Kuhn 1974: 16; Feyerabend 1981: 328f.): Wenn es keine zwingenden Falsifikationen gibt, was sagt der KR dann überhaupt aus, das sich von einem „Anything goes" unterscheidet? Was schlägt er an methodologischen Regeln vor, die der Wissenschaft zur Orientierung dienen können?

Da ist zunächst das Prinzip des *Fallibilismus*. Dieses wird durch die angeführte Kritik nicht in Frage gestellt, sondern eher unterstützt. Aber allein hat der Fallibilismus keine sehr spezifischen Konsequenzen, wenn er auch manche Urteile und Vorgehens-

weisen auszuschließen vermag, wie z.B.: „T stellt den endgültigen Stand unserer Wissenschaft dar und braucht nicht weiter überprüft werden." Der Fallibilismus bewahrt vor Dogmatismus, er regt an zu einem empfehlenswerten Maß an Skepsis und zu pluralistischem Denken (Albert 1968: 49). Dennoch ist er allein noch keine Methodologie.

Ein erster Baustein zu einer solchen ist allerdings das in *Logik* vorgeschlagene *hypothetisch-deduktive Prüfverfahren.* Es verlangt, dass 1. aus Theorien Beobachtungsaussagen abgeleitet, 2. Beobachtungen angestellt und 3. Widersprüche zwischen Theorie, Hilfsannahmen und Beobachtungsergebnissen beseitigt werden. Dies sind relativ allgemeine Anforderungen. Nichtsdestoweniger schließen sie ein „Anything goes" deutlich aus.

Darüber hinaus enthält der KR eine Reihe weiterer *methodologischer Regeln.* Jarvie (2001) zählt 15 solcher Regeln auf. Wir können sie hier nicht alle einer detaillierten Betrachtung unterziehen. Ich möchte aber eine bestimmte Regel diskutieren, die mir als die wichtigste erscheint. In *Logik* (1935/1994: 16) wird gefordert, ein zu überprüfendes Annahmensystem *in jeder Weise einer Falsifikation auszusetzen.* An anderer Stelle fordert Popper „ernsthafte" oder „strenge" Tests von Theorien. Was soll dies heißen? Es sind im groben Rahmen des deduktiven Prüfverfahrens verschiedene Strategien des Vorgehens denkbar, und der KR schlägt vor, bestimmte von ihnen zu bevorzugen. Man soll solche Folgerungen aus einer Theorie testen, die im Lichte des *Hintergrundwissens unerwartet* sind, als unwahrscheinlich erscheinen. Er spricht auch von *riskanten* Vorhersagen (1994: 350:

Popper versuchte, diesen Gedanken formal zu definieren (1935/1994: Neuer Anhang, Abschnitt *IX; 1997: 565ff.). Seine Vorschläge haben sich aber als nicht haltbar herausgestellt (Vetter 1967; Kutschera 1972). Eher lässt sich ein qualitativer Begriff der strengen Prüfung einführen. An anderer Stelle (Gadenne 2002) habe ich einen entsprechenden Vorschlag gemacht. Es empfiehlt sich, die Idee der strengen Prüfung in zwei methodologische Regeln aufzulösen. Bei näherer Analyse ergibt sich nämlich, dass man es mit zwei Forderungen zu tun hat, die in der Forschungspraxis eine unterschiedliche Rolle spielen.

Es geht *erstens* darum, dass die Prüfaussage P nicht aus von T verschiedenen Annahmen ableitbar sein sollte, die zum Hintergrundwissen HW gehören. Nur dann handelt es sich um einen *ernsthaften Test.* HW steht hierbei für die Gesamtmenge des zum betreffenden Zeitpunkt akzeptierten Wissens, das sowohl Gesetze als auch singuläre Tatsachen umfasst. Wenn P aufgrund von HW allein, ohne Verwendung von T, vorhergesagt werden kann, dann wird T hier *keinem echten Risiko* ausgesetzt. Dass so etwas zu vermeiden ist, gilt in den empirischen Wissenschaften als selbstverständlich. Unter anderem dienen Kontrolltechniken wie die Elimination, Konstanthaltung und Randomisierung von Störvariablen diesem Zweck. Eine Untersuchung, die sich um die Kontrolle von Störfaktoren nicht kümmert, würde man nicht als wissenschaftlich anerkennen.

Zweitens kann man *kritische Tests* durchführen: Man kann Theorien *gegeneinander* testen, sie empirisch miteinander *konfrontieren.* „Kritisch" soll hier zum Ausdruck bringen, dass eine Folgerung von T gesucht wird, die aus der Sicht einer anderen Theorie bzw. des etablierten Wissens falsch ist. Aus T wird eine Folgerung P gezogen, die einer

konkurrierenden Theorie T' widerspricht, eventuell einer älteren, etablierten Theorie. Ein kritischer Test ist der Idee des „experimentum crucis" verwandt. Es muss sich jedoch nicht um eine experimentelle Testsituation handeln. Außerdem ist das „experimentum crucis" von dem Gedanken getragen, eine der beteiligten Theorien zu beweisen, während beim kritischen Testen davon ausgegangen wird, dass beide Theorien falsch sein könnten.

Im Unterschied zur Durchführung eines adäquaten Untersuchungsplans, der den üblichen Standards genügt und damit einen ernsthaften Test darstellt, handelt es sich bei einem kritischen Test nicht um eine Routineangelegenheit, sondern um etwas, das Phantasie erfordert. In den empirischen Wissenschaften wird ein kritisches Testen nicht ausdrücklich gefordert, im Unterschied zur Durchführung ernsthafter Tests.

Die empirische Konfrontation von Theorien bringt einen großen Gewinn: Das empirische Resultat spricht entweder gegen T, und dies ist ein Gewinn, denn es kann nur von Vorteil sein, die Fehler innerhalb von neuen Annahmensystemen möglichst schnell herauszufinden. Oder das Ergebnis spricht für T und damit gegen T', und in diesem Fall haben wir herausgefunden, dass in einer anderen, eventuell gut etablierten Theorie ein Irrtum enthalten ist. Unter der Voraussetzung des Fallibilismus, dass wir stets mit Irrtümern rechnen müssen, und mit der Zielsetzung, diese aufzufinden und zu beseitigen, kann man ein solches Vorgehen nur positiv beurteilen. Wer überhaupt empirische Forschung betreibt und aus der Empirie lernen will, hat Grund, so vorzugehen.

Wie erscheinen diese Regeln aus der Perspektive der oben dargestellten Kritik am KR? Sie sind von den dort aufgeführten Einwänden nicht betroffen. Diese Regeln fordern nicht dazu auf, einen konträren empirischen Befund sofort zum Anlass zu nehmen, T aufzugeben. Sie zielen wohl darauf ab, konträre Tatsachen zu finden, sei es gegen T oder gegen ein altes Annahmensystem. Ob und wann T zu verwerfen ist, lässt sich aber nicht methodologisch fixieren.

Es zeigt sich, dass der KR auch ohne Kriterium einer gesicherten Falsifikation methodologische Regeln enthält, die der empirischen Forschung Empfehlungen geben können. Die deduktive Methode gibt einen weiten Rahmen vor. Sie muss stets innerhalb der von Duhem beschriebenen Problemsituation operieren, die viele Wege zulässt. Die beiden Regeln, ernsthafte Tests und, wenn möglich, kritische Tests durchzuführen, geben genauere Hinweise. Sie lenken die Phantasie der Forscher beim Ausdenken und Planen von empirischen Untersuchungen in eine bestimmte Richtung, sie sind also heuristischen Prinzipien ähnlich. Albert (1982: 52) fasst Methodologie dementsprechend als *rationale Heuristik* auf.

III. Die Akzeptanz von Theorien und das Induktionsproblem

Auch im Zusammenhang mit der Idee der *Bewährung* sind gegen den KR wichtige Einwände erhoben worden (z.B. Salmon 1968; Schurz 2002). Das Problem, um das es geht, lässt sich folgendermaßen entwickeln: Nehmen wir an, eine Theorie T ist zunächst noch ungeprüft, hat dann einen ersten Test bestanden und somit eine erste Bewährung erlangt, und zu einem noch späteren Zeitpunkt gilt T schließlich als hochgradig bewährt. Haben wir nun am Ende mehr Grund zu der Annahme, dass T wahr ist,

als zu Anfang? Und dürfen wir uns stärker darauf verlassen, dass T zutreffende Vorhersagen macht? Wenn der KR diese Fragen verneint, so muss man anzweifeln, dass die empirischen Tests irgendeinen Zuwachs an Erkenntnis erbracht haben, und die hohe Bewährung von T wäre dann auch ohne jeden praktischen Nutzen. Antwortet der KR aber mit ja, so erlaubt er ein Urteil, das aus der deduktiven Logik allein nicht folgt und offenbar eine Art *induktiven* Schließens darstellt, denn zutreffende Vorhersagen in vergangenen Tests sagen deduktiv-logisch nichts darüber aus, ob zukünftige Vorhersagen mit T zutreffen.

Der KR scheint in diesem Punkt in Schwierigkeiten zu geraten. Insbesondere in der Praxis stellt sich die Frage, ob man sich auf eine bestimmte, bewährte Theorie verlassen soll, oder, falls mehrere Theorien zur Auswahl stehen, an welcher man sich beim Handeln orientieren soll. Im praktischen Leben verlassen sich Menschen auf Hypothesen, die sich in der Vergangenheit bestätigt haben. Wenn wir z.B. nicht aus dem Fenster eines Hochhauses springen, um auf schnellstem Weg nach unten zu gelangen, so hat dies unter anderem den Grund, dass wir die Hypothese für wahr halten, dass solche Stürze lebensgefährlich sind. Wie kann der KR dieser Praxis Rechnung tragen, ohne Prinzipien der rationalen Rechtfertigung und Auswahl einzuführen?

Auch was die Methodologie und den Erkenntnisfortschritt angeht, gibt es ein Problem: Die Annahme, dass das Befolgen bestimmter methodologischer Regeln (wie kritisches Testen) den Fortschritt fördert, dass es zu dem Ziel beiträgt, im Laufe der Zeit unter den nichtfalsifizierten Theorien möglichst viele wahre Theorien zu haben, lässt sich nicht aus der deduktiven Logik ableiten. Welchen Status hat diese Annahme? Ist sie nicht einem induktiven Prinzip ähnlich?

Mit diesen Fragen konfrontiert, äußerte sich Popper folgendermaßen: Er erklärte es als rational gerechtfertigt, unter konkurrierenden Theorien die bestbewährte bis auf Weiteres *für die der Wahrheit nächste* zu halten (1973: 97). Außerdem formulierte er ein Prinzip der *pragmatischen Bevorzugung*: Vorziehen als Grundlage für unsere Handlungen sollten wir jeweils die bestbewährte Theorie (1973: 34). Er sagte weiterhin, Bewährung könne als „Anzeichen" für die „Annäherung an die Wahrheit" dienen (1973: 97). Und er machte auch einmal das Zugeständnis, dass im KR ein „Hauch" von Induktion enthalten sei (1974: 1192).

An anderer Stelle widersprach er dem jedoch und betonte, Bewährung sage „nicht das geringste über die zukünftigen Leistungen oder die ‚Verlässlichkeit' einer Theorie" (1973: 30). Und zu glauben, dass eine Theorie wahr sei, dazu gäbe es „ganz einfach *keinen* Grund" (1983: 67). Dies ist einigermaßen verwirrend. Betrachten wir, wie seine Nachfolger mit dieser Problematik umgehen.

Agassi äußerte als erster den Gedanken, dass die Idee der Bewährung einen Rest von „Rechtfertigungsdenken" darstellt, den es zu überwinden gelte. Auf jeden Fall sei Bewährung in der Grundlagenforschung verzichtbar, allenfalls in der Technologie könne sie eine Rolle spielen (Agassi 1985). Bartley (1987) befürwortete einen „pankritischen" Rationalismus, der die in der Tradition für grundlegend gehaltene Frage „Wie kann man diese Aussage begründen?" schlichtweg für nicht beantwortbar erklärt und durch die Frage ersetzt: „Wie kann man diese Aussage am besten kritisieren?" Rationalität sei mit *Offenheit für Kritik* und kritischem Vorgehen gleichzusetzen.

Am radikalsten wird dieser Lösungsansatz von David Miller (1994, 2002) vertreten. Nach dieser Auffassung gibt der kritische Rationalismus jeglichen Versuch auf, Aussagen zu begründen oder rechtfertigen zu wollen, sei es durch hinreichende oder nicht hinreichende Gründe. Es gibt *keine guten Gründe* für irgendeine Annahme. Auch singuläre (Beobachtungs-)Aussagen können nicht gerechtfertigt werden. Die Klassifikation von Aussagen in wahre und falsche wird nicht dadurch erreicht, dass man bestimmte Aussagen als gestützt oder irgendwie gerechtfertigt erweist. Sie wird erreicht, indem man bestimmte Aussagen ohne Rechtfertigung als falsch klassifiziert. Hypothesen und Theorien werden in den Bereich der empirischen Wissenschaften aufgenommen, wenn sie mit potentiellen Beobachtungsaussagen in Widerspruch geraten können. Wenn eine Hypothese geprüft wird und die Prüfung besteht, geschieht mit ihr nichts. Der positive Ausgang des Tests verändert an ihrem Status nichts. Sie verbleibt im Bereich der Wissenschaft, ohne so etwas wie Stützung oder Glaubwürdigkeit gewonnen zu haben. Wenn sie den Test nicht besteht, wird sie aus dem Bereich der Wissenschaft entfernt. Es ist hierbei wichtig, dass Verfahren zur Kritik und Verwerfung von Hypothesen bei jeder möglichen Gelegenheit ins Spiel gebracht werden: Die offene Aufnahme in den Bereich der Wissenschaft (ohne Forderung irgendeiner Art von Stützung) verlangt rigorose Eliminierungsverfahren. Es gibt durchaus „Wissen" oder „Erkenntnis" in Gestalt von Theorien, nur hat und behält dieses Wissen grundsätzlich den Status von *Ratenversuchen*.

Aber ist diese Lösung des Induktionsproblems zufrieden stellend? Ist sie von einem Skeptizismus noch zu unterscheiden? Sind hier nicht alle Theorien *gleich ungerechtfertigt?* Es gibt zwar angeblich den Unterschied zwischen falsifizierten und nicht falsifizierten Theorien, der die Möglichkeit eröffnet, nicht falsifizierte Theorien den falsifizierten vorzuziehen. Aber auch dieses „Vorziehen" scheint beliebig zu sein, da keine Möglichkeit vorgesehen ist, falsifizierende Beobachtungsaussagen rational zu rechtfertigen.

Und was sagt diese Methodologie zu den Problemen des praktischen Handelns? Man verlässt sich in einem mehr oder weniger hohen Grad auf die Theorie, von der man sein Handeln leiten lässt. Wer eine Brücke baut, einen Ofen beheizt, eine Mahlzeit zubereitet, verlässt sich bewusst oder unbewusst darauf, dass gewisse Gesetzmäßigkeiten weiter bestehen. Kann man dies bestreiten?

Miller (1994: 38ff.) geht diesen Einwand so an, dass er zunächst das seiner Meinung nach naive Bild menschlicher Praxis, von dem hier ausgegangen wird, bestreitet: Die Anwendung von Theorien besteht nicht darin, dass eine bestimmte Theorie als die glaubwürdigste, am besten gerechtfertigte, ausgewählt wird, um dann aus ihr Folgerungen für das Handeln abzuleiten. Wissenschaftliche Theorien geben überhaupt keine Handlungsanweisungen, sondern haben die logische Form von Verboten, sie sagen uns, welche Wege nicht funktionieren. Praxis geht in Wirklichkeit so vor sich: Es werden Vorschläge zur Lösung eines anstehenden Problems gemacht und kritisch diskutiert. Hierbei können wissenschaftliche Theorien in kritischen Argumenten vorkommen. Manche der Vorschläge werden am Ende „überleben", der Kritik besser standhalten als andere. Es ist rational, denjenigen Vorschlägen zu folgen, die der Kritik am besten standhalten. Bei diesem Verfahren werden keinerlei Theorien als „verlässlich" vorausgesetzt, und das Problem einer Rechtfertigung von Theorien stellt sich nicht. Miller hält

alle Einwände, wonach auch eine kritisch-rationale Position auf Induktion nicht völlig verzichten könnte, für ein grundlegendes Missverständnis. Dass so lange Zeit nach Hume manche immer noch diesem Missverständnis erliegen, hält er für einen „Skandal der Philosophie" (2002: 81).

Ich denke, dass es teilweise richtig ist, was Miller über Praxis sagt. Man findet zu einem praktischen Problem im Allgemeinen nicht gleich *die* passende Theorie und deduziert daraus die Lösung. Das Rechtfertigungsproblem taucht aber auch in der Millerschen Schilderung von Praxis wieder auf. Angenommen, jemand macht einen Vorschlag zur Lösung eines praktischen Problems, und dieser wird kritisch diskutiert. Gegen diesen Vorschlag A werden Einwände erhoben, z.B. dass ein geplantes Gebäude nicht stabil genug sei, oder dass eine erwogene Senkung bestimmter Steuern nicht den angestrebten Effekt hätte usw. Es wird ein Gegenvorschlag B gemacht, der diesen Einwänden nicht ausgesetzt ist, und am Ende wird B akzeptiert. In einer solchen Diskussion wird man nun immer gewisse gesetzesartige Annahmen in Anspruch nehmen, die zum Inhalt haben, dass A bestimmte unerwünschte Konsequenzen hat, und dass B diese Konsequenzen nicht hat. Ohne solche Annahmen könnte man kaum gegen A und für B argumentieren. Und auf eben diese Annahmen *verlässt* man sich, wenn man am Ende B als geeigneter befindet.

Alan Musgrave (1993, 1999) vertritt, gegen Bartley und Miller, folgende Auffassung: Wir können Theorien nicht begründen, wir können aber begründen oder rechtfertigen, dass wir einige Theorien glauben bzw. für wahr halten, nämlich jeweils die zu einer bestimmten Frage am besten bewährte Theorie. Nennen wir dies das kritisch-rationale *Rechtfertigungsprinzip* (kurz *R*) für Hypothesen bzw. Theorien: Es ist vernünftig, (zum Zeitpunkt t) diejenige Theorie für wahr zu halten (zu glauben), die (zum Zeitpunkt t) am besten bewährt ist.

Wenn es auf einem bestimmten Gebiet nur eine vorgeschlagene Theorie gibt, die ernsthafte Tests bestanden hat, dann ist sie die bestbewährte Theorie, und man darf sie vorläufig für wahr halten und zur Anwendung empfehlen. Dieses Urteil ist natürlich revidierbar, da sich Gegenbefunde einstellen können. Gibt es mehrere konkurrierende Theorien, so sollten diese kritischen Tests unterzogen werden, bis sich deutliche Unterschiede in der Bewährung zwischen ihnen ergeben. Hat T_2 einige kritische Tests bestanden, an denen T_1 gescheitert ist, so hat T_2 den höheren Bewährungsgrad. Wenn es zwei konkurrierende Theorien gibt, die beide eine Reihe von kritischen Tests bestanden haben, so sollte man ein Urteil aufschieben, bis es durch kritische Tests gelungen ist, falsifizierende Befunde zu einer der beiden zu gewinnen.

Musgrave meint, dass dies die zentrale Idee des kritischen Rationalismus ist und zugleich zusammen fasst, was Popper dazu gesagt hat. Letzteres ist jedoch nicht zutreffend. Die Aussage Poppers, die *R* am nächsten kommt, lautet, dass es gerechtfertigt ist, *die bestbewährte Theorie* für die *der Wahrheit nächste* zu halten. Aus heutiger Sicht erscheint diese Lösung allerdings eher problematischer als diejenige, die Musgrave vorschlägt (und Popper zuschreibt): Es gibt nämlich leider bisher keine überzeugende Theorie der Wahrheitsnähe (vgl. zu diesem Problem Kuipers 1987). Der Begriff der Wahrheit ist weniger problematisch als der der Wahrheitsnähe, und dies ist ein Grund, eher *R* als Poppers Rechtfertigungsprinzip zu akzeptieren.

Warum meinte Popper, dass es in diesem Zusammenhang besser sei, von Wahrheitsnähe als von Wahrheit zu sprechen? Er glaubte, dass sich auf lange Sicht jede Theorie als falsch erweisen wird: Keine Theorie stimmt vollkommen mit der Wirklichkeit überein. Popper (1983: 25f.) verwies darauf, dass selbst Einstein seine spezielle und seine allgemeine Relativitätstheorie nur als Annäherungen betrachtet habe. Wenn nun von keiner Theorie erwartet werden kann, dass sie im vollen Sinne des Wortes wahr ist, kann Fortschritt nur darin bestehen, dass Theorien der Wahrheit näher kommen.

In der Tat müsste man es als unvernünftig ansehen, selbst die bestbewährte Theorie vorläufig für wahr zu halten, wenn es Grund gäbe, jede Theorie für falsch zu halten. Aber wie könnte es hierfür a priori einen Grund geben? Der Fallibilismus impliziert einen solchen Grund nicht. Er besagt, dass sich jede Theorie als falsch herausstellen *kann*, aber er besagt *nicht*, dass sie es wird. Die Wissenschaft hat eine Vielzahl an Theorien entwickelt, die bis heute noch nicht falsifiziert worden sind und die es möglicherweise niemals werden. Auf jeden Fall spricht die bloße Möglichkeit, dass eine Theorie sich als falsch herausstellen könnte, nicht dagegen, eine gut bewährte Theorie vorläufig für wahr zu halten.

Ist *R* ein *induktives* Prinzip? Man mag *R* „induktiv" nennen, weil es Urteile über die Glaubwürdigkeit und Vorzugswürdigkeit von Theorien erlaubt, die nicht aus akzeptierten Beobachtungsaussagen deduktiv ableitbar sind. Es spielt aber genau genommen keine Rolle, welchen Namen man der Sache gibt. Wichtig ist vielmehr, dass es sich *nicht* um Induktion handelt, wie sie in der Erkenntnistheorie überwiegend verstanden und kritisiert worden ist. *R* gehört *nicht* zu den induktiven Verfahren zur Entdeckung von Gesetzen (an die J. S. Mill glaubte). Es ist *kein* induktives Verfahren zum Wahrheitsnachweis von Gesetzen (das es nach Hume auch nicht geben kann). Und es ist *kein* Kalkül einer induktiven Wahrscheinlichkeit (wie Carnaps induktive Logik).

R unterscheidet sich aber auch von Theorien deduktiver Bestätigung, die allein mit Hilfe deduktiv-logischer Beziehungen zu definieren versuchen, unter welchen Bedingungen eine Hypothese relativ zu bestimmten Beobachtungssätzen als bestätigt gelten soll. *R* enthält nämlich ein *pragmatisches* Element, das nicht auf logische Beziehungen zwischen Sätzen zurückführbar ist: Die Bewährung resultiert aus *aktiven Versuchen*, die entsprechende Hypothese oder Theorie zu falsifizieren. Wenn etwas einer Theorie überhaupt Glaubwürdigkeit verleihen kann, dann ist es am ehesten der Umstand, dass sie solchen Widerlegungsversuchen standgehalten hat.

Kann es für *R* noch eine weitere Begründung geben? Muss *R* eventuell ohne Rechtfertigung akzeptiert werden? Worrall (1989) scheint dieser Auffassung zu sein. Musgrave (1999: 330) vertritt, dass *R* zwar gerechtfertigt werden kann, jedoch nur auf zirkuläre Weise: *R* ist akzeptabel, weil *R* der Kritik besser standhält als konkurrierende (induktive) Prinzipien, die bisher vorgeschlagen worden sind. Dieser Rechtfertigungsversuch macht selbst von einem Prinzip Gebrauch, das *R* sehr ähnlich ist: Es ist vernünftig, unter konkurrierenden Vorschlägen (empirischen Annahmen, methodologischen Prinzipien, moralischen Normen usw.) jeweils denjenigen vorläufig zu akzeptieren, der bislang der Kritik am besten standgehalten hat. Dies ist nichts anderes als der oben erwähnte *methodische Rationalismus*, und *R* ist dessen Konkretisierung für die Be-

wertung empirisch-wissenschaftlicher Hypothesen. Musgrave hält auf dieser Ebene der Analyse eine zirkuläre Rechtfertigung für besser als eine dogmatische Akzeptanz.

Kann man *R* eventuell *pragmatisch* rechtfertigen? Eine Rechtfertigung, die auf Nützlichkeit abzielt, ist natürlich ebenfalls zirkulär. Um die zukünftige Nützlichkeit von *R* zu belegen, würde man wiederum ein Prinzip wie *R* benötigen. Aber eine andere pragmatische Rechtfertigung ist vielleicht überzeugender, eine, die zunächst darauf verweist, dass etwas, das praktisch unvermeidlich ist, nicht irrational sein kann: Es ist in der Wissenschaft wie allgemein im Leben unvermeidlich, Handlungsentscheidungen zu treffen, und im Zusammenhang mit solchen Handlungsentscheidungen haben Menschen stets bestimmte Gesetzeshypothesen darüber, was zu was führt. Die Möglichkeit, solche Gesetzeshypothesen nicht zu haben, scheint uns nicht zur Wahl zu stehen. Daher kann es nicht als irrational erklärt werden, sich überhaupt an irgendeiner Hypothese zu orientieren. Nun kommt der zweite Teil der Argumentation: Wenn es richtig ist, dass Menschen immer die eine oder andere Hypothese als die glaubwürdigste akzeptieren, dann erscheint es rational, die Glaubwürdigkeit davon abhängig zu machen, wie intensiv eine Theorie auf Irrtümer hin getestet worden ist. Dass eine Theorie falsche Folgerungen enthält, findet man schließlich im Großen und Ganzen eher heraus, wenn man nach ihnen aktiv sucht, als wenn man Theorien gar nicht testet oder sie nur mit den Beobachtungen konfrontiert, die sich zufällig ergeben. (Dies kann man zu bestreiten versuchen, wird dabei aber argumentativ einen schweren Stand haben.) Wenn wir ohnehin nicht vermeiden können, stets irgendetwas zu glauben, dann doch am besten das, was wir aktiv auf falsche Folgerungen hin überprüft und als haltbar befunden haben. Was sonst sollten wir vernünftigerweise tun?

IV. Braucht die Wissenschaft normative Wissenschaftstheorie?

Der KR schlägt methodologische Regeln vor. Was ist der Status solcher Regeln? Sind es Festsetzungen, Konventionen? Und welche Funktion und Bedeutung haben sie für die Wissenschaften, auf die sie sich beziehen? Brauchen diese Wissenschaften Regeln oder Normen? Brauchen sie normative Wissenschaftstheorie, wie z.B. den KR? In den späten 1960er Jahren neigte man in den Sozialwissenschaften dazu, die letzte Frage zu bejahen. Die Reflexion der Grundlagen einer Disziplin, wie etwa der Soziologie, und ihre wissenschaftstheoretische Fundierung wurden als notwendig und wichtig angesehen. Heute ist dies viel weniger selbstverständlich. Und als Konkurrenz zur normativen Wissenschaftstheorie hat sich, zusätzlich zur Wissenschaftsgeschichte, das interdisziplinäre Unternehmen der *Wissenschaftsforschung* entwickelt (vgl. Felt et al. 1995). Letztere hat sich unter anderem zum Ziel gesetzt, die sozialen Determinanten wissenschaftlicher Tätigkeit und ihrer Ergebnisse aufzuzeigen. Viele Vertreter dieser Disziplin betreiben diese Forschung auf der Grundlage einer *konstruktivistischen* Sichtweise: Wissenschaftler konstruieren ihre „Erkenntnisse“ auf eine Art, die man mit soziologischen oder ethnographischen Methoden erforschen und beschreiben kann. Von vielen wird diese Spielart des Konstruktivismus zum Anlass genommen, in Frage zu stellen, dass Wissenschaft einen Anspruch auf Rationalität und Wahrheitserkenntnis erheben könnte.

Beginnen wir mit der Frage, was der genaue Sinn methodologische Regeln ist. Sind sie den Regeln oder Normen der Moral vergleichbar? Fordern sie kategorisch Handlungen bzw. Unterlassungen wie: ‚Setze Hypothesen bzw. Theorien kritischen Prüfversuchen aus.' ‚Betrachte eine Theorie niemals als endgültig bewiesen.' ‚Falls empirische Ergebnisse einer Theorie T widersprechen, so verteidige T nicht durch unprüfbare Ad-hoc-Annahmen' usw.? Die entsprechenden Forderungen müssten noch auf bestimmte Arten von Problemsituationen bezogen werden, in diesem Fall auf die Planung von Hypothesentests und auf das weitere Vorgehen, wenn empirische Resultate einer Theorie widersprechen. Für eine derartige Situation wird eine bestimmte Vorgehensweise oder Methode M gefordert. Die Norm oder Regel hätte demnach die Form: Wähle in Situation S die Vorgehensweise M.

Es ist jedoch klar, dass diese Interpretation der Natur wissenschaftlicher Erkenntnissuche nicht gerecht würde. Eine Forderung wie ‚Setze Hypothesen kritischen Prüfversuchen aus' macht keinen Sinn, wenn man sie als kategorischen Imperativ versteht, in Analogie zu einer Norm wie ‚Sage stets die Wahrheit'. Warum soll man Hypothesen kritisch prüfen? Jemand, der nicht so verfährt, handelt keineswegs unmoralisch, er handelt allenfalls (aus der Sicht einer bestimmten wissenschaftstheoretischen Auffassung) nicht optimal im Hinblick auf bestimmte Ziele. Dies legt es nahe, methodologische Normen oder Regeln als *hypothetische Normen* aufzufassen, die sich auf eine bestimmte Zielsetzung beziehen. Angenommen, Wissenschaft hat ein Ziel oder eine Reihe von Zielen. Es kann etwa darum gehen, wahre Theorien von hoher Erklärungskraft zu erlangen, oder empirisch adäquate Theorien mit hohem Vorhersagewert. Bezeichnen wir eine derartige Zielsetzung als Z. Dann können wir eine methodologische Regel so auffassen: Wenn du das Ziel Z hast, dann wähle in Situation S die Vorgehensweise M.

Hans Albert (1982: 52ff.) interpretiert dementsprechend Methodologie als *Technologie*. Er betont hierbei (siehe oben), dass die Regeln dieser technologischen Disziplin keinen sehr strengen Charakter haben können. Sie sind *heuristischer* Natur: Sie geben Empfehlungen, zur Lösung des jeweiligen Problems eine bestimmte Richtung einzuschlagen, legen den Lösungsweg aber nicht im Detail fest.

Wer methodologische Regeln auf diese Weise auffasst, setzt voraus, dass bestimmte *methodologische Hypothesen* wahr sind. Die Empfehlung „Teste Theorien kritisch, um zu wahren Theorien zu gelangen" macht nur dann Sinn, wenn man annimmt, dass ceteris paribus kritisches Testen von Theorien dazu beiträgt, möglichst viele wahre Theorien zu erlangen. Allgemein gesprochen, wenn wir M als Mittel für Z vorschlagen, nehmen wir an, dass M zu Z beizutragen vermag. Falls es konkurrierende Methoden M_1 und M_2 gibt, die beide dazu entworfen wurden, zu Z beizutragen, schlagen wir M_1 genau dann vor, wenn wir M_1 für effektiver halten als M_2. Unser Vorschlag beruht dann auf der *methodologischen Hypothese*, dass M_1 im Hinblick auf Z effektiver ist als M_2. Eine solche Hypothese kann selbstverständlich falsch sein.

Methodologische Regeln bzw. Empfehlungen dieser Art beziehen sich auf *Vorgehensweisen*. Eine Wissenschaftstheorie kann nun nicht nur aus solchen Verfahrensregeln und den zugrunde liegenden Hypothesen bestehen. Sie benötigt auch Regeln zur *Bewertung der Ergebnisse*, z.B. ein Prinzip (wie das oben diskutierte), das etwas darüber sagt, wann es vernünftig ist, eine Aussage zu akzeptieren. Ohne solche Bewertungsprinzipien, d.h. nur mit Empfehlungen für das Vorgehen ausgestattet, würde eine wissen-

schaftstheoretische Position nichts darüber aussagen, unter welchen Bedingungen es *rational gerechtfertigt* ist, etwas zu glauben bzw. für wahr zu halten. Und damit bliebe auch völlig offen, wann jemand Wissen oder Erkenntnis besitzt. Aus methodologischen Regeln, die lediglich Vorschläge für das Vorgehen im Forschungsprozess machen, folgt nicht logisch, dass bestimmte Resultate einen Fortschritt bzw. eine Erkenntnis darstellen.

Damit ist grob skizziert, was hier unter *normativer Wissenschaftstheorie* als Disziplin verstanden werden soll: Sie befasst sich mit a) Regeln für das Vorgehen, die zweckrationalen und heuristischen Charakter haben, b) Regeln zur Bewertung von Ergebnissen und c) methodologischen Hypothesen, die den Regeln zugrunde liegen. Normative Wissenschaftstheorie schlägt entsprechende Regeln bzw. Hypothesen vor und unterzieht sie einer kritischen Diskussion. Hierzu werden unter anderem auch Erkenntnisse der empirischen Wissenschaften selbst herangezogen (z.B. über Wahrnehmungsforschung). Dass es möglich ist, methodologische Regeln bzw. die entsprechenden Hypothesen kritisch zu diskutieren und mit Argumenten zu befürworten, wurde oben anhand der Regel des kritischen Testens sowie anhand des Prinzips *R* demonstriert.

Was ist nun die Funktion von Methodologie für die empirischen Wissenschaften? „Was nützt mir Wissenschaftstheorie, um mein konkretes (fachwissenschaftliches) Problem zu lösen?“ So lautet eine Frage, die trotz ihrer offensichtlichen Unsinnigkeit nicht selten provokativ an die Wissenschaftstheorie gerichtet wird. Natürlich ist ihr Beitrag nicht auf dieser Ebene zu suchen. Aber worin besteht er? Kann sie die Wissenschaften unterstützen oder sie kritisieren und verbessern? In diesem letzten Abschnitt möchte ich einige Funktionen der normativen Wissenschaftstheorie aufzeigen.

1. In allen empirischen Wissenschaften werden *epistemische Bewertungen* vorgenommen. Beispiele für entsprechende Urteile (von Wissenschaftlern, nicht von Wissenschaftstheoretikern), die zum Tagesgeschäft gehören, sind etwa: Dieses Argument ist wenig überzeugend. Jene Theorie vermag nicht alle wichtigen empirischen Befunde zu erklären. Theorie T ist hochgradig bewährt. Theorie T_1 wird durch die empirischen Resultate besser gestützt als T_2. Diese Daten sind mit einem systematischen Messfehler behaftet. Dieser Versuchplan erlaubt nicht die Aussage, dass die Veränderung der unabhängigen Variablen die Ursache der Veränderung der abhängigen Variablen ist.

Es ist unverkennbar, dass solchen Urteilen bestimmte *methodologische Regeln* bzw. *Bewertungsprinzipien* zugrunde liegen, von eben der Art, wie sie in der Wissenschaftstheorie explizit diskutiert werden. In den Einzelwissenschaften ist es eher nicht üblich, grundlegende erkenntnis- oder wissenschaftstheoretische Aspekte zu reflektieren und zum Thema von Diskussionen zu machen. Sie haben dort den Status von Konventionen, die als selbstverständlich gelten. Beispielsweise wird die methodologische Annahme, dass Beobachtungen überhaupt notwendig und dazu geeignet sind, Hypothesen zu prüfen, nicht zum Problem erhoben, sondern einfach als Selbstverständlichkeit erachtet, die jeder automatisch anerkannt hat, der empirische Wissenschaft betreibt.

Normative Wissenschaftstheorie macht diese Grundlagen explizit. Sie präzisiert die entsprechenden Regeln und methodologischen Hypothesen. Hier und da korrigiert sie auch einzelne methodologische Annahmen, die heute als überholt gelten, sich in manchen Wissenschaftsbereichen aber erhalten haben. (Beispielsweise sind manche Wissenschaftler nach wie vor davon überzeugt, dass theoretische Begriffe operational definiert

werden müssen oder dass Theorien induktiv gewonnen werden.) Dadurch gibt sie den Resultaten der empirischen Wissenschaften ihre *Rechtfertigung* als *Erkenntnis*. Sie begründet deren Anspruch darauf, dass ihre Forschungsresultate Aussagen sind, die man vorläufig für wahr halten darf, wenn sie auch nicht als absolut sicher gelten können.

Die oben aufgeworfene provokante Frage ist also so zu beantworten: Wissenschaftstheorie löst zwar nicht die konkreten einzelwissenschaftlichen Probleme, aber sie gibt heuristische Prinzipien zu deren Lösung an, und sie begründet, inwiefern eine gefundene Lösung als eine Erkenntnis gelten kann.

2. Ein grundlegender Unterschied zwischen einem normativen und einem deskriptiven (z.B. historischen oder soziologischen) Zugang zur Wissenschaft besteht darin, dass Ersterer es vorsieht, auch *kritisch* zu dem Geschehen in den Wissenschaften Stellung zu nehmen. Was dort geschieht, kann rekonstruiert und beschrieben, aber auch der Kritik unterzogen werden. Diese Position wird allerdings von vielen als naiv oder anmaßend empfunden. Wie können Philosophen beurteilen, ob die konkrete Arbeit im Rahmen eines physikalischen oder ökonomischen Forschungsprogramms gut und erfolgversprechend verläuft, oder aber einer Korrektur bedarf? Ich denke in der Tat, dass sie es im Allgemeinen nicht können, sofern sie nicht zugleich auch Experten auf dem betreffenden Gebiet sind. Dies ist nicht die Ebene, auf der normative Wissenschaftstheorie sich sinnvoll engagieren kann. Sie kann nicht und sollte nicht versuchen, die Probleme der Einzelwissenschaften zu lösen.

Überzeugender ist es, wenn sie sich die *Institutionen* vornimmt, die in den Wissenschaften eingerichtet worden sind, Institutionen wie z.B. Fachzeitschriften, Kriterien der Annahme und Ablehnung von Manuskripten, die Organisation und Praxis der Begutachtung, Techniken wissenschaftlichen Arbeitens oder Veranstaltungen zur Ausbildung von Nachwuchswissenschaftlern. Solche Institutionen sind dazu da, den Erkenntnisfortschritt in den jeweiligen Disziplinen zu fördern. Und vermutlich dienen sie diesem Zweck auch im Großen und Ganzen, es ist aber die Frage, ob sie es immer tun und ob sie eventuell verbesserungsbedürftig sind.

Um zu illustrieren, wie aus wissenschaftstheoretischer Sicht eine Analyse von Institutionen auf ihre Zweckrationalität aussehen kann, sei ein Beispiel angeführt, wobei wir wiederum von der wissenschaftstheoretischen Position ausgehen, dass es für einen Erkenntnisfortschritt günstig ist, Hypothesen der Kritik zu unterziehen. Daraus folgt die Empfehlung, solche Institutionen einzurichten, die Kritik jeglicher Art fördern bzw. sie nicht erschweren.

Melton, Herausgeber der renommierten Fachzeitschrift „Journal of Experimental Psychology", legte vor längerer Zeit dar, nach welchen Kriterien die Annahme von Manuskripten zur Publikation erfolgt. Als Hauptkriterium hob er die statistische Signifikanz hervor, und es wird dabei klar, dass Arbeiten mit signifikanten Resultaten eine größere Chance erhalten, publiziert zu werden, als nicht signifikante (Melton 1962). Nun wird auf dem betrachteten Gebiet aus der Hypothese oder Theorie, die getestet werden soll, so gut wie immer die statistische Alternativhypothese abgeleitet, die der Nullhypothese widerspricht. Ein ‚signifikantes' Ergebnis ist eines, das es erlaubt, die Nullhypothese zurückzuweisen und die Alternativhypothese zu akzeptieren, ein ‚nicht signifikantes' ist eines, das es nicht gestattet, die Nullhypothese zurückzuweisen und die Alternativhypothese zu akzeptieren. Signifikante Ergebnisse sind daher solche, die

es gestatten, die getestete (psychologische) Theorie als bewährt anzusehen. Nicht signifikante sind solche, die dazu geeignet wären, diese Theorie in Frage zu stellen. Melton hielt es also für wissenschaftlich, signifikante, was bedeutet: die Theorie bestätigende, Ergebnisse für die Publikation zu bevorzugen (vgl. Bredenkamp 1972: 51f.).

Normative Wissenschaftstheorie kann dazu beitragen, die Institutionen und Praktiken der Wissenschaft auf ihre Zweckrationalität hin zu analysieren (vgl. dazu Max Albert 2002). Am erfolgversprechendsten ist es, wenn sie diese Aufgabe gemeinsam mit den Experten des betreffenden Faches angeht.

3. Kommen wir nochmals darauf zurück, dass die empirischen Wissenschaften ihre implizit vorausgesetzten methodologischen Prinzipien gewöhnlich nicht thematisieren. Dies ist zutreffend, jedoch ergänzungsbedürftig. Insbesondere die Sozialwissenschaften befassen sich in ihrer fachspezifischen Methodenlehre durchaus explizit mit Fragen, die zu den in der Wissenschaftstheorie behandelten Problemen in einem engen (leider viel zu wenig beachteten) Zusammenhang stehen. Es werden dort bereichsspezifische methodologische Fragen eingehend erörtert, und entsprechende Ergebnisse werden in Lehrbüchern präsentiert und in der Ausbildung vermittelt. Es wird z.B. gelehrt, dass bestimmte Untersuchungspläne eine *Kausalinterpretation* der Resultate erlauben und andere nicht. Es gibt eine Kontroverse über die methodologischen Vorzüge bzw. Nachteile von experimenteller gegenüber Feldforschung. Es wird diskutiert, unter welchen Bedingungen ein empirisches Resultat auf andere Personen bzw. Situationen *generalisierbar* ist und vieles mehr. Es ist unmittelbar einsichtig, dass diese Themen mit wissenschaftstheoretischen Fragen zusammenhängen, z.B. mit dem Problem, welches Verständnis von Kausalität zugrunde gelegt werden soll, und mit dem Induktionsproblem.

Ich denke nun, dass die normative Wissenschaftstheorie einen wichtigen Beitrag zu den bereichsspezifischen methodologischen Diskussionen leisten kann. Sie kann behilflich sein, die jeweiligen Probleme zu klären. Sie kann logische Inkonsistenzen ausfindig machen und problematische methodologische Voraussetzungen aufzeigen. Auch hierzu sei ein Beispiel angeführt. Viele Lehrbücher und Methodenkurse vermitteln das bekannte System von Gütekriterien für empirische Untersuchungen, zu dem die *interne* und *externe Validität* gehört. Diese Lehre wurde von Campbell und Stanley (1963) ausführlich dargestellt und später von Cook und Campbell (1979) erweitert. Es lässt sich nun erstens zeigen, dass die Forderung nach externer Validität unvereinbar mit anderen methodologischen Annahmen ist, von denen die genannten Autoren ausgehen. Zweitens beruht die Idee der externen Validität auf einer Form des Induktivismus, der für die Zielsetzung der betreffenden Wissenschaften kaum eine geeignete Methodologie darstellen dürfte (für Details dieser Argumentation vgl. Gadenne 1976; 1984: Kap. 10).

Meines Erachtens sollte Wissenschaftstheorie künftig verstärkt als *fachspezifische* Wissenschaftstheorie, etwa der Soziologie oder Ökonomie, betrieben werden. Wenn sie sich im betreffenden Gebiet kundig macht, kann sie ihre Analysemethoden dazu einsetzen, zur Lösung der gegenstandsbezogenen methodologischen Fragen einen Beitrag zu leisten. Die diskutierten methodologischen Regeln des KR sind hierfür ein geeigneter Ausgangspunkt.

Literatur

Adorno, Theodor W., Hans Albert, Ralf Dahrendorf, Jürgen Habermas, Harald Pilot und *Karl R. Popper*, 1969: Der Positivismusstreit in der deutschen Soziologie. Neuwied: Luchterhand.
Agassi, Joseph, 1985: Technology: Philosophical and Social Aspects. Dordrecht: Reidel.
Albert, Hans, 1968: Traktat über kritische Vernunft. Tübingen: Mohr-Siebeck.
Albert, Hans, 1978: Traktat über rationale Praxis. Tübingen: Mohr-Siebeck.
Albert, Hans, 1982: Die Wissenschaft und die Fehlbarkeit der Vernunft. Tübingen: Mohr-Siebeck.
Albert, Max, 2002: Der kritische Rationalismus und die Verfassung der Wissenschaft. S. 231–241 in: *Jan Böhm, Heiko Holweg* und *Claudia Hoock* (Hg.), Karl Poppers kritischer Rationalismus heute. Tübingen: Mohr-Siebeck.
Andersson, Gunnar, 1988: Kritik und Wissenschaftsgeschichte. Tübingen: Mohr-Siebeck.
Bartley III, William W., 1987: Flucht ins Engagement. Tübingen: Mohr-Siebeck.
Bredenkamp, Jürgen, 1972: Der Signifikanztest in der psychologischen Forschung. Frankfurt a.M.: Akademische Verlagsgesellschaft.
Campbell, Donald T., und *Julian C. Stanley* , 1963: Experimental and Quasi-Experimental Designs for Research on Teaching. S. 171–246 in: *Nathaniel L. Gage* (Hg.), Handbook of Research on Teaching. Chicago: Rand McNally.
Cook, Thomas D., und *Donald T. Campbell*, 1979: Quasi-Experimentation. Design and Analysis Issues for Field Settings. Chicago: Rand McNally.
Duhem, Pierre, 1908/1978: Ziel und Struktur der physikalischen Theorien. Originalausgabe Leipzig. Nachdruck Hamburg 1978: Meiner.
Felt, Ulrike, Helga Nowotny und *Klaus Taschwer*, 1995: Wissenschaftsforschung. Eine Einführung. Frankfurt a.M.: Campus.
Feyerabend, Paul, 1976: Wider den Methodenzwang. Frankfurt a.M.: Suhrkamp.
Feyerabend, Paul, 1981: Rückblick. S. 320–372 in: *Hans Peter Duerr* (Hg.), Versuchungen. Aufsätze zur Philosophie Paul Feyerabends. Frankfurt a.M.: Suhrkamp.
Gadenne, Volker, 1976: Die Gültigkeit psychologischer Untersuchungen. Stuttgart: Kohlhammer.
Gadenne, Volker, 1984: Theorie und Erfahrung in der psychologischen Forschung. Tübingen: Mohr Siebeck.
Gadenne, Volker, 1998: Spielarten des Duhem-Quine-Problems. Logos 5: 117–148.
Gadenne, Volker, 2002: Hat der kritische Rationalismus noch etwas zu lehren? S. 58–78 in: *Jan Böhm, Heiko Holweg* und *Claudia Hoock* (Hg.), Karl Poppers kritischer Rationalismus heute. Tübingen: Mohr-Siebeck.
Jarvie, Ian C., 2001: The Republic of Science. Amsterdam: Rodopi.
Kuhn, Thomas S., 1967: Die Struktur wissenschaftlicher Revolutionen. Frankfurt a.M.: Suhrkamp.
Kuhn, Thomas S., 1974: Logik der Forschung oder Psychologie der wissenschaftlichen Arbeit? S. 1–24 in: *Imre Lakatos* und *Alan Musgrave* (Hg.), Kritik und Erkenntnisfortschritt. Braunschweig: Vieweg.
Kuipers, Theo A. F. (Hg.), 1987: What Is Closer-to-the-Truth? Amsterdam: Rodopi.
Kutschera, Franz von, 1972: Wissenschaftstheorie, Bd. 2. München: Fink.
Lakatos, Imre, 1974: Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme. S. 89–189 in: *Imre Lakatos* und *Alan Musgrave* (Hg.), Kritik und Erkenntnisfortschritt. Braunschweig: Vieweg.
Melton, Arthur W., 1962: Editorial. Journal of Experimental Psychology 64: 553–557.
Miller, David, 1994: Critical Rationalism: A Restatement and Defence. Chicago: Open Court.
Miller, David, 2002: Induction: A Problem Solved. S. 81–106 in: *Jan Böhm, Heiko Holweg* und *Claudia Hoock* (Hg.), Karl Poppers kritischer Rationalismus heute. Tübingen: Mohr-Siebeck.
Musgrave, Alan, 1993: Popper on Induction. Philosophy of the Social Sciences 23: 516–527.
Musgrave, Alan, 1999: Essays on Realism and Rationalism. Amsterdam: Rodopi.
Popper, Karl R., 1935/1994: Logik der Forschung. 10. Aufl. Tübingen: Mohr-Siebeck.
Popper, Karl R., 1973: Objektive Erkenntnis. Hamburg: Hoffmann und Campe.
Popper, Karl R., 1974: Replies to My Critics. S. 961–1197 in: *Paul A. Schilpp* (Hg.), The Philosophy of Karl Popper. La Salle (Illinois): Open Court.

Popper, Karl R., 1983: The Postskript to the Logic of Scientific Discovery Vol. 1: Realism and the Aim of Science. Totowa (NJ): Rowman and Littlefield.

Popper, Karl R., 1994: Vermutungen und Widerlegungen, Teilband I: Vermutungen. Tübingen: Mohr-Siebeck.

Popper, Karl R., 1997: Vermutungen und Widerlegungen, Teilband II: Widerlegungen. Tübingen: Mohr-Siebeck.

Salmon, Wesley C., 1968: The Justification of Inductive Rules of Inference. S. 24–43 in: *Imre Lakatos* (Hg.), 1968: The Problem of Inductive Logic. Amsterdam: North Holland.

Schmid, Michael, 1993: Der Positivismusstreit in der Deutschen Soziologie – 30 Jahre danach. Logos 1: 35–81.

Schurz, Gerhard, 2002: Karl Popper, Deduktion, Induktion und Abduktion. S. 126–143 in: *Jan Böhm, Heiko Holweg* und *Claudia Hoock* (Hg.), Karl Poppers kritischer Rationalismus heute. Tübingen: Mohr-Siebeck.

Vetter, Hermann, 1967: Wahrscheinlichkeit und logischer Spielraum. Tübingen: Mohr-Siebeck.

Wettersten, John R., 1992: The Roots of Critical Rationalism. Amsterdam: Rodopi.

Worrall, John, 1989: Why both Popper and Watkins Fail to Solve the Problem of Induction. S. 257–296 in: *Fred d'Agostino* und *Ian C. Jarvie* (Hg.), Freedom and Rationality: Essays in Honor of John Watkins. Dordrecht: Kluwer.

STATISTIK: VOM GEBURTSHELFER ZUM BREMSER DER ERKENNTNIS IN DEN SOZIALWISSENSCHAFTEN?

Walter Krämer

Zusammenfassung: Die beschreibende und die schließende Statistik tragen unterschiedlich zur Erkenntnis in den Sozialwissenschaften bei. Während in der statistischen Lehre und Forschung die induktiven, mathematisch-statistischen Verfahren dominieren, wird in den Anwendungen vor allem Hilfe bei fundamentalen, vorgelagerten Definitions- und Messproblemen gebraucht. Dieses Ungleichgewicht hat dazu geführt, dass die herkömmliche Methodenwissenschaft von der Statistik die problemorientierten Sozialwissenschaften zuweilen eher bremst statt unterstützt.

I. Der Status quo

Die Statistik, und insbesondere die mathematische Statistik, ist heute als unentbehrliches Werkzeug auch in den Sozialwissenschaften anerkannt. Das war nicht immer so. Von Peter Hofstätter (1993) wissen wir, dass ein „junger Psychologe", das Lehramt an einer Universität des deutschen Sprachgebietes anstrebend, noch vor 60 Jahren von seiner Fakultät zunächst mit dem Argument abschlägig beschieden werden konnte, dass er gewisse „dem deutschen Geist völlig fremde [statistische] Methoden" in seinen Arbeiten verwendet hätte.

Der „junge Psychologe" war wohl niemand anders als Hofstätter selbst, und sein damaliger Examinator, die philosophische Fakultät der Universität Graz, wusste sich im Einklang mit einer durch zahlreiche Bibelstellen (1. Chronik, 21; 2. Samuel, 24; Matthäus 6,26) und Klassikerzitate, von Goethe über die Romantiker bis zu Thomas Mann, unterfütterten, das ganze deutschsprachige Geistesleben durchdringenden Einstellung, „wie gemein und uniform die Masse sei", deren nähere Betrachtung sich nicht lohne und die besser „der Teufel und die Statistik hole" (Nietzsche 1874).

Das ist heute anders. Zumindest im akademischen Leben, und insbesondere auch in den Sozialwissenschaften, ist es *gerade* das Verhalten der Massen, das uns interessiert, und statt den Teufel und die Statistik in einen Topf zu werfen, hat letztere fast schon den Status einer Heiligen erreicht, hat sich die einstmalige Zahlenfeindlichkeit in eine fast schon übertriebene, zumindest äußerliche Zahlenverehrung, und damit einhergehend in eine mögliche Überschätzung mathematisch-statistischer Methoden bei der Analyse ebendieser Zahlen umgekehrt.

Die folgende *Tabelle 1* wertet einmal alle Abhandlungen in der Kölner Zeitschrift für Soziologie und Sozialpsychologie der Jahre 2000 und 2001 nach der Vorgangsweise (empirisch vs. theoretisch) und der Art der dabei verwendeten Methoden aus. Von den insgesamt 49 in diesen Jahren erschienenen Abhandlungen wendeten 39 statistische

Tabelle 1: Abhandlungen in der Kölner Zeitschrift für Soziologie und Sozialpsychologie in den Jahren 2000–2001 nach Art der Datenverwendung

Abhandlung	
Typ	Anzahl
ohne Daten	10
deskriptiv statistisch argumentierend	11
induktiv statistisch argumentierend	28
Unter Verwendung von (Mehrfachnennungen möglich)	
– Linearen Regressionsmodellen	16
– Logit-Modellen	16
– Hazardratenmodellen	2
– Pfadmodellen	1
– LISREL-Modellen	1

Methoden und davon wiederum 28 mehr oder weniger aufwändige inferenzstatistische Verfahren an (t- und F-Tests, Regressions- und Hazardratenanalysen, Logit-Schätzungen oder Verfahren für LISREL- oder Pfadmodelle). Nur zehn Abhandlungen kamen ohne Daten „aus dem wahren Leben“ aus.

In benachbarten Wissenschaften ist die Lage ähnlich. Von den bislang 53 Nobelpreisträgern der Wirtschaftswissenschaften haben 26 auch als Statistiker geforscht, die vorläufig letzten beiden, Robert Engle und Clive Granger, kann man sogar als lupenreine mathematische Statistiker bezeichnen. In der Biologie, in der Psychologie, in der Linguistik sind statistische Verfahren heute nicht mehr wegzudenken, und sogar in die vormals eher zahlenfernen Rechts-, Geschichts-, und Politikwissenschaften zieht immer mehr Statistik ein (siehe etwa Beck 2000 oder Eisenberg 2000), von einer geradezu sklavischen Unterwerfung vieler Mediziner unter gar nicht oder falsch verstandene t- und F-Tests ganz zu schweigen (siehe etwa Beck-Bornholdt und Dubben 1996, 1997 und in diesem Band, oder Krämer und Gigerenzer 2005, wo man zahlreiche, vor allem unter Medizinern beliebte Fehlinterpretationen statistischer Signifikanztests aufgelistet findet).

Inzwischen ist das Pendel, so die These des vorliegenden Beitrags, vielleicht sogar über einen optimalen „Statistisierungsgrad“ vieler Wissenschaften hinausgeschossen (im Sinne der konventionellen mathematisch-statistischen Inferenzmethodologie), hat der Grenznutzen auf verschiedenen Gebieten schon begonnen, negativ zu werden. „Doch wenn wir auch noch so stolz auf die Zahlen sind, mit denen wir unsere Analysen betreiben und Prognosen untermauern, mir scheint trotzdem, daß wir uns durch die Fixierung auf Zahlen mehr täuschen als informieren lassen,“ schreibt Ernst-Peter Fischer (1996) in einem Aufsatz mit dem bezeichnenden Titel „Zahlen geben kein Verantwortungsgefühl“. „Wir ersticken in Zahlen und sehen nicht mehr weiter, als sie reichen.“ Wir geben quasi die Verantwortung für das Erkennen und für die Erkenntnis an die Zahlen und an die statistischen Methoden zur Auswertung dieser Zahlen ab. Statt zu fragen: „Wie kann mir die Statistik helfen, dieses oder jenes Rätsel aufzuklären?“ sieht man die Welt nur noch durch die Brille einer Computersoftware und passt nach dem Maslowschen Hammer-Prinzip Modelle an, nicht deshalb, weil sie den Daten angemes-

sen wären, sondern weil gerade die ganze Fachwelt davon spricht oder weil sie im Rechner so bequem verfügbar sind.[1]

Der vorliegende Beitrag versucht, hier Merk- und Marksteine zu setzen und die Statistisierung der Sozialwissenschaften im Hinblick auf deren Nutzen für ebendiese Wissenschaften zu bewerten. Analog der Nietzsche-Schrift „Vom Nutzen und vom Nachteil der Historie für das Leben", der auch das obige Zitat entnommen ist, könnte ihr Titel auch „Vom Nutzen und vom Nachteil der Statistik für die Wissenschaft vom menschlichen Zusammenleben" lauten. Nur wenn auch die Nachteile eines allzu blinden Vertrauens in statistische Modelle und Verfahren deutlich werden, können sich deren Vorteile weiter zur Vermehrung unseres Wissens auf effiziente Weise nutzen lassen.

II. Beschreibende vs. schließende Statistik

Ein Großteil der zur Zeit in verschiedenen Wissenschaften beobachtbaren Debatten zum Sinn und Unsinn der Statistik erklärt sich aus der verwaschenen Bedeutung des Begriffs „Statistik" selbst. Bis zur Wende vom 19. zum 20. Jahrhundert war damit – entsprechend der Herkunft aus dem italienischen statista = Staatsmann – „ursprünglich soviel wie Staatskunde, worunter die systematische Darstellung der Verfassung, der Organisation, der Bevölkerungsverhältnisse, der militärischen und wirtschaftlichen Hilfsquellen und der sonstigen bemerkenswerten Einrichtungen eines oder mehrerer Staaten" zu verstehen (Brockhaus Konversationslexikon 1895), zunehmend ergänzt durch eine „Klarlegung der Zustände und Erscheinungen des gesellschaftlichen Lebens, soweit solche in statistisch erfassbaren, sozialen Maßen zum Ausdruck kommen" (Der große Brockhaus 1934).

Man kann es nur einen unglücklichen Zufall nennen, dass um die gleiche Zeit, da der Brockhaus dieses schrieb, in Großbritannien und anderswo ein Gruppe kreativer Mathematiker um Karl und Egon Pearson, Jerzy Neyman und Ronald Fisher eine Wissenschaft begründete, die sie ebenfalls „Statistik" nannten, denn diese induktive mathematische Statistik hat mit der Staatenkunde gleichen Namens soviel gemein wie die Geologie mit der Geographie; es handelt sich um grundverschiedene Untersuchungsgegenstände und Herangehensweisen, allein durch diesen Zufall eines gemeinsamen Namens und zum großen Erstaunen aller Studierenden, die in einer so benannten Vorlesung beide Wissenschaften nebeneinander wiederfinden und diesen Bruch sehr wohl erkennen, üblicherweise in einer einzigen Lehrveranstaltung vereint.

Die relative Bedeutung dieser beiden – gleich benannten, aber dennoch grundverschiedenen – Methodenlehren für den Erkenntnisfortschritt in den Wissenschaften entspricht in keiner Weise ihrer relativen Bedeutung in der akademischen Lehre und Forschung. Hier dominiert die mathematisch-induktive Statistik, während die Sachwissenschaften durchweg mehr Bedarf an Hilfestellung bei elementaren Problemen der Messung und Erhebung haben. Bekannte Beispiele aus den Wirtschaftswissenschaften sind

1 „... it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail" (Maslow 1966).

die Messung der Arbeitslosigkeit oder die adäquate Bestimmung durchschnittlicher Preisänderungen. Glaubt man hier etwa dem sog. „Boskin-Komitee“ (1996), sind die in westlichen Industrienationen seit dem Zweiten Weltkrieg ausgewiesenen Inflationsraten wegen Nichtberücksichtigung von Qualitätsverbesserungen um bis zu einem Prozentpunkt jährlich überhöht. Mit anderen Worten, das „reale“ Sozialprodukt ist heute um mehr als die Hälfte höher als in den amtlichen Statistiken angegeben. Allein durch die mangelhafte Berücksichtigung der Mobilfunktechnologie sind etwa nach Hausman (1999) die amtlich ausgewiesenen Preisanstiege für die Telekommunikation in den USA von 1988 bis 1996 um rund 50 Prozentpunkte zu hoch.

Geradezu legendär sind auch die Fallstricke bei der Messung der Arbeitslosigkeit. Wann immer es das alte Vorurteil zu bestätigen gilt, dass mit Statistik alles zu beweisen sei, muss in der Regel die Arbeitslosigkeit als Zeuge her. Nach offizieller Lesart zählt in Deutschland als arbeitslos, wer (I) mindestens 15 Stunden in der Woche gegen Entgelt arbeiten will, (II) nicht nur vorübergehend Arbeit sucht, (III) älter als 15 und jünger als 66 Jahre ist, (IV) dem Arbeitsmarkt unmittelbar zur Verfügung steht sowie (V) bei einem Arbeitsamt offiziell als arbeitsuchend gemeldet ist. Wer nur eine Teilzeitarbeit von weniger als 15 Stunden oder eine Ferienstelle sucht, wer wegen Krankheit oder Umschulung dem Arbeitsmarkt vorübergehend nicht zur Verfügung steht oder die Suche per Arbeitsamt ganz einfach aufgegeben hat, der oder die ist damit auch nicht arbeitslos. So gesehen sind die amtlichen Zahlen tatsächlich, wie viele glauben, viel zu klein.

Auf der anderen Seite sind in den „offiziellen“ Arbeitslosenzahlen aber auch viele Menschen enthalten, die in Wahrheit keine Arbeit suchen und nur zum Abschöpfen verschiedener Vergünstigungen beim Arbeitsamt gemeldet sind. Nach Brixy et al. (2002) sucht z.B. nur die Hälfte aller beim Arbeitsamt gemeldeten Personen tatsächlich eine Stelle. Die wichtigsten wahren Gründe für den Gang zum Arbeitsamt waren (Mehrfachnennungen möglich): „Ich möchte sicherstellen, daß die Zeit der Arbeitslosigkeit später bei der Rentenberechnung berücksichtigt wird“ (83 Prozent). „Ich bin auf Leistungen des Arbeitsamtes angewiesen“ (76 Prozent). „Ich überbrücke die Zeit bis zum Beginn meiner neuen Stelle bzw. Ausbildung“ (57 Prozent). „Ich überbrücke die Zeit bis zum Ruhestand“ (21 Prozent). „Das Sozialamt verlangt, daß ich mich arbeitslos melde“ (16 Prozent). „Ich muß mich arbeitslos melden, um ausreichend Unterhalt von meinem geschiedenen Ehepartner zu bekommen“ (4 Prozent) und „Ich überbrücke die Zeit bis zum Zivil- oder Wehrdienst“ (2 Prozent). Mit anderen Worten, das Arbeitsamt ist für viele eine Sozialbehörde und keine Arbeitsplatzvermittlungsstelle.

Die Auflösung dieses Dilemmas soll hier nicht interessieren (siehe dazu etwa Sorrentino 2000 oder Krämer 2003, 2004). Worauf es ankommt, ist: Nicht ausgefeilte mathematisch-statistische Methoden, sondern fundamentale Probleme des Messens und Vergleichens stehen hier im Rampenlicht.

In den Sozialwissenschaften sind es vor allem die Definition und Messung von Gesundheit und Krankheit, Analphabetentum oder Armut, auch das Ziehen unverzerrter Stichproben und das Formulieren neutraler Fragebögen, welche die Anwender in erster Linie interessieren. Die Fehler, die hier möglich sind und auch gemacht werden, stellen alle Nachteile weit in den Schatten, die durch Verwendung ineffizienter oder nicht dem Stand der Technik entsprechender Schätz- und Testverfahren entstehen oder ent-

stehen können. So ist etwa seit langem wohlbekannt, wie stark die Ergebnisse von Umfragen aller Art von den Begleitumständen der Umfrage abhängen. Aus den USA weiß man, dass mündliche Interviews zu den Themen Abtreibung, Todesstrafe oder Sozialhilfe andere Ergebnisse haben, je nachdem, ob der Interviewer ein Schwarzer oder ein Weißer ist, und auch die Reihenfolge der Fragen und natürlich die konkrete Formulierung sind für das Ergebnis von erheblicher Bedeutung. Im März 1955 etwa fragte Elisabeth Noelle-Neumann eine repräsentative Stichprobe von Arbeitern: „Finden Sie, dass in einem Betrieb alle Arbeiter in der Gewerkschaft sein sollten?“ mit folgendem Resultat:

Dafür	44%
Dagegen	20%
Unentschieden	36%

Dann legte sie einer anderen, gleich großen und ebenfalls repräsentativen Stichprobe die gleiche Frage vor, nur mit dem Zusatz „... oder muss man es jedem einzelnen überlassen, ob er in der Gewerkschaft sein will oder nicht?“ mit dem Ergebnis:

Dafür	24%
Dagegen	70%
Unentschieden	6%

Dieser scheinbar unschuldige Zusatz halbiert damit die Anhängerschaft der Gewerkschaften von vorher 44 auf hinterher nur noch 24 Prozent; zugleich lässt er die Gegner von vorher 20 auf hinterher 70 Prozent anwachsen – eine mehr als dreifach größere Opposition nur wegen eines kleinen Nebensatzes.

Einen großen Unterschied macht auch, ob man etwas verbieten oder „nicht erlauben“ soll. 54 Prozent der Befragten in einer amerikanischen Umfrage glaubten, dass die USA öffentliche Angriffe auf die Demokratie verbieten sollten. Erheblich mehr, nämlich 75 Prozent dagegen glaubten, die USA sollten öffentliche Angriffe auf die Demokratie nicht erlauben. Diese Fallstricke bei Umfragen sind seit langem wohl bekannt (siehe etwa Diekmann 1995 oder Sirken et al. 1999), werden aber oft zugunsten ausgefeilter Verfahren bei der Auswertung der Umfragen übersehen.

Ebenfalls seit langem wohlbekannt, aber bei weitem noch nicht gelöst, sind auch in den Sozialwissenschaften Definitionsprobleme aller Art. So wird derzeit in Deutschland laut über die Zunahme des Analphabetentums geklagt. Aber wer ist ein Analphabet? Zu Kaiser Wilhelms Zeiten wurde als Analphabet gezählt, wer bei der Musterung statt mit seinem Namen mit einem Kreuz unterschrieb. Heute ist nach dem „Bericht zur Bekämpfung des Analphabetismus in der Bundesrepublik“ bereits ein Analphabet, wer „sich nicht beteiligen kann an all den zielgerichteten Aktivitäten [seiner] Gruppe und ... Gemeinschaft, bei denen Lesen, Schreiben und Rechnen erforderlich ist“. Ein großer Teil der in letzter Zeit so beklagten Zunahme der Analphabeten in Deutschland auf heute angeblich über vier Millionen ist damit ein reines Artefakt einer Ausweitung der Definition.

Oder Krankheit. Bäuerinnen kommen in der Regel nicht auf den Gedanken, Rückenschmerzen als Krankheit zu bewerten und deshalb zum Arzt zu gehen, während Stadtfrauen das viel öfter tun. Fortgesetztes Husten ist nach E.K. Koos (1970) für 77

Prozent der Oberschicht Grund für einen Arztbesuch, in der Unterschicht dagegen nur für 23 Prozent. Bei chronischer Müdigkeit ist das Verhältnis 80:19, bei Blut im Stuhl 98:60 und bei Schmerzen im Brustkorb 80:31. Ob man als „krank" gilt, hängt also auch davon ab, wo man wohnt und wieviel Geld man hat.

Auch die „Armut" ist alles andere als einfach zu messen. Die politische Linke hierzulande nimmt als Maß dafür gern den Anteil der Familien mit einem Einkommen unterhalb der Hälfte des Durchschnitts. Und wundert sich, dass trotz wachsenden Wohlstands die Armut in Deutschland nicht verschwindet. Nach den Maßstäben der Vereinten Nationen dagegen gilt als arm, wer weniger als einen Dollar täglich zum Überleben zur Verfügung hat. Damit ist in Deutschland niemand, weltweit aber jeder fünfte – zusammen über eine Milliarde Menschen – arm. Auch hier soll die Auflösung dieses Dilemmas nicht weiter interessieren (siehe dazu Krämer 2000) – wichtig ist, dass in der modernen Armutsdiskussion weniger inferenzstatistische Methodenfortschritte als fundamentale, vorgelagerte Definitions- und Messkonzepte weiterhelfen.

Dieses Ungleichgewicht spiegelt sich aber weder in der Lehre (siehe etwa Schmerbach und von der Lippe 2003) noch in den Forschungsleistungen der Wissenschaft von der Statistik wider, und das ist der erste Grund, warum deren Beitrag zum Fortschritt in manchen Substanzwissenschaften heute zunehmend kritischer gesehen werden kann.

III. Nutzen vs. Grenznutzen mathematisch-statistischer Verfahren

Der zweite Grund ist: Man muss den gesamten Nutzen, den kumulativen Beitrag der mathematischen Statistik zum Erkenntnisfortschritt in einer Substanzwissenschaft von dem *zusätzlichen* Nutzen unterscheiden, den eine weitere Verbesserung der mathematisch-statistischen Methoden mit sich bringt. Auch wenn niemand bestreitet, dass etwa ein t-Test oder eine Varianzanalyse wertvolle Einsichten erlauben oder zumindest erlauben können, die ohne dergleichen Methoden niemals möglich wären, so darf man doch an dem Nutzen*zuwachs* durch immer weitere Verfeinerungen dieser Methoden durchaus zweifeln. War etwa die Erfindung der gewöhnlichen Kleinst-Quadrate-Methode durch C.F. Gauss noch ein wahres Göttergeschenk für viele Wissenschaften, so hat die zweistufige Kleinst-Quadrate-Methode von H. Theil (1953) nur noch die Ökonometriker bewegt, und die weitere Verfeinerung zur dreistufigen Kleinst-Quadrate-Methode durch Zellner und Theil (1962) vielleicht sogar mehr Unheil als Gewinn gebracht. Wie etwa in Krämer (1980) nachgewiesen, schmelzen die Effizienzvorteile dieses ausgefeilten Verfahrens bei nicht ganz korrekt spezifizierten Modellen – und welches statistische Modell entspricht genau der Wirklichkeit? – sehr schnell, und zwar schneller als bei einfacheren Verfahren dahin, so dass letztere netto zu besseren Anpassungen der Modelle und zu besseren Prognosen führen. Oder um mit Wirtschafts-Nobelpreisträger Haavelmo (1958) zu sprechen „The concrete results of our efforts at quantitative measurement often seem to get worse the more refinement of tools ... we call into play."

In Krämer et al. (1985) werden versuchsweise ein gutes Dutzend empirisch geschätzter, in verschiedenen Journalen publizierter Regressionsmodelle einer Batterie von

neu entwickelten Tests auf korrekte funktionale Form, Konstanz der Regressionskoeffizienten oder vergessene Regressoren unterworfen, mit dem Ergebnis, dass kein einziges dieser Modelle diese Prüfung unbeschadet übersteht. Der wahre funktionale Zusammenhang zwischen Regressanden und Regressoren war nicht, wie fälschlich angenommen, linear, es wurden relevante Regressoren vergessen, die Regressionskoeffizienten waren nicht für alle Datenpunkte in der Stichprobe identisch usw. Man kann wohl ohne Übertreibung unterstellen, dass ein großer Teil der in den Sachwissenschaften, insbesondere auch in den Sozialwissenschaften, mit zum Teil aufwendigen Methoden geschätzten Modelle die wahren Sachverhalte auch nicht näherungsweise richtig wiedergibt (im Sinne von: eine nach welchen Kriterien auch immer akzeptable Approximation darstellt).

Was soll aber dann der Einsatz hochkomplizierter Schätz- und Testverfahren? Diese lenken doch nur von den wahren Schwierigkeiten ab. Außerdem lassen Sie oft vergessen, dass jeder statistische Signifikanztest immer eine *Sammelhypothese* testet: Die „offizielle" Nullhypothese, und dann noch die Unterstellung, das jeweilige Modell wäre korrekt spezifiziert. Nur wenn auch letzteres gilt, deutet ein signifikantes Testresultat tatsächlich auf „signifikante" Effekte hin. Ist letzteres dagegen nicht erfüllt, kann eine „signifikante" Teststatistik auch ganz andere Ursachen haben als das statistische Verfahren suggeriert, und genau das scheint auch oft in Anwendungen der Fall zu sein. So gehen etwa „signifikante" Koeffizienten in linearen Regressionsmodellen oft auf positiv korrelierte Störgrößen zurück, welche die Schätzung der Störgrößenvarianz nach unten verzerren und damit die t-Statistik aufblasen (siehe etwa Krämer und Michels 1997 und die dort zitierte Literatur zu heteroskedastie- und autokorrelationsrobusten t-Tests), oder haben signifikante Durbin-Watson-Statistiken ihre Ursache nicht in autokorrelierten Störgrößen, zu deren Entlarvung der Test ursprünglich entworfen worden ist, sondern in nichtlinearen funktionalen Abhängigkeiten oder vergessenen Regressoren (Krämer und Sonnberger 1986), um nur zwei der häufigsten Fehlalarme zu erwähnen.

Selbst bei korrekt spezifizierten Modellen werden die Ergebnisse statistischer Signifikanztests oft nicht richtig interpretiert bzw. als Kronzeuge für Aussagen benutzt, die auf falsche Fährten führen und auf diese Weise den Erkenntnisfortschritt hemmen. Ein zum Niveau von 5 Prozent signifikantes Testresultat sagt ja nur: Wenn die getestete Hypothese zuträfe, käme ein solches Resultat mit weniger als 5 Prozent Wahrscheinlichkeit zustande. Das gilt aber für viele andere Null-Hypothesen auch. Daraus ist keinesfalls zu schließen, das Gegenteil der Nullhypothese wäre wahr. Hier scheint mir oft das Poppersche Prinzip verletzt zu werden, als Nullhypothese immer nur die aktuell bewährte, anerkannte Theorie zu wählen, und dieser die Fakten gegenüberzustellen. Nur so, durch das Ersetzen des aktuell besten Modells durch ein vergleichsweise besseres, kommt man der „Wahrheit" näher. Wem aber nützt das Verwerfen eines ohnehin schon nicht als adäquat betrachteten Modells?

Nochmals irreführender ist oft die Interpretation eines *nicht* signifikanten Tests. Auch das bedeutet nicht, dass die Nullhypothese zutrifft, sondern nur, dass sie nicht verworfen werden kann, und auch das trifft bei einem gegebenen Datensatz für viele Nullhypothesen zu. Besonders bei kleinen Stichproben ist fast jede Modellannahme mit den Daten kompatibel. Bei größeren Stichproben dagegen führen schon kleinste

Abweichungen von der Nullhypothese zu signifikanten Testresultaten, mit zum Teil seltsamen Konsequenzen: In den notorisch prozessfreudigen USA, wo sich Arbeitnehmer gerne gerichtlich gegen eine vermeintliche oder tatsächliche Benachteiligung wehren, fallen große Firmen schon bei minimalen Abweichungen von einer wie auch immer gewählten Norm als „signifikante" Diskriminierer auf, während Kleinbetriebe diskriminieren können wie sie wollen – es ist alles nicht signifikant.

Auch andernorts richtet das „Signifikanztest-Ritual" (Sedlmeier 1996; Krämer 2002; Gigerenzer et al. 2004) mehr Schaden als Nutzen an. Haller and Kraus (2002) z.B. legten 44 deutschen Psychologiestudenten, die gerade einen Kurs in induktiver Statistik hinter sich hatten, 30 ihrer Lehrer (Professoren, Mitarbeiter und studentische Hilfskräfte) sowie 39 weiteren Wissenschaftlern den folgenden Fragebogen zur Bedeutung eines t-Tests vor, mit den Antwortalternativen ja oder nein.[2] Es waren zwei Stichproben vom Umfang 20 auf identische Erwartungswerte zu testen, die t-Statistik war zu einem Niveau von 1 Prozent signifikant:

1. Die Hypothese, es gäbe keine Unterschiede, ist falsch.
2. Die Wahrscheinlichkeit, dass die Nullhypothese zutrifft, liegt unter 1 Prozent.
3. Die Vermutung, es könnte Unterschiede geben, ist richtig.
4. Man kann immerhin die Wahrscheinlichkeit dafür angeben, dass diese Vermutung richtig ist.
5. Die Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Fehler zu machen, ist kleiner als 1 Prozent.
6. Wenn das Experiment sehr oft wiederholt würde, käme in 99 Prozent der Fälle eine signifikante Prüfgröße zustande.

Alle diese Behauptungen sind falsch. Nur 20 Prozent der Statistik-Lehrer, 10 Prozent der übrigen Wissenschaftler und keiner der Studenten sah das aber so. Die meisten kreuzten eine oder mehrere der obigen Falschaussagen als richtig an. Selbst die Dozenten hielten im Mittel 1,9 der oben aufgeführten Falschaussagen für wahr. „Teaching statistics to psychology students in German Universities does not include effective enlightenment about the meaning of significance: four out of every five methodology instructions show misconception about this concept with their students" (Haller und Kraus 2002: 8).

Mit anderen Worten, das Signifikanztest-Ritual trägt kaum zur weiteren Erkenntnis, mehr zur Vernebelung des Erkenntnisfortschritts bei. Statt den Blick für die in den Daten enthaltene Botschaft zu schärfen, lenkt es sogar eher davon ab. Das routinemäßige Durchführen von Signifikanztests ohne Rücksicht auf die Bedingungen, unter denen solche Tests überhaupt sinnvoll sind, unter gleichzeitiger Fehlinterpretation der erzielten Resultate, lenkt auf falsche Fährten, macht blind für wichtige Signale, engt die Sicht auf die soziale Umwelt ein und ist netto eine Bremse des Erkennens in so mancher Wissenschaft.

2 Haller und Kraus geben die Fragen in der englischen Übersetzung; die folgenden Alternativen sind ins Deutsche rückübersetzt.

IV. Fazit

Falsch verstandene klassische Inferenzstatistik ist schlechter als gar keine Inferenzstatistik. Durch die Betonung klassischer Schätz- und Testverfahren und durch die Kanalisierung von Lehr- und Forschungsaufwand in deren Effizienzverbesserung hat sich die konventionelle Statistik von den Bedürfnissen der Anwender entfernt, sie lenkt diese sogar ganz im Gegenteil vom weiteren Erkennen ab.

Gefragt sind stattdessen robuste, datenorientierte Methoden. Statt nach der optimalen Methode zu suchen, gegeben ein restriktiv spezifiziertes statistisches Modell, sollte man sich mehr um modellunabhängige Verfahren kümmern, und wie ein Blick in die Statistik-Fachbereiche weltweit zeigt, ist hier ein Umdenken tatsächlich schon im Gange. Möge es recht bald seinen Weg auch in die Lehrbücher und in die Anwendungen finden.

Literatur

Beck, Nathaniel L., 2000: Political Methodology: A Welcoming Discipline. Journal of the American Statistical Association 95: 651–654.

Beck-Bornholdt, Hans-Peter, und *Hans-Hermann Dubben,* 1996: Is the Pope an Alien? Nature 381: 3.

Beck-Bornholdt, Hans-Peter, und *Hans-Hermann Dubben,* 1997: Der Hund, der Eier legt – Erkennen von Fehlinformationen durch Querdenken. Reinbek: Rowohlt.

Boskin, Michael J., Ellen R. Dulberger, Robert J. Gordon, Zvi Grilliches und *Dale W. Jorgenson,* 1996: Toward a More Accurate Measure of the Cost of Living. Abschlußbericht für das Senate Finance Committee, 4. Dezember.

Brixy, Udo, Rainer Gilbery, Doris Hess und *Helmut Schröder,* 2002: Wie nah am Arbeitsmarkt sind die Arbeitslosen? IAB Kurzbericht Nr. 2 vom 21. Januar.

Diekmann, Andreas, 1996: Empirische Sozialforschung – Grundlagen, Methoden, Anwendungen. Reinbek bei Hamburg: Rowohlt.

Eisenberg, Theodore, 2000: Empirical Methods and the Law. Journal of the American Statistical Association 95: 665–669.

Fischer, Ernst Peter, 1996: Zahlen geben kein Verantwortungsgefühl. Die Weltwoche, 11. April: 59.

Gigerenzer, Gerd, Stefan Krauss und *Oliver Vitouch,* 2004: The Null Ritual: What You always Wanted to Know about Significance Testing but were Afraid to Ask. In: *D. Kaplan* (Hg.), Handbook on Quantitative Methods in the Social Sciences. London: Sage.

Hausman, Jerry, 1999: Cellular Telephones, New Products, and the CPI. Journal of Business and Economic Statistic 17: 188–194.

Haavelmo, Trygve, 1958: The Role of the Econometrician in the Advancement of Economic Theory. Econometrica 26: 351–357.

Haller, Heiko, und *Stefan Krauss,* 2002: Misinterpretation of Significance: A Problem Students Share with their Teachers? Methods of Psychological Research – Online 7: 1–20.

Hofstätter, Peter R., 1993: Von Zahlen wegschauen. Die Welt, 7. Juni: 7.

Koos, Erich L., 1970: Krankheit in Regionville. In: *Alexander Mitscherlich* et al. (Hg.), Der Kranke in der modernen Gesellschaft. Köln: Kiepenheuer und Witsch.

Krämer, Walter, 1980: Eine Rehabilitation der Gewöhnlichen Kleinst-Quadrate-Methode als Schätzverfahren in der Ökonometrie. Frankfurt a.M.: Haag und Herchen.

Krämer, Walter, 2000: Armut in der Bundesrepublik – Zur Theorie und Praxis eines überforderten Begriffs. Frankfurt a.M./New York: Campus.

Krämer, Walter, 2002: Denkste! Trugschlüsse aus der Welt des Zufalls und der Zahlen (2. Taschenbuchauflage). München: Piper.

Krämer, Walter, 2003: Der Nürnberger Trichter, oder: wer zählt die Arbeitslosen. Kursbuch 152: 93–102.

Krämer, Walter, 2004: So lügt man mit Statistik (5. Taschenbuchauflage). München: Piper.

Krämer, Walter; Harald Sonnberger, Johann Maurer und *Peter Havlik,* 1985: Diagnostic Checking in Practice. Review of Economics and Statistics 118–123.

Krämer, Walter, und *Harald Sonnberger,* 1986: The Linear Regression Model under Test. Heidelberg: Physika.

Krämer, Walter, und *Sonja Michels,* 1997: Autocorrelation- and Heteroskedasticity-consistent t-values with Trending Data. Journal of Econometrics 76: 141–147.

Krämer, Walter, und *Gerd Gigerenzer,* 2005: How to Confuse with Statistics. Statistical Science, im Erscheinen.

Maslow, Abraham H., 1966: The Psychology of Science. New York: Harper & Row.

Nietzsche, Friedrich, 1874: Unzeitgemäße Betrachtungen. Zweites Stück: Vom Nutzen und Nachteil der Historie für das Leben. Leipzig: Verlag E.W. Fritzsch.

Schmerbach, Sybille, und *Peter von der Lippe,* 2003: Mehr Wirtschaftsstatistik in der Statistikausbildung für Volks- und Betriebswirte. Allgemeines Statistisches Archiv 87: 335–344.

Sedlmeier, Peter, 1996: Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. Methods of Psychological Research – Online 3: 39–42.

Sirken, Monroe G., Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur und *Roger Tourangeau,* 1999: Cognition and Survey Research. New York: Wiley.

Sorrentino, Constance, 2000: International Unemployment Rates: How Comparable are They? Monthly Labor Review, Juni: 3–19.

Theil, Henry, 1953: Repeated Least Squares Applied to Complete Equation Systems. Den Haag: Central Planning Bureau.

Zellner, Arnold, und *Henry Theil,* 1962: Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. Econometrica 30: 54–78.

DIE BEDEUTUNG DER STATISTISCHEN SIGNIFIKANZ

Hans-Hermann Dubben und Hans-Peter Beck-Bornholdt*

Zusammenfassung: Zunächst die gute Nachricht: es besteht Hoffnung, dass in naher Zukunft die gewaltige und üblicherweise ungelesene Flut von mehr als 5000 biomedizinischen Publikationen mit so genannten „signifikanten" Ergebnissen sich zu einem übersichtlichen und lesenswerten Bach gesundschrumpft. Die schlechte Nachricht: die vorherrschende Interpretation von p-Werten ist unkorrekt und begünstigt die Produktion „statistisch signifikanter", aber falscher Ergebnisse. Stellen Sie sich eine perfekt durchgeführte klinische Studie vor. Die experimentelle Therapie zeigt eine höhere Überlebensrate als die Standardtherapie. Sie wissen, dass dieser Vorteil auch in methodisch perfekten Studien durch zufällige Schwankungen der Behandlungsergebnisse zustande gekommen sein kann. Deshalb stellen Sie einem Statistiker die klinisch relevante Frage: „Wie wahrscheinlich ist es, dass ich mich irre, wenn ich die experimentelle Therapie für besser halte?" Nach ein paar Berechnungen kommt die Antwort: „Der beobachtete Unterschied ist statistisch signifikant (p=0,03)." Damit ist ein wichtiges Kriterium für eine Publikation erfüllt, aber: Was bedeutet dieser Satz? Die Antwort des Statistikers bedeutet *nicht,* was die meisten glauben: „Wenn ich die experimentelle Therapie für besser als die Standardtherapie halte, beträgt die Irrtumswahrscheinlichkeit 3 Prozent." *Der p-Wert ist nicht die Antwort auf Ihre Frage!* Der p-Wert besagt lediglich: „Wenn beide Therapien in Wirklichkeit gleichwertig sind, dann kann die beobachtete (oder eine größere) Differenz mit 3 Prozent Wahrscheinlichkeit zufällig auftreten." Der Unterschied dieser beiden Sätze ist keine Haarspalterei, sondern möglicherweise der häufigste und schwerste Irrtum der biomedizinischen Forschung. Der zugrunde liegende Irrtum wird an einem alltäglichen Beispiel und anhand „statistisch signifikanter" Ergebnisse erläutert. Es wird gezeigt, dass p-Werte allein kein Maß für wissenschaftliche Evidenz sind.

Signifikanztests haben eine sehr große Bedeutung in unserer gegenwärtigen wissenschaftlichen Kultur. Kaum eine Publikation mit quantitativen Resultaten kommt ohne Statistik und ohne Signifikanztest aus. Sie werden als Ausdruck von Wissenschaftlichkeit angesehen und sind meist auch Eingangsvoraussetzung dafür, dass ein Manuskript zur Publikation überhaupt angenommen wird. Das Ziel statistischer Erwägungen ist es, wahre Ergebnisse möglichst zuverlässig von Zufallstreffern zu unterscheiden. Inwieweit dies tatsächlich auch erfüllt wird, soll in dieser Arbeit hinterfragt werden.

Alltägliche Beispiele werden zeigen, dass der Grundgedanke von Signifikanztests sehr einfach ist. Auch wird höhere Mathematik nicht von Nöten sein; vielleicht mal ein Dreisatz oder ein Blick in eine Abbildung oder eine Tabelle.

* Institut für Allgemeinmedizin. Universitäts-Klinikum Hamburg-Eppendorf, D-20246 Hamburg.

I. Fußball-Statistik

Letzte Woche traten der FC Aheim und der SC Beheim gegeneinander an. Die beiden Mannschaften sind wie zwei Klone. Exakt gleich gut, sagen die Kenner der Szene. Gestern hat der FC Aheim gewonnen. Niemanden hat es gewundert. Neulich beim Freundschaftsspiel war es andersrum.

Heute spielte der SC Beheim gegen TSV Versager. Der TSV ist unbestritten sehr viel schlechter als der SC, aber heute gab es immerhin ein Unentschieden. Das hat auch niemanden gewundert. Hätten sie länger gespielt, hätte der Bessere sich sehr wahrscheinlich durchgesetzt.

Bemerkenswert ist, dass das Ergebnis eines einzelnen Spieles nicht unbedingt repräsentativ dafür ist, welche der beiden Mannschaften die bessere ist. Falls es überhaupt eine bessere gibt. Etwas anders sieht es bei einem sehr hohen Sieg von 6:0 oder extremer aus. Dann ist das Ergebnis sogar statistisch signifikant (Beck-Bornholdt und Dubben 2001).

II. Was will der Verbraucher von Studienergebnissen wissen?

In einer klinischen Studie hat ein Arzt die Wirksamkeit zweier Behandlungen, Therapie A und Therapie B, miteinander verglichen. In der Gruppe, die Therapie A erhalten hat, ist die 5-Jahres-Überlebensrate um 10 Prozentpunkte höher als in der anderen Gruppe. Können wir nun sicher sein, dass Therapie A tatsächlich besser ist als Therapie B?

Nein, das können wir nicht. Dazu müssen wir nur an die Fußballergebnisse denken. Therapie A und Therapie B sind ein einziges Mal gegeneinander angetreten. Das Studienergebnis gibt nicht mit Sicherheit wieder, welche Therapie die bessere ist, falls es überhaupt eine bessere gibt. Eine Studie kann man nicht so einfach wiederholen wie ein Fußballspiel und die nächsten Patienten sitzen auch bereits im Wartezimmer. Im Interesse seiner Patienten möchte unser Arzt daher sofort so viel wie möglich aus seiner Studie folgern. Er weiß, dass zufällige Schwankungen den Unterschied auch dort vortäuschen können, wo keiner ist. Der Arzt möchte daher wissen: *„Wie groß ist die Wahrscheinlichkeit, dass ich mich irre, wenn ich Therapie A für besser als Therapie B halte?“*

Mit dieser Frage wendet er sich an einen Experten, beispielsweise an einen Biomathematiker oder Statistiker. Nach Durchführung üblicher, aber für Normalsterbliche undurchsichtige Berechnungen, erhält der Arzt folgende Antwort: *„Therapie A ist statistisch signifikant besser als die Standardbehandlung (p=0,03).“* Dieser Satz wird wörtlich in die nun fällige Publikation übernommen und das Ergebnis mit dem Gütesiegel *„p=0,03“* versehen.

Damit wurde einem üblichen, aber nicht unbedingt sinnvollen Ritual Folge geleistet. Die Frage des Arztes wurde damit nicht beantwortet. Der Satz *„Therapie A ist statistisch signifikant besser als die Standardbehandlung (p=0.03)“* bedeutet nicht das, was die meisten glauben. Er bedeutet *nicht*: *„Wenn ich Therapie A für besser als Therapie B halte, dann irre ich mich mit einer Wahrscheinlichkeit von 3 Prozent“*. Letztere Interpretation ist schlicht falsch, auch wenn es in ansonsten grundsoliden Statistikbüchern so

stehen mag. Häufig werden Signifikanzniveau und Irrtumswahrscheinlichkeit verwechselt.

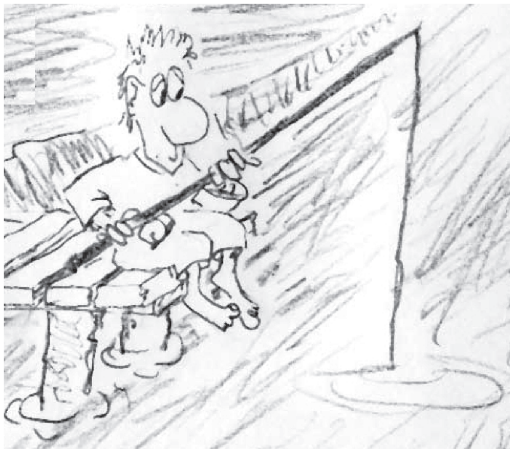
Die wahre Bedeutung des fraglichen Satzes ist eine ganz andere: *„Falls die beiden Therapien gleich gut sind, dann beträgt die Wahrscheinlichkeit 3 Prozent, dass die beobachteten oder noch extremere Ergebnisse auftreten"*. Der Unterschied zwischen der falschen Interpretation und der wahren Bedeutung ist keine Haarspalterei, sondern möglicherweise der häufigste und folgenreichste Irrtum der modernen internationalen medizinischen Forschung. Woher weiß ich, ob die Therapien gleich gut sind? Und was ist, wenn sie nicht gleich gut sind (was ich ja vielleicht sogar hoffe)? Im Moment stehen wir noch im Dunkeln. Um ein erstes Licht in die Sache zu bringen, machen wir einen virtuellen Kurzurlaub im fernen Syldavien.

III. Der Fischtest: Angeln in Syldavien

Syldavien lockt mit endlosen Wäldern und Seen. Einzig und allein in der Syldavischen Seenplatte kann man die schmackhaften Leckerellen angeln (*Abbildung 1*). Außer diesen gibt es hier nur noch eine einzige andere Fischart: die Ekelitzen. Die schmecken wie sie heißen, und selbst hungrige Katzen greifen da lieber zum Dosenöffner. Damit Sie möglichst viele Leckerellen angeln, haben Sie sich einen zertifizierten und entsprechend teuren Köder beschafft. Dieser Köder verspricht, dass 90 Prozent der Leckerellen anbeißen, die sich dem Köder auf weniger als einen Meter genähert haben. Die meisten Ekelitzen hingegen verschmähen den Köder. Wenn sie sich dem Köder auf weniger als einen Meter genähert haben, beißen sie nur in 5 Prozent der Fälle an.

Sie genießen die Ruhe, den Blick in den wolken- und mückenlosen syldavischen Himmel gerichtet ... und plötzlich beißt ein Fisch an. Wie groß ist nun die Wahrscheinlichkeit, dass das eine Ekelitze ist? Bitte sehen Sie sich den Fisch nicht an und lesen Sie nicht weiter, ohne über die Frage nachgedacht zu haben!

Abbildung 1: Angeln in Syldavien



Richtig! Die Frage ist so gar nicht beantwortbar. Die Wahrscheinlichkeit, dass eine Ekelitze am Haken hängt, hängt davon ab, wie groß der Leckerellen-Anteil am Fischbestand des Sees ist. Betrachten wir doch einfach einmal die Extreme. Wenn es keine Ekelitzen mehr gibt, dann zappelt immer eine Leckerelle am Haken. Das andere Extrem ist, dass es keine Leckerellen mehr gibt. Dann nützt Ihnen der beste Köder nichts und Sie haben nie eine Leckerelle gefangen, wenn es am Haken zieht. Nur mit den Anbißwahrscheinlichkeiten auf der Gebrauchsanleitung des Köders können sie Ihre Frage offensichtlich nicht beantworten. Sie benötigen noch eine dritte Größe: den Anteil der Leckerellen.

IV. Testeigenschaften eines Feuermelders

Etwas technischer, aber immer noch sehr alltäglich ist die Arbeitsweise eines Feuermelders. Sie ist in *Tabelle 1* schematisch dargestellt. Ein Haus kann für diese Betrachtung genau zwei relevante Dinge tun: es brennt oder es brennt nicht. Der Feuermelder selbst bietet auch keine große Vielfalt an Aktivitäten: entweder er schlägt Alarm oder er lässt es bleiben.

Wenn es brennt, wünschen wir uns, dass er Alarm gibt. Ein schlechter Feuermelder kann an dieser Stelle versagen, und trotz Feuer keinen Alarm geben. Das kann sehr gefährlich und sehr teuer werden. Eine hohe Ansprechwahrscheinlichkeit, die im Falle eines Brandes möglichst zuverlässig zu einem Alarm führt, ist also wünschenswert und zeichnet einen guten Feuermelder aus.

Wenn es nicht brennt, wollen wir keinen Alarm. So ein Fehlalarm kann ebenfalls gefährlich und teuer werden. Wenn die ganze Belegschaft eines Bürohauses durch das Treppenhaus auf die Straße stürmt, sind blaue Flecken und umgeknickte Füße nicht auszuschließen. Hinzu kommt der Arbeitsausfall und eventuell die Kosten für einen unnötigen Feuerwehreinsatz. Kurzum: Eine geringe Fehlalarmwahrscheinlichkeit ist wünschenswert und zeichnet ebenfalls einen guten Feuermelder aus.

Tabelle 1: Testeigenschaften eines Feuermelders (Erläuterung im Text)

	Feuermelder	
	Alarm	Kein Alarm
Das Haus brennt	Richtiger Alarm *(Ansprechwahrscheinlichkeit)*	falsch
Das Haus brennt nicht	Falscher Alarm *(Fehlalarmwahrscheinlichkeit)*	richtig
Wahrscheinlichkeit, dass ein Alarm richtig ist:	Anzahl richtiger Alarme/ Summe aller Alarme	

Wichtig zu merken: Ein Feuermelder hat zwei Qualitätsmerkmale. Wenn es ein guter Feuermelder ist, hat er 1. eine geringe Fehlalarmwahrscheinlichkeit (d.h. selten Alarm, wenn es nicht brennt) und 2. eine hohe Ansprechwahrscheinlichkeit (d.h. sehr oft Alarm, wenn es tatsächlich brennt).

Der Gebrauchsanleitung ihres gerade erstandenen Feuermelders entnehmen Sie folgendes: Die Fehlalarmwahrscheinlichkeit beträgt 5 Prozent. An 100 feuerfreien Tagen gibt es also 5 Fehlalarme (es gibt der Einfachheit halber maximal einen Alarm pro Tag). Die Ansprechwahrscheinlichkeit beträgt laut Gebrauchsanleitung 90 Prozent. Wie häufig werden Sie in hundert Tagen (es gibt der Einfachheit halber maximal einen Brand pro Tag) durch Fehlalarm aufgeschreckt und wie häufig ist es ein echter Alarm? Bitte denken Sie darüber nach!

Richtig, man muss wissen, wie häufig es dort, wo der Feuermelder steht, überhaupt brennt. Steht er völlig allein auf einem unbrennbaren Betonsockel im friesischen Watt fernab etwaiger Öl- oder Erdgasfelder, dann ist es wohl immer ein Fehlalarm, wenn das Gerät Alarm schlägt. Je mehr die Umgebung jedoch feuergefährdet ist, umso wahrscheinlicher wird es ein echter Alarm sein. Um die obige Frage zu beantworten, werden, wie bei den Leckerellen insgesamt, drei Größen benötigt: die Fehlalarm- und die Ansprechwahrscheinlichkeit, die beide ausschließlich Eigenschaften des Feuermelders sind, sowie eine Aussage darüber, wie häufig es in der Umgebung des Feuermelders überhaupt brennt.

V. Signifikanztests in klinischen Studien

Prof. McResult hat eine neue Therapie entwickelt (wofür oder wogegen spielt hier keine Rolle) und an 50 Patienten ausprobiert. Es wird ein etwas größerer Anteil Patienten geheilt als mit der Standardtherapie, die seit Jahren im ganzen Lande eingesetzt wird. Nur ein Patient hatte bei Prof. McResult eine Nebenwirkung. Das entspricht 2 Prozent. Die Standardtherapie, an tausenden von Patienten erprobt, hat ebenfalls 2 Prozent derselben Nebenwirkung. Auf einer eilig einberufenen Pressekonferenz verkündet Prof. McResult, dass die neue Therapie besser wirkt bei gleich bleibender Rate an Nebenwirkungen. Wer jetzt noch nach der alten Therapie verfahre, der beginge einen Kunstfehler und mache sich strafbar.

Vorsicht, Vorsicht! Zwei Dinge können Prof. McResult in seiner Euphorie entgangen sein. Erstens: die Ergebnisse seiner Therapie waren in dieser einen Studie nur zufällig besser. Wenn zwei gleich starke Fußballmannschaften gegeneinander spielen, ist das Ergebnis auch nicht immer unentschieden. Zweitens: Hätte nur einer mehr von McResults Patienten eine Nebenwirkung gehabt, dann hätte die Nebenwirkungsrate bereits 4 Prozent betragen und somit mehr (das Doppelte!) als in der Standardtherapie. Mit anderen Worten, es kann gut sein, dass Prof. McResult etwas wichtiges übersehen hat. Wenn zwei ungleiche Fußballmannschaften aufeinander treffen, kann es eben trotzdem ein „unentschieden“ geben. Mit welcher Wahrscheinlichkeit der eine oder der andere Fehler begangen werden, sind Fragen, deren Beantwortung Sache der Statistik ist. Damit sind wir beim Signifikanztest angelangt.

Erstaunlicherweise haben ein Signifikanztest in einer klinischen Studie und ein Feuermelder sehr viel gemeinsam (*Tabellen 1* und *2*). In einer Studie werden zwei Therapien A und B verglichen. Die beiden Therapien können in diesem Zusammenhang nur zwei relevante Eigenschaften haben: sie sind tatsächlich unterschiedlich oder sie

sind es nicht. Der Signifikanztest selbst bietet wie der Feuermelder nur zwei Varianten: Er gibt dem Ergebnis das Prädikat „statistisch signifikant“ oder er tut es nicht.

Wenn Therapie A und Therapie B tatsächlich unterschiedlich sind, möchten wir, dass der Signifikanztest uns dies mit dem Gütesiegel „statistisch signifikant“ möglichst zuverlässig anzeigt. Wie beim Feuermelder wünschen wir uns eine möglichst hohe Ansprechwahrscheinlichkeit. Sie hat in diesem Zusammenhang einen anderen Namen; sie wird *Power* genannt. Wie fast alle Vergleiche hinkt auch dieser, allerdings ohne die Analogie zu zerstören. Die Ansprechwahrscheinlichkeit des Feuermelders hängt nicht unbedingt von der Größe des Brandes oder dem brennenden Material ab. Die *Power* einer Studie ist um so größer, je größer der Unterschied der Wirkung von Therapie A und Therapie B ist. Schließlich ist es ja auch viel wahrscheinlicher, ein im Heuhaufen verlorenes Hufeisen wieder zu finden, als eine Stecknadel. Die *Power* steigt auch mit der Anzahl der Patienten, die in einer Studie ausgewertet werden. Dies sagt vielen bereits die Intuition. Einer Studie mit 1000 Patienten trauen wir es eher zu, einen Unterschied von fünf Prozentpunkten, z.B. zwischen zwei Nebenwirkungsraten, aufzudecken, als einer Studie mit nur 100 Patienten oder gar nur 10 Patienten.

Von den drei Größen *Power*, *Unterschied* und *Anzahl* der Patienten muss man zwei kennen, um die dritte berechnen zu können. Die für die Studienplanung wichtige Frage „Wie viele Patienten brauche ich für eine Studie mit 80 Prozent Power?“ scheint damit auf den ersten Blick unbeantwortbar, weil der Unterschied ja nicht bekannt ist. Würde ich ihn kennen, müsste und dürfte ich die Studie gar nicht durchführen. Entscheidend ist in dieser Situation nicht der tatsächlich vorhandene Unterschied, sondern wie groß der Unterschied sein müsste, damit er relevant ist und auch praktische Konsequenzen aus dem Ergebnis gezogen werden würden. Ein Hustensaft, der bei 6.200 von 10.000 Patienten wirkt, statt nur bei 6000 wie das Vergleichspräparat, hat rein mathematisch zunächst die Nase vorn. Aber wohl kaum ein Praktiker wird in diesem Fall die Verbesserung um zwei Prozentpunkte als relevant erachten. Man muss sich vor der Studie im klaren sein über den „minimalen klinisch relevanten Unterschied“, den man als eine Verbesserung ansehen würde.

Für die weiteren Belange dieser Abhandlung ist wichtig zu merken: Die *Power* ist die Wahrscheinlichkeit, den minimalen klinisch relevanten Unterschied aufzuzeigen, sofern er tatsächlich vorhanden ist. Wünschenswert ist natürlich eine möglichst hohe *Power*. In klinischen Studien wird meistens eine *Power* von 80 oder 90 Prozent angestrebt.

Eine *Power* von 80 Prozent bedeutet gleichzeitig, mit 20 Prozent Wahrscheinlichkeit (= 100 Prozent – *Power*) den vorhandenen Unterschied zu übersehen. Dies entspricht dem Versagen des Feuermelders bei einem Brand. Das Übersehen hat in der Statistik den Namen „Fehler zweiter Art“. Einen Unterschied zu übersehen, kann bedeuten, dass eine zum Greifen nahe liegende Therapieverbesserung doch nicht zum Einsatz kommt oder dass die Erhöhung einer Nebenwirkungsrate nicht bemerkt wird.

Tabelle 2 zeigt im Vergleich mit *Tabelle 1* die Analogie zum Feuermelder. Bisher wurde der Fall erläutert, in dem Therapie A besser ist als Therapie B.

Was ist, wenn die beiden Therapien sich tatsächlich nicht unterscheiden? Um diesen Fall richtig zu verstehen, ist es günstig, sich zunächst in ein Würfelspiel hineinzudenken.

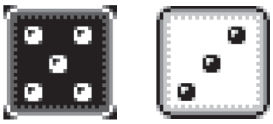
Tabelle 2: Die Tabelle und darin enthaltene Begriffe werden nach und nach im Text erläutert

	Signifikanztest	
	Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	Richtig signifikantes Ergebnis *(Power)*	Falsch (Fehler 2. Art)
Therapie A ist nicht besser als Therapie B	Falsch signifikantes Ergebnis oder Fehler 1. Art *(p-Wert)*	Richtig
Positiver prädiktiver Wert:	Anzahl richtig signif. Ergebnisse/ Summe aller signif. Ergebnisse	

Tabelle 3: Mögliche Ergebnisse beim Werfen von zwei Würfeln

	Schwarz	Weiß	Differenz		Schwarz	Weiß	Differenz
1	6	1	5	19	3	3	0
2	6	2	4	20	2	2	0
3	5	1	4	21	1	1	0
4	6	3	3	22	5	6	–1
5	5	2	3	23	4	5	–1
6	4	1	3	24	3	4	–1
7	6	4	2	25	2	3	–1
8	5	3	2	26	1	2	–1
9	4	2	2	27	4	6	–2
10	3	1	2	28	3	5	–2
11	6	5	1	29	2	4	–2
12	5	4	1	30	1	3	–2
13	4	3	1	31	3	6	–3
14	3	2	1	32	2	5	–3
15	2	1	1	33	1	4	–3
16	6	6	0	34	2	6	–4
17	5	5	0	35	1	5	–4
18	4	4	0	36	1	6	–5

Abbildung 2: Ergebnisse einer spielerischen Studie, in der ein schwarzer Würfel mit einem weißen Würfel verglichen wurde



Sie haben einen schwarzen und einen weißen Würfel. Sie werfen beide, schwarz zeigt eine „5“, weiß eine „3“ (*Abbildung 2*). Schwarz hat jetzt zwar gewonnen, aber würden Sie folgern, dass der schwarze Würfel tatsächlich besser (d.h. so präpariert ist, dass er systematisch höhere Zahlen zeigt) ist als der weiße? Wahrscheinlich nicht, denn ob ein tatsächlicher Unterschied zwischen den beiden Würfeln besteht, bekommen Sie nur heraus, indem sie mehrmals werfen. In der Forschung hieße dies, eine ganze Studie zu

wiederholen, was natürlich weitaus schwieriger (u.U. sogar unmöglich) und zeitaufwändiger ist als einfach noch einmal zu würfeln.

In *Tabelle 3* sind alle Ergebnisse aufgelistet, die bei der Begegnung „Schwarzer Würfel gegen Weißer Würfel" möglich sind. Wenn beide Würfel „gleichermaßen gut" sind, also keiner irgendwie präpariert („gezinkt") ist, dann sind all diese Ergebnisse gleich wahrscheinlich. Es gibt 36 Möglichkeiten. Sie sind hier nach der Differenz Schwarz – Weiß sortiert. In 10 von 36 Fällen, also 10/36 = 0,28 = 28 Prozent der Fälle, hat schwarz zwei oder mehr Punkte Vorsprung vor weiß. Diese 28 Prozent sind der p-Wert des 5:3 Ergebnisses. Die Frage: „Wie wahrscheinlich ist es, dass schwarz zufällig 2 oder mehr Punkte Vorsprung hat, obwohl beide Würfel gleich gut sind?" wird durch diesen p-Wert beantwortet.

Übertragen auf eine klinische Studie lautet die Frage: „Wie wahrscheinlich ist es, dass die Ergebnisse von Therapie A und Therapie B sich so, wie beobachtet, oder extremer unterscheiden, obwohl beide Therapien gleichermaßen wirksam sind?" Die Antwort ist der p-Wert. Dessen Berechnung ist allerdings selten so übersichtlich wie bei unserem Würfelspiel. Der grundlegende Gedankengang ist derselbe und nur um den soll es hier gehen.

Wann ist denn das Ergebnis nun statistisch signifikant? Die Antwort ist kurz und einfach: wenn der p-Wert weniger als 5 Prozent beträgt. Dieses so genannte 5-Prozent-Signifikanzniveau fällt vom Himmel. Es ist reine Konvention. Ob diese 5 Prozent sinnvoll sind, werden wir später näher betrachten. Aber wir sollten es hier schon einmal auf die Würfel anwenden. Hätte schwarz mit einer „6" gegen eine weiße „1" gewonnen, hätten wir einen p-Wert von 1/36 = 0,028 = 2,8 Prozent[1] erhalten, denn nur in einem von 36 Fällen ist schwarz um 5 Punkte (oder mehr) besser als weiß. In der medizinischen Forschung würde man nun nach einem einzigen Wurf sagen können, dass der schwarze Würfel „statistisch signifikant" besser ist als der weiße und meistens daraus folgern, dass er tatsächlich und systematisch besser ist.

Nun kehren wir zurück zu der Frage, was ist, wenn die beiden Therapien sich tatsächlich nicht unterscheiden. Mit der gerade wiedergegebenen Definition der statistischen Signifikanz und des p-Wertes werden mit 5 Prozent Wahrscheinlichkeit die Ergebnisse von Therapie A und Therapie B sich soweit unterscheiden, dass sie als statistisch signifikant gelten. Mit 5 Prozent Wahrscheinlichkeit wird daher ein Fehlalarm ausgelöst. In der Statistik wird der Fehlalarm „Fehler erster Art" genannt.

Eine Studie zeichnet sich also durch zwei Angaben aus. Erstens, die *Power*, die angibt, mit welcher Wahrscheinlichkeit ein tatsächlich vorhandener Unterschied aufgezeigt wird (tatsächlich vorhandene Leckerellen zum Beißen verführt werden oder bei einem tatsächlichen Brand der Feuermelder Alarm schlägt) und, zweitens, das Signifikanzniveau, das angibt, mit welcher Wahrscheinlichkeit bei tatsächlich nicht vorhandenem Unterschied Zufallsergebnisse für bare Münze genommen werden (eine Ekelitze am Haken hängt oder ein Fehlalarm ausgelöst wird).

Um *Tabelle 2* mit Leben zu füllen, benötigen wir so etwas wie die Häufigkeit der Leckerellen oder der Brände. Sonst können wir nichts rechnen. Die Wahrscheinlichkeit, dass in einer Studie tatsächlich etwas untersucht wird, das besser ist als die Stan-

1 Für Insider: Dies gilt für einen einseitigen Signifikanztest, der voraussetzt, dass schwarz nur besser, aber nicht schlechter als weiß sein kann.

dardbehandlung („A ist besser als B") soll im Folgenden als „Wahrscheinlichkeit guter Ideen" bezeichnet werden. In der nächsten Tabelle beträgt sie 10 Prozent. Die Zahl fällt vom Himmel, aber mit irgendetwas muss man anfangen. Dieser Punkt wird später noch diskutiert werden.

Unter insgesamt 1000 Studien (*Tabelle 4*) ist Therapie A entsprechend einer Wahrscheinlichkeit guter Ideen von 10 Prozent in 100 Studien tatsächlich besser als in Therapie B. Auf Grund der Power von 80 Prozent werden davon 80 richtig erkannt und erhalten das Prädikat „statistisch signifikant". In 20 Fällen wird der Unterschied zwischen A und B übersehen. In 900 Studien ist Therapie A besser als B. Entsprechend der Definition des Signifikanzniveaus sind die Unterschiede in den Ergebnissen in 5 Prozent der Fälle zufällig so groß, dass sie fälschlicherweise als „statistisch signifikant" bezeichnet werden. In insgesamt 45 Studien fallen wir also auf ein Zufallsergebnis herein.

Wir haben nun 125 Studien mit signifikantem Ergebnis. Davon gehören 80 zu einem tatsächlichen „A ist besser als B". Die Wahrscheinlichkeit, dass ein signifikantes Ergebnis ein tatsächliches „A ist besser als B" signalisiert, beträgt somit 80/125 = 64 Prozent. Dies ist der so genannte „positive prädiktive Wert" der Studie. In 100 Prozent – 64 Prozent = 36 Prozent der Fälle irren wir uns jedoch.

Wenn wir unseren Forschern unterstellen, dass sie in jeder zehnten Studie eine wirklich gute Idee verfolgen, dass sie eine methodisch perfekte Studie durchführten und dass diese Studie eine Power von 80 Prozent bei einem Signifikanzniveau von 5 Prozent hatte, dann ist die Aussage etwa einer von drei derartiger Studien trotzdem ein Irrtum!

Tabelle 4: Positiver prädiktiver Wert einer einwandfreien klinischen Studie (Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B: 10 Prozent („Wahrscheinlichkeit guter Ideen"). Power: 80 Prozent. Signifikanzniveau: 5 Prozent)

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	100	80% richtig signifikant: 80 Studien	Falsch nicht signifikant: 20 Studien
Therapie A ist nicht besser als Therapie B	900	5% falsch signifikant: 45 Studien	Richtig nicht signifikant: 855 Studien
Summe:	1000	125	
Positiver prädiktiver Wert:		80/125 = 64%	
Irrtumswahrscheinlichkeit:		100% – 64% = 36%	

Wie bei den Leckerellen und dem Feuermelder benötigen wir drei Größen, um den positiven prädiktiven Wert unserer Schlussfolgerung „A ist besser als B" abschätzen zu können. Zwei Größen, die Eigenschaften des Tests sind (p-Wert, power), und eine, die das Gewässer charakterisiert, in dem wir fischen. Letztere, die Wahrscheinlichkeit für gute Ideen, ist unbekannt, da kein verlässliches Verfahren bekannt ist, mit dem man sie abschätzen könnte. Die Eingangs gestellte Frage „*Wie groß ist die Wahrscheinlichkeit,*

dass ich mich irre, wenn ich Therapie A für besser als Therapie B halte?" kann daher nicht beantwortet werden. Erst recht wird sie nicht beantwortet, wenn man zusätzlich noch die Power unberücksichtigt lässt und nur den p-Wert bzw. das Signifikanzniveau kennt.

1. Anzahl der Patienten und *Power*

Sogar bei den in der methodischen Qualität hoch angesiedelten randomisierten kontrollierten Studien erreicht nur ein kleiner Teil tatsächlich eine Power von 80 Prozent (Freiman et al. 1992). Im Mittel (Median) haben real existierende Studien eine Power von lediglich 25 Prozent (hier gerechnet für einen Unterschied von 20 Prozentpunkten zwischen den verglichenen Therapien). Der Grund dafür ist einfach, dass die Studien viel zu wenig Patienten enthalten. Oft hat man das Patientenaufkommen überschätzt und daher nicht genug Patienten zusammenbekommen und einfach aufgehört oder man hat sich überhaupt gar keine Gedanken darüber gemacht.

2. Signifikanzniveau: Es gibt viele Arten, die 5-Prozent-Hürde zu nehmen

Auch mit dem Signifikanzniveau wird es in der klinischen Forschung nicht so genau genommen. Mit Absicht oder aus Unwissenheit, das sei dahingestellt. Wenn wir in unserem Würfelspiel nicht einmal gewürfelt hätten, sondern zweimal, dann wäre das Ergebnis nicht mehr statistisch signifikant, denn die Wahrscheinlichkeit für einen „6:1-Sieg" beim ersten oder zweiten Wurf ist natürlich viel größer als ein „6:1-Sieg" auf Anhieb. In der klinischen Forschung ist es nicht unüblich, zunächst einmal viele Daten zu sammeln, und dann zu schauen, ob nicht irgendetwas Signifikantes dabei ist. Es werden dutzendweise Einflussparameter und Endpunkte auf Signifikanz getestet. Definitionsgemäß ist im Mittel jedes zwanzigste Ergebnis statistisch signifikant, entsprechend 5 Prozent. Und entsprechend vielfältig und widersprüchlich fallen die damit verbundenen klinischen Empfehlungen aus. Ein weiteres probates Mittel ist es, eine Studie zu starten, alle zwei Wochen einen Signifikanztest zu machen und die Studie in dem Augenblick für beendet zu erklären, wenn die favorisierte Therapie signifikant vorne liegt. Zum Vergleich: Beim Fußball steht von vornherein fest, wie lange gespielt wird. Der Trainer darf das Spiel nicht abpfeifen, sowie seine Mannschaft den Führungstreffer erzielt hat. Pocock et al. (1987) fanden, dass in randomisierten Studien im Mittel über 6 Endpunkte berichtet und 4 Signifikanztests gemacht werden. Damit steigt die Wahrscheinlichkeit für einen Fehler erster Art um das vierfache. Das tatsächliche Signifikanzniveau beträgt jetzt nicht mehr 5 Prozent, sondern 20 Prozent.[2] Pococks Feststellung stammt aus dem Jahre 1987. Die mittlere Anzahl an Signifikanztests pro Studie dürfte bis zum Jahre 2004 ganz enorm angestiegen sein, weil mittlerweile verfügbare Statistikprogramme es immer einfacher machen, innerhalb von Minuten Dutzende von Variablen und Subgruppen auf Signifikanz zu testen.

2 Genau genommen beträgt es: $1 - 0{,}95 \times 0{,}95 \times 0{,}95 \times 0{,}95 = 1 - 0{,}81 = 0{,}19 = 19\%$.

Tabelle 5: Positiver prädiktiver Wert einer ansonsten einwandfreien klinischen Studie mit einer Power von 25 Prozent (diese geringe Power ist der Normalfall in klinischen Studien; es wurden vier Signifikanztests durchgeführt; das effektive Signifikanzniveau beträgt daher 20 Prozent; Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B: 10 Prozent)

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	100	25% richtig signifikant: 25 Studien	Falsch nicht signifikant: 75 Studien
Therapie A ist nicht besser als Therapie B	900	20% falsch signifikant: 180 Studien	Richtig nicht signifikant: 720 Studien
Summe:	1000	205	
Positiver prädiktiver Wert:		25/205 = 12%	
Irrtumswahrscheinlichkeit:		100% – 12% = 88%	

3. Eine schon etwas realistischere Studie

Wir nehmen den realistischen Fall an, dass die Power einer Studie nur 25 Prozent beträgt. Ferner werden vier Signifikanztests statt einem gemacht, wodurch das effektive Signifikanzniveau 20 Prozent beträgt. Die einzelnen Zahlenwerte sind in *Tabelle 5* eingetragen. Der positive prädiktive Wert beträgt in diesem Fall nur noch 12 Prozent. Nur jede achte Studie mit einem signifikanten Ergebnis zeigt eine tatsächliche Verbesserung an. Von 8 Publikationen ist nur eine brauchbar – aber welche?

4. Ein pragmatischer Verbesserungsvorschlag

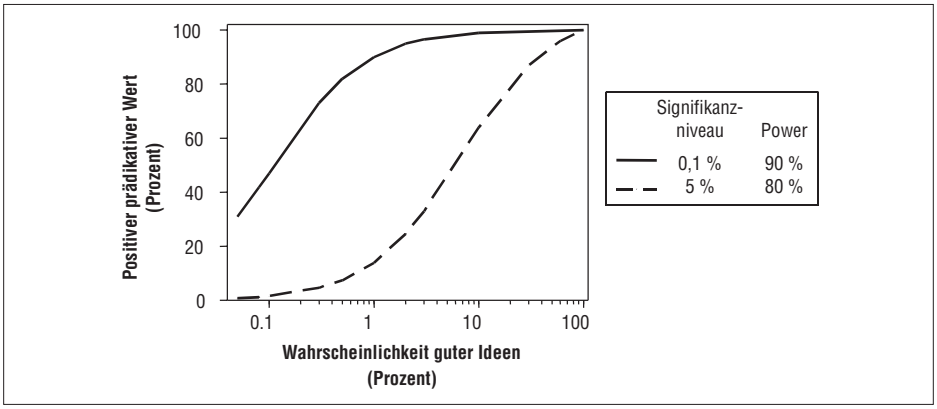
Offenbar führt der gegenwärtige Umgang mit Signifikanztests zu einer unvertretbar hohen Fehlerquote. Glücklicherweise gibt es einen sehr einfachen und wirksamen Verbesserungsvorschlag (Sterne und Smith 2001): eine Power von 90 Prozent und ein Signifikanzniveau von 0,1 Prozent. Im Vergleich zum bislang angestrebten (90 Prozent Power und 5 Prozent Signifikanzniveau) erfordert dieser Vorschlag etwa doppelt so viele Patienten in einer Studie. Was es uns einbringt, können wir mit *Tabelle 6* ausrechnen. Damit wir bei ganzen Zahlen bleiben können, wird jetzt von 10.000 Studien ausgegangen. 10 Prozent davon, also 1000, untersuchen eine tatsächliche Verbesserung. Diese wird auf Grund der Power von 90 Prozent in 900 Studien auch festgestellt, indem das Ergebnis das Gütesiegel „statistisch signifikant“ erhält. Mit dem neuen Signifikanzniveau bringen die 9000 Studien, in denen Therapie A und Therapie B gleichwertig sind, 9 Fehlalarme ein. Der resultierende positive prädiktive Wert beträgt somit 900/909 = 0,99 = 99 Prozent.

Da macht das Lesen von Publikationen wieder Spaß, wenn man mit einer so hohen Gewissheit etwas Dauerhaftes dazu lernen kann. Auch wenn die Wahrscheinlichkeit guter Ideen geringer ausfallen sollte als hier angenommen, ist der prädiktive Wert noch

Tabelle 6: Positiver prädiktiver Wert von Studien mit 90 Prozent Power und einem Signifikanzniveau von 0,1 Prozent. Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B: 10 Prozent („Wahrscheinlichkeit guter Ideen")

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	1.000	90% richtig signifikant: 900 Studien	Falsch nicht signifikant: 100 Studien
Therapie A ist nicht besser als Therapie B	9.000	0,1% falsch signifikant: 9 Studien	Richtig nicht signifikant: 8991 Studien
Summe:	10.000	909	
Positiver prädiktiver Wert:		900/909 = 99%	
Irrtumswahrscheinlichkeit:		100% – 99% = 1%	

Abbildung 3: Positiver prädiktiver Wert von Studien mit 90 Prozent Power und einem Signifikanzniveau von 0,1 bzw. 80 Prozent Power und 5 Prozent Signifikanzniveau bei unterschiedlicher Wahrscheinlichkeit guter Ideen (Der positive prädiktive Wert gibt an, mit welcher Wahrscheinlichkeit das Ergebnis einer Studie richtig ist. Die gegenwärtige medizinische Forschung spielt sich weit unterhalb der gestrichelten Kurve ab.)



immer recht gut (*Abbildung 3*). Mit 1 Prozent guten Ideen beträgt er 90 Prozent. Sollte nur jede tausendste Idee gut sein, sinkt der positive prädiktive Wert auf 40 Prozent. Immerhin wäre noch ungefähr jede zweite Studie es wert, sie zu lesen.

Sehr viel geringer fällt der positive prädiktive Wert für eine Power von 80 Prozent bei dem üblichen Signifikanzniveau von 5 Prozent aus. Da die genannte Power und das Signifikanzniveau wie oben besprochen in den wenigsten Studien tatsächlich erreicht wird, spielt sich die gegenwärtige medizinische Forschung weit unterhalb der gestrichelten Kurve ab.

5. Ein Gedanke zur Wahrscheinlichkeit guter Ideen

Zehn Prozent gute Ideen? Das ist doch viel zu wenig, oder? Greifen wir zu den Sternen: Hundert Prozent gute Ideen sind es ganz gewiss nicht. Dann könnte man sich gleich alle Studien sparen und müsste Wissenschaftler nur noch wie ein Orakel befragen. Wie wäre es mit 50 Prozent? Das bedeutet, dass ein Experte gleichfalls 50 Prozent Irrtumswahrscheinlichkeit hätte. Wenn sich fünf Experten einig sind über eine neue Behandlung, dann beträgt die Irrtumswahrscheinlichkeit nur noch 0,5 x 0,5 x 0,5 x 0,5 x 0,5 = 0,03125 = 3,125 Prozent. Das ist weniger als 5 Prozent. Eine derartige Studie wäre also bereits statistisch signifikant, bevor sie überhaupt begonnen hat.

VI. Geschichtliches: ein alter Hut

Die in diesem Artikel dargestellten Gedanken basieren im Wesentlichen auf den Überlegungen des englischen Geistlichen Thomas Bayes (1702–1761). Sein wichtigstes Werk wurde erst nach seinem Tode von John Canton in den Jahren 1763 und 1764 in den *Philosophical Transactions of the Royal Society of London* (Band 53, S. 376–399 und Band 54, S. 298–310) unter dem Titel „An Essay Towards Solving a Problem in the Doctrine of Chances" veröffentlicht. Diese Arbeit markiert einen wichtigen Wendepunkt in der Entwicklung der Wahrscheinlichkeitsrechnung.

VII. Was wird für gute Studienplanung benötigt?

Ein Signifikanzniveau von 0,1 Prozent statt der üblichen 5 Prozent ist wünschenswert. Das liefert erheblich weniger falsch positive Ergebnisse und damit auch weniger Literatur über falsch positive Ergebnisse. Die derzeit gigantische Verschwendung von forscherischem Enthusiasmus und materiellen Ressourcen für sinnlose Zufallsergebnisse wird dadurch entscheidend reduziert.

Eine Studie sollte mindestens eine Power von 90 Prozent für den untersuchten Endpunkt haben, um relevante Unterschiede überhaupt aufdecken zu können, und weil eine hohe Power maßgeblich zum prädiktiven Wert einer Studie beiträgt. Je weniger Patienten in einer Studie sind, um so geringer ist deren Power und um so geringer ist der prädiktive Wert der Studie. Und das auch bei positivem Ergebnis! Es ist für die Planung also unerlässlich, die benötigte Patientenzahl vor Beginn einer Studie sachlich abzuschätzen und mit dem tatsächlich vorhandenen Patientenaufkommen zu vergleichen.

Die Planer einer klinischen Studie sollten Experten auf dem medizinischen Gebiet sein, auf dem mit der Studie geforscht wird. Das erhöht die Wahrscheinlichkeit, dass in der Studie tatsächlich eine gute Idee verfolgt wird. Es ist unbedingt ratsam, bereits bei der Planung einen Statistiker mit einzubeziehen.

Der Einsatz korrekter Forschungsmethodik ist unerlässlich. Hier konnten nur wenige Punkte erwähnt werden. Man kann ganze Bücher darüber schreiben (z.B. Beck-Bornholdt und Dubben 2001, 2003). Die zu messenden Endpunkte (Heilungsrate oder Blutdruck oder Überlebenszeit oder ...) müssen eindeutig und vor Beginn der

Studie definiert werden. Damit dies für Außenstehende nachvollziehbar ist, sollte es ein Studienprotokoll geben, das nach Möglichkeit bei neutraler Stelle vor Studienbeginn hinterlegt wird.

VIII. Konsequenz für Studieninterpretation

Dies ist der Abschnitt für alle, die Publikationen zu klinischen Studien lesen (müssen). Ein Großteil dessen, was in „wissenschaftlichen" Journalen gedruckt wird, ist nicht valide. Die Kunst des Lesens besteht daher maßgeblich darin, möglichst frühzeitig zu erkennen, ob die Publikation, mit der man sich gerade beschäftigt, es überhaupt wert ist.

Wichtig ist, dass der p-Wert *nicht* die Irrtumswahrscheinlichkeit ist. In vielen Statistikbüchern steht dies anders. Die gegenwärtig normale Praxis, ausschließlich den p-Wert zu berücksichtigen, ist nicht geeignet, wahre Ergebnisse von Zufallstreffern hinreichend zuverlässig zu unterscheiden. Sie ist aber geeignet zu „beweisen", dass der Papst ein Außerirdischer ist (Beck-Bornholdt und Dubben 1996).

Vorsicht ist geboten bei Studien mit geringer Power, sprich: bei Studien mit nur wenigen Patienten. Der Erkenntnisgewinn durch solche Studien ist gering. Auch bei positivem Ergebnis ist der prädiktive Wert sehr wahrscheinlich sehr gering. Dies wird von Statistikern häufig heruntergespielt oder gar verneint. Studien mit zu geringer Power haben geringe Aussagekraft.

Achten Sie darauf, dass ein primärer Endpunkt *a priori* festgelegt wurde. Bei mehreren und bei unklar definierten Endpunkten ist der wahre p-Wert eines Ergebnisses u.U. sehr viel größer als der angegebene p-Wert bzw. das angegebene Signifikanzniveau. Das Ergebnis ist dann nicht interpretierbar.

Last not least sollte man sich fragen, ob man das, was in einer Studie untersucht wurde, überhaupt für eine gute Idee hält. Und noch ein last-not-least: Es gibt unermesslich viele weiterer Fehler und Irrtümer, mit denen die Aussagekraft teurer und aufwendiger Studien zunichte gemacht werden kann. Falls Sie ein paar davon kennen lernen möchten, und auch erfahren möchten, welches Unwesen sie im täglichen Leben treiben, empfehlen wir unsere beiden Bücher.

Literatur

Beck-Bornholdt, Hans-Peter, und *Hans-Hermann Dubben*, 2003: Der Schein der Weisen – Irrtümer und Fehlurteile im täglichen Denken. Reinbek: Rowohlt Verlag.

Beck-Bornholdt, Hans-Peter, und *Hans-Hermann Dubben*, 2001: Der Hund, der Eier legt – Erkennen von Fehlinformation durch Querdenken. Reinbek: Rowohlt Verlag.

Beck-Bornholdt, Hans-Peter, und *Hans-Hermann Dubben*, 1996: Is the Pope an Alien? Nature 381: 730.

Freiman, J.A., T.C. Chalmers, H. Smith und *R.R. Kuebler*, 1992: The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Controlled Trial. S. 357–373 in: *John C. Bailar III* und *Frederick Mosteller* (Hg.), Medical Uses of Statistics. Boston, MA: New England Journal of Medicine Books.

Pocock, S.J., M.D. Hughes und *R.J. Lee*, 1987: Statistical Problems in the Reporting of Clinical Trials. A Survey of Three Medical Journals. New England Journal of Medicine 317: 426–32.

Sterne, J.A.C., und *G.D. Smith*, 2001: Sifting the Evidence – What's Wrong with Significance Tests? British Medical Journal 322: 226–231.

ZÄHLEN ODER MESSEN?

Jürgen Rost

Zusammenfassung: In den empirischen Sozialwissenschaften werden Indikatoren ausgezählt und Itemantworten aufaddiert und die Anzahlen und Summen als Skalenwerte bestimmter Variablen bezeichnet. Das Rasch-Modell ist das einzige Modell für Fragebogendaten, bei dessen Geltung die ungewichtete Summe der Itemantworten die gesamte statistische Information in den Daten über den Ausprägungsgrad einer metrischen latenten Variable ausschöpft und die Items unterschiedliche Schwierigkeit besitzen können. Der Beitrag stellt den Ablauf einer Testanalyse nach dem derzeitigen Stand der Messtheorie von Rasch dar, dies jedoch unter der Voraussetzung, dass es keine präzise a priori-Hypothese über die psychometrische Struktur der analysierten Testitems gibt. Dabei kommen die folgenden Modelle bzw. Modelltypen zur Anwendung: Rasch-Modell, partial credit model, rating scale model, latent class analysis, Klassenmodelle für ordinale Daten, mixed Rasch-Modell und Hybrid-Modelle. Zur Illustration der Methoden einer Skalenanalyse, die quantifizierende wie klassifizierende Modelle einbezieht, wurden die neun Items zur Verhaltensbeurteilung aus dem ALLBUS 2002 ausgewählt. Diese Skalenanalyse hat zu vier Gruppen von Personen geführt, innerhalb derer das Auszählen von Summenscores eine sinnvolle Messoperation darstellt. Jede dieser Gruppen zeichnet sich durch ein spezifisches Antwortprofil aus, das gleichsam den Referenzrahmen für die Interpretation des Summenscores darstellt.

I. Einleitung

Messen heißt in den empirischen Sozialwissenschaften und speziell in der Umfrageforschung oft Auszählen und Addieren. Eine Variable wird gemessen, indem man eine Reihe von Indikatoren erhebt oder einige Fragen stellt und die Indikatorwerte oder die Antworten auf die Fragen gleichsinnig codiert und auszählt bzw. aufaddiert. Natürlich ist die Addition eine zulässige und häufig gebrauchte Messoperation. So wird der Verschmutzungsgrad der Hautoberfläche durch Auszählen der Bakterien pro Flächeneinheit und das Gewicht eines Goldklumpens auf einer altertümlichen Balkenwaage durch Addieren der zum Aufwiegen benötigten Gegengewichte gemessen.

Die Messoperation ist dann sinnvoll, wenn man weiß, *welche* Bakterien man auszählen muss, um ein sinnvolles Konstrukt von „Verschmutzung" zu messen. Die Addition ist dann sinnvoll, wenn man weiß, wie schwer die einzelnen Normgewichte sind und die Waage so konstruiert ist, dass wirklich nur deren Gewicht das Messergebnis beeinflusst.

In den empirischen Sozialwissenschaften werden Indikatoren ausgezählt und Itemantworten aufaddiert und die Anzahlen und Summen als Skalenwerte bestimmter Variablen bezeichnet. Wozu braucht man dann noch *Methoden* der Messung und Skalierung? Es könnte ja sein, dass diese Summenwerte gar nichts messen, sondern nur Zah-

len sind – Hausnummern, wie man auch sagt (obwohl Hausnummern im Referenzsystem einer bestimmten Straße sehr wohl etwas messen).

Es gibt drei Wege zu begründen, warum ein Summenwert einen Messwert darstellt. Erstens, die theoretische Begründung, die darin besteht, dass man eine gute Theorie über jedes Item hat. Diese Theorie muss begründen, warum die aufsummierte Itemantwort additiv zum Ausprägungsgrad der gemessenen Variable beiträgt. Je nach Anspruch, den man an eine Theorie stellt, könnte man sagen, dass es solche Theorien in den Sozialwissenschaften sehr selten gibt, oder dass es derer zu viele gibt. Tatsächlich gibt es in den Sozialwissenschaften für jede Variable eine große Zahl möglicher Operationalisierungen oder zumindest die Möglichkeit, sie auf verschiedene Weise in Items umzusetzen. Als „theoretische Begründung", warum ein Test nun gerade so aussieht, wie er ist, reichen unsere theoretischen Annahmen im Allgemeinen nicht aus.

Der zweite Weg ist die post-hoc Legitimation, d.h. man erklärt eine Zahlenreihe zu Messwerten und schaut im Nachhinein, mit welchen anderen Variablen sie zusammenhängt. Nach dieser Sichtweise ist kein Test jemals ein schlechter Test, da nie mit Sicherheit ausgeschlossen werden kann, dass er doch noch mit einer interessanten Variable korreliert. Obwohl dieser „Validitätspragmatismus" illustre Vertreter hat, z.B. Frederik Lord (vgl. Rost 2004), ist er wohl mit verantwortlich für die unüberschaubare Flut sozialwissenschaftlicher Forschungsergebnisse.

Der dritte Weg besteht darin, eine Theorie darüber zu haben, wie ein empirischer Indikator oder die einzelne Itemantwort mit der zu messenden Variable funktional zusammenhängt, und diese Theorie einem ernsthaften Falsifikationsversuch auszusetzen (dem sie standhält). Eine Legitimation für die Summenbildung resultiert daraus natürlich nur, wenn im Rahmen dieser Theorie abgeleitet werden kann, dass die Summenwerte die besten Schätzer für die zu messende Variable sind. Es gibt verschiedene formale Theorien (zur Unterscheidung von formalen Modellen und Theorien siehe Rost 2004), die den Zusammenhang von Itemantwort und zu messender Variable so formalisieren, dass der Summenwert zugleich der Messwert für die Personen ist. Das ist bei der (deterministischen) Guttman-Skala so, die annimmt, dass die Itemantworten in Abhängigkeit vom Messwert der Person mit der Wahrscheinlichkeit 0 oder 1 gegeben werden; aber auch bei der Mokken-Skala, die überschneidungsfreie, aber sonst undefinierte Itemfunktionen annimmt; und beim Binomial-Modell, das konstante Itemschwierigkeiten annimmt (Rost 2004). Gilt eines dieser Modelle für die Items, die eine bestimmte Variable messen sollen, so ist tatsächlich die Summe der Itemantworten einer Person ihr *Messwert.*

Die Messwerte unter diesen drei Modellen haben jedoch unterschiedliche Skalenqualität. Gilt die Gutmann- oder die Mokken-Skala, so haben die Summenwerte nur die Qualität einer Ordinalskala, dürfen also z.B. nicht mit korrelationsstatistischen Modellen weiter verarbeitet werden. Unter Geltung des Binomial-Modells haben die Summenwerte zwar das Niveau einer Intervallskala, dafür impliziert dieses Modell jedoch, dass alle Items gleich schwierig sind.

Es gibt ein viertes Modell, bei dem die Items unterschiedlich schwierig sein dürfen, und die Summenwerte zwar nicht selbst Intervallskalenniveau haben, es aber eine einfache, streng monotone und schon „fast" lineare Transformation der Summenwerte in intervallskalierte Messwerte gibt. Das ist das Rasch-Modell (Rasch 1960).

Das Rasch-Modell ist das einzige Modell für Fragebogendaten (darunter wird eine Personen × Items-Matrix mit nominalen oder ordinalen Itemantworten verstanden), bei dessen Geltung die *ungewichtete* Summe der Itemantworten die gesamte statistische Information in den Daten über den Ausprägungsgrad einer metrischen quantitativen latenten Variable ausschöpft und die Items unterschiedliche Schwierigkeit besitzen können.

Dieses Modell eignet sich deshalb nicht nur für den „dritten" Weg der Legitimation von Summenwerten, sondern kann auch eingesetzt werden, den „ersten" Weg zu kontrollieren. Wird postuliert, dass die Items gemeinsam additiv zum Ausprägungsgrad der zu messenden Variable beitragen, so muss das Rasch-Modell auf die betreffenden Testdaten passen. Wenn dies nicht der Fall ist, stimmt die Theorie (für diese Daten) nicht. Diese etwas ausführliche Einleitung soll ausdrücken, dass das Rasch-Modell nicht irgendein Modell ist, das man einsetzen kann oder – zugunsten eines besseren Modells – auch nicht. Vielmehr modelliert das Rasch-Modell genau das, was wir täglich in der empirischen Sozialforschung tun: Summen für Messwerte halten.[1]

Der Beitrag stellt den Ablauf einer Testanalyse nach dem derzeitigen Stand der Messtheorie von Rasch dar, dies jedoch unter der Voraussetzung, dass es keine präzise a priori-Hypothese über die psychometrische Struktur der analysierten Testitems gibt. Idealerweise sollte es so eine a priori-Theorie geben. Dann würde sich die Testanalyse auf die Prüfung des einen betreffenden Modells beschränken. Im vorliegenden Fall wurde bewusst das explorative Vorgehen gewählt, um die Vielfalt der Modellvarianten vorzustellen und die damit verbundene Logik der Auswertungsschritte darzustellen. Dabei kommen die folgenden Modelle bzw. Modelltypen zur Anwendung: Rasch-Modell, partial credit model, rating scale model, latent class analysis, Klassenmodelle für ordinale Daten, mixed Rasch-Modell und Hybrid-Modelle. Sie sind im Detail in Rost (2004) beschrieben.

Ein Nachweis der *externen* Validität sind die vier latenten Klassen, die bei dieser Testanalyse „herauskommen", also der Nachweis eines Zusammenhangs mit anderen Variablen, wird nicht geführt. Vielmehr wird der Test so ausgewertet, dass die Testergebnisse *interne* Validität (Modellgeltung) besitzen und somit die notwendige *Voraussetzung* für eine mögliche externe Validität. Im Unterschied zu Lienert (1961), der die (externe) Validität als (oberstes) Gütemerkmal von Tests auffasst, geht Fischer (1974) davon aus, dass die externe Validität nicht mehr zum Gegenstandsbereich der Testtheorie gehört. Wenn man einen erwarteten Zusammenhang mit einer externen Variable *nicht* findet, so hat man vielleicht eine irrelevante Variable gemessen – aber man *hat* eine gemessen.

1 Einführungen in die Messtheorie von Rasch finden sich in Rost und Spada (1978), Rost und Strauss (1992) sowie in Rost (1999a, 2004), vertiefende Darstellungen in Rost (1999b, 2002), Langeheine und Rost (1988), Rost und Langeheine (1997), Fischer und Molenaar (1995), Hambleton und van der Linden (1997) und Boomsma et al. (2001). Zu einer aktuellen Diskussion der Rasch-Messtheorie siehe Rost (2005).

II. Das Rasch-Modell

Dass sich das Rasch-Modell nach 44 Jahren noch nicht als „die Methode" der Messung und Skalierung in den Sozialwissenschaften etabliert hat, liegt wohl an einigen schwer auszuräumenden Vorurteilen: Das Rasch-Modell
– sei nur für Leistungstests geeignet,
– beziehe sich nur auf dichotome (0–1) Daten,
– könne man nur auf Items anwenden, die speziell dafür konstruiert worden seien,
– kenne keine Item-Trennschärfen,
– sei auf eindimensionale Messungen beschränkt,
– passe in der Regel nicht auf die Daten,
– nötige zur Selektion unpassender Items,
– setze große Stichproben voraus.
Das alles ist falsch. Ebenso falsch wie die Behauptung, das Rasch-Modell müsse erst einmal in Methodenvergleichs-Studien nachweisen, dass es bessere Ergebnisse hervorbringt als das herkömmliche Vorgehen der Skalierung. Man kann nicht eine Theorie mit einer Praxeologie vergleichen: Für die gängige Praxis des Auszählens und Addierens in der Sozialforschung gibt es nur *eine* Theorie, das Rasch-Modell.

Das Auszählen ist zwar die am weitesten verbreitete Skalierungsmethode, jedoch wird sie oft nur mangels geeigneter Alternativen angewandt. Auszählen verbietet sich dann, wenn man an der Interpretation von Antwortprofilen interessiert ist. Bei einer Profilanalyse geht es darum, welche Antworten auf eine Frage typischerweise mit welchen Antworten auf andere Fragen verknüpft sind. Das Auswertungsziel besteht nicht darin, die befragten Personen in die Menge der möglichen Summenscores abzubilden, sondern in eine Menge von typischen Profilverläufen. Während die Menge der Summenscores bekannt ist, müssen die typischen Profilverläufe erst anhand von Daten ermittelt werden.

Auszählen und Profile typisieren sind zwei komplementäre Wege der Analyse von Umfragedaten. Wo die eine Methode angewendet wird, verbietet sich die andere, und es scheint außer zahlreichen Hybridvarianten (die hoch-interessant sein können) keinen dritten Weg neben Auszählen und Profilanalysen zu geben. Es ist der Gegensatz von Quantifizieren und Klassifizieren, der Unterschied zwischen der Messung einer ordinalen oder metrischen Variable und der Erfassung einer nominalen Personenvariable, man kann sogar sagen, es ist die Dichotomie von Quantität (Summenwerte) und Qualität (Profile).

So wie das Rasch-Modell als Urform von Modellen über das Auszählen gleichsinniger Itemantworten angesehen werden kann, gilt die latent class analysis (LCA) von Paul Lazarsfeld als Urform der Klassifikation und Profilanalyse von Itemantworten. Man kann darüber streiten, ob klassifizierende Methoden der Fragebogenanalyse unter die Überschrift „Messen und Skalieren" gehören, schließlich steckt in dem Begriff der Skalierung das (lateinische) Wort für Treppe, was suggeriert, dass Messen erst von der Ordinalskala an aufwärts stattfindet. Von den Naturwissenschaften wissen wir, dass Messungen sogar möglichst auf einer Verhältnisskala liegen sollten. Klassifizieren ist dagegen „nur" Messen auf Nominalniveau.

Interessanter noch als die Frage, ob Klassifizieren schon Messen ist, ist die Frage, was von beiden einen Spezialfall des anderen darstellt. Klassifizieren ist eine schwache Form des Messens, könnte man sagen, bei der es nur wenige unterscheidbare Messwerte gibt und diese noch nicht einmal geordnet sind. Man könnte auch sagen, Messen ist ein Spezialfall von Klassifizieren, bei dem die Klassen eine Ordnung oder eine Metrik haben. Betrachtet man die mathematischen Modelle der beiden Verfahrensgruppen, so wird die Frage eindeutig beantwortbar: Das Rasch-Modell lässt sich als ein entsprechend restringiertes latentes Klassenmodell darstellen (und berechnen), aber die latent class analysis nicht als Rasch-Modell. Insofern ist Klassifikation die allgemeinere, übergeordnete Methode. Die Prüfung, ob eine Menge von Items ein quantitatives Konstrukt erfasst, sollte daher auch über einen Vergleich von quantitativen und klassifikatorischen Modellen erfolgen. Dies werden auch die ersten Schritte in der folgenden Auswertung eines Datenbeispiels aus dem Allbus sein. Alle hier referierten Skalenanalysen wurden mit dem Programm WINMIRA von Davier (2001) vorgenommen.

III. Die Skalenanalyse der Items zur Verhaltensbeurteilung aus dem ALLBUS 2002

Zur Illustration der Methoden einer Skalenanalyse, die quantifizierende wie klassifizierende Modelle einbezieht, wurden die neun Items zur Verhaltensbeurteilung aus dem ALLBUS 2002 ausgewählt. In diesem Aufgabenblock bekommen die befragten Personen die folgenden Verhaltensweisen vorgelegt:

1 *Ein Mann schlägt sein 10-jähriges Kind, weil es ungehorsam war.*
2 *Eine Frau lässt einen Schwangerschaftsabbruch vornehmen, weil sie keine Kinder haben möchte.*
3 *Ein Arzt gibt einem unheilbar kranken Patienten auf dessen Verlangen hin ein tödliches Gift.*
4 *Ein Arbeitnehmer macht absichtlich beim Lohnsteuerjahresausgleich falsche Angaben und erhält dadurch 500 Euro zuviel Lohnsteuerrückerstattung.*
5 *Jemand fährt mit öffentlichen Verkehrsmitteln, ohne einen gültigen Fahrausweis zu besitzen.*
6 *Ein Mann zwingt seine Ehefrau zum Geschlechtsverkehr.*
7 *Jemand raucht mehrmals in der Woche Haschisch.*
8 *Ein Mann hat homosexuelle Beziehungen zu einem anderen Mann.*
9 *Ein verheirateter Mann hat mit einer anderen Frau ein Verhältnis.*

Die befragten Personen sollten zu jeder Verhaltensweise angeben, ob sie persönlich das beschriebene Verhalten für sehr schlimm, ziemlich schlimm, weniger schlimm oder für überhaupt nicht schlimm halten. Es wurden diejenigen 2.673 Fälle aus dem ALLBUS 2002 in die Analysen einbezogen, die bei keiner der neun Fragen eine fehlende Antwort aufweisen. Im Folgenden wird bewusst darauf verzichtet, die Auswertung von einer Theorie über das Antwortverhalten bei diesen neun Fragen leiten zu lassen, obwohl dies der wissenschaftstheoretisch befriedigendere Weg wäre. Die Entscheidung über das

am besten passende Modell wird vielmehr einem globalen Gütekriterium überlassen, nämlich dem Bayes information criterion, BIC.[2]

1. Erfassen die Items der Verhaltensbeurteilung ein quantitatives oder typologisches Konstrukt?

Wenn eine quantitative latente Variable das Antwortverhalten hinsichtlich der neun Items bestimmt, dürfen sich die Personen nur hinsichtlich ihres Zustimmungsgrades bezüglich *aller* Items unterscheiden, d.h. es darf keine Interaktion zwischen den Items und der Personeneigenschaft geben. Die latente Variable würde in diesem Fall die Tendenz der Personen ausdrücken, die genannten Verhaltensweisen als mehr oder weniger schlimm zu beurteilen. Auch wenn die Annahme eines derartigen eindimensionalen latenten Konstruktes nicht besonders sinnvoll ist, lässt sich die Skala natürlich unter dieser Prämisse analysieren.

Die Anwendung des Rasch-Modells (Abk. RM in *Tabelle 2*) auf die Daten ergibt das folgende Bild. Die Skala hat zwar eine mittelhohe Reliabilität von 0,65 bis 0,74 je nach Berechnungsmethode,[3] weist aber sehr schlechte Item-fit-Werte auf. In der folgenden *Tabelle 1* ist der Q-Index[4] wiedergegeben (Rost und von Davier, 1994), der insbesondere die Items 1 und 6 als extrem schlecht zum Rasch-Modell passend ausweist. Diese beiden Items sind zugleich die „leichtesten", d.h. diejenigen Verhaltensweisen, die man am ehesten als „sehr schlimm" einschätzt (die Antwortvariablen wurden

Tabelle 1: Itemparameter und Item-fit-Maße im Rasch-Modell. Als „Schwierigkeit ist der logistische Itemparameter des Rasch-Modells bezeichnet (der Z-Wert erlaubt die Prüfung der statistischen Signifikanz des Q-Index)

Item	Schwierigkeit	Q-Index	Z-Wert
1	–0.97293	0.3587	5.9112
2	0.42982	0.1643	–1.6888
3	0.81590	0.1994	0.8750
4	0.20254	0.1927	–1.3848
5	0.25264	0.1697	–2.8022
6	–1.36812	0.3451	2.3817
7	–0.23602	0.1388	–2.6831
8	1.05578	0.1786	0.8639
9	–0.17961	0.1727	–2.2782

2 Die Informationskriterien, zu denen auch der BIC gehört, setzen den Wert der Likelihoodfunktion für ein Modell mit der Anzahl unabhängiger Parameter des Modells und der Stichprobengröße in Beziehung. Sie drücken damit aus, wie gut die Daten zum Modell passen (je kleiner der BIC, desto besser passt das Modell).

3 Die Reliabilität ist als varianzabhängiges Konzept nicht im Rahmen der Rasch-Messtheorie definiert. Man kann aber die klassische Definition übernehmen und die beiden benötigten Varianzen (wahre und beobachtete Varianz) im Rahmen des Rasch-Modells schätzen. Hierfür gibt es verschiedene Möglichkeiten (Rost 2004).

4 Der Q-Index ist ein auf den bedingten Patternwahrscheinlichkeiten beruhendes Abweichungsmaß für den Itemfit. Je größer der Q-Wert, desto schlechter passt das Item.

umgepolt, so dass ein hoher Wert stets eine Einschätzung als „schlimm" bezeichnet). Es handelt sich um die beiden Items, die Gewalt in der Familie ansprechen.
Die Items 5, 7 und 9 haben laut Z-Wert einen statistisch signifikanten overfit, d.h. die Items passen „zu gut" zum Modell. Um die Gründe für diese Modellabweichungen zu klären, wurden die 1- bis 5-Klassenlösungen des Modells der Latent-Class-Analyse (LCA1 bis LCA5 in *Tabelle 2*) berechnet. *Tabelle 2* gibt in jeder Zeile nach dem Kürzel für das betreffende Modell (die Kürzel werden jeweils im Text eingeführt) die Werte der logarithmierten Likelihoodfunktion wieder. Diese Werte steigen mit zunehmender Klassenanzahl an, da mit jeder neuen Klasse weitere Modellparameter hinzukommen. Informationskriterien wie der BIC setzen den Wert der Likelihoodfunktion mit der Anzahl der Modellparameter Parameteranz.) in Beziehung und erlauben somit eine Abwägung zwischen Aufwand (Anzahl der Modellparameter) und Ertrag (Wert der Likelihoodfunktion).

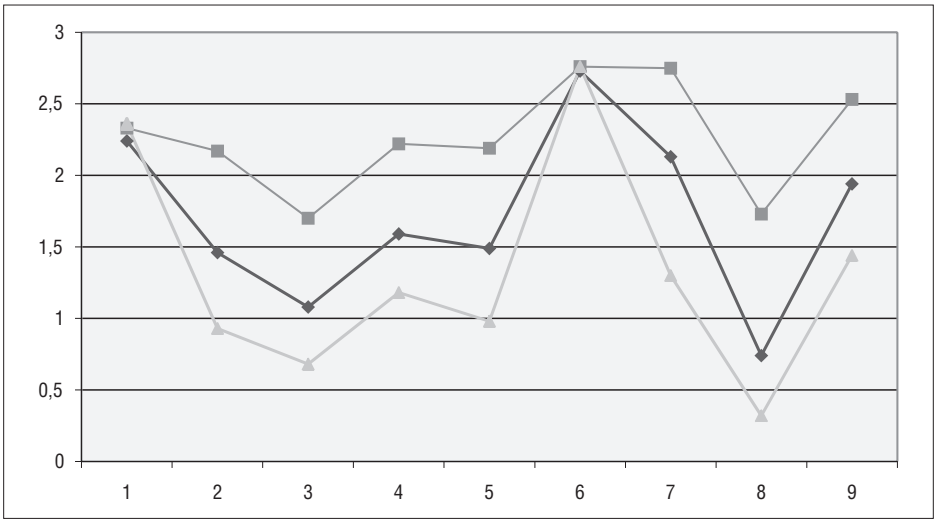
Tabelle 2: Modell-fit-Statistiken

Modell	Log-Likelihood	Parameteranz.	BIC
RM	–27117	53	54652
LCA1	–28058	27	56329
LCA2	–26978	55	54391
LCA3	–26727	83	54110
LCA4	–26635	111	54147
LCA5	–26517	139	54131
HYB3RM1SV	–26685	84	54033
MRM3SV	–26558	86	53795
MRM4SV	–26446	115	53800
HYB4DIS2RS1	–26498	83	53652

Nach dem BIC-Kriterium passt die 3-Klassen-Lösung der LCA (LCA3) am relativ besten, für jede weitere latente Klasse steigt der BIC wieder an. Nach diesem Ergebnis ist es hinreichend, nur drei Gruppen (Klassen) von Respondenten zu unterscheiden, wenn man innerhalb der Klassen stochastische Unabgängigkeit annimmt. *Abbildung 1* zeigt die mittleren Itemantworten in diesen drei latenten Klassen.

Es zeigt sich, dass es sich um drei geordnete Klassen handelt, d.h. Klassen, die sich lediglich in ihrem Antwort*niveau* hinsichtlich der Items unterscheiden. Eine Ausnahme bilden wiederum die Items 1 und 6, die überhaupt nicht zwischen den drei Klassen diskriminieren. Während die sieben übrigen Items geordnete Klassen anzeigen und somit eine quantitative Personenvariable suggerieren, zerstören die beiden Gewalt-Items, nämlich das Schlagen des eigenen Kindes bzw. die Vergewaltigung in der Ehe dieses Bild. Eine Eliminierung dieser beiden Items stellt keine Lösung dar, denn es soll ein Modell für die Verhaltensbeurteilung *aller* neun Items gefunden werden. Vielleicht spiegeln die beiden Gewalt-Items wichtige Unterschiede im Urteil der befragten Personen wider. Offensichtlich tun sie dies aber nicht im Rahmen eines eindimensionalen Rasch-Modells und auch nicht im Rahmen der normalen Latent-Class-Analyse.

Abbildung 1: Die Profile der erwarteten Itemantworten für die 3-Klassen LCA



2. Gibt es eine skalierbare Teilpopulation?

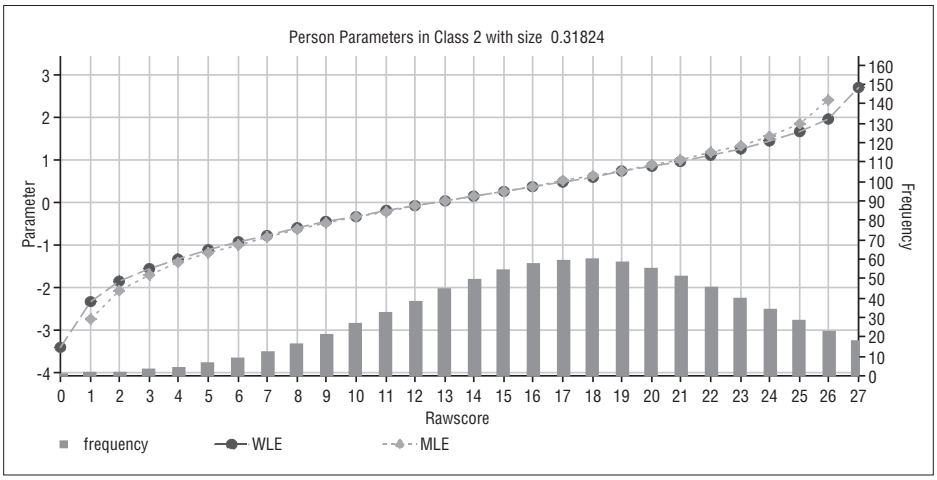
Obwohl die bisherigen Ergebnisse dafür sprechen, dass zumindest zwei Items nicht zu der Skala passen, besteht die Möglichkeit, dass diese vermeintliche Heterogenität der *Items* in Wirklichkeit eine Heterogenität der *Personen* ausdrückt. Vielleicht gibt es eine Subpopulation der befragten Personen, in der das eindimensionale Rasch-Modell gilt, während für den Rest der Population ein so genanntes „unskalierbares Antwortverhalten" vorliegt. Diese Hypothese kann mit Hilfe eines so genannten Hybrid-Modells getestet werden, bei dem für eine Subpopulation (latente Klasse) die Gültigkeit des Rasch-Modells angenommen wird, während eine unbekannte Anzahl von weiteren latenten Klassen dazu dient, die unskalierbaren Personen abzubilden.

Da sich in den Analysen sehr schnell zeigte, dass ein 2-Klassenmodell – welches auch immer – nicht auf die Daten passt (die entsprechenden Werte sind in *Tabelle 2* nicht aufgeführt), wurde zur Identifikation einer Klasse von Skalierbaren ein 3-Klassen Hybridmodell gerechnet, bei dem in einer Klasse die Geltung des Rasch-Modells angenommen wird (HYB3RM1SV).[5] Um die Parameteranzahl gering zu halten, aber auch weil sich eine skalierbare Teilpopulation idealerweise durch eine annähernd normalverteilte Personenvariable auszeichnet, wurde die Scoreverteilung in dieser Klasse durch eine 2-parametrige Funktion angepasst (v. Davier 2001; Rost 2004). Die anderen beiden Klassen sind „normale" Klassen mit konstantem Antwortverhalten für alle Personen. Der BIC-Wert dieses Modells macht im Vergleich zu den bisher behandelten Modellen einen großen Sprung (s. *Tabelle 2*), das Modell könnte also passen.

5 Das Kürzel bedeutet, dass ein Hybrid-Modell mit 3 Klassen (HYB3) angenommen wurde, in dem in einer Klasse das Rasch-Modell gilt (RM1), bei welchem die Scoreverteilung restringiert wurde (SV).

In der Klasse, in der das Rasch-Modell gilt, hat der Fragebogen eine Reliabilität von 0,69 bis 0,76 (je nach Berechnungsmethode, s.o.), also einen etwas höheren Wert als in der Gesamtpopulation. Das hängt damit zusammen, dass die Scoreverteilung sehr breit ist, d.h. nennenswerte Häufigkeiten (> 10) von Score 7 bis 27 aufweist (s. *Abbildung 2*). Diese Klasse scheint tatsächlich die Personen zu enthalten, die eine sehr starke quantitative Variation in ihrem Antwortverhalten aufweisen.

Abbildung 2: Die Scoreverteilung in der Rasch-Klasse des Hybrid-Modells



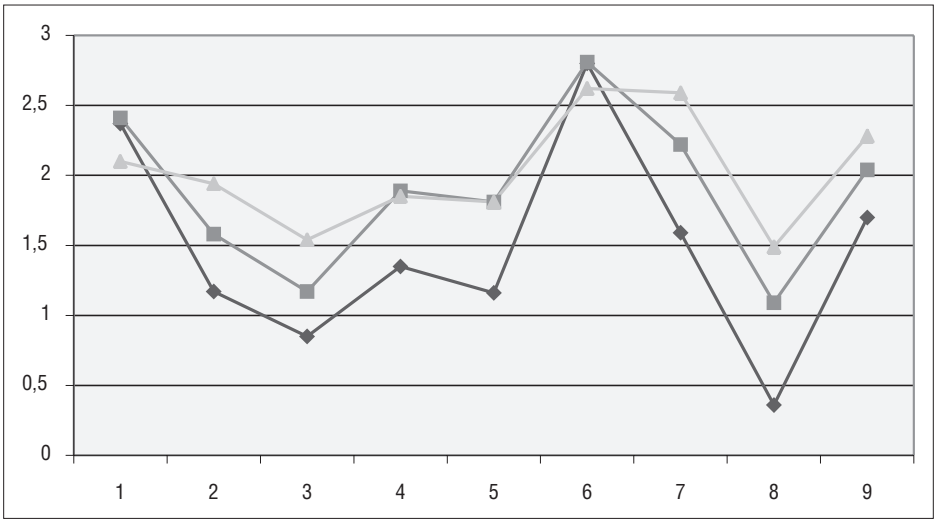
Allerdings umfasst diese Klasse nur 32 Prozent der befragten Personen, so dass dieses Modell 68 Prozent als unskalierbar ausweist, was eine schwer zu akzeptierende Größenordnung ist. Weitere Hinweise auf die Interpretationswürdigkeit dieses Modells geben die in den Klassen erwarteten Itemantworten (s. *Abbildung 3*).

Auffällig ist hier, etwa im Vergleich zu *Abbildung 1*, dass zwei Profile bei den Items 4 und 5 übereinander liegen. Das Profil mit dem Dreieck hat bei den beiden Gewalt-Items eine geringfügig niedrigere Einschätzung. Da das am besten passende Modell aber noch nicht gefunden ist, wird auf die weitere Interpretation dieses Hybridmodells verzichtet.

3. Gibt es Teilpopulationen, in denen die Items etwas unterschiedliches messen?

Im Hybridmodell wurde nur für eine latente Klasse angenommen, dass sie im Sinne des Rasch-Modells skalierbar sei. Da diese Klasse jedoch nicht größer war als die beiden „normalen" Klassen, liegt es nahe, auch für die beiden restlichen Klassen die Geltung des Rasch-Modells anzunehmen. Diese Modelle, bei denen in jeder Klasse die Geltung des Rasch-Modells angenommen wird, nennt man Mixed-Rasch-Modelle (MRM). Tatsächlich sinkt der BIC-Wert für das 3-Klassen mixed Rasch-Modell (mit Glättung der Scoreverteilung in allen drei Klassen) noch einmal deutlich ab (s. *Tabelle 2,* MRM3SV).

Abbildung 3: Die Erwartungswertprofile des 3-Klassen Hybridmodells



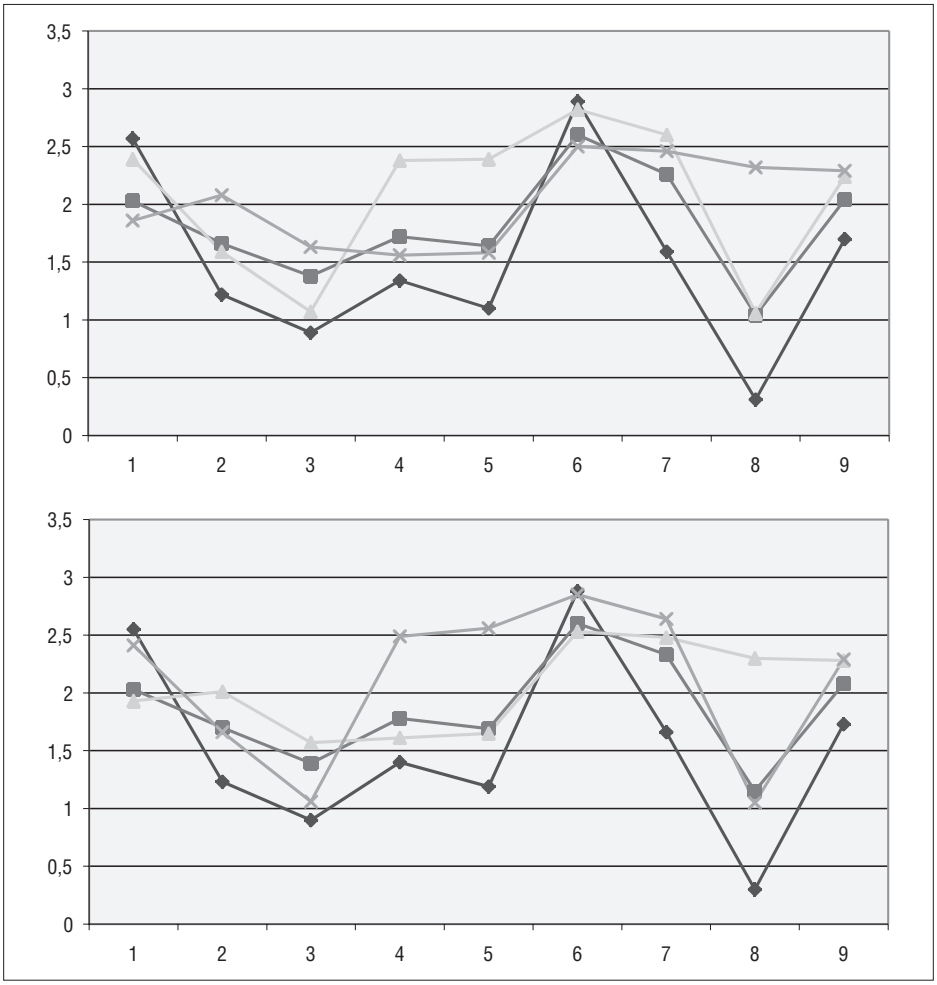
Die 4-Klassenlösung bringt dagegen wieder einen leichten Anstieg des BIC-Wertes (s. *Tabelle 2*, MRM4SV). Der Anstieg der Log-Likelihood um 112 Punkte wird mit 29 zusätzlichen Parametern „erkauft", was sich im BIC-Wert entsprechend negativ niederschlägt. Da es sich um mehrkategorielle, genauer gesagt um 4-stufige Antwortvariablen handelt, werden pro Item 3 Schwellenparameter geschätzt, ohne dass diese für die Interpretation der Ergebnisse von ausschlaggebender Bedeutung wären. Es können also noch weitere Parameter eingespart werden, indem man die Schwellenabstände restringiert.

Das Ratingskalen-Modell parametrisiert nur den Mittelpunkt der Schwellen als Parameter der Itemschwierigkeit und mit 2 weiteren Parametern die für alle Items als konstant angenommenen Schwellenabstände (Andrich 1978). Das Ratingskalen-Modell wurde für eine der 4 Klassen des mixed Rasch-Modells angenommen. Das Dispersionsmodell verbreitert oder schmälert mit einem zusätzlichen (Dispersions-)Parameter pro Item die Schwellendistanzen des Ratingskalen-Modells. Das Dispersionsmodell wurde für 2 weitere Klassen des mixed Rasch-Modells angenommen, was das gesamte mixed Rasch-Modell ebenfalls zu einer Art Hybridmodell macht, da unterschiedliche Modelle in den latenten Klassen spezifiziert werden.

Obwohl die Modellierung der Schwellendistanzen auch ein Gegenstand der präexperimentellen Hypothesenbildung sein sollte, wurde in diesem Fall das (laut BIC-Wert) am besten passende 4-Klassenmodell durch Ausprobieren gefunden. Hier geht es zunächst nur um die Einsparung von Modellparametern, auf die Interpretation der Schwellendistanzen wird im nächsten Abschnitt eingegangen.

Von ausschlaggebender Bedeutung ist jedoch die Frage, ob mit beiden 4-Klassenlösungen dasselbe gemessen wird, d.h. ob die Itemprofile in den 4 Klassen unter beiden Modellen vergleichbar sind. *Abbildung 4* zeigt, dass das in sehr hohem Ausmaß der Fall ist. Die Profilverläufe sind fast identisch.

Abbildung 4: Die Erwartungswertprofile unter zwei unterschiedlich restringierten mixed Rasch-Modellen; im oberen 4-Klassenmodell wurden nur die Scoreverteilungen geglättet (MRM4SV); im unteren 4-Klassenmodell wurde zusätzlich in 2 Klassen das Dispersionsmodell und in einer Klasse das Ratingskalenmodell angenommen (HYB4DIS2RS1)



Das in diesen Abbildungen zum Ausdruck kommende Ergebnis wird hier als das endgültige Resultat der Skalenanalyse angesehen, da kein Modell mit einem besseren BIC-Wert in dieser Modellfamilie gefunden wurde. Es hat eine sehr interessante Interpretation, so dass man dieses Modell auch aus inhaltlichen Gründen gegen alternative Modelle favorisieren und verteidigen könnte.

Die größte Klasse 1, die im restringierten mixed Rasch-Modell (*Abbildung 4*, unten) 49 Prozent der befragten Personen umfasst (s. das Profil mit den Karos), hat im Mittel eine eher tolerante Einstellung gegenüber den erfragten Verhaltensweisen, mit

Ausnahme der beiden Gewalt-Items 1 und 6, die diese Hälfte der Bevölkerung sehr schlimm findet. Die zweitgrößte Klasse mit 26 Prozent der Personen (s. das Profil mit den Quadraten) findet im Vergleich zur Klasse 1 die meisten Verhaltensweisen etwas schlimmer, die Gewalt gegen Kinder oder die Ehefrau jedoch etwas weniger schlimm. Ihr Profil liegt im Vergleich zu den drei anderen Profilen eher in der Mitte.

Interessant sind die beiden kleineren Klassen drei und vier. 15 Prozent der Bevölkerung finden es schlimm oder sehr schlimm, wenn zwei Männer eine homosexuelle Beziehung zueinander haben, d.h. sie finden das in etwa genauso schlimm, wie wenn die Ehefrau vergewaltigt wird und schlimmer noch als das eigene Kind zu schlagen. Hier spiegelt sich offenbar Adornos Konstrukt der autoritären Persönlichkeit wider.

10 Prozent der Bevölkerung (das Profil mit den Kreuzen) haben nicht das Problem mit der Homosexualität, halten es aber für sehr schlimm, wenn man sich beim Lohnsteuerjahresausgleich 500 Euro erschwindelt oder im Bus oder der Bahn schwarz fährt. Auch Haschischkonsum wird als schlimm angesehen. Diese Art von Kleinkriminalität ruft hier offenbar eine intensivere moralische Verurteilung hervor als es Fragen der Sterbehilfe oder des Schwangerschaftsabbruchs tun.

Soviel zur Interpretation der „qualitativen" Personenunterschiede, die anhand der Mittelwertsprofile der vier latenten Klassen vorgenommen wurde. Diese Profile lassen leicht vergessen, dass es innerhalb jeder Klasse eine starke Variation im Niveau der Antwortprofile gibt, da ja das Rasch-Modell in jeder Klasse gilt.

4. Sind die Personen in den Klassen Rasch-skalierbar?

Das mixed Rasch-Modell hat vier qualitativ unterschiedliche Typen von Respondenten identifiziert, über die angenommen wird, dass innerhalb jeder dieser Subpopulationen das Rasch-Modell gilt. *Tabelle 3* gibt einige Kennwerte wieder, anhand derer sich die Qualität der Skalierung in den Klassen beurteilen lässt.

Tabelle 3: Klassengröße (Größe), mittlere Zuordnungswahrscheinlichkeit (p(g|x)), Reliabilität (Rel.), Standardabweichung der Messwerte (SD) und die Items mit signifikantem Item(mis)fit

Klasse	Größe	p(g|x)	Rel.	SD	Item Q
1: die Toleranten	.47	.84	.70	.70	8: 3.5
2: die Mittleren	.26	.73	.75	.88	keiner sign.
3: keine Homosexualität	.15	.73	.72	.73	keiner sign.
4: die Kleinbürgerlichen	.12	.73	.70	.70	8: 1.7; p=0.4

Die Respondenten der großen Klasse („die Toleranten") sind mit p=.84 am besten zu identifizieren. Die mittlere Zuordnungswahrscheinlichkeit (p(g|x)) gibt das arithmetische Mittel der Wahrscheinlichkeiten aller Personen an, aufgrund des gezeigten Antwortvektors × der Klasse g anzugehören. Die Skala in dieser Klasse hat aber eine etwas geringere Reliabilität als die zweite Klasse, vermutlich, weil auch die Varianz der Messwerte in dieser Klasse, verglichen mit der zweiten, eher eingeschränkt ist. Hinzu kommt, dass das Homosexualitätsitem 8 in dieser Klasse einen schlechten Itemfit hat,

was an der extremen Schwierigkeit dieses Items in Klasse 1 liegt. 96 Prozent der Personen dieser Klasse finden Homosexualität überhaupt nicht oder weniger schlimm. Vermutlich sind die 4 Prozent, die Homosexualität ziemlich oder sehr schlimm finden, so zufällig über die Skala verteilt, dass eine niedrige Trennschärfe resultiert. In Klasse 4 ist dieses Item „hart an" der Signifikanzgrenze.

Insgesamt gesehen sind die Kennwerte hinreichend gut, um von einer Skalierbarkeit innerhalb der vier Klassen sprechen zu können, d.h. die klassenspezifischen Personenparameter des Rasch-Modells dürfen interpretiert werden. Es fragt sich allerdings, ob sie auch *zwischen* den Klassen verglichen werden können. Generell ist hierzu anzumerken, dass die Itemparameter innerhalb jeder Klasse summennormiert werden, so dass die Personenparameter die „Fähigkeit" einer Person hinsichtlich des rechnerisch mittleren Items ausdrücken. Entsprechend hat auch der Mittelwert der Personenparameter einer Klasse eine Bedeutung, d.h. er drückt die mittlere Eigenschaftsausprägung in dieser Klasse aus (s. *Tabelle 4*, mittlere Spalte). Demzufolge weist die zweite Klasse die höchsten Eigenschaftsausprägungen auf. Diese Überlegung spricht für eine Vergleichbarkeit zwischen den Klassen, denn alle Personenparameter haben dasselbe Bezugssystem, nämlich die mittlere Schwierigkeit aller Items.

Fernerhin sollte man sich vergegenwärtigen, dass jede Person nur einen Summenscore hat, egal welcher latenten Klasse sie einmal zugeordnet wird. Innerhalb einer latenten Klasse sagt der Summenscore alles über die Personeneigenschaft aus, zwischen den Klassen hängt der Personenparameter natürlich auch von den anderen Parametern in dieser Klasse ab, insbesondere den Schwellendistanzen (s. Rost 2002). *Tabelle 4* zeigt in der letzten Spalte am Beispiel des Summenscores 20, dass demselben Score in jeder Klasse deutlich unterschiedliche Personenparameterwerte zugeordnet werden können, im vorliegenden Beispiel der Wert 0.79 für eine Person in der dritten Klasse und 1.85 in der zweiten Klasse.

Tabelle 4: Personenparameter in den Klassen

Klasse	mittlerer Personenparameter	Personenparameter des Scores 20
1	0.228	1.426
2	1.106	1.846
3	0.697	0.789
4	0.865	0.937

In der dritten und vierten Klasse sind die Werte deutlich geringer als in den ersten beiden Klassen. Das liegt daran, dass die Schwellen der Antwortskala in Klasse 3 und 4 sehr dicht beieinander liegen, was Extremantworten eher erleichtert. Als Schwelle bezeichnet man den Wert der gemessenen Variable, für den die Wahrscheinlichkeiten zweier benachbarter Antwortkategorien gleich groß sind. Ein Antwortformat mit fünf geordneten Antwortkategorien hat also vier Schwellen, deren Parameter anhand der Daten ermittelt werden. Liegen die Schwellen dicht beieinander, so sind Antworten in den beiden Extremkategorien wahrscheinlich. Das ist bei den Klassen 3 und 4 des vorliegenden Beispiels der Fall.

Personen mit einem extrem hohen Summenscore erhalten in diesen beiden Klassen einen niedrigeren Personenparameter als Personen mit demselben Score in einer Klasse

mit großen Schwellendistanzen, da ihr Extremscore nur zum Teil auf eine extreme Eigenschaftsausprägung zurückzuführen ist. Zum anderen Teil ist er durch den Gebrauch der Ratingskala bedingt, in diesem Fall durch die Tendenz zum extremen Urteil. Auch diese Überlegung spricht für eine Vergleichbarkeit der Personenparameter zwischen den Klassen, werden doch die Personenparameter gleichsam korrigiert hinsichtlich des Gebrauchs der Antwortskala in der betreffenden Klasse.

Ob es aber wirklich Sinn macht, die Personenparameter zwischen den Klassen zu interpretieren, hängt letztlich von der Semantik der latenten Klassen ab und muss daher in jedem Einzelfall entschieden werden. Schreibt man im vorliegenden Beispiel den latenten Klassen die Bedeutung unterschiedlicher moralischer Referenzsysteme zu, vor deren Hintergrund Verhaltensweisen als „schlimm" beurteilt werden, so stellt sich tatsächlich die Frage, wie man einen Messwert im Referenzsystem des Bildungsbürgertums mit dem Messwert einer autoritären Persönlichkeitsstruktur im Sinne Adornos vergleichen soll. Was bedeutet überhaupt ein Messwert über die Tendenz, Verhaltensweisen „schlimm" zu finden? Ist es so etwas wie die Härte oder Strenge des moralischen Urteils? Eine wichtige Rolle bei der Frage, ob ein Test überhaupt etwas misst und was er misst, spielen die Schwellenparameter, auf die im Folgenden eingegangen wird.

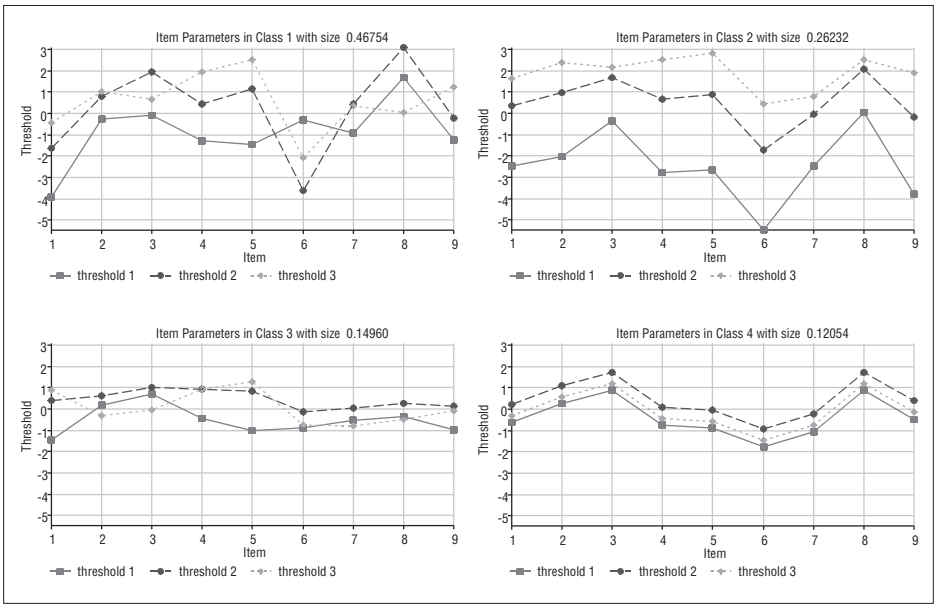
5. Was bedeuten die Schwellenparameter?

An der Frage nach der Interpretation der Schwellenparameter scheiden sich die Meinungen der Experten in zwei Lager. Dieser Streit beginnt bei der Frage, ob das partial credit-Modell überhaupt die sinnvollste Verallgemeinerung des dichotomen Rasch- Modells für ordinale Daten ist (und nicht ein Ansatz wie das so genannte graded response-Modell (Samejima 1969), das nicht Schwellen, sondern Kategoriengrenzen parametrisiert) und geht weiter mit der Auseinandersetzung um die so genannte joining assumption im Rahmen der Jansen-Roskam Debatte (Roskam 1995; Andrich 1995). Schließlich stellt es eine bislang nicht einheitlich beantwortete Frage dar, ob denn die Schwellenparameter geordnet sein müssen, um von „Messung" sprechen zu dürfen, oder ob ihre Anordnung völlig irrelevant ist.

Wirklich praxisrelevant ist die letztgenannte Facette dieser Auseinandersetzung, die Ordnung der Schwellenparameter. Wenn sie für alle Items eines Fragebogens geordnet sind, so hat das die Konsequenz, dass sie eine (für jedes Item unterschiedliche) Partitionierung des latenten Kontinuums definieren. Die Ordnung der manifesten Antwortskala „überhaupt nicht – weniger – ziemlich – sehr schlimm" wird auf das latente Kontinuum projiziert und verleiht diesem die Bedeutung, dass ein höherer Messwert ab einem bestimmten Punkt (der Schwellenlokation) die Wahl der nächst höheren Antwortkategorie impliziert (zumindest mit hoher Wahrscheinlichkeit). Sind die Schwellen nicht geordnet, gibt es nicht für jede Antwortkategorie ein Segment der latenten Variable, in dem die Wahrscheinlichkeit diese Kategorie anzukreuzen maximal ist.

Im vorliegenden Datenbeispiel sind nur in Klasse 2 alle Schwellenparameter geordnet (s. *Abbildung 5*). In Klasse 1 durchbricht die oberste Schwelle bei den Items 3 und

Abbildung 5: Die Schwellenparameter im 4-Klassen mixed Rasch-Modell (die untersten Schwellen (zwischen überhaupt nicht schlimm und wenig schlimm) sind mit Quadraten gekennzeichnet und einer durchgehenden Linie verbunden; die obersten Schwellen (zwischen ziemlich schlimm und sehr schlimm) sind mit Karos gekennzeichnet und mit einer fein gestrichelten Linie verbunden)



8 die Ordnung und die unterste Schwelle bei Item 6 (bedingt durch die hohe Schwierigkeit, Vergewaltigung in der Ehe *überhaupt nicht schlimm* zu finden). In Klasse 3 sind die beiden ersten Schwellen geordnet und Schwelle 3 ist bei vier Items sogar leichter als die erste Schwelle. In Klasse 4, in der das Ratingskalen-Modell gilt, liegen die Schwellen eng beieinander, wobei die oberste Schwelle in der Mitte liegt. Die Anomalien hinsichtlich der obersten Schwelle könnten daran liegen, dass die Kategorie „sehr schlimm" kein deutlich härteres Urteil bedeutet als „ziemlich schlimm". Tatsächlich wird in der Umgangssprache der Ausdruck „ziemlich schlimm" bisweilen noch als Steigerung von „sehr schlimm" benutzt. Eine derart problematische Beschriftung von Ratingkategorien könnte die Erklärung dafür sein, dass die letzte Schwelle nicht den erforderlichen Schwierigkeitsgrad hat, der der höchsten Kategorie einer Ratingskala zukommt.

Der Modellfit wird durch eine Unordnung der Schwellenparameter nicht beeinträchtigt – das ist das Hauptargument derjenigen Experten, die die Ordnung der Schwellen als Kriterium der Skalierbarkeit ablehnen. Insofern spricht nicht viel dagegen, das präferierte 4-Klassen mixed Rasch-Modell trotz ungeordneter Schwellen zu akzeptieren und zu interpretieren. Sollten sich aber Vermutungen über die Ursachen der Unordnung erhärten lassen, spricht viel dafür, die Labels der Ratingskala in weiteren Datenerhebungen zu ändern.

Das Beispiel zeigt, dass in ordinalen Rasch-Modellen nicht nur Personen und Items auf derselben Metrik skaliert werden, sondern auch die Antwortkategorien in Form der zwischen ihnen liegenden Schwellen. Die Schwellenparameter können Hinweise auf die Interpretation der latenten Variable, auf bestimmte Antwortstile (response sets) und auf die Etikettierung der Antwortkategorien geben. Im vorliegenden Beispiel stellt sich die Frage, ob ein moralisches Urteil mit der lapidaren Formulierung „etwas schlimm finden" valide erfasst werden kann. Insbesondere wenn es um ein stark negatives Urteil geht, scheinen „ziemlich schlimm" und „sehr schlimm" nicht hinreichend zu diskriminieren.

6. Gibt es das am besten passende Modell (the best fitting model)?

Die Zeiten, in denen man glaubte, man könne von einem Modell statistisch nachweisen, ob es gilt oder nicht, sind wohl endgültig vorbei. Aber man unterscheidet gerne, ob ein Modell auf die Daten *passt* oder *nicht passt.* Der Unterschied zwischen beiden Redeweisen liegt darin, dass die *Geltung* eines Modells suggeriert, es könne kein anderes Modell gelten. *Passt* ein Modell lediglich auf die Daten, so kann es natürlich andere Modelle geben, die auch passen. Dies ist auch oft der Fall. Bei der Messung, oder genauer, bei der Suche nach einem Testmodell, das eine bestimmte Messoperation legitimiert, sucht man nach einem Modell, das passt, eingedenk der Möglichkeit, dass es andere Modelle gibt, die auch passen. Dies ist Ausdruck einer konstruktivistischen Wissenschaftstheorie.

Der Anspruch, entscheiden zu können, ob ein Modell passt oder nicht, ist schon weniger konstruktivistisch – impliziert er doch ein statistisches Entscheidungskriterium. Bei Testmodellen geht es darum, die Häufigkeitsverteilung der möglichen Antwortmuster zu modellieren, so dass ein Modelltest die Übereinstimmung von beobachteten und erwarteten Patternhäufigkeiten testet. Prüfstatistiken hierfür gibt es zahlreich, Likelihoodquotienten (LR), Pearson-Chi-Quadrat (X2), Cressie&Read-Index (CR) etc., jedoch sind die asymptotischen Voraussetzungen für die Prüfverteilungen in der Regel nicht gegeben (Read und Cressie 1988). Hierzu bieten Bootstrapverfahren eine gute Alternative (v. Davier 1997).

Die Anwendung des Bootstrap auf das 4-Klassen mixed Rasch-Modell ergab folgendes Bild. Von 100 resimulierten Datensätzen wiesen 10 Datensätze einen größeren LR als den beobachteten auf, kein Datensatz eine größere CR-Statistik und 6 Datensätze einen größeren Pearson-Chi-quadrat. Die statistische Entscheidung, ob das interpretierte mixed Rasch-Modell passt, bewegt sich also um die übliche 5 Prozent Marke herum. Es ist zwar beruhigend zu wissen, dass das Modell zumindest annähernd passt, aber eine Gewähr dafür, dass es nicht andere, gleich gut oder besser passende Modelle gibt, ist es nicht.

In der vorangehenden Skalenanalyse wurde konsequenterweise davon gesprochen, welches Modell *relativ* besser passt als ein anderes. Bei dem für diesen Zweck benutzten BIC-Index handelt es sich um eine einfache Straffunktion, die den Wert der Log-Likelihoodfunktion mit der Anzahl der Modellparameter eines Modells in Beziehung setzt. Arbeitet man mit Algorithmen, die die Möglichkeit zu Parametereinsparungen

bieten, so kann man das Ergebnis derartiger Modellvergleiche etwas beeinflussen, wovon in der oben dargestellten Skalenanalyse auch Gebrauch gemacht wurde.

Das Fazit dieser Überlegungen geht dahin, dass man zum Zwecke einer Messung oder Skalierung ein auf die Daten passendes Modell suchen und dabei nach dem am besten passenden streben sollte, auch wenn es das best-passende Model, the best fitting model nur im Rahmen einer wohldefinierten Modellfamilie gibt.

IV. Fazit

Eingangs wurde die Frage erörtert, was eine einfache Messoperation wie das Auszählen von Itemantworten zu einem Vorgang macht, bei dem wirklich Messwerte und keine „Hausnummern" heraus kommen. Es wurde argumentiert, dass die Geltung des Rasch-Modells die Voraussetzung dafür darstellt, dass der Summenscore alles über die Person aussagt, d.h. das vollständige Antwortmuster einer Person keine darüber hinausgehende Information enthält. Oft passt jedoch das Rasch-Modell nicht auf ein Set von Fragen, die eine bestimmte Eigenschaft der Respondenten erfassen sollen. Dies trifft auch für das hier verwendete Datenbeispiel aus dem Allbus zu. Offensichtlich besagt das Antwortprofil doch etwas, was durch den Summenscore nicht ausgedrückt wird.

Der im vorangehenden Abschnitt beschriebene Weg der Skalenanalyse hat zu vier Gruppen von Personen geführt, innerhalb derer das Auszählen von Summenscores eine sinnvolle Messoperation darstellt. Jede dieser Gruppen zeichnet sich durch ein spezifisches Antwortprofil aus, das gleichsam den Referenzrahmen für die Interpretation des Summenscores darstellt. Fernerhin wird demselben Summenscore in jeder Gruppe ein etwas anderer Messwert zugeordnet, d.h. die Bedeutung des Summenscores wird durch die Zugehörigkeit zu den 4 Typen *moderiert.* Man könnte auf die Frage, ob es sich hierbei überhaupt noch um Messung handelt, antworten, dass es eine Art bedingte Messung ist, Messung unter der Bedingung eines bestimmten Personentyps.

Es zeigte sich aber auch, dass die Profilverläufe der latenten Klassen sehr interessante *qualitative* Unterschiede zwischen den Personen beschreiben, während der *quantitative* Messwert *nur* die Intensität des (moralischen?) Urteils ausdrückt. Tatsächlich ist die Zugehörigkeit zu einer latenten Klasse oft die diagnostisch interessantere Information im Vergleich zu dem klassenspezifischen Messwert. Insofern macht es auch Sinn, bei der Klassenzugehörigkeit von einem zweiten Messwert auf Nominalskalenniveau zu sprechen. Das mixed Rasch-Modell misst gewissermaßen zwei Variablen gleichzeitig: Eine qualitative und eine quantitative Variable.

In der durchgeführten Skalenanalyse kamen die direkte Verallgemeinerung des dichotomen Rasch-Modells für ordinale Daten, das so genannte partial credit-Modell, und dessen Restriktionen für Ratingskalen vor; fernerhin das mixed Rasch-Modell, die latent class-Analyse sowie deren Hybrid-Varianten. Daneben gibt es noch eine Reihe weiterer Verallgemeinerungen, die ebenfalls in der Umfrageforschung breite Einsatzmöglichkeiten haben. Diese Verallgemeinerungen lassen sich den drei Richtungen zuordnen: Mehrdimensionale Rasch-Modelle, linear logistische Modelle und multimodale Modelle (s. Rost 1999, 2001). Der Vorteil, dass man sich mit diesen Modellen innerhalb einer Modellfamilie bewegt, liegt darin, dass alle Modelle miteinander verglichen

werden können und man somit das am besten passende Modell in einer großen Menge von Alternativen ermitteln kann.

Das weit verbreitete Vorgehen, die Summe von Antworten auf bestimmte Items „per fiat" zu Messwerten zu erklären, sollte eigentlich der Vergangenheit angehören, auch in der professionellen Umfrageforschung.

Literatur

Andrich, David, 1978: A Rating Formulation for Ordered Response Categories. Psychometrika 43 (4): 561–573.

Andrich, David, 1995: Models for Measurement, Precision, and the Nondichotomization of Graded Responses. Psychometrika 60 (1): 7–26.

Davier, Matthias von, 1997: Bootstrapping Goodness-of-fit Statistics for Sparse Categorical Data: Results of a Monte Carlo Study. Methods of Psychological Research 2 (2), www.mpr-online.de.

Davier, Matthias von, 2000: WINMIRA – A Program System for Analyses with the Rasch Model, with the Latent Class Analysis and with the Mixed Rasch Model. Progamma.

Fischer, Gerhard, 1974: Einführung in die Theorie psychologischer Tests. Bern: Huber.

Fischer, Gerhard, und *Ivo W. Molenaar,* 1995: Rasch-models – Foundations, Recent Developments, and Applications. New York: Springer.

Langeheine, Rolf, und *Jürgen Rost,* 1988: Latent Trait and Latent Class Models. New York: Plenum.

Lienert, Gustav, 1961: Testaufbau und Analyse. Weinheim: Beltz.

Rasch, Georg, 1960: Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Nielsen & Lydiche (expanded edition: Chicago: University of Chicago Press 1980).

Read, TRC, und *Cressie, NAC,* 1988: Goodness-of-fit Statistics for Discrete Multivariate Data. Berlin/Heidelberg/New York: Springer.

Roskam, Edward E., 1995: Graded Responses and Joining Categories: A Rejoinder to Andrich' „Models for measurement, precision, and nondichotomization of graded responses". Psychometrika 60 (1): 27–35.

*Rost, Jürgen,*1997: Logistic Mixture Models. S. 449–463 in: *Wim van der Linden* und *Ronald Hambleton* (Hg.), Handbook of Modern Item Response Theory. Berlin/Heidelberg/New York: Springer.

Rost, Jürgen, 1999a: Test- und Fragebogenanalysen. S. 455–480 in: *Bernd Strauß, Herbert Haag* und *Michael Kolb* (Hg.), Datenanalyse in der Sportwissenschaft. Schorndorf: Hofmann.

Rost, Jürgen, 1999b: Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau 50 (3): 140–156.

Rost, Jürgen, 2001: The Growing Family of Rasch Models, in: *Anne Boomsma, Marc van Duijn* und *Tom Snijders* (Hg.), Essays on Item Response Theory. Berlin/Heidelberg/New York: Springer.

Rost, Jürgen, 2002: When a Questionnaire Fails to be Unidimensional. Psychologische Beiträge 44: 108–125.

Rost, Jürgen, 2004: Lehrbuch Testtheorie – Testkonstruktion. 2. Aufl. Bern: Huber.

Rost, Jürgen, 2005: Die Messtheorie von Rasch in Psychologie und Pädagogik. Lengerich: Pabst Science Publishers.

Rost, Jürgen, und Matthias von Davier, 1994: A Conditional Item Fit Index for Rasch Models. Applied Psychological Measurement 18 (2): 171–182.

Rost, Jürgen, und *Rolf Langeheine* (Hg.), 1997: Applications of Latent Trait and Latent Class Models in the Social Sciences. New York*:* Waxmann.

Rost, Jürgen, und *Hans Spada,* 1978: Probabilistische Testtheorie. S. 59–97 in: *Karl Josef Klauer*, Handbuch für pädagogische Diagnostik, Bd. I. Düsseldorf: Schwann.

Rost Jürgen, und *Bernd Strauß,* 1992: Psychometrics and Test Theory – Recent Developments. German Journal of Psychology 16 (2): 91–119.

Samejima, Fumiko, 1969: Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika, Special monograph.

III. Stichproben, Gewinnung von Daten, Datenqualität

MIKRODATEN DER AMTLICHEN STATISTIK – IHR POTENZIAL IN DER EMPIRISCHEN SOZIALFORSCHUNG

Heike Wirth und Walter Müller

Zusammenfassung: Mikrodaten der amtlichen Statistik stellen eine wesentliche Datenquelle für die empirische Sozial- und Wirtschaftsforschung dar. Ihr Analysepotenzial wird in Deutschland bislang nur wenig ausgeschöpft, weil Kenntnisse darüber in der Scientific Community nur sehr selektiv vorhanden sind und weil eine wissenschaftsadäquate Nutzung dieser Daten bis vor kurzem kaum möglich war. Mit Unterstützung durch das Bundesministerium für Bildung und Forschung werden gegenwärtig erhebliche Anstrengungen unternommen, um die amtlichen Datenbestände für die Forschung zu erschließen und den Zugang zu den Daten zu erleichtern. Der Beitrag gibt einen Überblick über Mikrodaten verschiedener amtlicher Einrichtungen in Deutschland, die für die empirische Wirtschafts- und Sozialforschung besonders bedeutsam sind. Er diskutiert ihre Stärken und Restriktionen und illustriert mit ausgewählten Beispielen aus der Forschungsliteratur das Analysepotenzial einzelner Datenbestände. Der Beitrag schließt mit einem Ausblick auf zukünftige Perspektiven der Nutzung von amtlichen Mikrodaten.

I. Einleitung

Spätestens seit Durkheims Studie über den Selbstmord, welche durch den systematischen Vergleich von Selbstmordstatistiken aus verschiedenen europäischen Ländern soziale Regelmäßigkeiten des Freitodes aufzeigte und hierüber die Notwendigkeit der Soziologie als eigenständiger wissenschaftlicher Disziplin begründete, haben Daten der amtlichen Statistik einen zentralen Platz in der empirisch orientierten Soziologie. Geigers (1932) Untersuchung zur sozialen Schichtung des deutschen Volkes auf Basis von aggregierten Berufs- und Arbeitsstättenzählungsdaten ist ein weiteres klassisches Beispiel in dieser langen Tradition. Nach dem Zweiten Weltkrieg flossen Daten der amtlichen Statistik vorwiegend in die damaligen Analysen und Lehrbücher zur Sozialstruktur der Bundesrepublik ein (Bolte und Aschenbrenner 1964; Bolte und Kappe 1964; Claessens et al. 1965; Peisert 1967). Anfang der 1970er Jahre wurden amtliche Daten zunehmend für die Generierung sozialer Indikatoren genutzt. Der große Stichprobenumfang, die regelmäßigen Erhebungszyklen und die langen Zeitreihen beinhalten ein

hohes Potenzial für die in der wissenschaftsbasierten Sozialberichterstattung angestrebte kontinuierliche Beobachtung des sozialen Wandels. Sowohl in der Soziale Indikatoren-Bewegung als auch bei den frühen Sozialstrukturanalysen wurde hierbei vorwiegend auf Aggregatdaten zurückgegriffen, da die für eine effiziente Analyse von amtlichen Massendaten erforderlichen Ressourcen in der empirischen Sozialforschung zum damaligen Zeitpunkt noch kaum verfügbar waren.

Wenn hier und im Folgenden von amtlichen Daten gesprochen wird, sind damit alle Daten gemeint, deren Generierung – entweder in Form einer Direkterhebung oder prozessproduziert – durch rechtliche oder verwaltungsmäßige Grundlagen geregelt ist. Hierbei kann es sich um förmliche Gesetze,[1] Rechtsverordnungen, spezifische Rechtsvorschriften oder Verwaltungsvorschriften handeln. Damit beinhaltet der Begriff „amtliche Daten" nicht nur die von der Bundesstatistik oder Landesstatistik generierten Daten, sondern auch die bspw. von der Bundesagentur für Arbeit, von der Bundesbank oder von dem Verband deutscher Rentenversicherungsträger produzierten Daten haben im genannten Sinne amtsmäßige Grundlagen. Gleichfalls basieren die von den statistischen Ämtern durchgeführten Bevölkerungsumfragen mit sozialwissenschaftlichem Charakter, z.B. die Zeitbudgeterhebung oder das Europäische Haushaltspanel, auf einer rechtlichen Grundlage[2] und sind in diesem Sinne „amtliche" Daten.

Mit diesem Beitrag wollen wir einen Überblick geben zum gegenwärtigen Stand der Nutzung amtlicher Mikrodaten in der empirischen Sozialforschung. Hierzu skizzieren wir zunächst die wesentlichen Phasen im Zugang zu amtlichen Mikrodaten. Anschließend zeigen wir die Nutzungspotenziale wie auch Restriktionen zunächst allgemein auf, um dann am Beispiel ausgewählter Scientific-Use-Files das Forschungspotenzial dieser Daten darzulegen. Der Beitrag Papier schließt mit einem Ausblick auf zukünftige Perspektiven in der Nutzung amtlicher Mikrodaten.

II. Von Aggregatdaten zu Mikrodaten – Von der absoluten zur faktischen Anonymität von Mikrodaten[3]

Mit der Verbreitung multivariater Analyseverfahren in der empirischen Sozialforschung wie auch den stetigen Verbesserungen der technischen Möglichkeiten nahm seit den 1970er Jahren der Bedarf zu, amtliche Daten für wissenschaftliche Untersuchungen sozialer oder ökonomischer Sachverhalte zu nutzen. In der Soziologie wuchs das Interesse, diese Daten vergleichbar zu sozialwissenschaftlichen Daten als Mikrodaten mit Bezug zu in der Soziologie gebräuchlichen theoretischen Konzepten zu analysieren oder zur Prüfung von entsprechenden Theorien zu nutzen. Parallele Entwicklungen gab es

1 So ist im Bundesstatistikgesetz festgelegt, dass jede Bundesstatistik durch ein spezifisches Gesetz angeordnet werden muss. Beispiele für spezifische Gesetze sind etwa das Mikrozensusgesetz, das Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstands, das Wohnungsstatistikgesetz oder das Gesetz über das Ausländerzentralregister.

2 Hierbei handelt es sich um Erhebungen für besondere Zwecke, die auf Basis von § 7 BStatG (1987) durchgeführt werden.

3 Teilweise enge Anlehnung an Müller (1999a: 8ff.).

in der empirischen Ökonomie zusammen mit dem Bestreben, über Mikrodatensimulationen ökonomische und soziale Auswirkungen von politischen Entscheidungen transparenter zu machen, um so diese Entscheidungsprozesse auf eine rationalere Grundlage zu stellen (Krupp 1968, 2002). Ein bedeutender Akteur in dieser Phase der Erschließung amtlicher Mikrodaten für die Forschung war die sozialpolitische Forschergruppe Frankfurt/Mannheim (Krupp und Zapf 1977). Soziologen und Ökonomen entwickelten gemeinsam auf Basis von Daten des Mikrozensus und der Einkommens- und Verbrauchsstichprobe ein Sozialpolitisches Indikatoren- und Entscheidungssystem (SPES) für die Untersuchung des sozialstrukturellen Wandels in Deutschland (Brennecke 1975; Galler und Steger 1978; Kortmann 1978; Klanberg 1978; Helberger 1982). Eine andere Projektgruppe konzentrierte sich auf die Analyse der Mikrozensus-Zusatzbefragung von 1971, die mit ihrem detaillierten Erhebungsprogramm zur sozialen Herkunft und zum beruflichen Werdegang der Befragten einzigartige Möglichkeiten für die Analyse der „berufliche[n] und soziale[n] Umschichtung der Bevölkerung" in Folge des Zweiten Weltkrieges und in den Nachkriegsjahrzehnten bot (Handl et al. 1977). Eine Vielzahl von Publikationen dokumentiert die Fruchtbarkeit dieser Daten für die wissenschaftliche Analyse des sozialstrukturellen Wandels (vgl. z.B. Mayer 1977a, 1977b; Blossfeld 1983; Handl 1978, 1988; Müller et al. 1983). Diese Innovationsphase war dadurch gekennzeichnet, dass die für die Forschung benötigten Mikrodaten[4] von der amtlichen Statistik problemlos und nahezu kostenlos zur Verfügung gestellt wurden.

Schon die Anschlussprojekte – der Sonderforschungsbereich 3 „Mikroanalytische Grundlagen der Gesellschaftspolitik" (1979–1990) und das VASMA-Projekt (Vergleichende Analysen der Sozialstruktur mit Massendaten, 1979–1984)[5] – fanden hinsichtlich des Datenzugangs erheblich ungünstigere Randbedingungen vor. Infolge der Ende der 1970er Jahre einsetzenden Datenschutzdiskussion sah sich der Gesetzgeber veranlasst, den Zugang zu amtlichen Mikrodaten stärker zu reglementieren. Das Bundesstatistikgesetz von 1980 erlaubte die Weitergabe von Einzelangaben nur noch unter der Bedingung, dass die Möglichkeit einer Reidentifikation absolut ausgeschlossen war. In der Folge wurden die weitergegebenen Daten teilweise bis zur Unbrauchbarkeit anonymisiert, und die 1980er Jahre führten zu einem erheblichen Rückschritt hinter bereits Erreichtes. In mehreren Konferenzen und Beiträgen war der defizitäre Datenzugang Thema kontroverser Diskussionen zwischen Wissenschaftlern und Vertretern der amtlichen Statistik (Kaase et al. 1980; Müller und Hauser 1987; Statistisches Bundesamt 1985; Statistisches Bundesamt 1987). Mit der Novellierung des Bundesstatistikgesetzes 1987 wurde dann das bei der Übermittlung von Mikrodaten generell geltende Prinzip der „absoluten Anonymität" für den Fall der wissenschaftlichen Nutzung durch das Prinzip der „faktischen Anonymität" ergänzt. Danach darf die amtliche Statistik für die Durchführung wissenschaftlicher Vorhaben Einzelangaben weitergeben, sofern eine Reidentifikation der Befragten nur mit einem unverhältnismäßig großen Aufwand an Zeit,

4 Dies umfasste die Mikrozensen 1962 bis 1969, die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 sowie die Einkommens- und Verbrauchsstichprobe 1963, 1969.

5 Vgl. hierzu Haller und Müller (1983), Willms-Herget (1985), Blossfeld (1985), Lüttinger (1989), Haller (1989). Ein knapper Überblick zur Zielsetzung dieses Projektes findet sich bei Müller (1999a: 10f.).

Kosten und Arbeitskraft möglich ist. Der Gesetzgeber hat jedoch nicht präzisiert, wie die faktische Anonymität konkret umzusetzen ist. Eine Klärung dieses Problems hat sich Ende der 1980er Jahre durch eine umfassende Untersuchung ergeben (Müller et al. 1991). Die für den Wissenschaftsbereich denkbaren „Angriffsszenarien" und vorhandenen Reidentifikationsrisiken[6] wurden im Detail untersucht. Am Beispiel von Mikrozensus sowie Einkommens- und Verbrauchsstichprobe wurden empirisch fundierte Empfehlungen für die Umsetzung der faktischen Anonymisierung und entsprechende Anonymisierungsmaßnahmen erarbeitet, die spezifischen Risikokonstellationen bei möglichst schonendem Eingriff in das Analysepotenzial der Daten gezielt entgegenwirken.[7] Das Bundesministerium für Bildung und Forschung hat diese Untersuchungen unterstützt und später durch eine zentrale Finanzierung der Datenbereitstellungskosten für eine Reihe von Scientific-Use-Files ermöglicht, dass diese Daten der Wissenschaft nun zu einem erschwinglichen Unkostenbeitrag zur Verfügung stehen. Der Erfolg dieser Bemühungen zeigt sich in der sprunghaft angestiegenen Nutzung von Scientific-Use-Files durch die empirische Sozialforschung. Allein mit dem Scientific-Use-File des Mikrozensus arbeiten gegenwärtig rund 130 Institute (Lüttinger et al. 2003a).

Diese auf einem sehr langen Weg erzielten Fortschritte im Zugang der Forschung zu amtlichen Mikrodaten fanden ihre Fortsetzung in der vom BMBF 1999 eingerichteten „Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik" (KVI). Die von der KVI erarbeiteten Empfehlungen reichen von institutionellen Regelungen der Kooperation, Empfehlungen zu verschiedenen Datenerhebungen, Methodenforschung, Methodenausbildung, Datenschutz bis hin zur Datenarchivierung (KVI 2001; Müller 2001). In einem nächsten Schritt wurde ein Ausschuss zur Gründung eines „Rates für Sozial- und Wirtschaftsdaten" berufen. Dieser hatte die Aufgabe, das BMBF bei der Prioritätensetzung von Methodenprojekten für eine bessere Datenversorgung zu beraten, gab Empfehlungen zur Einrichtung von Forschungsdatenzentren und Datenservicezentren und bereitete die Einrichtung des Rates für Sozial- und Wirtschaftsdaten (RatSWD) vor. Letzterer wird im Herbst 2004 berufen und dient vor allem dem Ziel, die Situation der empirischen Sozial- und Wirtschaftsforschung nachhaltig zu verbessern.[8] Als eine unmittelbare Folge des Wirkens des Gründungsausschusses hat eine intensive Diskussion darüber eingesetzt, wie die an zahlreichen Stellen von verschiedenen amtlichen Datenproduzenten erfassten Daten der Forschung in einer systematischen Weise als Scientific-Use-Files zur Verfügung gestellt werden können. Durch die Etablierung von so genannten Forschungsdatenzentren[9] bei

6 Bei der Analyse wurde deutlich, dass der Nutzen einer Reidentifikation innerhalb des Wissenschaftsbereichs als sehr gering einzuschätzen ist. Allerdings bereitet eine empirische Quantifizierung eines solchen subjektiven Nutzens Schwierigkeiten. Deshalb erfolgte die Bewertung im Wesentlichen durch die Gegenüberstellung des für eine Reidentifikation zu betreibenden Aufwands im Vergleich zu den Kosten einer alternativen Informationsbeschaffung.

7 Dieses Projekt bezog sich auf amtliche Bevölkerungserhebungen. Etwas andere Risikokonstellationen ergeben sich bei Betriebs- und Wirtschaftsdaten. In welcher Form diese Daten als Scientific-Use-Files zur Verfügung gestellt werden können, wird gegenwärtig in einem Forschungsprojekt überprüft (Ronning und Gnoss 2003).

8 Für eine detaillierte Beschreibung der Zielsetzungen und der Struktur des RatSWD vgl. www.ratswd.de.

9 Neben den Forschungsdatenzentren wurde von der KVI auch die Förderung von in der For-

den wichtigsten Produzenten öffentlicher Daten werden sich die Rahmenbedingungen der Nutzung von amtlichen Mikrodaten für die empirische Sozialforschung in Zukunft deutlich verbessern.

Ein solches Forschungsdatenzentrum wurde beim Statistischen Bundesamt eingerichtet. Hier sollen Grundsatzfragen des Datenzugangs – unter Wahrung von Statistikgeheimnis und Datenschutz und in Kooperation mit der Wissenschaft – geklärt werden. Zugleich soll hier für die Forschung vor allem auch ein Zugang zu solchen Mikrodaten geschaffen werden, bei welchen eine Anonymisierung nur mit einem gravierenden Informationsverlust möglich ist. Weiterhin wurde ein gemeinsames Forschungsdatenzentrum der statistischen Landesämter etabliert, in welchem bislang nur über die einzelnen Landesämter zugängliche Mikrodaten zusammengeführt und der Forschung zukünftig länderübergreifend für Analysen zugänglich gemacht werden (Zühlke et al. 2003). Ebenfalls im Aufbau ist ein Forschungsdatenzentrum bei der Bundesagentur für Arbeit (BA) mit dem Ziel, die umfangreichen prozessproduzierten Datenbestände wie auch Eigenerhebungen der BA und des IABs für die Forschung zu erschließen (Oertel et al. 2004). Vielversprechend ist auch die Einrichtung eines Forschungsdatenzentrums bei der gesetzlichen Rentenversicherung (Rehfeld 2004), welches hoffen lässt, dass die bei den Sozialversicherungsträgern bestehenden und von der Forschung noch kaum genutzten Datenbasen mittelfristig für Analysen zur Verfügung stehen werden.

III. Potenziale und Restriktionen der Sekundäranalyse amtlicher Mikrodaten

Die Nutzung von amtlichen Mikrodaten durch die empirische Sozialforschung ist ein typisches Beispiel für Sekundäranalysen. Es sind Sekundäranalysen, weil die für die Bearbeitung von deskriptiven, hypothesenprüfenden oder methodisch orientierten Fragestellungen herangezogenen Daten nicht von den jeweiligen Forschern selbst generiert werden. Häufig werden sie auch für Untersuchungsvorhaben genutzt, die von den Primärerhebern nicht oder nur am Rande angedacht waren. Sekundäranalysen dienen nicht nur dem ökonomischen Einsatz der verfügbaren Ressourcen, sondern sind ein Kernelement für die Erfüllung der zentralen wissenschaftlichen Norm der Kritikoffenheit und der intersubjektiven Nachprüfbarkeit von Aussagen. Oft kann auch nur durch die Reanalyse von Daten die Robustheit von Befunden überprüft oder zwischen unterschiedlichen Interpretationen entschieden werden.[10] Da amtliche Daten primär für politisch-administrative Zwecke erhoben werden und eine wesentliche Informationsgrundlage für politische Entscheidungen bilden, wobei einzelnen Befunden eine weitreichende gesellschaftliche Bedeutung zukommen kann, ist es bei ihnen besonders wichtig, sie mit verschiedenen Methoden sowie unter verschiedenen theoretischen Perspektiven zu reanalysieren und konkurrierenden Ansätzen auszusetzen (Wagner 2000).

schung bereits verankerten Servicezentren angeregt mit dem Ziel, das Analysepotenzial amtlicher Mikrodaten weiter zu optimieren. Ein solches Servicezentrum wird bei ZUMA in Form des German Microdata Lab (GML) und am Forschungsinstitut zur Zukunft der Arbeit (IZA) in Form eines Internationalen Datenservicezentrums vom BMBF finanziell unterstützt (Lüttinger et al. 2003b).

10 Allerdings kommt diese Form der Sekundäranalyse in der empirischen Sozialforschung bislang eher zu selten vor (Diekmann 1998: 59f.).

Die Analyse amtlicher Daten teilt aber natürlich auch den zentralen Nachteil vieler Sekundäranalysen: Die verfügbaren Daten sind nicht immer in optimaler Weise für die Bearbeitung der spezifischen Forschungsfrage geeignet (Hyman 1972; Hakim 2000, Kiecolt und Nathan 1985). Bei der inhärenten Ausrichtung dieser Daten auf administrative Zwecke und oft auch ihrer Genese aus administrativen Vorgängen sind die damit verbundenen Probleme wahrscheinlich ausgeprägter als bei sozialwissenschaftlichen Daten.

1. Spezifische Potenziale amtlicher Mikrodaten

Die spezifischen Stärken von amtlichen Daten lassen sich im Wesentlichen durch vier Punkte charakterisieren (Alba et al. 1994b): 1. Amtliche Daten weisen – soweit es sich nicht sogar um Vollerhebungen handelt – häufig einen sehr großen Stichprobenumfang auf, der den sozialwissenschaftlicher Daten typischerweise um ein Vielfaches übersteigt. 2. Amtliche Daten sind oft nur in einem geringen Umfang von Unit-Non-Response betroffen. Zum einen besteht für manche Erhebungen, wie z.B. beim Mikrozensus oder bei Betriebs- und Unternehmenserhebungen eine gesetzliche Teilnahmepflicht, die die Nichtbeteiligungsquoten niedrig hält. Zum andern resultieren z.B. prozessproduzierte Daten (wie Sozialhilfestatistik, Arbeitslosenstatistik, Beschäftigtenstatistik, Lohn- und Einkommenssteuerstatistik, Kriminalstatistik) direkt aus administrativen Vorgängen und stellen entsprechend Vollerhebungen[11] dar. 3. Amtliche Erhebungen weisen häufig einen Kontextbezug auf, d.h. sie bilden zumindest in Teilen den sozialen Kontext der individuellen Akteure ab. Bei den Bevölkerungserhebungen trifft dies beispielsweise für den Mikrozensus, die Einkommens- und Verbrauchsstichprobe sowie die Zeitbudgeterhebung zu. Hier werden jeweils Daten für alle im Haushalt lebenden Personen erhoben. Durch die Klumpeneigenschaft der Mikrozensusstichprobe sind äquivalente Daten über das nahräumliche nachbarschaftliche Umfeld verfügbar, da alle Haushalte, die in einem räumlich definierten Auswahlbezirk leben, in die Erhebung einbezogen werden. Ein anderes Beispiel ist die Beschäftigtenstatistik, bei der im Prinzip die Informationen über einzelne Beschäftigte im Kontext aller Beschäftigten eines Betriebes analysiert werden können. Ähnliches gilt für das „linked Employer-Employee" Datenfile des IAB, welches Betriebsdaten mit Informationen aus der Beschäftigtenstatistik verknüpft (Bender et al. 2000). Hierdurch können etwa individuelle Erwerbsbiographien unter Kontrolle von betrieblichen Randbedingungen untersucht werden. 4. Amtliche Erhebungen weisen eine hohe Kontinuität auf. Vielfach werden die Daten auf Basis derselben Konzepte in identischer oder allenfalls geringfügig veränderter Form über lange Zeiträume gesammelt. Zugleich beruhen eine Reihe der in der amtlichen Statistik genutzten Konzepte auf internationalen Standards (Hahlen 2002). Wenngleich hierdurch auch nicht alle Probleme der komparativen Forschung gelöst sind, verbessern

11 Natürlich nur soweit die interessierenden Informationen meldepflichtig sind oder gemeldet werden. Die Sozialstatistik z.B. umfasst nicht alle Personen, die sozialhilfebedürftig sind, sondern nur solche, die diesen Anspruch formal angemeldet haben. Gleiches gilt auch für die Arbeitslosenstatistik und andere Registerdaten. Zu den Restriktionen der Kriminalstatistik vgl. Heinz (2004).

diese Standards doch in mancher Hinsicht die Chancen sinnvoller Vergleiche (Hoffmeyer-Zlotnik und Wolf 2003).

In welcher Weise lassen sich diese spezifischen Stärken amtlicher Mikrodaten von der empirischen Sozialforschung nutzen?

Bedingt durch den in der Regel sehr großen Stichprobenumfang weisen amtliche Daten einen geringen Stichprobenfehler auf, was in mehrfacher Hinsicht vorteilhaft ist. So lassen sich mit diesen Daten auch für spezielle Bevölkerungsgruppen wie Migranten, Alleinerziehende, nichteheliche Lebensgemeinschaften, Geringqualifizierte, geringfügig Beschäftigte, Selbstständige oder einzelne Berufsgruppen relativ präzise Punktschätzungen erzielen. Dies ist vor allem dann von Bedeutung, wenn das Forschungsinteresse auf praxisorientierte Fragestellungen gerichtet ist, wie etwa die Abschätzung der Kosten einer auf eine bestimmte Gruppe abzielenden sozialpolitischen Maßnahme. Ein Beispiel hierfür wäre etwa die Frage, welche Folgen die Einführung der Pflegeversicherung Mitte der 1990er Jahre auf die quantitative Entwicklung der Pflegeberufe hatte (Voges 2002). Ein niedriger Stichprobenfehler ist auch für die Analyse von Veränderungen im Zeitverlauf wichtig. Die Identifizierung von Trends setzt voraus, dass der Stichprobenfehler kleiner ist als die Veränderung selbst. Insbesondere wenn Veränderungen über kurze Zeitintervalle interessieren (Beispiel Arbeitslosigkeitsentwicklung), stößt man bei sozialwissenschaftlichen Erhebungen nicht selten an Grenzen, da die Stichprobenschwankungen die erwartbaren realen Veränderungen häufig um ein Vielfaches übertreffen. Auch für die Beobachtung von sich eher langsam entwickelnden Trends des sozialstrukturellen Wandels und vor allem, wenn solche bei kleinen Bevölkerungsgruppen erkannt werden sollen, sind die regelmäßig durchgeführten, großen amtlichen Bevölkerungserhebungen unverzichtbar.

Auch bei hypothesenüberprüfenden Untersuchungen für spezifische Teilpopulationen erweisen sich die großen Fallzahlen amtlicher Daten als hilfreich: Häufig muss weniger stark aggregiert werden als bei Verwendung sozialwissenschaftlicher Daten, so dass differenziertere Analysen möglich sind, wie etwa zum Bildungs- oder Erwerbsverhalten unterschiedlicher ethnischer Gruppen (Alba et al. 1994a; Baumann 1999; Granato und Kalter 2001; Kalter und Granato 2002).[12] Abgesehen von inhaltlichen Interessen an einer möglichst differenzierten Betrachtung kann dies auch von gesellschaftspolitischer Relevanz sein. Sollen bspw. die Auswirkungen der Aktiven Arbeitsmarktpolitik in Deutschland betrachtet werden, ergibt sich aus den bis Ende der 1990er Jahre durchgeführten Evaluationsstudien ein eher diffuses Bild von positiven und negativen Effekten. Dies scheint entscheidend dadurch bedingt zu sein, dass für die jeweiligen Analysen kleine Ausschnitte aus sozialwissenschaftlichen Erhebungen genutzt und auf Grund geringer Fallzahlen auch sehr heterogene Maßnahmen zusammengefasst wurden. Für eine angemessenere Abbildung und Überprüfung der verschiedenen arbeitspolitischen Maßnahmen kommt man um die Nutzung von prozessproduzierten amtlichen Daten nicht umhin (Fitzenberger und Speckesser 2000).

Wenngleich kein deterministischer Zusammenhang zwischen der Höhe der Ausschöpfungsquote und dem Ausmaß der systematischen Stichprobenverzerrungen besteht, nimmt das Risiko eines systematischen Bias mit sinkenden Ausschöpfungsquoten

12 Dieser Vorteil kommt allerdings nur zum Tragen, wenn die für die Analyse zentralen Indikatoren auch in den amtlichen Daten erfasst sind.

zu (Esser et al. 1989; Goyder 1987; Schnell 1997). In diesem Kontext werden amtliche Daten mit ihren nach wie vor hohen Ausschöpfungsquoten häufig als Referenzstatistik herangezogen, um die Qualität der sozialwissenschaftlichen Stichprobenerhebungen zu überprüfen, deren Ausschöpfungsquoten inzwischen auf ein Niveau zwischen 50 und 60 Prozent[13] gesunken sind. Auf Basis von vergleichbar verfügbaren soziodemografischen Grundmerkmalen wie Alter, Geschlecht, Bildungsabschluss, Erwerbsstatus und Familienstand lässt sich überprüfen, ob und welche Bevölkerungsgruppen in den Stichproben in systematischer Weise unterrepräsentiert sind (Hartmann und Schimpl-Neimanns 1992; Koch 1997, 1998). Einer neueren Studie zufolge scheinen sinkende Ausschöpfungsquoten im Falle von melderegistergestützten Stichprobenziehungen[14] nicht notwendigerweise mit zunehmenden Verzerrungen in den Anteilswerten der zentralen Grundmerkmale einherzugehen (Schneekloth und Leven 2003). Allerdings sind die meisten dieser Untersuchungen eher deskriptiv orientiert. Ob sich eine systematische Unterrepräsentation spezifischer Gruppen bspw. von Geringqualifizierten oder Singles auf die Schätzung von Zusammenhängen auswirkt, ist hingegen bislang kaum untersucht. Verzerrte Schätzergebnisse wären etwa dann zu erwarten, wenn die Ausfälle in den unterrepräsentierten Gruppen nicht zufällig, sondern wiederum systematisch sind.

Die regelmäßige Periodizität der amtlichen Statistiken mit weitgehend identisch angelegten Frageprogrammen ist insbesondere für Untersuchungen zu Fragen des sozialen und wirtschaftlichen Wandels interessant. Faktisch sind amtliche Daten bislang die einzige Basis, um zentrale sozialstrukturelle Trends wie etwa den Wandel in den Berufs-, Wirtschafts- und Arbeitsmarktstrukturen, die Integration von Frauen in die Arbeitswelt oder die Veränderungen in den Haushalts- und Familienstrukturen für sehr lange Zeitreihen zu analysieren (vgl. z.B. Haller und Müller 1983; Müller et al. 1983; Willms-Herget 1985; Stockmann 1986; Handl 1988). Ein aktuelles Beispiel hierfür ist eine Untersuchung zu der Entwicklung der Top-Einkommen in Deutschland zwischen 1891 und 1995 auf Basis der Steuerstatistik (Dell 2003). Auch für die international vergleichende Forschung kommt amtlichen Daten eine nicht zu unterschätzende Bedeutung zu (Brauns et al. 2003; Luber 2003; Müller und Gangl 2003; Schettkat 2003), da in der empirischen Sozialforschung bislang Datenbasen fehlen, die mit ähnlicher Periodizität, Vergleichbarkeit und ähnlichem Stichprobenumfang einen umfangreichen Kanon an sozialstrukturellen Merkmalen in einer Vielzahl von Ländern erheben.

13 Auch in der empirischen Wirtschaftsforschung wird auf das Problem einer rückläufigen Antwortbereitschaft hingewiesen (Wagner 1999).

14 In diesem Kontext sei darauf hingewiesen, dass das Melderegister, welches ebenfalls eine amtliche Datenquelle ist, vielfach die beste Basis für eine kontrollierte Ziehung von Zufallsstichproben für Bevölkerungserhebungen darstellt. Wenngleich erfahrungsgemäß circa 15 Prozent der gezogenen Melderegisteradressen nicht mehr aktuell sind (Schneekloth und Leven 2003: 24), lassen sich diese stichprobenneutralen Ausfälle in kontrollierter Weise durch Reserveadressen ersetzen. Bislang zählt die Melderegisterstichprobe auf Grund des damit einhergehenden relativ hohen Zeit- und Kostenaufwands jedoch nicht zu den Standardverfahren in der empirischen Sozialforschung.

2. Spezifische Restriktionen amtlicher Mikrodaten

Diesen skizzierten Stärken amtlicher Mikrodaten stehen allerdings auch einige spezifische Restriktionen gegenüber.

Wie oben erwähnt, leiden Sekundäranalysen generell darunter, dass die Daten nicht immer optimal den Anforderungen für die jeweilige Fragestellung entsprechen. Oft wird mit ziemlich unbefriedigenden Proxy-Lösungen gestrickt. Bei der Verwendung amtlicher Daten ist diese Problematik besonders ausgeprägt. Auf Grund des spezifischen Generierungsprozesses innerhalb von Verwaltungsprozeduren und in der Nähe des politischen Prozesses sowie der damit einhergehenden Orientierung an gesetzlich definierten Tatbeständen sind amtliche Daten in mancher Hinsicht durch eine buchhalterische Zählmentalität und paragraphenorientierte Dürre gekennzeichnet. Die Abbildung der sozialen Realität bleibt häufig auf die Ordnungsperspektive legal definierter Sachverhalte und administrativer Aufzeichnungen beschränkt (obwohl natürlich auch diese als solche von Interesse sind). Auf die Erfassung von Einstellungen, Überzeugungen, Handlungszielen oder Präferenzen wird weitgehend verzichtet. Dies bedingt, dass das Analysepotenzial von amtlichen Daten i.d.R. engeren inhaltlichen Grenzen unterliegt als das sozialwissenschaftlicher Daten.[15] Der Wissenschaftler, der amtliche Daten nutzt, muss sich deshalb in jedem einzelnen Fall klar darüber werden, ob sie seinen theoretischen oder methodischen Anforderungen entsprechen oder welche Verzeichnungen der sozialen Wirklichkeit damit verbunden sein mögen, und dass er Sachverhalte analysiert, wie sie der Staat oder die amtliche Statistik definieren. Weiterhin konzentriert sich die amtliche Statistik bislang vorwiegend auf Querschnittserhebungen. Verlaufsdaten, ob in Form von Retrospektivfragen oder echten Panelerhebungen, stellen – vor allem bei amtlichen Bevölkerungserhebungen – bislang eher eine Ausnahme dar.[16] Dies führt dazu, dass die amtliche Statistik ein eher statisches Bild der gesellschaftlichen Wirklichkeit zeichnet, bei welchem dynamische Prozesse weitgehend ausgeblendet bleiben. Veränderungen können nur auf der Aggregatebene des Vergleichs von Querschnittserhebungen, nicht jedoch auf der Ebene des Geschehens bei individuellen Akteuren nachvollzogen werden. Ein typisches Beispiel hierfür ist etwa die Analyse von Arbeitslosenquoten versus dynamischen Prozessen des Eintritts in und des

15 Hinzu kommt, dass inhaltliche Anregungen aus der Wissenschaft für eine verbesserte Datenerhebung zwar durchaus auf Interesse stoßen, im Verlaufe des Gesetzgebungsverfahrens dann jedoch häufig anderen Interessen untergeordnet werden. Ein aktuelles Beispiel hierfür ist das Mikrozensusgesetz ab 2005. Nach dem Gesetzes-Entwurf des Bundesinnenministeriums sollten beispielsweise Angaben zur Einbürgerung, zu Gründen eines Berufswechsels, zu Lebenspartnern außerhalb des Haushaltes und Anzahl der geborenen Kinder erfasst werden. Alle diese wichtigen Informationen fielen neben weiteren wie Fachrichtung der Schul- oder Hochschulausbildung, Besuch von Kindergarten, -krippe, -hort; Eheschließungsjahr etc., dem Willen des Bundesrates zum Opfer (Bundestag Drucksache 15/3060, 2004).

16 Eine solche Ausnahme ist beispielsweise der Mikrozensus. Der Mikrozensus ist ein rotierendes Panel, bei welchem die Haushalte eines Auswahlbezirks in vier aufeinander folgenden Jahren befragt werden. Allerdings dürfen die Statistischen Ämter erst seit 1996 die für die Zusammenführung der Panelwellen benötigten Ordnungsnummern speichern. Gegenwärtig wird vom BMBF ein Projekt gefördert mit dem Ziel, die Paneleigenschaften des Mikrozensus in Form eines Scientific-Use-File für die Forschung nutzbar zu machen (Lüttinger et al. 2003b: 162).

Austritts aus der Arbeitslosigkeit (Gangl 2003). Insbesondere für die Untersuchung von Wirkungsmechanismen ist die Verwendung von Verlaufsdaten unabdingbar.

Nicht unproblematisch ist auch das bislang unzureichende Wissen hinsichtlich der Reliabilität und Validität von amtlichen Daten. Insofern die amtliche Statistik eine wichtige Datenquelle für die empirische Sozial- und Wirtschaftsforschung ist, werden nicht nur Dokumentationen zur Datenaufbereitung benötigt, sondern es ist auch unabdingbar, eine breite Informationsbasis darüber zu haben, was die amtliche Statistik mit ihren Verfahren misst und wie zuverlässig die Messungen sind (Wagner 2000; Schimpl-Neimanns 2002). Ähnlich wie das bei den großen, regelmäßig erhobenen sozialwissenschaftlichen Studien – insbesondere dem ALLBUS und dem SOEP – gebräuchlich ist, wäre es sehr wünschenswert, wenn auch die amtliche Statistik ihre Instrumente und Verfahren kontinuierlich durch spezifische Methodenstudien prüfen und einer Evaluation unterziehen würde, bzw. vorhandene Studien öffentlich zugänglich machen würde. Dies umso mehr, als „amtlich" erzeugte Daten die zentrale Grundlage für politische Entscheidungen sind.

Alle Daten sind mehr oder weniger fehlerbehaftet, hiervon sind auch amtliche Daten nicht ausgenommen. Die Aussagekraft lässt sich aber nur dann adäquat einschätzen, wenn das Ausmaß und die Art möglicher Verzerrungen bekannt sind. Ein Schritt in diese Richtung ist die kürzlich von Stauder und Hüning (2003) vorgelegte Analyse zur Reliabilität der Einkommensangaben im Mikrozensus anhand der Einkommens- und Verbrauchsstichprobe. Die Einkommensinformationen im Mikrozensus wurden in der Literatur bislang eher kritisch bewertet, da einerseits das individuelle Gesamteinkommen pauschal erfragt wird, andererseits keine spitzen Einkommenswerte, sondern nur Einkommensklassen erfasst werden. In ihrer Arbeit untersuchen Stauder und Hüning (2003), wie sich diese Faktoren auf Maßzahlen der zentralen Tendenz der relativen Einkommensverteilung und der relativen Einkommensarmut auswirken. Im Ergebnis ist festzuhalten, dass die Pauschalabfrage von Einkommen eine Vernachlässigung kleinerer Einkommensbestandteile mit sich bringt. Dies führt zu einer tendenziellen Unterschätzung des durchschnittlichen Einkommensniveaus im Mikrozensus. Da dieser Effekt jedoch in der gesamten Population nahezu gleichmäßig auftritt, sind die Relationen der Verteilungen hiervon nicht betroffen.

Eine gewisse Analyserestriktion ergibt sich auch aus der faktischen Anonymisierung der Daten, da die Anonymisierung immer mit einer Verringerung des Analysepotenzials einhergeht. Im Unterschied zur absoluten Anonymisierung zielt die faktische Anonymisierung jedoch darauf ab, das Analysepotenzial der Daten weitmöglichst zu erhalten. Empirischen Untersuchungen zufolge ist das Reidentifikationsrisiko bei einer gleichzeitig tiefen sachlichen und regionalen Differenzierung der Daten am größten. Deshalb werden Scientific-Use-Files im Allgemeinen mit einer möglichst differenzierten sachlichen Tiefengliederung bei einer gleichzeitig nur sehr groben Regionalgliederung weitergegeben.[17] Allerdings wurden auch Anonymisierungskonzepte entwickelt, die

17 Beim Mikrozensus (Vergleichbares gilt für andere amtliche Bevölkerungserhebungen) werden z.B. für die sehr differenziert erfassten und leicht erfahrbaren Merkmale Beruf, Wirtschaftszweig, Nationalität und Alter im Scientific-Use-File keine Ausprägungen ausgewiesen, die weniger als 50.000 Personen in der Grundgesamtheit umfassen. Für alle anderen Merkmale liegt das Minimum bei 5.000 Personen in der Grundgesamtheit. Wie man sich durch einen Blick

eine Regionalisierung von Scientific-Use-Files ermöglichen und den Bedarf der Regionalforschung abdecken sollen. Beispielsweise ist gegenwärtig die Generierung eines Mikrozensus-Regionalfile in Bearbeitung, in welchem Regionaleinheiten ab 100.000 Einwohnern ausgewiesen werden sollen (Wirth und Zühlke 2004). Für die IAB-Beschäftigtenstichprobe existiert schon seit längerer Zeit ein solches Regionalfile (Haas 2001).

IV. Das Analysepotenzial von amtlichen Mikrodaten am Beispiel ausgewählter Scientific-Use-Files

Wie ein Blick in das Gutachten der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI 2001: 52ff.) zeigt, werden amtliche Daten zu einer Vielzahl gesellschaftlicher Teilbereiche erhoben. Die in dem KVI-Gutachten betrachteten Themenfelder umfassen „Bevölkerungsstand, Bevölkerungsstruktur und Bevölkerungsentwicklung", „Staatsverfassung und Staatsorgane", „Wahlen und Parteien", „Rechtspflege, Gerichtsbarkeit, innere und äußere Sicherheit", „Erwerbstätigkeit und Einkommen", „Vermögen, privater Verbrauch und Wohnen", „Gesundheits- und Sozialwesen", „Bildung und Kultur", „Zeitverwendung", „Freizeit und Erholung", „Wirtschaft", „Umwelt, Rohstoffe und Energie", „Wissenschaft, Forschung und Entwicklung" sowie „subjektive Indikatoren". Wobei bislang jedoch noch keine systematische Übersicht existiert, welche Informationen zu welchen Themenfeldern von welchen Institutionen generiert bzw. gesammelt werden. Insbesondere scheint es auch für die in den einzelnen Themenfeldern ausgewiesenen Experten nicht ganz einfach zu sein, sich einen Überblick über das vorhandene Datenpotenzial zu verschaffen (KVI 2001: 53). Ein Grund hierfür ist sicherlich die Vielzahl von Institutionen, die als Datenproduzenten für ein spezifisches Themenfeld zu berücksichtigen sind. Beispielsweise sind für den soziologisch interessanten Bereich Gesundheits- und Sozialwesen folgende Datenquellen in Betracht zu ziehen: 1. Die allgemeinen Systeme der sozialen Sicherung (z.B. Gesetzliche Rentenversicherung, Altersversicherungssysteme, Krankenversicherung, Kindergeld, Arbeitsförderung); 2. Sondersysteme (z.B. Versorgungswerke); 3. Leistungssysteme des öffentlichen Dienstes; 4. Arbeitergeberleistungen; 5. Soziale Hilfen und Dienste (z.B. Sozialhilfestatistik, Wohnhilfestatistik, Jungendhilfestatistik); 6. Indirekte Leistungen (z.B. Familienlastenausgleich); 7. systemübergreifende Daten (z.B. Sozialbudget, Volkswirtschaftliche Gesamtrechung). Angesichts dieser vielfältigen Informationsquellen kann auch im vorliegenden Beitrag keine Bestandsaufnahme amtlicher Datenbasen vorgenommen werden, zumal viele dieser Daten via RatSWD erst noch für die Wissenschaft erschlossen werden müssen. Stattdessen werden wir am Beispiel von drei amtlichen Statistiken – der Beschäftigtenstatistik, der Einkommens- und Verbrauchsstichprobe und dem Mikrozensus –, die schon seit längerer Zeit in Form von Scientific-Use-Files zur Verfügung stehen, das spezifische Potenzial dieser Daten für die empirische Wirtschafts- und Sozialforschung skizzieren. Eine

auf die Randverteilungen des Mikrozensus leicht überzeugen kann, bedeutet dies keine nennenswerte Einschränkung des Analysepotenzials. Deutlich stärkere Einschränkungen sind für die Regionalforschung gegeben, da das Scientific-Use-File nur das Bundesland und die politische Gemeindegrößenklasse enthält.

Übersicht der gegenwärtig zur Verfügung stehenden und geplanter Scientific-Use-Files findet sich im Anhang.

1. Die Beschäftigtenstatistik als Scientific-Use-File

Eine der wichtigsten Datenquellen der empirischen Arbeitsmarktforschung ist die prozessproduzierte Beschäftigtenstatistik der Bundesagentur für Arbeit. Die Beschäftigtenstatistik basiert auf einem 1973 in Westdeutschland eingeführten integrierten Meldeverfahren zur Kranken-, Renten- und Arbeitslosenversicherung, welches die Arbeitgeber verpflichtet, alle sozialversicherungspflichtig beschäftigten Arbeitnehmer[18] an die Sozialversicherungsträger zu melden (Bender et al. 1999). Anzugeben sind der Beginn, das Ende wie auch jede Veränderung des Beschäftigungsverhältnisses. Wann immer eine Beschäftigung über das Jahresende hinweg besteht, erfolgt eine so genannte Jahresmeldung. Daher ist jedes Jahr, in welchem eine Person sozialversicherungspflichtig erwerbstätig war, durch mindestens einen Record in der Beschäftigtenstatistik abgebildet. Im Vergleich zu Umfragedaten ist bei diesen prozessproduzierten Daten insbesondere hinsichtlich des sozialversicherungspflichtigen Entgelts und der Beschäftigungszeiten von einer hohen Zuverlässigkeit auszugehen, da Fehlerquellen wie z.B. Erinnerungsfehler oder Item-nonresponse entfallen. Da die Meldungen mit den Versicherungsbeitragszahlungen korrespondieren und Ansprüche vor allem an spätere Rentenzahlungen bestimmen, kann eine hohe Genauigkeit unterstellt werden (jedenfalls im Rahmen der legalen Beschäftigungsverhältnisse). Auch das für Verlaufsdaten typische Problem der Panelmortalität ist allenfalls minimal, da Ausfälle nur dann entstehen, wenn Erwerbstätige vorübergehend oder endgültig aus der sozialversicherungspflichtigen Beschäftigung ausscheiden. Ebenfalls ein Plus dieser prozessproduzierten Daten ist, dass mit den Personen, die zum ersten Mal oder erneut versicherungspflichtig erwerbstätig sind, kontinuierlich neue Fälle in die Beschäftigtenstatistik aufgenommen werden (Koch und Meinken 2003).

Seit Mitte der 1990er Jahre steht eine faktisch anonymisierte Ein-Prozent-Stichprobe der Beschäftigtenstatistik, die so genannte IAB-Beschäftigtenstichprobe, als Scientific-Use-File für die Forschung zur Verfügung, die gegenwärtig den Zeitraum 1975 bis 1997 umfasst.[19] Das File enthält sowohl verschiedene soziodemographische Merkmale der Beschäftigten, z.B. Geschlecht, Alter, Familienstand, Nationalität, Schul- und Ausbildung, wie auch für jedes Beschäftigungsverhältnis einer Person den Beginn und das Ende der Beschäftigung, den Beruf, die Stellung im Beruf, die Arbeitszeit, das sozialversicherungspflichtige Bruttoentgelt sowie den Wirtschaftszweig des Beschäftigungsbetriebs. Darüber hinaus wurden auf individueller Ebene weitere Informationen zu Lohn-

18 Die Beschäftigtenstatistik umfasst alle Arbeiter, Angestellten und Auszubildenden soweit sie sozialversicherungspflichtig beschäftigt sind. Nicht enthalten sind Beamte, Selbstständige und mithelfende Familienangehörige. In Westdeutschland werden damit circa 80 Prozent aller Erwerbstätigen erfasst, in Ostdeutschland circa 86 Prozent (Stand 1995) (Bender et al. 1999: 106).

19 Darüber hinaus stehen inzwischen eine Regionalstichprobe (Haas 2001) und das BA-Beschäftigtenpanel (Koch und Meinken 2003) als Scientific-Use-Files zur Verfügung.

ersatzleistungen (z.B. Zeiträume, in welchen Arbeitslosengeld, Arbeitslosenhilfe oder Unterhaltsgeld für Fortbildung oder Umschulung bezogen wurde) aus der so genannten Leistungsempfängerdatei der Bundesanstalt für Arbeit zugespielt. Die individuellen Erwerbsbiographien können damit sowohl in Hinblick auf sozialversicherungspflichtige Beschäftigungsverhältnisse als auch für Phasen der registrierten Arbeitslosigkeit abgebildet werden. Für jedes Beschäftigungsverhältnis sind zusätzlich aggregierte Informationen zur Betriebsgröße, Bildungs- und Ausbildungsstruktur der Betriebe enthalten, so dass auch der betriebliche Kontext in den Analysen berücksichtigt werden kann (Bender et al. 1999; Bender und Haas 2002).

Selbst wenn der Merkmalsumfang der IAB-Beschäftigtenstichprobe gemessen an typischen sozialwissenschaftlichen Erhebungen eher gering ist, beinhaltet das Scientific-Use-File ein hohes Analysepotenzial für die Arbeitsmarktforschung. Es wird bislang allerdings stärker von den empirischen Wirtschaftswissenschaften als von den Sozialwissenschaften genutzt. Insbesondere die exakten Angaben über den Bruttolohn der sozialversicherungspflichtig Beschäftigten[20] zählen zu den Stärken dieser Daten. Entsprechend bilden Analysen der unterschiedlichen Facetten des Zusammenhangs von Lohn und Arbeitsmarktstrukturen einen Anwendungsschwerpunkt (Mavromaras und Rudolph 1997; Baltagi und Blien 1998; Baltagi et al. 2000; Dustmann und Meghier 1999; Fitzenberger 1999a, 1999b; Puhani 2003). Gleichfalls lassen sich mit diesen Daten vielfältige Forschungsfragen hinsichtlich beruflicher und betrieblicher Mobilität wie auch zu Berufseinstiegsprozessen bearbeiten (Bender et al. 2000; Mertens 1998; Konietzka 2002), sowie die Verweildauern in der Arbeitslosigkeit (Fahrmeier et al. 2000) oder auch Verteilungseffekte der Arbeitslosenversicherung (Mavromaras und Rudolph 1998) untersuchen. Darüber hinaus ermöglicht die hohe Fallzahl detaillierte Arbeitsmarktanalysen für kleine Subpopulationen wie Migranten (Bauer et al. 1998) oder Auszubildende (Euwals und Winkelmann 2001).

Das Potenzial dieser Daten für die Untersuchung wirtschaftspolitisch wie auch gesamtgesellschaftlich relevanter Fragen soll am Beispiel einer Studie skizziert werden, die sich mit der so genannten „Krugman Hypothese" beschäftigt. Dieser Hypothese zufolge ist die rapid zunehmende Einkommensungleichheit bei tendenziell sinkenden Arbeitslosenquoten in den USA und Großbritannien einerseits und andererseits die zunehmende Arbeitslosenquote bei im Zeitverlauf vergleichsweise stabilen Einkommensungleichheiten in Deutschland auf die gleiche Ursache zurückzuführen: eine deutlich rückläufige Nachfrage nach geringqualifizierten Arbeitskräften (Puhani 2003). Während dieser Nachfragerückgang in den USA und Großbritannien über Lohnabstriche reguliert werden würde, führe er in Deutschland auf Grund von institutionalisierten, starren Lohnstrukturen zu einem Anstieg der Arbeitslosigkeit. Empirische Unterstützung der Krugman-Hypothese wäre dann gegeben, wenn in Deutschland ein überdurchschnittlicher Anstieg der Arbeitslosen in der Gruppe der Geringqualifizierten zu beobachten wäre. In den USA und Großbritannien hingegen sollten die Arbeitslosenquoten für diese Gruppe relativ stabil sein bzw. abnehmen, die Einkommensabstände zu anderen Qualifikationsgruppen hingegen zunehmen.

20 Bedingt durch die Konzentration auf sozialversicherungspflichtig Beschäftigte sind am oberen Ende der Einkommensskala circa 10 Prozent der Einkommen nicht erfasst (Ammermüller und Weber 2003: 3).

Die Überprüfung dieser These erfolgt für Deutschland hauptsächlich auf Basis der IAB-Beschäftigtenstichprobe, da diese Datenquelle bezüglich der Genauigkeit der Bruttolohnangaben für Deutschland einzigartig ist und zugleich ausreichend hohe Fallzahlen für robuste Schätzungen vorliegen.[21] In der Tendenz lässt sich festhalten, dass die empirische Evidenz für die Krugman-Hypothese eher schwach ist. Zunächst ist für den Beobachtungszeitraum (Anfang der 1990er bis Anfang der 2000er Jahre[22]) in allen drei Ländern eine steigende Nachfrage nach hochqualifizierten Arbeitskräften zu verzeichnen. Der Einfluss auf das Einkommen und die Arbeitslosenquote der Geringqualifizierten entspricht jedoch nur teilweise den theoretischen Erwartungen. Für die Gruppe der am niedrigsten Qualifizierten[23] zeigt sich in Westdeutschland bei rückläufigen relativen Einkommen ein Anstieg in der Arbeitslosenquote, was als Indikator für eine noch immer zu wenig flexible Lohnstruktur interpretiert werden könnte. In den USA sind relative Einkommensverluste dieser Gruppe bis in die späten 1990er Jahre mit stabilen Arbeitslosenquoten, seit Ende der 1990er Jahre mit sinkenden Arbeitslosenquoten verknüpft. Im Unterschied hierzu sind in Großbritannien die Arbeitslosenquoten und die Einkommen der minimal qualifizierten Arbeitskräfte über den Beobachtungszeitraum relativ konstant. Dies könnte jedoch auch dadurch bedingt sein, dass der Anteil der Geringqualifizierten im gleichen Zeitraum in Großbritannien erheblich zurückgegangen ist (Puhani 2003: 24) und diese Arbeitskräfte damit möglicherweise in geringerem Umfang als benötigt auf dem Arbeitsmarkt verfügbar waren[24].

Auch hinsichtlich der zweitniedrigsten Qualifikationsgruppe[25] zeigen sich Inkonsistenzen zur Krugman-Hypothese. In den USA ist diese Gruppe durch sinkende Löhne bei relativ stabiler Arbeitslosenrate charakterisiert. In Großbritannien hingegen nimmt die Arbeitslosenrate trotz sinkender Löhne zu. Im Vergleich hierzu deuten die Befunde für diese Gruppe in Deutschland ungeachtet rigiderer Lohnstrukturen auf positive Arbeitsmarktchancen hin (Puhani 2003: 25). Letzteres könnte bedeuten, dass die duale Ausbildung in Deutschland, wenngleich durchaus verbesserungsfähig, im Vergleich zu äquivalenten Qualifikationsniveaus in den USA oder Großbritannien zu einer höheren Humankapitalausstattung führt, mit entsprechend besseren Arbeitsmarkt- und Einkommenschancen.

2. Die Einkommens- und Verbrauchsstichprobe als Scientific-Use-File

Eine im Hinblick auf Einkommenssituation, Vermögenslage und Verbrauchssituation privater Haushalte in Deutschland einzigartige Datenquelle ist mit der Einkommens- und Verbrauchsstichprobe (EVS) gegeben. Die EVS wird seit 1962/63 in einem Tur-

21 Parallel hierzu werden auch Schätzungen auf Basis des Mikrozensus und des SOEPs durchgeführt. Zu den spezifischen Stärken und Schwächen der jeweiligen Daten für die hier interessierende Fragestellung siehe Puhani (2003: 6f.).

22 Datenbedingt erstreckt sich das Beobachtungszeitfenster für Deutschland nur bis 1997.

23 Entspricht für Deutschland der Gruppe, die keine berufliche Qualifikation aufweist; USA: Highschool-dropout; Großbritannien: unterhalb O-Level-Äquivalent.

24 Zu ähnlichen kritischen Einschätzungen der Krugman-Hypothese und zu Auswirkungen der Globalisierung vgl. Atkinson (1997) und Goldthorpe (2003: 34ff.).

25 Deutschland: Lehrabschluss; USA: Highschool; Großbritannien: O-Level-Äquivalent.

nus von fünf Jahren erhoben. Im Zentrum der Befragung stehen Einkommen, Ausgaben sowie ausgewählte Vermögensarten der privaten Haushalte. Weiterhin werden auch direkte Steuern, Sozialversicherungsbeiträge, die Ausstattung mit Gebrauchsgütern und die Wohnsituation erfasst. Bis 1993 wurden die Einkommen und Ausgaben für jeweils ein ganzes Jahr dokumentiert. Mit der Erhebung 1998 wurde die EVS auf ein Rotationsverfahren umgestellt, bei welchem jeweils ein Viertel der befragten Haushalte für ein Quartal ihre Einnahmen und Ausgaben dokumentieren.[26]

Die Teilnahme an der EVS ist freiwillig. Um die auf Grund der Freiwilligkeit zu erwartenden Ausfälle insbesondere bei Niedrig- und Hocheinkommensbeziehern sowie Einpersonenhaushalten zu kompensieren, wird die EVS als Quotenstichprobe durchgeführt (Kühnen 1999). Angestrebt wird ein durchschnittlicher Auswahlsatz von 0,2 Prozent der Privathaushalte, was einem Stichprobenumfang von mehr als 70.000 Haushalten entspricht. Quotierungsmerkmale sind das Haushaltsnettoeinkommen, der Haushaltstyp,[27] die soziale Stellung des Haupteinkommensbeziehers[28] und das Bundesland. Erhebungseinheit der EVS ist der private Haushalt, wobei Haushalte mit sehr hohem Einkommen – nicht zuletzt auf Grund der gering ausgeprägten Teilnahmebereitschaft – nicht einbezogen werden.[29] Ebenfalls nicht berücksichtigt werden Wohnsitzlose und die Anstaltsbevölkerung. Einkommens- und Vermögenswerte am oberen und unteren Rand sind daher in der EVS unvollständig erfasst, so dass die Ungleichheit der Vermögensverteilung vermutlich tendenziell unterschätzt wird (Hauser und Stein 2001). Haushalte mit einer ausländischen Bezugsperson werden seit 1993 in der EVS berücksichtigt. Einschränkend ist weiterhin anzumerken, dass die EVS bislang nur einen geringen Kanon an soziodemographischen Merkmalen enthält. Insbesondere werden für die Analyse sozialer Differenzierungsprozesse relevante Merkmale wie Beruf, Stellung im Betrieb oder Schulbesuch der Kinder nicht erfasst, so dass manch soziologisch interessierende Fragestellungen nicht oder nur eingeschränkt untersuchbar sind.[30]

---

26 Der Umstieg von der einjährigen auf die dreimonatige Anschreibedauer geht mit einem Genauigkeitsverlust einher. Insbesondere zeitbezogene Einmalzahlungen (Urlaubsgeld, Weihnachtsgeld, Gratifikationen) bleiben unter Umständen unberücksichtigt. Allerdings sind derartige Periodeneffekte nach wie vor geringer als bei Verwendung eines Stichtagskonzeptes (Stauder und Hüning 2003: 12f.; Münnich 2000: 683).

27 Differenziert nach Einpersonenhaushalten, Ehepaaren und nichtehelichen Lebensgemeinschaften ohne Kinder; Alleinerziehenden; Ehepaaren und nichtehelichen Lebensgemeinschaften mit Kindern, höchstens ein Partner erwerbstätig; Ehepaaren und nichtehelichen Lebensgemeinschaften mit Kindern, beide Partner erwerbstätig; sonstigen Haushalten (Kühnen 1999: 111).

28 Selbstständige und Landwirte, Beamte, Angestellte, Arbeiter, Rentner/Pensionäre; Sonstige Nichterwerbstätige.

29 Die obere Abschneidegrenze für das Einkommen lag 1973 bei monatlich 15.000 DM, 1978 bei 20.000 DM, 1983 bis 1988 bei 25.000 DM, seit 1993 liegt sie bei 35.000 DM (2003: 18.000 Euro) (Hauser und Stein 2001: 160). Im Mikrozensus werden die Nettoeinkommen erstmals seit 2000 auch in den oberen Einkommensklassen etwas differenzierter erfasst. Die nach oben offene Randklasse liegt nun bei „35.000 DM (bzw. 17.895 Euro) und mehr". Eine der wenigen amtlichen Datenquellen, in der auch Top-Einkommensbezieher differenziert erfasst werden, ist die alle drei Jahre durchgeführte Lohn- und Einkommenssteuerstatistik, die seit kurzem auch als Scientific-Use-File für Forschungsvorhaben genutzt werden kann (http://www.forschungsdatennetzwerk.de/datenangebot.asp).

30 Für eine kritische Diskussion der Vermögenserfassung in der EVS siehe Hauser und Stein (2001), für Anmerkungen zu methodischen Besonderheiten der EVS siehe Bundesregierung

Trotz der erwähnten Einschränkungen ist die EVS hinsichtlich der Einnahmen- und Ausgabensituation von privaten Haushalten in Deutschland die wohl umfassendste Datenquelle, die der empirischen Sozial- und Wirtschaftsforschung gegenwärtig zur Verfügung steht. Keine andere Erhebung enthält eine solche Vielfalt an Informationen über die Zusammensetzung und Verteilung von Haushaltseinkommen, Haushaltsvermögen und Schulden, sozialen Transferleistungen, Haushaltsausgaben sowie zum Besitz von langlebigen Konsumgütern. Gegenwärtig stehen die EVS 1993, 1998 und 2003 als Scientific-Use-File für die Forschung zur Verfügung. Darüber hinaus werden in absehbarer Zeit auch die Mikrodaten der EVS für den Zeitraum vor 1993 in dieser Weise für die Forschung nutzbar sein.[31]

In der empirischen Sozial- und Wirtschaftsforschung wird die EVS bislang vor allem für die Analyse von Einkommens- und Vermögensverteilungen sowie Armutsstudien genutzt.[32] So war sie unter anderem eine wesentliche Datenquelle für den ersten Armuts- und Reichtumsbericht der Bundesregierung (2001a, 2001b; Hauser und Becker 2001a). Darüber hinaus wird sie insbesondere für die Untersuchung sozialpolitischer Fragestellungen herangezogen (Hauser 2003; Hauser und Becker 2001b). Vor dem Hintergrund der im Zuge der Rentenreform vorgenommenen Einführung einer zusätzlichen privaten Alterssicherung auf freiwilliger Basis gehen etwa Hauser und Stein (2001) auf Basis der EVS der Frage nach, inwieweit Familienhaushalte zwischen Ende der 1980er und Ende der 1990er Jahre in der Lage waren, privates Vermögen zu kumulieren, das für die Altersvorsorge genützt werden könnte. Danach ist in Westdeutschland die relative Vermögensposition von familiären Haushalten, d.h. Lebensformen mit Kindern, im Vergleich zum Gesamtdurchschnitt rückläufig. Paare mit einem Kind verfügen über ein Nettovermögen von durchschnittlich 85 Prozent, Paare mit zwei und mehr Kindern über 66 Prozent des Durchschnittsvermögens aller Haushalte. Stark unterdurchschnittlich ist die Vermögensposition von Alleinerziehenden, deren Vermögen Ende der 1990er Jahre nur 46 Prozent des durchschnittlichen Haushaltsvermögens beträgt. Zwischen 1993 und 1998 sinken bei ihnen auch die absoluten Vermögenswerte. Wenig überraschend nehmen kinderlose Paare mit 142 Prozent eine überdurchschnittliche Vermögensposition ein, gefolgt von alleinstehenden Männern mit 128 Prozent und „sonstigen" Haushalten mit 117 Prozent. Zusammenfassend lassen

(2001b: 26ff.). Vorschläge für eine stärker am Forschungsbedarf orientierte Datenerhebung finden sich in Bundesregierung (2001b: 35ff.).

31 Diese alten EVS-Daten sind ein schönes Beispiel für nicht-intendierte und doch positive Effekte der Übermittlung von amtlichen Mikrodaten an die Wissenschaft vor der Verschärfung des Statistikgesetzes in den 1980er Jahren. Denn es handelt sich hierbei um Daten, die z.T. noch zu Zeiten des ehemaligen Sonderforschungsbereichs 3 der Universitäten Frankfurt und Mannheim von der amtlichen Statistik an die Forschung übermittelt und von der Forschungsgruppe um Richard Hauser kontinuierlich genutzt und gepflegt wurden. Die im Rahmen dieses EVS-Projekts aufgebaute Datenbank umfasst die Erhebungen (bzw. Merkmalsauszüge hieraus) ab 1962/63 bis einschließlich 1988. Insbesondere die Daten aus den 1960er Jahren waren nur noch in diesem Projekt, nicht mehr jedoch bei der amtlichen Statistik vorhanden. Auf Initiative von Richard Hauser werden diese Daten künftig als Scientific-Use-Files verfügbar sein.

32 Auch für familiensoziologische Fragestellungen bietet die EVS ein durchaus interessantes Analysepotenzial, z.B. in Hinblick auf den Zusammenhang zwischen bildungsspezifischen Haushaltsressourcen und Familiengründung. Allerdings wird dieses Potenzial unseres Wissens von der empirischen Sozialforschung bislang noch kaum genutzt.

die Analysen erkennen, dass familiäre Haushalte während des Beobachtungszeitraums stärkere Probleme bei der Vermögensbildung hatten als andere Haushaltstypen. Dabei erweist sich die Vermögenslage von Alleinerziehenden als besonders ungünstig. Für Hauser und Stein (2001) stellt sich daher die Frage, ob Familien mit Kindern überhaupt in der Lage sein werden, die Sparbeträge für die zusätzliche Alterssicherung zu erbringen. Sie vermuten viel eher, dass die staatlichen Anreize zur privaten Altersvorsorge in der Tendenz von den oberen Vermögensgruppen genutzt werden und die ungleiche Vermögensverteilung zwischen den verschiedenen Haushaltstypen wenn nicht vergrößert, so auch nicht verringert wird. Wenngleich mit den Daten auch noch Fragen offen bleiben,[33] so ist die vorliegende Analyse sozialpolitisch sicher von erheblicher Bedeutung.

In der politischen wie auch der wissenschaftlichen Diskussion entsteht gelegentlich der Eindruck, dass ein individueller Niedriglohn gleichzusetzen ist mit einer prekären Einkommenssituation. Bei dieser Argumentation bleibt sowohl der Haushaltskontext wie auch das staatliche Umverteilungssystem ausgeblendet, beide Faktoren beeinflussen jedoch die Einkommensposition. Zusätzliche Einkommen von Haushaltsmitgliedern können Armut verhindern, umgekehrt kann aber auch der Haushaltskontext erst zu einer prekären Einkommensposition führen. Das staatliche Transfersystem kann Einkommensarmut mindern, es ist aber auch nicht ausgeschlossen, dass das Einkommen von Individuen durch Steuer- und Sozialabgaben unter die Armutsgrenze sinkt (Strengmann-Kuhn 2000). Diese Zusammenhänge werden in einer Studie von Becker (2000) auf Basis von EVS-Daten einer näheren Betrachtung unterzogen. Als wesentlicher Befund lässt sich festhalten, dass das Armutsrisiko von Niedriglohnbeziehern zwar höher ist als bei Erwerbstätigen im Durchschnitt, aber kein deterministischer Zusammenhang zwischen Niedriglohn und Einkommensarmut vorliegt. Von den im Niedriglohnbereich im engeren Sinne[34] Erwerbstätigen (circa 10 Prozent aller Erwerbstätigen) befindet sich – nach Einbeziehung des Haushaltskontextes – nur noch knapp jeder Dritte weiterhin in prekären Einkommenssituationen.[35] Beim Rest wird das geringe individuelle Einkommen durch das Arbeitseinkommen der anderen Haushaltsmitglieder kompensiert. Umgekehrt sinken rund vier Prozent der Personen mit höherem Einkommen – nach Berücksichtigung des Haushaltskontextes – unter die Niedrigeinkommensgrenze. Gleichfalls armutsreduzierend wirkt das staatliche Transfersystem. Mehr als die Hälfte der auch im Haushaltskontext unterhalb der Armutsschwelle verbleibenden Erwerbstätigen erreicht durch staatliche Transfers ein über der Armutsgrenze liegendes Einkommen. Interessant in diesem Kontext ist, dass ein überdurchschnittlicher Verdienst nicht immer vor einer prekären Einkommenssituation schützt: Über ein Drittel der Haushalte, deren Haupteinkommensbezieher ein relatives Bruttoeinkommen zwischen 150 und 200 Prozent des durchschnittlichen Nettoäquivalenzeinkommens be-

33 Zum Beispiel: Handelt es sich bei den kinderlosen Paaren tatsächlich um solche ohne Kinder oder auch um solche, bei denen die Kinder den Familienhaushalt schon verlassen haben?

34 Niedriglohngrenze im engen Sinne: Bruttomonatslohn entspricht maximal der Hälfte des durchschnittlichen Nettoäquivalenzeinkommens. Niedriglohngrenze im weiten Sinne: Bruttomonatslohn in Höhe des durchschnittlichen Nettoäquivalenzeinkommens (Becker 2000: 2).

35 Prekäre Einkommenssituationen im Haushaltskontext sind definiert als Nettoäquivalenzeinkommen unterhalb von 75 Prozent des durchschnittlichen Nettoäquivalenzeinkommens (Becker 2000: 3).

zieht, lebt in unzulänglichen Einkommensverhältnissen. Hierbei handelt es sich vorwiegend um Familien mit mehreren Kindern, bei welchen der Familienlastenausgleich nicht ausreicht, ein Absinken in prekäre Einkommenssituationen zu verhindern (Becker 2000). Abgesehen von der gesellschaftspolitischen Relevanz dieser Befunde verdeutlichen sie auch, dass die Erfassung des individuellen Bruttoeinkommens als Indikator für die Einkommenssituation im Haushaltskontext nicht ausreichend ist.

3. Der Mikrozensus als Scientific-Use-File

Gemeinsam mit der Beschäftigtenstatistik und der Einkommens- und Verbrauchsstichprobe stellt der Mikrozensus eine der wichtigsten amtlichen Datenquellen für die empirische Wirtschafts- und Sozialforschung dar. Der Mikrozensus, der seit 1957 jährlich erhoben wird, ist eine Ein-Prozent-Haushaltsstichprobe und umfasst über 320.000 Haushalte mit circa 720.000 Personen. Für den Mikrozensus besteht eine gesetzliche Teilnahmepflicht.[36] Die Erhebung ist als eine Mehrzweckstichprobe angelegt. Erwerbsstatistische und soziodemographische Merkmale bilden das Kernprogramm. Darüber hinaus werden Informationen über die soziale Absicherung, Quellen des Lebensunterhalts und Einkommenshöhe, zum allgemeinen und beruflichen Ausbildungsabschluss, zur Aus- und Weiterbildung, zur Gesundheit und zur Wohnsituation erhoben (Lüttinger und Riede 1997). Das Erhebungsprogramm ist zwar primär am Informationsbedarf von Verwaltung und Politik konzipiert. Auf Grund seiner thematischen Breite, der Erfassung aller Personen im Haushalt, seiner langfristigen Periodizität in Kombination mit einem sehr großen Stichprobenumfang und einer hohen Ausschöpfungsquote weist der Mikrozensus aber ein beträchtliches Potenzial für sozialwissenschaftliche Analysen auf.

Nach einer Durchsicht der bislang auf Basis von Mikrozensen durchgeführten Studien (Schimpl-Neimanns 1998, 2002) wird der Mikrozensus vor allem für die Untersuchung von Fragestellungen herangezogen, bei welchen ein großer Stichprobenumfang hilfreich ist. Die Daten werden vorwiegend für Untersuchungen kleiner Bevölkerungspopulationen genutzt, die bei sozialwissenschaftlichen Daten oftmals an zu geringen Fallzahlen scheitern, ob es sich nun um atypische Beschäftigungsformen (Groß 2001), Selbstständige (Jungbauer-Gans 1999; Pfeiffer 1999b), Migranten (Baumann 1999; Granato und Kalter 2001; Granato 2003), differenzierte Bildungsgruppen (Schimpl-Neimanns 2000), Nichteheliche Geburten (Konietzka und Kreyenfeld 2001) oder spezifische Berufsgruppen (Voges 2002) handelt.

Beispielhaft soll hier auf eine Arbeit von Granato und Kalter (2001)[37] eingegangen werden, die sich mit der Frage nach den Ursachen der andauernden ethnischen Un-

36 Allerdings sind bestimmte Merkmale von der Auskunftspflicht befreit. Hierzu gehören u.a. die Angaben zum Eheschließungsjahr, zur Wohn- und Lebensgemeinschaft, zur Aufenthaltsdauer (für Ausländer) sowie zum allgemeinen und beruflichen Ausbildungsabschluss (für Personen ab dem 51. Lebensjahr).

37 Weitere Analysen zu ethnischen Ungleichheiten auf Basis des Mikrozensus beschäftigen sich etwa mit dem Bildungssystem (Alba et al. 1994a; Dinkel et al. 1999; Kristen und Granato 2004) sowie mit demographischen und familiensoziologischen Dimensionen (Kalter 2001;

gleichheit auf dem deutschen Arbeitsmarkt beschäftigt. Ausgangspunkt der Untersuchung ist der in einer Reihe von neueren Studien konstatierte Sachverhalt, nach welchem die arbeitsmarktsspezifische Ungleichheit zwischen Migranten und Deutschen im Zeitverlauf zwar zurückgegangen ist, nach wie vor jedoch erhebliche Unterschiede erkennbar sind. Diese äußern sich in einer im Vergleich zu deutschen Erwerbstätigen deutlich schlechteren Arbeitsmarktpositionierung von Migranten der zweiten Generation. Die für die Benachteiligung der Migranten der ersten Generation zentralen Erklärungsfaktoren der sozial selektiven Migration, einer als Folge der Migration partiellen Entwertung von Humankapital sowie der spezifisch konjunkturellen Situation mit einer hohen Nachfrage nach gering qualifizierten Arbeitskräften, treffen für die zweite Generation nicht mehr zu. Der Bedarf an niedrigqualifizierten Arbeitskräften ist rapide zurückgegangen, die zweite Generation ist entweder in Deutschland geboren oder vor dem siebten Lebensjahr nach Deutschland gezogen. Die für den Arbeitsmarkt relevanten schulischen und beruflichen Qualifikationen wurden in Deutschland erworben. Eine naheliegende Erklärung wäre das Vorliegen von ethnischen Diskriminierungsmechanismen auf dem Arbeitsmarkt. Allerdings greift diese Erklärung dann zu kurz, wenn nicht zugleich auch die Möglichkeit nach wie vor ethnisch unterschiedlicher Humankapitalausstattungen berücksichtigt wird. In ihrer Arbeit gehen Granato und Kalter dieser Frage nach, indem sie die Arbeitsmarktpositionierung von Deutschen und sechs Migrantengruppen (Griechenland, Italien, Ex-Jugoslawien, Portugal, Spanien und Türkei) auf Basis des Mikrozensus empirisch untersuchen. Berücksichtigt wird hierbei sowohl die Qualifikation wie auch der Generationenstatus. Als wesentliches Ergebnis der multivariaten Analyse lässt sich festhalten, dass die Schlechterstellung von Migranten der ersten Generation auf dem Arbeitsmarkt zu einem beträchtlichen Teil durch die geringere Ausstattung mit formalem Bildungskapital erklärbar ist, aber bei weitem nicht vollständig. Auch bei gleichem Bildungsniveau liegen die relativen Chancen der ersten Generation, eine Angestelltenposition zu erreichen, nur zwischen dem 0,13-fachen für Türken bis 0,35-fachen für Portugiesen der deutschen Vergleichspopulation. Für die zweite Generation hingegen sind mit Ausnahme der türkischen Population diese Nationalitäteneffekte nahezu verschwunden. Dies spricht dafür, dass die für Migranten der zweiten Generation beobachtbare ungünstigere Arbeitsmarktpositionierung – abgesehen von der türkischen Population – nicht primär durch diskriminierende Allokationsmechanismen auf dem Arbeitsmarkt erzeugt wird, sondern vor allem aus der unterschiedlichen Humankapitalausstattung der ethnischen Gruppen resultiert. Letzteres kann jedoch durchaus die Folge anderer Arten von Diskriminierungsmechanismen – etwa im Bildungssystem – sein (Granato und Kalter 2001: 515).

Eine weitere Stärke des Mikrozensus ergibt aus der Anlage als Haushaltsstichprobe, die eine Abbildung von Haushalts- und Familienkontexten erlaubt. Befragt werden alle in einem Haushalt bzw. in einer Familie lebenden Personen, woraus sich für familiensoziologische Studien ein vielfältiges Forschungspotenzial ergibt (Wirth 1996; Dathe 1999; Roloff 1999; Wirth 2000; Konietzka und Kreyenfeld 2001; Huinink 2002; Kreyenfeld 2004). Der Haushalt wird im Mikrozensus definiert über das Kriterium des

---

Weick 2001; Kalter und Granato 2002). Für eine Darstellung von Stärken und Schwächen des Mikrozensus bei der Untersuchung von Migrations- und Integrationsfragen siehe Schimpl-Neimanns (2002).

gemeinsamen Wohnens und Wirtschaftens von Personen. Der Familienbegriff beinhaltet Ehepaare mit und ohne Kinder, sowie – unabhängig vom Familienstand – alleinerziehende Personen mit ihren im Haushalt lebenden ledigen Kindern. Seit 1996 werden auch nichteheliche Lebensgemeinschaften direkt erfasst, die zuvor nur indirekt abgebildet werden konnten. Analyserestriktionen ergeben sich vor allem durch die Art der Erfassung von Verwandtschaftsbeziehungen sowie der Beschränkung auf den Befragungshaushalt. So beziehen sich die erfragten Verwandtschaftsverhältnisse nur auf die Bezugsperson des Haushalts und seit 1996 – sofern vorhanden – auch auf den Lebenspartner bzw. die Lebenspartnerin der Haushaltsbezugsperson. Sofern mehrere Familien in einem Haushalt leben, ist daher nicht genau bekannt, ob und in welchen Verwandtschaftsbeziehungen die einzelnen Personen zueinander stehen. Stief-, Adoptiv- oder Pflegekinder werden der Kategorie leibliche Kinder zugeordnet, was in der Konsequenz dazu führt, dass die in der Tendenz zunehmenden Patchwork-Familien und deren spezifischer Problemlagen mit dem Mikrozensus nicht untersucht werden können. Durch die Beschränkung auf die im Haushalt lebenden Personen (Koresidenzprinzip) stehen weiterhin keine Informationen über außerhalb des Haushalts lebende Kinder oder Partner zur Verfügung. Im Hinblick auf die anhaltende gesellschaftspolitische Debatte über die zunehmende Kinderlosigkeit hochqualifizierter Frauen (Engstler 1998), stellt das Nichtwissen über die tatsächlich vorhandenen Kinder familiensoziologisch ein erhebliches Analysedefizit des Mikrozensus dar.[38] Auch die wachsende Zahl der „living-apart-together"-Beziehungen, d.h. Paare mit getrennter Haushaltsführung und ihre spezifischen Arbeits- und Lebensbedingungen können mit dem Mikrozensus nicht untersucht werden. Der Versuch, diese modernen Partnerschaftsformen zukünftig ebenfalls im Mikrozensus zu erfassen, fand auf Seiten des Bundesrats keine Zustimmung (Bundesrat Drucksache 195/04: 1).

Ungeachtet der erwähnten Einschränkungen bietet die Abbildung des Haushaltskontextes im Mikrozensus ein erhebliches Forschungspotenzial, was bspw. auch von der Bildungsforschung genutzt wird. Hier bietet sich die Möglichkeit, Merkmale von Jugendlichen in schulischer oder beruflicher Ausbildung mit Informationen zur sozioökonomischen Lage der Familie zu verknüpfen. Als Beispiel hierfür sei eine methodisch aufwändige Untersuchung von Schimpl-Neimanns (2000) genannt. Diese geht der Frage nach, inwieweit der Ausbau des Bildungswesens seit den 1960er Jahren mit einem Abbau der herkunftsspezifischen Bildungsungleichheiten einhergegangen ist. Betrachtet wird die Bildungsbeteiligung (Schulbesuch oder erreichter Bildungsabschluss) 14- bis 18-jähriger deutscher Jugendlicher über einen Zeitraum von fast 40 Jahren (1950 – 1989) in Abhängigkeit des Bildungsniveaus des Elternhauses sowie der beruflichen Stellung des Familienvorstands. Ein wesentliches Ergebnis dieser Studie ist, dass im betrachteten Zeitraum nur geringfügige Veränderungen des Zusammenhangs zwischen sozialer Herkunft und Bildungsbeteiligung stattgefunden haben. Diese spiegeln sich ins-

38 Es mutet etwas ironisch an, dass die für das Mikrozensusgesetz ab 2005 neu vorgeschlagene Frage nach der Anzahl der von einer Frau geborenen Kinder durch den Bundesrat (Drucksache 195/04) abgelehnt wurde, mit der Begründung: „dass diese Frage auch in der Bevölkerungsstatistik mit Rücksicht auf den Adoptionsschutz und andere Gegebenheiten nicht einmal im Zusammenhang mit der Geburtenmeldung erhoben wird". Denn der Kern des Problems besteht ja eben darin, dass diese Information nirgendwo in der amtlichen Statistik erhoben wird.

besondere in einem abgeschwächten Einfluss der beruflichen Stellung des Familienvorstands auf den Schulbesuch der Kinder wider. Am deutlichsten zeigt sich dieser Effekt in Bezug auf die Chancen des Besuchs einer weiterführenden Schule versus der Hauptschule. Hier ist eine deutliche Reduzierung der sozialen Selektivität im Zeitverlauf erkennbar. Noch immer stark selektiv wirkt die sozioökonomische Herkunft hingegen hinsichtlich der Chance von Jugendlichen, ein Gymnasium oder eine Realschule zu besuchen. In Hinblick auf den Einfluss des Bildungsniveaus der Eltern auf den Schulbesuch der Jugendlichen deutet sich ein Ungleichheitsabbau bezüglich der relativen Chance eines Realschul- versus Hauptschulbesuchs an. Im Unterschied hierzu scheinen die nach Bildungsherkunft ungleichen Chancen von Kindern, ein Gymnasium versus eine Realschule zu besuchen, im Zeitverlauf konstant, wobei sich Ende der 1980er Jahre sogar eine Tendenz zur Ungleichheitszunahme andeutet. Im Kontext der PISA-Studie, welche die öffentliche Diskussion über das Ausmaß und die Folgen sozial selektiver Bildungschancen wiederbelebt hat, wäre eine Aktualisierung der oben beschriebenen Untersuchung mit neueren Daten sicherlich von hohem Interesse. Allerdings ist eine Fortsetzung der Zeitreihe in dieser Form nicht mehr bzw. nur noch sehr eingeschränkt möglich, da das Merkmal Schulbesuch im Mikrozensus seit 1991 nach der Internationalen Standardklassifikation für das Bildungswesen (ISCED), d.h. in Klassenstufen, erhoben wird. Dies erleichtert zwar internationale Vergleiche, wird aber den Besonderheiten des dreigliedrigen Bildungssystems in Deutschland nicht gerecht.

Neben den bereits erwähnten Studien dokumentiert sich das Analysepotenzial des Mikrozensus in einer Vielzahl weiterer Publikationen. Zu erwähnen sind Untersuchungen, die die jährliche Periodizität nutzen, um den langfristigen sozialen und wirtschaftlichen Wandel zu untersuchen. Diese Analysen unterliegen zwar gewissen Einschränkungen, da der Mikrozensus gegenwärtig erst ab dem Erhebungsjahr 1989 in Form von Scientific-Use-Files zur Verfügung steht.[39] Für den Zeitraum bis 1989 muss sich die Forschung bislang noch mit Aggregatdaten behelfen. Dennoch liegen inzwischen eine Reihe von Längsschnittstudien zu verschiedenen thematischen Fragestellungen vor, bspw. zum Wandel des Zusammenhangs zwischen Bildungsniveau und Berufsstatus (Handl 1996), zur Entwicklung der Einkommensungleichheit (Steiner und Wagner 1998), zu der geschlechtsspezifischen Erwerbstätigkeit nach Bildung (Fitzenberger et al. 2004), zu Veränderungen der Arbeitszeitstruktur (Wolf 1999) oder der beruflichen Selbstständigkeit (Luber 2003).

Ergänzend zu den jährlich erhobenen Fragen des Mikrozensus-Grundprogramms werden weitere thematische Felder in Zusatz- und Ergänzungsprogrammen in mehrjährigem Abstand, teilweise ohne Auskunftspflicht und mit reduziertem Auswahlsatz, abgedeckt (Lotze und Breiholz 2002). Solche Fragemodule betreffen beispielsweise spezielle Aspekte der Gesundheit (Helmert und Borgers 1998; Warschburger und Lohre 1998), der Wohnsituation oder der sozialen Sicherung. Bulmahn (1999) zeigt auf Basis von Mikrozensusdaten, dass die private Altersvorsorge (in Form einer Lebensversicherung) stark von der Höhe des Einkommens und der beruflichen Position beeinflusst ist. Insbesondere die private Altersvorsorge von un- und angelernten Arbeitern und

39 Zu den Möglichkeiten das Prinzip der faktischen Anonymisierung auf die Mikrozensus-Erhebungen vor 1989 anzuwenden vgl. Lüttinger und Wirth (2004).

einfachen Angestellten, deren gesetzliche Rentenansprüche eher gering ausfallen werden und die eher selten auf eine betriebliche Altersvorsorge zurückgreifen können, ist im Vergleich zu anderen sozialen Gruppen deutlich geringer ausgeprägt.

Wie bereits erwähnt, weisen die im Mikrozensus erhobenen Einkommensinformationen im Vergleich zur Einkommens- und Verbrauchsstichprobe einen erheblich geringeren Informationsgehalt auf. Gefragt wird nach dem gruppierten Monats-Nettoeinkommen (April), bezogen auf alle Einkommensquellen. Hierin enthalten sind Einkommen aus selbstständiger Tätigkeit und abhängiger Erwerbstätigkeit, staatliche Transferleistungen (Sozialhilfe, Arbeitslosenunterstützung, Kindergeld, Renten etc.), private Unterstützungen, Einkommen aus Vermietung, Verpachtung und Kapitalvermögen sowie Sachbezüge.[40] Mit der Konzentration auf das Gesamt-Nettoeinkommen und dem Fehlen einer direkten Angabe zum Erwerbseinkommen ist das Analysepotenzial der Daten für arbeitsmarktökonomische Fragestellungen stark begrenzt. Ungeachtet dieser Einschränkung wird der Mikrozensus vermehrt auch zur Schätzung von Einkommensfunktionen genutzt (Pfeiffer und Brade 1995; Jungbauer-Gans 1999; Pfeiffer 1999a; Freeman und Schettkat 2000; Groß 2001; Butz 2001; Machin und Puhani 2003).

4. Internationale Perspektive

Für international vergleichende Studien erweist es sich als Vorteil, dass die europäische Arbeitskräftestichprobe in Form einer Substichprobe in den Mikrozensus integriert ist. Hinzu kommt, dass inzwischen verschiedene, auf den internationalen Vergleich ausgerichtete sozialwissenschaftliche Konzepte, z.B. die CASMIN-Skala für Bildung, die EGP-Skala für die sozioökonomische Lage oder ISEI für den sozioökonomischen Status, für Mikrozensusdaten operationalisiert wurden und der Forschung allgemein zugänglich zur Verfügung stehen. Thematische Schwerpunkte der bisher mit dem Mikrozensus durchgeführten international vergleichenden und zum Teil auch längsschnittlich ausgerichteten Studien bilden Fragen zum Zusammenhang zwischen Bildungsqualifikation und sozio-ökonomischer Position bzw. Arbeitsmarktlage (Brauns et al. 1999, 2000, 2003; Haas und Lutz 1999; Lohmann und Luber 2000; Fitzenberger und Wunderlich 2002; Luber 2003). Müller und Gangl (2003), Müller et al (2000) und Kogan und Müller (2003) haben mit den Daten der Europäischen Arbeitskräftestichprobe gezeigt, wie und weshalb an der Schnittstelle des Übergangs vom Bildungssystem ins Erwerbssystem in den unterschiedlichen Ländern Europas die Bedingungen der Integration von Ausbildungsabsolventen in das Erwerbssystem variieren. Haisken-DeNew und Schmidt (1999) betrachten branchenspezifische Einkommensunterschiede in den USA und in Deutschland. Machin und Puhani (2003) untersuchen geschlechtsspezifische Einkommensunterschiede von Akademikern in Deutschland und Großbritannien in Abhängigkeit vom Studienfach. Wenig überraschend stehen die relativ höheren Einkommen von Männern in beiden Ländern in einem engen Zusammenhang mit ihrer im Vergleich zu Frauen deutlich stärkeren Konzentration auf naturwissenschaftliche

40 Bei selbstständigen Landwirten und mithelfenden Familienangehörigen wird das Einkommen nicht erhoben.

oder technisch orientierten Studienabschlüsse, wobei dieser Effekt in England stärker ausgeprägt ist als in Deutschland.

Als eine Konsequenz der europäischen Integrationsprozesse werden komparative Ansätze für die empirische Forschung wichtiger. Wenngleich die meisten Untersuchungen nach wie vor am nationalen Kontext orientiert sind, rückt der europäische Integrationsprozess den Vergleich von Randbedingungen und Trends in einer Nation mit jenen in anderen europäischen Nationen in den Vordergrund. Eine unabdingbare Voraussetzung hierfür sind zuverlässige, differenzierte und vergleichbare Daten zur Sozialstruktur, wie sie derzeit in Form der europäischen Arbeitskräftestichprobe zum Teil vorliegen. Obwohl verstärkt am Aufbau international vergleichbarer sozialwissenschaftlicher Erhebungen (z.B. ISSP oder ESS) gearbeitet wird, wird es noch einige Zeit in Anspruch nehmen, bis die empirische Sozialforschung einen, amtlichen Mikrodaten vergleichbar umfangreichen, Datenbestand für Europa verfügbar haben wird (Müller 1999b).

Der europäische Integrationsprozess wird nicht nur zu einem Bedeutungsverlust von nationalen Grenzen führen. Je mehr die nationalen Einheiten an Macht verlieren und die supranationale Einheit an Form gewinnt, desto wichtiger werden auf Grund des „Subsidaritätsprinzips" subnationale Regionen werden. Die Herausbildung von regionalen Identitäten und die Unterstützung einer mehr oder weniger balancierten Entwicklung der verschiedenen Regionen innerhalb von Europa sind zentrale Punkte der politischen Agenda der Europäischen Union. Es ist absehbar, dass regionale Belange – gerade in Hinblick auf die Osterweiterung – die politische Diskussion in den nächsten Jahren erheblich beeinflussen werden. In der Folge wird auch die empirische Sozialforschung der Regionalisierung ihrer Forschungsbefunde mehr Gewicht als bisher beimessen. Sowohl in Deutschland wie auch in Europa als Ganzem wird die Forschung für differenzierte Regionalanalysen vor allem auf amtliche Mikrodaten angewiesen sein.

Legt man diese Trends zugrunde, wird die Bedeutung von amtlichen Mikrodaten für die international vergleichende Sozialforschung mittelfristig zunehmen. Bislang ist der Zugang zu internationalen amtlichen Mikrodaten allerdings noch sehr unbefriedigend. Die Nutzung der faktisch anonymisierten Mikrodaten ist in Deutschland aus rechtlichen Gründen auf Wissenschaftler und Forschungsinstitutionen im Inland beschränkt. Die Daten werden weder an deutsche noch an ausländische Forscher im Ausland weitergegeben. Es gibt auch noch keinen EU-weiten Scientific-Use-File für den European Labour Force Survey. Allerdings wurde kürzlich in einer EU-Verordnung[41] der Zugang der Forschung zu Gemeinschaftsstatistiken geregelt. Danach werden sich in Zukunft zwei Zugangswege anbieten: Erstens durch Analysen vor Ort, d.h. in den Räumen von Eurostat; zweitens durch die Weitergabe von anonymisierten Daten an die Wissenschaft. Bei einer näheren Betrachtung der zweiten Regelung deuten sich jedoch schon zukünftige Probleme an. So dürfen selbst anonymisierte Daten dann nicht weitergegeben werden, wenn ein Staat dies nicht wünscht. Darüber hinaus sieht die Definition für anonyme Daten eine Minimierung des Reidentifikationsrisikos in Übereinstimmung mit den zum jeweiligen Zeitpunkt besten verfügbaren Verfahren vor.

41 Verordnung (EG) Nr. 831/2002 der Kommission vom 17. Mai 2002 zur Durchführung der Verordnung Nr. 322/97 des Rats über die Gemeinschaftsstatistiken – Regelung des Zugangs zu vertraulichen Daten für wissenschaftliche Zwecke.

Dies kann bei einer engen Auslegung im Sinne von absoluter Anonymität interpretiert werden. Die verheerenden Folgen für das Analysepotenzial der Daten wurden eingangs charakterisiert. Es bleibt zu hoffen, dass sich Eurostat für eine wissenschaftsfreundliche Interpretation dieser Regel entschließt, d.h. sich für eine Anonymisierung analog zu den heutigen Standards für die Scientific-Use-Files von Mikrozensus, EVS oder Beschäftigungsstichprobe entscheidet.

V. Ausblick

Initiiert durch die Aktivitäten der KVI (2001) haben sich die allgemeinen Rahmenbedingungen für die Sekundäranalyse von amtlichen Mikrodaten in der empirischen Wirtschafts- und Sozialforschung in jüngster Zeit deutlich verbessert. Dies zeigt sich sowohl in der Etablierung der erwähnten Forschungsdatenzentren bei verschiedenen öffentlichen Datenproduzenten, die der Forschung einen Zugang vor Ort zu nur schwer anonymisierbaren amtlichen Einzeldaten bieten sollen, wie auch in der Förderung von wissenschaftsnahen Servicezentren. Letzteren kommt unter anderem die Aufgabe zu, das Analysepotenzial der Daten durch wissenschaftliche Dienstleistungen (Lüttinger et al. 2003b) zu erweitern und bei der Operationalisierung der faktischen Anonymität der unterschiedlichen Datenbestände mitzuwirken. Der enorme Fortschritt kommt jedoch vor allem auch darin zum Ausdruck, dass gegenwärtig in einer Reihe von Projekten überprüft wird, in welcher Form für die Forschung bislang nicht oder nur sehr schwer zugängliche Datenbestände als Scientific-Use-Files verfügbar gemacht werden können. Hierzu gehören wirtschaftsstatistische Daten (Ronning und Gnoss 2003), Sozialversicherungsdaten (Rehfeld 2004) und steuerstatistische Daten (Zwick 1999). Ähnliche Entwicklungen sind für andere Bereiche wie Gesundheit oder Umwelt zu erwarten. In hoffentlich naher Zukunft werden der Forschung damit zunehmend mehr amtliche Mikrodaten für unterschiedlichste inhaltliche Bereiche in ähnlich flexibler Weise wie die von den Sozialwissenschaften generierten Daten für die Bearbeitung von inhaltlichen und methodischen Fragestellungen zur Verfügung stehen.

Der einfache, möglichst unbürokratische und kostengünstige Zugang ist eine zentrale Voraussetzung für die Nutzung dieser Daten für Sekundäranalysen, allerdings nicht die einzige. Welche Bedeutung der Analyse von amtlichen Mikrodaten zukünftig für den Erkenntnisgewinn zukommen wird, ist auch maßgeblich davon beeinflusst, inwieweit die öffentlichen Datenproduzenten im Vorfeld der Datengenerierung Tendenzen des wissenschaftlichen Datenbedarfs, die auf Grundlage der neueren Theorie- und Methodenentwicklung absehbar sind, zur Kenntnis nehmen werden.

Von erheblicher Bedeutung ist hierbei der zunehmende Bedarf an Panel- und Ereignisdaten, die tiefergehende Erkenntnisse über das soziale Geschehen ermöglichen als Querschnittsdaten. Ursache-Wirkungsmechanismen können nur mit Paneldaten in ihrer zeitlichen Reihenfolge theoretisch angemessen untersucht werden. Bei Querschnittsdaten ist man für die Abbildung von kausalen Mechanismen auf Hilfsannahmen angewiesen, die nicht immer der Realität entsprechen. Bspw. ist in ländlichen Regionen der Anteil der verheirateten Population höher, der Anteil der ledigen Population niedriger als in großstädtischen Regionen. Ist dies nun die Folge einer stärkeren Wertschätzung

der Ehe in ländlichen Regionen oder von Stadt-Land-Wanderungsprozessen im Anschluss an die Eheschließung und insbesondere bei einer Familiengründung? Die kausale Abfolge dieser sozialen Sachverhalte lässt sich mit Querschnittsdaten nicht adäquat untersuchen, hierzu werden Paneldaten benötigt. Auch in Hinblick auf die Untersuchung des sozialen Wandels sind Querschnittsdaten oftmals nicht hinreichend, da sie Veränderungen nur auf der Aggregatebene abbilden. Die der Bildung der Aggregate zugrunde liegende Dynamik entzieht sich dabei der Beobachtung. Wie etwa aus der Armutsforschung bekannt ist, stellen Armutsquoten nur eine Facette der Armut dar. Analysen auf Basis von Paneldaten zeigen, dass nur ein kleiner Teil der von monetärer Armut Betroffenen über einen längeren Zeitraum arm ist. Die Untersuchung solcher dynamischen Prozesse ebenso wie das Verstehen der auslösenden Mechanismen setzt die Verfügbarkeit von Paneldaten voraus. Während die Forschung seit geraumer Zeit verstärkt auf das Potenzial von Paneldaten für den Erkenntnisfortschritt setzt, weisen die amtlichen Datenbasen hier noch erheblichen Nachholbedarf auf. Nur wenige Datenbestände, wie z.B. die IAB-Beschäftigtenstichprobe, die laufenden Wirtschaftsrechnungen oder der Mikrozensus weisen bislang Panelcharakter auf. Wie oben (vgl. Fußnote 16) erwähnt, läuft gegenwärtig ein vom BMBF gefördertes Projekt mit dem Ziel, ein Mikrozensus-Panel als Scientific-Use-File zu generieren. Bedingt durch das Erhebungsdesign enthält der Mikrozensus allerdings nur für solche Haushalte Paneldaten, die innerhalb des vierjährigen Beobachtungszeitraum nicht umgezogen sind. Die sich hieraus ergebenden Selektivitätsprobleme sind evident. Die Bedeutung des Mikrozensus für die Forschung wäre noch um ein Vielfaches größer, wenn das Erhebungsdesign zu einem methodisch einwandfreien Paneldesign umgestaltet werden würde. Auch für die EVS wäre ein Übergang zu einer Panelbefragung anzustreben (Hauser und Becker 2001a).

Ein weiterer Trend, der in den kommenden Jahren verstärkt an Bedeutung gewinnen wird, ist die Verknüpfung von unterschiedlichen Datenbasen auf individueller Ebene. Vorreiter ist hierbei gegenwärtig die Arbeitsmarktforschung, die für eine angemessene Analyse der wechselseitigen Wirkung von Angebots- und Nachfragefaktoren sowohl Informationen über Betriebe wie auch über Individuen benötigt und diesen Bedarf entsprechend artikuliert. In Deutschland wird die Verknüpfung von Daten aus verschiedensten Quellen vor allem vom IAB vorangetrieben. Ein Beispiel hierfür ist das Employer-Employee-File, bei welchem Paneldaten für mehrere tausend Betriebe (Betriebspanel) mit den Daten der Arbeitskräfte aus der Beschäftigtenstatistik auf Mikroebene verknüpft werden (Bellmann et al. 2002). Die Generierung weiterer Datenfiles mittels integrierter Registerdaten, die dann etwa für die Evaluation arbeitsmarktpolitischer Maßnahmen herangezogen werden können, wird gegenwärtig überprüft. Es gibt eine Vielzahl weiterer Bereiche (z.B. Bildung oder Gesundheit), in welchen Datenbasen aus unterschiedlichen amtlichen Quellen vorliegen. Es wäre zu prüfen, inwieweit die vorliegenden amtlichen Erhebungen z.B. mit Registerdaten ergänzt werden könnten. Die Verknüpfung solcher Daten auf Mikroebene würde völlig neue Forschungsperspektiven eröffnen. Aus Sicht der Forschung wäre es daher ein enormer Fortschritt, wenn diese bislang brachliegenden Datenpotenziale systematisch verfügbar gemacht werden würden. In gleicher Weise sind jedoch auch die Interessen der datengebenden Personen oder Organisationen zu berücksichtigen. Eine unabdingbare Voraussetzung für jede

Zusammenführung von Datenbeständen auf Mikroebene ist die Wahrung des Datenschutzes und des Rechts auf informationelle Selbstbestimmung: Die Anonymität der „Datengeber" muss in jedem Fall gesichert sein. Zu denken wäre hier etwa an die Einrichtung von Treuhandstellen oder Ähnlichem. Ein Weg in diese Richtung ist auch die kürzlich vom Gründungsauschuss des Rates für Sozial- und Wirtschaftsdaten ausgesprochene Empfehlung zur gesetzlichen Verankerung eines gesetzlichen Forschungsgeheimnisses. Dieses zielt primär darauf ab, Forschungsdaten vor einer Beschlagnahmung zu schützen und gleichzeitig die Forschung mit einem Zeugnisverweigerungsrecht auszustatten. Im Gegenzug müssten Forscher bei einem Verstoß gegen den Datenschutz mit empfindlichen Strafen[42] rechnen. In Bezug auf amtliche Daten würde ein solches Forschungsdatengeheimnis dann einen Fortschritt darstellen, wenn es dazu beitragen würde, den Zugang zu diesen Daten und insbesondere die Verknüpfung von unterschiedlichen Datenquellen in wesentlicher Form zu erleichtern.

Die Erschließung dieser nur angedeuteten zusätzlichen Potenziale amtlicher Daten wird sicherlich nicht in unmittelbarer Zukunft realisiert werden. Neben der Klärung von rechtlichen Randbedingungen und der Finanzierungsfrage sind hier auch umfangreiche methodische Arbeiten zu leisten. Insbesondere bei der Erschließung von Registerdaten, die bekanntermaßen eine Reihe von Fallstricken aufweisen, muss die Datenqualität sorgfältig überprüft und dokumentiert werden. Dies kann auch beinhalten, dass die Inhalte der Register überdacht werden. Sowohl in Bezug auf die anstehende Methodenforschung wie auch in Hinblick auf inhaltliche Überlegungen wäre eine fortgesetzte Intensivierung der Zusammenarbeit zwischen Statistik und Forschung sicher ein großer Gewinn.

Literatur

Alba, Richard, Johann Handl und *Walter Müller*, 1994a: Ethnische Ungleichheit im deutschen Bildungssystem. Kölner Zeitschrift für Soziologie und Sozialpsychologie 46: 209–237.

Alba, Richard, Walter Müller und *Bernhard Schimpl-Neimanns*, 1994b: Secondary Analysis of Official Microdata. S. 55–78 in: *Ingwer Borg* und *Peter Ph. Mohler* (Hg.), Trends and Perspectives in Empirical Social Research. New York: de Gruyter.

Ammermüller, Andreas, und *Andrea Weber*, 2003: Education and Wage Inequality in Germany – A Review of the Empirical Literatur. ZEW Discussion Paper No. 03–29, Mannheim.

Atkinson, A.B., 1997: Bringing Income Distribution in from the Cold. Economic Journal 107: 297–321.

Baltagi, B. H., und *U. Blien*, 1998: The German Wage Curve: Evidence from the IAB Employment Sample. Economics Letters 61: 135–142.

Baltagi, B. H., U. Blien und *K. Wolf*, 2000: The East German Wage Curve: 1993–1998. Economics Letters 69: 25–31.

42 Wobei eine Verletzung des Statistikgeheimnisses von Seiten der Forschung auch schon in der Gegenwart mit einer Geld- oder Freiheitsstrafe bis zu zwei Jahren geahndet werden kann (Statistisches Bundesamt 1988: 25). Diese Sanktionsmöglichkeit ergibt sich aus der in § 16 (6) BStatG enthaltenen Eingrenzung des Empfängerkreises von faktisch anonymen Daten auf Amtsträger und für den öffentlichen Dienst besonders Verpflichtete. Diese Einschränkung der potenziellen Datenempfänger wurde damals explizit mit dem Ziel eingeführt, bei einer „unbefugten Offenbarung" eine strafrechtliche Belangung zu ermöglichen (Deutscher Bundestag 1986: 21f.).

Bauer, T., P. Pereira, M. Vogler und *K. F. Zimmermann*, 1998: Portuguese Migrants in the German Labor Market: Performance and Self-Selection. IZA Discussion Paper No. 20, Bonn.

Baumann, Carsten, 1999: Self-employment as an Indicator of Assimilation? Evidence from Foreign Entrepreneurs in Germany. S. 235–270 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6. Mannheim: ZUMA.

Becker, Irene, 2000: Niedrige Lohneinkommen: Ein Indiz für prekäre Einkommenssituation im Haushaltskontext? (II). S. 1–16 in: *Jürgen Schupp* und *Heike Solga,* Niedrig entlohnt = niedrig qualifiziert? Chancen und Risiken eines Niedriglohnsektors in Deutschland. Dokumentation der Tagungsbeiträge auf CD-ROM. Berlin: DIW/MPIfB.

Bellmann, Lutz, Stefan Bender und *Arnd Kölling*, 2002: Der Linked Employer-Employee-Datensatz aus dem IAB-Betriebspanel und Beschäftigtenstatistik der Bundesanstalt für Arbeit. S. 21–30 in: *Gerhard Kleinhenz* (Hg.), IAB-Kompendium Arbeitsmarkt- und Berufsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung 250.

Bender, Stefan, Annette Haas und *Christoph Klose*, 1999: Die IAB-Beschäftigtenstichprobe 1975–1995. ZA-Information 45: 104–115.

Bender, Stefan, Dirk Konietzka und *Peter Sopp*, 2000: Diskontinuität im Erwerbsverlauf und betrieblicher Kontext. Kölner Zeitschrift für Soziologie und Sozialpsychologie 52: 475–499.

Bender, Stefan, und *Annette Haas*, 2002: Die IAB-Beschäftigtenstichprobe. S. 3–12 in: *Gerhard Kleinhenz* (Hg.), IAB-Kompendium Arbeitsmarkt- und Berufsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung 250.

Blossfeld, Hans-Peter, 1983: Höherqualifizierung und Verdrängung – Konsequenzen der Bildungsexpansion in den Siebziger Jahren. S. 184–240 in: *Max Haller* und *Walter Müller* (Hg.), Beschäftigungssystem im gesellschaftlichen Wandel. Frankfurt a.M.: Campus.

Blossfeld, Hans-Peter, 1985: Bildungsexpansion und Berufschancen. Frankfurt a.M./New York: Campus.

Bolte, Karl Martin, und *Dieter Kappe*, 1964: Struktur und Entwicklung der Bevölkerung. Beiträge zur Wirtschafts- und Sozialkunde. Reihe B. Band 2. Opladen: Leske + Budrich.

Bolte, Karl Martin, und *Katrin Aschenbrenner*, 1964: Die gesellschaftliche Situation der Gegenwart: ausgewiesene Eigenarten der heutigen Gesellschaftsstruktur und der Stellung des Menschen in der Gesellschaft. Beiträge zur Wirtschafts- und Sozialkunde. Reihe B. Band 1. Opladen: Leske + Budrich.

Brauns, Hildegard, Markus Gangl und *Stefani Scherer*, 2003: Education and Unemployment Risks among Market Entrants: A Comparison of France, the United Kingdom and West Germany. S. 328–345 in: *Jürgen H. P. Hoffmeyer-Zlotnik* und *Christof Wolf* (Hg.), Advances in Cross-National Comparison. An European Working Book for Demographic and Socio-Economic Variables. New York.

Brauns, Hildegard, Susanne Steinmann und *Dietmar Haun*, 2000: Die Konstruktion des Klassenschemas nach Erikson, Goldthorpe und Portocarero (EGP) am Beispiel nationaler Datenquellen aus Deutschland, Großbritannien und Frankreich. ZUMA-Nachrichten 46: 7–63.

Brauns, Hildegard, Susanne Steinmann, Annick Kieffer und *Catherine Marry*, 1999: Does Education Matter? France and Germany in Comparative Perspective. European Sociological Review 15: 61–89.

Brennecke, Ralph, 1975: Die Konstruktion von sozioökonomischen Großsystemen. Verknüpfung von Modellen zur Analyse wirtschaftspolitischer Prozesse. Frankfurt a.M.: Campus.

Bulmahn, Thomas, 1999: Altersvorsorge im Wandel. S. 300–322 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6. Mannheim: ZUMA.

Bundesregierung, 2001a: Lebenslagen in Deutschland. Der erste Armuts- und Reichtumsbericht der Bundesregierung.

Bundesregierung, 2001b: Lebenslagen in Deutschland – Daten und Fakten. Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung.

Butz, Marcus, 2001: Lohnt sich Bildung noch? Ein Vergleich der bildungsspezifischen Nettoeinkommen 1982 und 1995. S. 95–117 in: *Peter A. Berger* und *Dirk Konietzka* (Hg.), Die Erwerbsgesellschaft. Neue Ungleichheiten und Unsicherheiten. Opladen: Leske + Budrich.

Claessens, Dieter, Arno Klönne und *Armin Tschoepe*, 1965: Sozialkunde der Bundesrepublik Deutschland. Düsseldorf: Diederichs.

Dathe, Dietmar, 1999: Familienlebenszyklus und Erwerbsbeteiligung der Ehepartner in Ehen mit Kindern in West- und Ostdeutschland. S. 149–170 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6. Mannheim: ZUMA.

Dell, Fabien, 2003: Top Incomes in Germany over the Twentieth Century: 1891–1995. Mimeo. http://pythie.cepremap.ens.fr/~piketty/Papers/Dell2003.pdf.

Deutscher Bundestag, 1986: Gesetzentwurf der Bundesregierung. Entwurf eines Gesetzes über die Statistik für Bundeszwecke. Drucksache 10/5345.

Diekmann, Andreas, 1998: Empirische Sozialforschung. Reinbek: Rohwolt.

Dinkel, Reiner H., Marc Luy und *Uwe Lebok*, 1999: Die Bildungsbeteiligung deutscher und ausländischer Jugendlicher in der Bundesrepublik. S. 354–375 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6. Mannheim: ZUMA.

Dustmann, C., und *C. Meghir*, 1999: Wages, Experience und Seniority. CEPR Discussion Paper 2077.

Engstler, Heribert, 1998: Die Familie im Spiegel der amtlichen Statistik. Herausgegeben vom Bundesministerium für Familie, Senioren, Frauen und Jugend. Bonn.

Esser, Hartmut, Heinz Grohmann, Walter Müller und *Karl-August Schäffer*, 1989: Mikrozensus im Wandel. Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung. Forum der Bundesstatistik 11. Stuttgart: Metzler-Poeschel.

Euwals, Rob, und *Rainer Winkelmann*, 2001: Why do Firms Train? – Empirical Evidence on the First Labour Market Outcomes of Graduated Apprentices. IZA Discussion Paper No. 319, Bonn.

Fahrmeir, Ludwig, Stefan Lang, Joachim Wolff und *Stefan Bender*, 2000: Semiparametric Bayesian Time-Space Analysis of Unemployment Duration. SFB 386, München: Discussionpaper 211.

Fitzenberger, Bernd, 1999a: Wages and Employment Across Skill Groups – An Analysis for West Germany. ZEW Economic Studies, Band 6.

Fitzenberger, Bernd, 1999b: Heterogenous Labor in a Structural Empirical Model of Wage Bargaining in West Germany. Dresdner Beiträge zur Volkswirtschaftslehre. Nr. 6/99. Technische Universität Dresden. Fakultät Wirtschaftswissenschaften.

Fitzenberger, Bernd, Reinhold Schnabel und *Gaby Wunderlich*, 2004: The Gender Gap in Labor Market Participation and Employment: A Cohort Analysis for West Germany. Journal of Population Economics 17: 83–116.

Fitzenberger, Bernd, und *Gaby Wunderlich*, 2002: The Changing Life Cycle Pattern in Female Employment: A Comparison of Germany and the UK. ZEW Discussion Paper No. 02–70, Mannheim.

Fitzenberger, Bernd, und *Stefan Speckesser*, 2000: Zur wissenschaftlichen Evaluation der Aktiven Arbeitsmarktpolitik in Deutschland: Ein Überblick. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 33: 357–371.

Freeman, Richard, und *Ronald Schettkat*, 2000: Skill Compression, Wage Differentials and Employment: Germany vs. the US. NBER Working Paper No. 7610. http://www.nber.org/papers/w7610.

Galler, H. P., und *A. Steger*, 1978: Mikroanalytische Bevölkerungssimulation als Grundlage sozialpolitischer Entscheidungen. S. 237–275 in: *Hans-Jürgen Krupp* und *Wolfgang Glatzer* (Hg.), Umverteilung im Sozialstaat. Empirische Einkommensanalysen für die Bundesrepublik. Frankfurt a.M.: Campus.

Gangl, Markus, 2003: Unemployment Dynamics in the United States and West Germany. Heidelberg: Pysica/Springer.

Geiger, Theodor, 1932: Die Soziale Schichtung des deutschen Volkes. Soziographischer Versuch auf statistischer Grundlage. Stuttgart: Enke.

Goldthorpe, John H., 2003: Globalisierung und soziale Klasse. In: *Walter Müller* und *Stefani Scherer* (Hg.), Mehr Risiken – Mehr Ungleichheit? Abbau von Wohlfahrtsstaat, Flexibilisierung von Arbeit und die Folgen. Frankfurt a.M.: Campus.

Goyder, J., 1987: The Silent Minority. Nonrespondents on Sample Surveys. Cambridge: Polity Press.

Granato, Nadia, 2003: Ethnische Ungleichheit auf dem deutschen Arbeitsmarkt. Schriftenreihe des Bundesinstituts für Bevölkerungsforschung, Band 33. Opladen: Leske + Budrich.

Granato, Nadia, und *Frank Kalter*, 2001: Die Persistenz ethnischer Ungleichheit auf dem deutschen Arbeitsmarkt. Diskriminierung oder Unterinvestition in Humankapital? Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 497–520.

Groß, Martin, 2001: Auswirkungen des Wandels der Erwerbsgesellschaft auf soziale Ungleichheit. Effekte befristeter Beschäftigung auf Einkommensungleichheit. S. 119–155 in: *Peter A. Berger* und *Dirk Konietzka* (Hg.), Die Erwerbsgesellschaft. Neue Ungleichheiten und Unsicherheiten. Opladen: Leske + Budrich.

Haas, Anette, 2001: Die IAB-Regionalstichprobe 1975–1997. ZA-Information 48: 128–141.

Haas, Joachim, und *Burkart Lutz,* 1999: L'Expansion de L'Education En Allemagne de L'Ouest. Projektbericht „Educational Expansion and Labour Market". Zentrum für Sozialforschung Halle e.V., Halle.

Hahlen, Johann, 2002: Die internationale Vergleichbarkeit von Statistiken – Allgemeine Aspekte und Beispiele aus dem Bereich der Haushaltsstichproben. S. 353–373 in: *Wolfgang Glatzer, Roland Habich* und *Karl Ulrich Mayer* (Hg.), Sozialer Wandel und gesellschaftliche Dauerbeobachtung. Opladen: Leske + Budrich.

Haisken-DeNew, John P., und *Christoph M. Schmidt*, 1999: Industry Wage Differentials Revisited: A Longitudinal Comparison of Germany and USA (1984–1996). IZA Discussion Paper No. 98. Bonn: IZA.

Hakim, Catherine, 2000: Research Design: Successful Designs for Social and Economic Research. London: Routledge.

Haller, Max, 1989: Klassenstrukturen und Mobilität in fortgeschrittenen Gesellschaften. Eine vergleichende Analyse der Bundesrepublik Deutschland, Österreichs, Frankreichs und den Vereinigten Staaten von Amerika. Frankfurt a.M.: Campus.

Haller, Max, und *Walter Müller,* 1983: Beschäftigungssystem im gesellschaftlichen Wandel. Frankfurt a.M./New York: Campus.

Handl, Johann, 1978: Ausmaß und Determinanten der Erwerbsbeteiligung von Frauen. Beiträge zur Arbeitsmarkt- und Berufsforschung 31. Nürnberg: 189–256.

Handl, Johann, 1988: Berufschancen und Heiratsmuster von Frauen. Empirische Untersuchungen zu Prozessen sozialer Mobilität. Frankfurt a.M.: Campus.

Handl, Johann, 1996: Hat sich die berufliche Wertigkeit der Bildungsabschlüsse in den achtziger Jahren verringert? Kölner Zeitschrift für Soziologie und Sozialpsychologie 48: 249–273.

Handl, Johann, Karl Ulrich Mayer und *Walter Müller*, 1977: Klassenlagen und Sozialstruktur. Empirische Untersuchungen für die Bundesrepublik Deutschland. Frankfurt a.M.: Campus.

Hartmann, Peter H., und *Bernhard Schimpl-Neimanns*, 1992: Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. Kölner Zeitschrift für Soziologie und Sozialpsychologie 44: 315–340.

Hauser, Richard, 2003: Sozialstaat und amtliche Statistik in der sozialpolitischen Forschung – der mikroanalytische Ansatz. S. 55–64 in: *Heiner Meulemann* (Hg.), Der Sozialstaat in der amtlichen Statistik. Sozialwissenschaftliche Tagungsberichte Band 6. Bonn: IZ.

Hauser, Richard, und *Holger Stein*, 2001: Die Vermögensverteilung und die Reform der gesetzlichen Rentenversicherung in der Bundesrepublik Deutschland. Arbeitspapier Nr. 27. Johann-Wolfgang Goethe-Universität. Frankfurt a.M.

Hauser, Richard, und *Irene Becker*, 2001a: Einkommensverteilung in Querschnitt und im Zeitverlauf 1973–1988. Lebenslagen in Deutschland. Der erste Armuts- und Reichtumsbericht der Bundesregierung. Bonn: Bundesministerium für Arbeit und Sozialordnung.

Hauser, Richard, und *Irene Becker*, 2001b: Lohnsubvention und verbesserter Familienlastenausgleich als Instrumente zur Verringerung der Sozialhilfeabhängigkeit. S. 293–312 in: *Hans-Christian Mager, Henry Schäfer* und *Klaus Schrüfer* (Hg.), Private Versicherung und Soziale Sicherung. Marburg: Metropolis.

Heinz, Wolfgang, 2004: Soziale und kulturelle Grundlagen der Kriminologie – der Beitrag der Kriminalstatistik. Wirtschaft und Statistik 2004(3): 346–364.

Helberger, Christof, 1982: Auswirkungen öffentlicher Bildungsausgaben in der Bundesrepublik Deutschland auf die Einkommensverteilung der Ausbildungsgeneration. Stuttgart: Kohlhammer.

Helmert, Uwe, und *Dieter Borgers*, 1998: Rauchen und Beruf. Eine Analyse von 100.000 Befragten des Mikrozensus 1995. Bundesgesundheitsblatt 41: 102–107.

Hoffmeyer-Zlotnik, Jürgen, und *Christof Wolf*, 2003: Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer Academic/Plenum Publisher.

Hyman, Herbert H., 1972: Secondary Analysis with Sample Surveys: Principles, Procedures and Potentialities. New York: Wiley.

Johannes Huinink, 2002: Polarisierung der Familienentwicklung in europäischen Ländern im Vergleich. Zeitschrift für Familienforschung. Sonderheft 2: 49–74.

Jungbauer-Gans, Monika, 1999: Der Lohnunterschied zwischen Frauen und Männern in selbständiger und abhängiger Beschäftigung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51: 364–390.

Kaase, Max, Hans-Jürgen Krupp, Manfred Pflanz, Erwin K. Scheuch und *Spiros Simitis* (Hg.), 1980: Datenzugang und Datenschutz. Königstein/Ts.: Athenäum.

Kalter, Frank, 2001: Die Kontrolle von Drittvariablen bei der Messung von Segregation. Ein Vorschlag am Beispiel der familialen Assimilation von Migranten. Zeitschrift für Soziologie 30: 452–464.

Kalter, Frank, und *Nadia Granato,* 2002: Demographic Change, Educational Expansion und Structural Assimilation of Immigrants. The Case of Germany. European Sociological Review 18: 199–216.

Kiecolt, K. Jill, und *Laura E. Nathan*, 1985: Secondary Analysis of Survey Data. Beverly Hills: Sage.

Klanberg, Frank, 1978: Armut und ökonomische Ungleichheit in der Bundesrepublik Deutschland. Frankfurt a.M.: Campus.

Koch, Achim, 1997: Teilnahmeverhalten beim ALLBUS 1994. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 98–122.

Koch, Achim, 1998: Wenn ‚mehr' nicht gleichbedeutend mit ‚besser' ist. Ausschöpfungsquoten und Stichprobenverzerrungen in allgemeinen Bevölkerungsumfragen. ZUMA-Nachrichten 49: 98–122.

Koch, Iris, und *Holger Meinken*, 2003: Das Beschäftigtenpanel der Bundesanstalt für Arbeit. ZA-Information 52: 159–170.

Kogan, Irena, und *Walter Müller* (Hg.), 2003: School-to-Work Transitions in Europe: Analyses of the EU LFS 2000 Ad Hoc Module. Mannheim: MZES.

Kohlmann, Annette, und *Holger Alda*, 2004: Datenzugang und Datenverfügbarkeit im FDZ der BA im IAB. Vortrag auf der 2. Konferenz für Wirtschafts- und Sozialdaten, Forum „Bildung und Arbeit", Wiesbaden, 4.–5. Juni 2004.

Konietzka, Dirk, 2002: Die soziale Differenzierung der Übergangsmuster in den Beruf. Die „zweite Schwelle" im Vergleich der Berufseinstiegskohorten 1976–1995. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 645–673.

Konietzka, Dirk, und *Michaela Kreyenfeld*, 2001: Non-Marital Births in East-Germany after Unification. Working Paper 2001–027. Rostock: Max-Planck-Institut für demografische Forschung.

Kortmann, Klaus, 1978: Die Generierung einer geschlossenen Mikrodatenbasis für die Bundesrepublik Deutschland. S. 193–235 in: *Hans-Jürgen Krupp* und *Wolfgang Glatzer* (Hg.), Umverteilung im Sozialstaat. Empirische Einkommensanalysen für die Bundesrepublik. Frankfurt a.M.: Campus.

Kreyenfeld, Michaela, 2004: Sozialstruktur und Kinderbetreuung. MPIDR Working Paper 2004–009.

Kristen, Cornelia, und *Nadia Granato,* 2004: Bildungsinvestitionen in Migrantenfamilien. IMIS-Beiträge 2004(23): 123–142.

Krupp, Hans-Jürgen, 1968: Theorie der personellen Einkommensverteilung: allgemeine Grundzüge und verteilungspolitische Simulationen. Volkswirtschaftliche Schriften 127. Berlin: Duncker & Humblot.

Krupp, Hans-Jürgen, 2002: Sozialberichterstattung und informationelle Infrastruktur. S. 391–403 in: *Wolfgang Glatzer, Roland Habich* und *Karl Ulrich Mayer* (Hg.), Sozialer Wandel und gesellschaftliche Dauerbeobachtung. Opladen: Leske + Budrich.

Krupp, Hans-Jürgen, und *Wolfgang Zapf*, 1977: Sozialpolitik und Sozialberichterstattung. Frankfurt a.M.: Campus.

Kühnen, Carola, 1999: Das Stichprobenverfahren der Einkommens- und Verbrauchsstichprobe 1998. Wirtschaft und Statistik 2: 111–115.

KVI (Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik) (Hg.), 2001: Wege zu einer besseren informationellen Infrastruktur. Baden-Baden: Nomos.

Lohmann, Henning, und *Silvia Luber*, 2000: Die Folge beruflicher Segregation für weibliche und männliche Selbständigkeit: Ein Vergleich zwischen der Bundesrepublik und dem Vereinigten Königreich. 2. Mikrozensus – Nutzerkonferenz: „Forschung mit dem Mikrozensus. Analysen zur Sozialstruktur und zum Arbeitsmarkt" am 12.–13. Oktober 2000 in Mannheim.

Lotze, Sabine, und *Holger Breiholz*, 2002: Zum neuen Erhebungsdesign des Mikrozensus. Teil 1. Wirtschaft und Statistik 5: 359–366.

Luber, Silvia, 2003: Berufliche Selbständigkeit im Wandel. Ein empirischer Vergleich der sozialen und wirtschaftlichen Struktur Selbständiger in Deutschland und Großbritannien. Frankfurt a.M.: Peter Lang.

Lüttinger, Paul, 1989: Integration der Vertriebenen. Frankfurt a.M./New York: Campus.

Lüttinger, Paul, Bernhard Schimpl-Neimanns, Heike Wirth und *Georg Papastefanou*, 2003b: Mikrodaten (German Microdata Lab): Das Servicezentrum für amtliche Mikrodaten bei ZUMA. ZUMA-Nachrichten 52: 153–172.

Lüttinger, Paul, Holger Breiholz und *Yvonne Lechert*, 2003a: Ergebnisse der zweiten Befragung von Nutzern der Mikrozensus-Scientific-Use-Files. ZUMA-Nachrichten 53: 128–135.

Lüttinger, Paul, und *Heike Wirth*, 2004: Zur Weitergabe von älteren amtlichen Mikrodaten an die Wissenschaft. Allgemeines Statistisches Archiv 4: 473–486.

Lüttinger, Paul, und *Thomas Riede*, 1997: Der Mikrozensus. Amtliche Daten für die Sozialforschung. ZUMA-Nachrichten Nr. 41: 19–44.

Machin, Stephen, und *Patrick Puhani*, 2003: Subject of Degree and the Gender Wage Differential: Evidence from the UK and Germany. Economic Letters 79: 393–400.

Mavromaras, Kostas G., und *Helmut Rudolph*, 1997: Wage Discrimination in the Re-Employment Process. Journal of Human Resources 32: 812–860.

Mavromaras, Kostas G., und *Helmut Rudolph*, 1998: Verteilungseffekte der Arbeitslosenversicherung in Deutschland in den 80er Jahren. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 31: 93–107.

Mayer, Karl Ulrich, 1977a: Statushierarchie und Heiratsmarkt. S. 155–232 in: *Johann Handl, Karl Ulrich Mayer* und *Walter Müller* (Hg.), Klassenlagen und Sozialstruktur. Frankfurt a.M.: Campus.

Mayer, Karl Ulrich, 1977b: Ungleiche Chancen und Klassenbildung. Soziale Welt 28: 466–493.

Mertens, Antje, 1998: Labor Mobility and Wage Dynamics – An Empirical Study for Germany in Comparison with the United States. Berlin/Aachen: Shaker Verlag.

Müller, Walter, 1998: Daten der amtlichen Statistik als Basis und Eckdaten für die sozialwissenschaftliche Forschung. S. 151–170 in: *Werner Haug, Klaus Armingeon, Peter Farago* und *Markus Zürcher* (Hg.), Statistik im Dienste der Öffentlichkeit. Bern: Bundesamt für Statistik.

Müller, Walter, 1999a: Der Mikrozensus als Datenquelle sozialwissenschaftlicher Forschung. S. 7–27 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten. Spezial Band 6. Mannheim: ZUMA.

Müller, Walter, 1999b: Amtliche Statistik und empirische Forschung: Wege in eine kooperative Zukunft. S. 17–35 in: *Statistisches Bundesamt* (Hg.), Kooperation zwischen Wissenschaft und amtlicher Statistik. Praxis und Perspektiven. Schriftenreihe Forum der Bundesstatistik, 34. Stuttgart: Metzler-Poeschel.

Müller, Walter, 2001: Konsequenzen für die Methodenausbildung aus dem Gutachten der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI). ZUMA-Nachrichten 49: 81–99.

M*üller, Walter, Angelika Willms-Herget* und *Johann Handl*, 1983: Strukturwandel der Frauenarbeit 1880–1980. Frankfurt a.M.: Campus.

Müller, Walter, Hildegard Brauns und *Susanne Steinmann*, 2000: Expansion und Erträge tertiärer Bildung in Deutschland, Frankreich und im Vereinigten Königreich. Berliner Journal für Soziologie 12: 37–62.

Müller, Walter, und *Markus Gangl* (Hg.), 2003: Transitions from Education to Work in Europe: the Integration of Youth on EU Labour Markets. Oxford: Oxford University Press.

Müller, Walter, und *Richard Hauser, R.*, 1987: Der Bedarf der Wissenschaft an anonymisierten Daten. S. 61–78 in: *Statistisches Bundesamt* (Hg.), Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten. Wiesbaden: W. Kohlhammer.

Müller, Walter, Uwe Blien, Peter Knoche, Heike Wirth et al., 1991: Die faktische Anonymität von Mikrodaten. Forum der Bundesstatistik 19. Stuttgart: Metzler-Poeschel.

Münnich, Margot, 2000: Einkommens- und Geldvermögensverteilung privater Haushalte in Deutschland – Teil 1. Ergebnis der Einkommens- und Verbrauchsstichprobe 1998. Wirtschaft und Statistik 9: 679–691.

Örtel, Martina, Jürgen Passenberger und *Markus Janser*, 2004: Forschungsdatenzentrum der BA geht an den Start. IAB-Kurzbericht 2004/9.

Peisert, Hansgert, 1967: Soziale Lage und Bildungschancen in Deutschland. München: Piper.

Pfeiffer, Friedhelm, 1999a: Tätigkeitsspektrum und Verdienst von Ingenieuren und Naturwissenschaftlern. S. 74–92 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalyse mit dem Mikrozensus. ZUMA-Nachrichten Spezial Band 6. Mannheim: ZUMA.

Pfeiffer, Friedhelm, 1999b: Existenzgründerpotenziale unter Arbeitssuchenden: Empirische Evidenz auf der Basis des Mikrozensus. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 32(3): 300–314.

Pfeiffer, Friedhelm, und *Joachim Brade*, 1995: Weiterbildung, Arbeitszeit und Lohneinkommen. S. 289–326 in: *Viktor Steiner* und *Lutz Bellmann* (Hg.), Mikroökonomik des Arbeitsmarktes. Beiträge zur Arbeitsmarkt- und Berufsforschung 192. Nürnberg: IAB.

Puhani, Patrick, 2003: A Test of the ‚Krugman Hypothesis' for the United States, Britain, and Western Germany. Bonn: IZA Discussion Paper No. 764.

Rehfeld, Uwe, 2004: Datenangebot und Informationsbedarf im Bereich der Alterssicherung – Forschungsdatenzentrum der Rentenversicherung im Aufbau. Deutsche Rentenversicherung 1–2: 63–75.

Roloff, Juliane, 1999: Die Einkommenslage deutscher und ausländischer Ehen mit Kindern in Westdeutschland 1997. Zeitschrift für Bevölkerungswissenschaft 24: 281–300.

Ronning, Gerd, und *Roland Gnoss*, 2003: Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik 42. Wiesbaden: Statistisches Bundesamt.

Schettkatt, Ronald, 2003: Differences in US-German Time-Allocation. Why do Americans Work Longer Hours than Germans? Discussion Papers FS I 02–212. Berlin: Wissenschaftszentrum.

Schimpl-Neimanns, Bernhard, 1998: Analysemöglichkeiten des Mikrozensus. ZUMA-Nachrichten Nr. 42: 91–119.

Schimpl-Neimanns, Bernhard, 2000: Soziale Herkunft und Bildungsbeteiligung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 52: 636–669.

Schimpl-Neimanns, Bernhard, 2002: Anwendungen und Erfahrungen mit dem Scientific-Use-File des Mikrozensus. ZUMA-Arbeitsbericht 2002/01. Mannheim: ZUMA.

Schneekloth, Ulrich, und *Ingo Leven*, 2003: Woran bemisst sich eine ‚gute' allgemeine Bevölkerungsumfrage? ZUMA-Nachrichten 53: 16–57.

Schnell, Rainer, 1997: Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Opladen: Leske + Budrich.

Statistisches Bundesamt, 1988: Das Arbeitsgebiet der Bundesstatistik 1988. Stuttgart: Kohlhammer.

Statistisches Bundesamt (Hg.), 1985: Datennotstand und Datenschutz. Die amtliche Statistik nach dem Volkszählungsurteil. Wiesbaden: W. Kohlhammer.

Statistisches Bundesamt (Hg.), 1987: Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten. Wiesbaden: W. Kohlhammer.

Stauder, Johannes, und *Wolfgang Hüning*, 2003: Die Messung von Äquivalenzeinkommen und Armutsquoten auf der Basis des Mikrozensus. Beitrag zur 3. Mikrozensusnutzerkonferenz. Mannheim: ZUMA.

Stein, Holger, 2001: Die Entwicklung des aggregierten Privatvermögens und seine Verteilung in Deutschland seit 1970. Arbeitspapier Nr. 30. Frankfurt a.M.: Johann-Wolfgang Goethe-Universität.

Steiner, Viktor, und *Kirsten Wagner*, 1998: Has Earnings Inequality in Germany Changed in the 1980's? Zeitschrift für Wirtschafts- und Sozialwissenschaften 118: 29–60.

Stockmann, Reinhard, 1986: Gesellschaftliche Modernisierung und Betriebsstruktur. Frankfurt a.M.: Campus.

Strengmann-Kuhn, Wolfgang, 2000: Niedrige Lohneinkommen: Ein Indiz für prekäre Einkommenssituation im Haushaltskontext? (I). S. 1–12 in: *Jürgen Schupp* und *Heike Solga*, Niedrig entlohnt = niedrig qualifiziert? Chancen und Risiken eines Niedriglohnsektors in Deutschland. Dokumentation der Tagungsbeiträge auf CD-ROM. Berlin: DIW/MPIfB.

Voges, Wolfgang, 2002: Pflege alter Menschen als Beruf. Wiesbaden: Westdeutscher Verlag.

Wagner, Gert, 2000: Selbstorganisation des Wissenschaftssystems würde Datenschutz vereinfachen und Re-Analysen befördern. ZUMA-Nachrichten 47: 75–88.

Wagner, Joachim, 1999: Nutzung von betrieblichen Einzeldaten au*s der* amtlichen Statistik durch externe Wissenschaftler. S. 9–17 in: *Statistisches Bundesamt* (Hg.), Möglichkeiten einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik. Stuttgart: Metzler Poeschel.

Warschburger, Sabine, und *Michael Lohre,* 1998: Wie verlässlich sind die Verbrauchsangaben im Mikrozensus? Eine Untersuchung alternativer Quellen zum deutschen Rauchverhalten. Discussion Papers in Economics Nr. 98-10. Universität Dortmund.

Weick, Stefan, 2001: Bei höherer Schulbildung neigen ausländische Männer eher zur Ehe mit deutscher Partnerin. Informationsdienst Soziale Indikatoren 25: 12–14.

Willms-Herget, Angelika, 1985: Frauenarbeit. Zur Integration der Frauen in den Arbeitsmarkt. Frankfurt a.M.: Campus.

Wirth, Heike, 1996: Wer heiratet wen? Die Entwicklung der bildungsspezifischen Heiratsmuster in Westdeutschland. Zeitschrift für Soziologie 25(5): 371–394.

Wirth, Heike, 2000: Bildung, Klassenlage und Partnerwahl. Eine empirische Analyse zum Wandel der bildungs- und klassenspezifischen Heiratsbeziehungen. Opladen: Leske + Budrich.

Wirth, Heike, und *Silvia Zühlke*, 2004: Mikrozensus-Regionalfile. Faktische Anonymisierung und Datenzugang. Vortrag. Workshop Deutschland regional. Sozialwissenschaftliche Daten im Forschungsverbund 09.07.–10.07.2004. Berlin: DIW.

Wolf, Elke, 1999: Dynamik der Arbeitszeitstruktur. Welche Rolle spielt der Strukturwandel? S. 119–148 in: *Paul Lüttinger* (Hg.), Sozialstrukturanalysen mit dem Mikrozensus. ZUMA-Nachrichten Spezial Nr. 6. Mannheim: ZUMA.

Zühlke, Silvia, Markus Zwick, Sebastian Scharnhorst und *Thomas Wende*, 2003: Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. Wirtschaft und Statistik 10: 906–911.

Zwick, Markus, 1999: Steuerstatistische Einzeldaten und ihre Auswertungsmöglichkeiten für die Wissenschaft. Allgemeines Statistisches Archiv 83: 248–253.

Anhang

Übersicht 1: Übersicht amtliche Mikrodaten, die gegenwärtig als Scientific-Use-Files, via Ferndatenanalyse oder Gastaufenthalt in den Forschungsdatenzentren für Forschungsvorhaben genutzt werden können (Stand: Juli 2004)

Scientific-Use-Files: ***Nutzung beschränkt auf Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger Forschung bzw. den dort beschäftigten Wissenschaftlern; bislang werden diese Daten nicht ins Ausland weitergeben.[a]***	
Mikrozensus[1a, 1b] *Verfügbar:* 1989; 1991; 1993; 1995; 1996; 1997; 1998; 1999; 2000 *In Vorbereitung*: – 2001; 2002ff.; Panelfile 1996 – 1999; Regionalfile; Rückwärtserschließung 1987; 1985ff.	
Einkommens- und Verbrauchsstichprobe[3] *Verfügbar:* 1993; 1998; 2003. – *In Vorbereitung:* 1988; 1983; 1978; 1973; 1969; 1962/63	
Zeitbudgeterhebung[1a, 2] *Verfügbar*: 1991/92. – *In Vorbereitung*: 2001/02	
Arbeitskräfteerhebung EU[2] (European Labour Force, Survey; ELFS) – In Deutschland werden die Angaben des ELFS als Unterstichprobe des Mikrozensus erhoben.	– Nutzung der deutschen Daten via Mikrozensus-Scientific-Use-File – Eurostat ermöglicht künftig Wissenschaftlern den Zugang zu den Daten aus allen teilnehmenden Ländern in ihrer Dienststelle oder der einzelstaatlichen Dienststelle. Anonymisierte Mikrodaten können auch für die wissenschaftliche Nutzung freigegeben werden. Beide Zugangsmöglichkeiten können nicht gewährt werden, wenn die jeweilige einzelstaatliche Stelle die Nutzung verweigert. Näheres dazu in der Verordnung Nr. 831/2002 der Kommission der Europäischen Gemeinschaften vom 17.5.2002. Die Bereitstellung von Scientific-Use-Files wird seit längerer Zeit angestrebt
Europäisches Haushaltspanel 1994–2001(ECHP)[2] – Deutsche Stichprobe 1994–1996, danach Weiterführung im Rahmen von SOEP	– Zugang zu den gesamteuropäischen Daten 1994–2001 – zu allerdings beträchtlichen Kosten – via Eurostat[8]
Erhebung über die berufliche Weiterbildung[2]	– Datenzugang bislang nur über Eurostat (siehe Arbeitskräfteerhebung)
Innovationserhebung[2]	– Datenzugang bislang nur über Eurostat (siehe Arbeitskräfteerhebung)
Niedrigeinkommenspanel (NIEP)[2]	
Lohn- und Einkommenssteuerstatistik 1998[1a]	– Regelmäßige Aktualisierung der Daten
Sozialhilfestatistik** 1997, 1998, 1999; 2000[1a, 2]	– Regelmäßige Aktualisierung der Daten
IAB-Beschäftigtenstichprobe (IABS)[2, 4] – Basisfile (1975–1995); Regionalfile (1975– 1997)	– Aktualisierung Regionalfile (1975–2002) – Aktualisierung Basisfile noch offen

BA-Beschäftigtenpanel[2, 4] 1998–2002	– Regelmäßige Aktualisierung der Daten
Erhebung zur Inanspruchnahme kultureller und sozialer Leistungen in Arbeiter- und Angestelltenhaushalten der DDR 1988[2a]	Diese amtlichen Bevölkerungserhebungen der DDR-Statistik sind für die Forschung via Bundesdatenarchiv als Scientific-Use-Files verfügbar
Einkommensstichprobe in Arbeiter- und Angestelltenhaushalten der DDR 1988[2a]	
Statistik des Haushaltsbudgets 1988/89 (DDR)[2a]	
Zeitbudgetstudien der DDR 1985, 1990	
Amtliche Mikrodaten, die von der Forschung gegenwärtig nur via Forschungsdatenzentren über Gastwissenschaftleraufenthalte oder Ferndatenanalyse genutzt werden können	
IAB-Betriebspanel[6]	
Linked Employer-Employee-Datei 1993–2001 (LIAB)[4]	
Gewerbesteuer-; Umsatzsteuer-, Körperschaftsstatistik[1a]	
Mittel- bis längerfristige geplante Datenerschließung, wobei die Art des Datenzugangs (via Scientific-Use-File oder Forschungsdatenzentren) noch offen ist	
Forschungsdatenzentrum der gesetzlichen Rentenversicherung (im Aufbau)[7]	Bereitstellung prozessproduzierter Mikrodaten zu – Rentenzugang/-wegfall – Rentenbestand – Anwartschaften – Versicherte – Rehabilitation
Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder[1a]	Das Datenangebot der Forschungsdatenzentren der amtlichen Statistik soll ständig erweitert werden. So werden zukünftig u.a. Mikrodaten der Einzelhandelsstatistik, der Strafverfolgungsstatistik, der Diagnosestatistik oder der Gehalts- und Lohnstrukturerhebung zur Verfügung stehen.
Forschungsdatenzentrum der Bundesagentur für Arbeit (im Aufbau)[4]	– Integrierte Erwerbsbiographien (IEB)[4] – Stellenangebotsstatistik (STEA)[5] – Arbeitssuchende/Arbeitslosenstatistik (AST)[5] – Maßnahmestatistik[5]

[a] Für einzelne Erhebungen (z.B. Mikrozensus, Sozialhilfestatistik, Zeitbudgeterhebungen) sind auch so genannte Campus-Files verfügbar, die in der Lehre eingesetzt werden dürfen und bei den Forschungsdatenzentren der Statistischen Ämter kostenlos downloadbar sind. [1a] Für Informationen siehe: http://www.forschungsdatennetzwerk.de/; http://www.destatis.de/fdz/ [1b] http://www.gesis.org/dauerbeobachtung/Mikrodaten/daten/index.htm [2] Für Informationen siehe: http://www.gesis.org/dauerbeobachtung/Mikrodaten/Informationsquellen/weitere_Mikrodaten.htm [2a] Für Informationen siehe: http://www.gesis.org/dauerbeobachtung/Mikrodaten/Daten/Abteilungsdaten/DDR/index.htm [3] Für Informationen vgl.: http://www.destatis.de/micro/d/micro_c2.htm; http://www.gesis.org/dauerbeobachtung/Einkommen/Wegweiser/serviceguide.htm [4] Vgl. Kohlmann und Ahlda (2004). [5] Vgl. Oertel et al. (2004). [6] http://betriebspanel.iab.de [7] http://fdz.iab.de (im Aufbau). [8] http://www.eu-datashop.de/service/EN/spezial/spezial.htm

* Die Sozialhilfestatistik wird in Form eines Public-Use-Files für jedermann zur Verfügung gestellt, d.h. diese Daten sind absolut anonymisiert.

GEWICHTUNGSPROBLEME IN DER DATENANALYSE

Siegfried Gabler

Zusammenfassung: In vielen Auswertungen empirischer Datensätze aus Umfragen wird das zugrunde liegende komplexe Stichprobenverfahren vernachlässigt und so getan, als wäre uneingeschränkt zufällig ausgewählt worden. Dies kann zu falschen Schlüssen führen. Neben den Punktschätzern werden insbesondere Varianzschätzungen und damit verbunden Konfidenzintervalle falsch berechnet. Die Verwendung von Gewichten erfordert allerdings Information über den Auswahlprozess und Annahmen über den Ausfallprozess. Die darauf beruhenden Design- und Anpassungsgewichte müssen für den Nutzer sinnvoll sein. Extremgewichte etwa sind zu vermeiden. Bei der Berücksichtigung des Auswahlverfahrens in multivariaten Modellen oder Tests erhält man statistisch gesehen richtige Aussagen, die auch in Bezug auf Abweichungen des Modells robust sind. Die inzwischen immer mehr gegebenen Möglichkeiten, in Analysen mit statistischen Softwareprogrammen das Design einzubinden, geben der Gewichtung eine vielversprechende Zukunft.

I. Einführung

Leslie Kish (1990, 2003) stellte die Frage „Gewichten: Warum, wann und wie?", und diese Frage wird auch heute noch häufig aufgeworfen, wenn es etwa um die Schätzung von Kennzahlen aus endlichen Gesamtheiten geht. In der Regel werden zwei Arten von Gewichtung unterschieden. Ist die Gewichtung durch das Auswahlverfahren bedingt, spricht man von Designgewichtung. Wird beispielsweise disproportional geschichtet ausgewählt, ergibt sich die Notwendigkeit von Gewichten bei der Schätzung zwangsläufig. Dies gilt allgemein bei Auswahlen mit ungleichen Inklusionswahrscheinlichkeiten für die Elemente einer Grundgesamtheit.

Neben der Designgewichtung gibt es auch die so genannte Zellgewichtung. Sie wird dann verwendet, wenn sich nach Durchführung der Erhebung zeigt, dass Verteilungen gewisser für wesentlich erachteter Variablen in der Stichprobe von denen in der Gesamtheit abweichen. Man versucht nun über eine Gewichtung der Elemente in der Stichprobe, die Verteilung von Variablen wie Alter, Geschlecht, Bundesland usw. an die Verteilung in der Gesamtheit anzupassen. Dies wird in der Umfragepraxis oft als Heilmittel für verzerrte Stichproben angesehen. Notwendig für eine Adaption an solche Kontroll- oder Anpassungsvariablen ist die Verfügbarkeit externer Daten, etwa aus der amtlichen Statistik.

Ein Grund dafür, warum Verteilungen von Variablen in der Stichprobe und Gesamtheit sich signifikant unterscheiden, kann im Non-response liegen. Dieser ist bei den heutigen sozialwissenschaftlichen Umfragen ein ernstzunehmender Faktor. Verzerrte Stichproben können die Folge sein. Dies offenbart einen Notstand, den man gerne durch Gewichtung reparieren würde. Ob dies erfolgversprechend ist oder nicht, hängt

sicher auch davon ab, inwieweit der Ausfallprozess in die Gewichte adäquat implementiert werden kann.

Die Verwendung von Gewichten stellt sich aber nicht nur bei der Schätzung von Kennzahlen der Grundgesamtheit. Eine große Unsicherheit besteht bei vielen Nutzern von Umfragen darin, ob Gewichte bei multivariaten Analyseverfahren, die auf Modellen beruhen, verwendet werden sollen oder nicht.

Die folgenden Überlegungen sollen dem Forscher helfen, sich bei der Frage nach Gewichtung auf sicherem Boden zu bewegen und damit die Entscheidung, ob und wenn ja, welche Gewichte verwendet werden sollen, wissenschaftlich fundiert getroffen wird. Grundlegende Arbeiten zur Gewichtung sind bei Bethlehem (1999, 2002), Kalton und Flores-Cervantes (2003) sowie Pfeffermann (1993) zu finden. Dass auch die amtliche Statistik in Deutschland komplexe Gewichtungsmethoden verwendet, geht am Beispiel des Zensustests aus Bihler (2002) hervor.

II. Gewichtung bei endlicher Population

Was heißt gewichten? Ein Gewicht ist ein multiplikativer Faktor, der auf einige oder alle Befragten in einer Umfrage angewendet wird. Die einzelnen Gewichte können kleiner, gleich oder größer als eins sein und werden i.d.R. auf die Fallzahl normiert. Sind die Gewichte alle gleich eins, spricht man von ungewichteten Daten. Die Anwendung einer Gewichtung ändert die relative Wichtigkeit der Befragten. Das angestrebte Ziel des Gewichtens ist es zu erreichen, dass das Stichprobenprofil einer interessierenden Variablen nach der Gewichtung dem tatsächlichen Profil in der Untersuchungsgesamtheit ähnlicher als im ungewichteten Fall ist. Stichprobentheoretisch würde man dann von einem besseren Schätzer sprechen, wobei besser z.B. erwartungstreu, kleinere Varianz oder robuster bedeuten kann.

Das folgende einfache Beispiel zeigt, dass die Verwendung von Gewichten bei der deskriptiven Beschreibung von Kennzahlen der Gesamtheit problemlos möglich ist. Angenommen, es interessiert die durchschnittliche Haushaltsgröße von acht Haushalten. Die einzelnen Haushaltsumfänge Y_i (i = 1, ..., 8) seien durch 1, 4, 1, 2, 2, 3, 2, 1 gegeben. Dann ist das arithmetische Mittel der Haushaltsgrößen

$$\bar{Y} = \frac{1}{8}\sum_{i=1}^{8} Y_i = \frac{1+4+1+2+2+3+2+1}{8} = 2.$$

Man kann die Berechnung auch anders vornehmen, wenn man nur die unterschiedlichen Haushaltsgrößen betrachtet und die Häufigkeit ihres Auftretens notiert. In unserem einfachen Beispiel wäre das

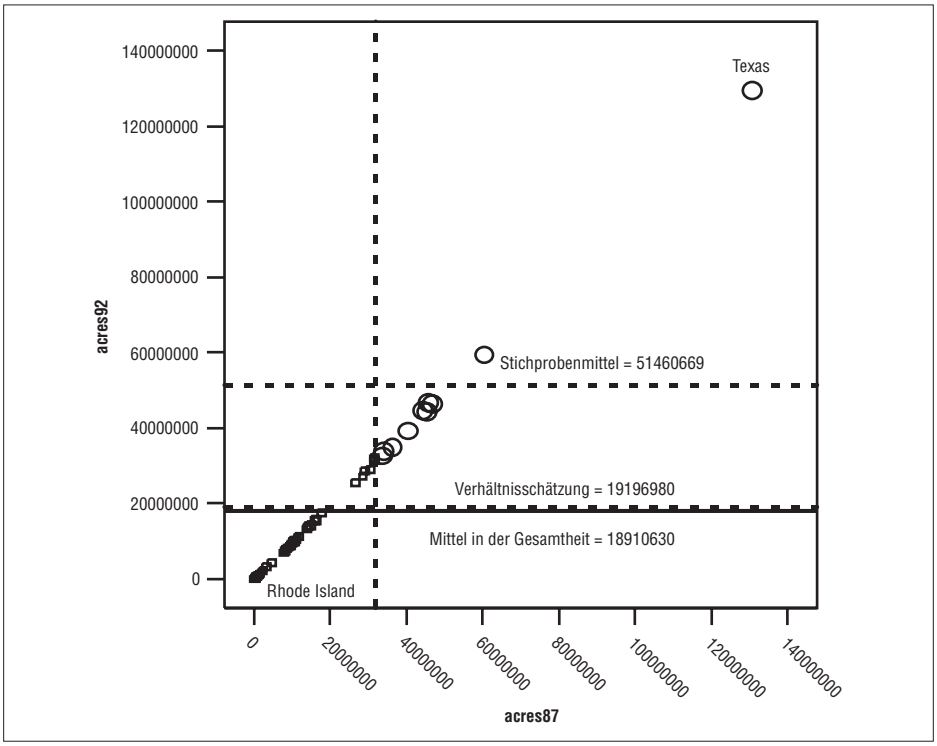
$$\bar{Y} = \frac{3\cdot 1+3\cdot 2+1\cdot 3+1\cdot 4}{8} = \frac{3}{8}\cdot 1+\frac{3}{8}\cdot 2+\frac{1}{8}\cdot 3+\frac{1}{8}\cdot 4,$$

also ein gewogenes arithmetisches Mittel der verschiedenen Haushaltsgrößen. Es leuchtet ein, dass die Gewichtung der verschiedenen Haushaltsgrößen notwendig ist, um zum richtigen Ergebnis zu kommen. Das einfache arithmetische Mittel der verschiede-

nen Haushaltsgrößen wäre 2,5 und von der tatsächlichen mittleren Haushaltsgröße 2 verschieden.

Das folgende Beispiel aus Lohr (1999) zeigt die Anbauflächen der 50 amerikanischen Staaten gemessen in acres [1 *acre* = 4047 m^2] in den Jahren 1987 und 1992. Angenommen, aus der Gesamtheit der 50 Staaten werden diejenigen 10 ausgewählt, die 1987 die größten Anbauflächen hatten. *Abbildung 1* zeigt das Streuungsdiagramm mit Kreisen als Symbole für die ausgewählten Staaten und den Rechtecken für die Staaten, die nicht in die Stichprobe gelangten. Das Stichprobenmittel der Anbauflächen 1992 ist offensichtlich ein schlechter Schätzer für das Gesamtmittel. Der Verhältnisschätzer als gewichteter Schätzer liefert deutlich bessere Ergebnisse, da zwischen den Anbauflächen in den beiden Jahren ein fast proportionaler Zusammenhang besteht.

Abbildung 1: Stichprobenmittel versus Gesamtmittel und Verhältnisschätzung



Das zu Beginn dieses Abschnitts beschriebene Vorgehen können wir auch als Schichtung nach Haushaltsgröße ansehen. Wir betrachten nun die bei Bevölkerungsumfragen in Deutschland häufig vorzufindende Tatsache, dass die Befragten aus Ost- und Westdeutschland in der Stichprobe bewusst disproportional vertreten sind.

Sind für ein Merkmal Y die (irgendwie berechneten) Werte t_{West} und t_{Ost} *gute* Schätzungen für die Mittel (etwa Anteile) $\bar{Y}_{West}$ und $\bar{Y}_{Ost}$ in den Gesamtheiten West und Ost, so ist

$$\frac{N_{West}}{N} t_{West} + \frac{N_{Ost}}{N} t_{Ost}$$

eine gute Schätzung für das interessierende Gesamtmittel

$$\bar{Y} = \frac{N_{West}}{N} \bar{Y}_{West} + \frac{N_{Ost}}{N} \bar{Y}_{Ost}.$$

Diese Gewichtung ist unabhängig von der Art und Weise, wie die Schätzungen t_{West} und t_{Ost} bestimmt werden. Insbesondere können in t_{West} und t_{Ost} bereits Gewichtungen eingegangen sein. Der eben betrachtete Fall ist ein Spezialfall der geschichteten Auswahl, wenn also in den einzelnen Schichten (im Beispiel Ost-West) einer Gesamtheit unabhängig voneinander Stichproben ausgewählt werden und Aussagen über Kennzahlen der Gesamtheit gemacht werden sollen. Von der Stichprobentheorie her gesehen ist die Gewichtung in solchen Fällen designbedingt nicht nur sinnvoll, sondern notwendig. Zwei einfache Beispiele für Ost-West Gewichte bei ALLBUS-Daten von 1992 sind bei Gabler (1992) nachzulesen, von denen wir das Ergebnis der Gewichtung bei der Variablen V198 „Familienstand" mit den Ausprägungen

E = Verheiratet und leben mit ihrem Ehepartner zusammen,
T = Verheiratet und leben getrennt,
W = Verwitwet,
G = Geschieden,
L = Ledig

in *Tabelle 1* angeben.

Tabelle 1: V198 FAMILIENSTAND, BEFRAGTE in Prozent

	WEST		OST		GESAMT		
	ungewichtet	Designgewichtet	ungewichtet	Designgewichtet	ungewichtet	Ost/West gewichtet	Ost/West Designgewichtet
E	58,8	67,3	67,1	74,3	61,5	60,4	68,6
T	1,2	0,7	0,5	0,4	1,0	1,0	0,6
W	11,3	6,9	8,1	4,6	10,2	10,7	6,5
G	6,1	4,4	8,4	6,2	6,9	6,6	4,8
L	22,7	20,7	15,9	14,6	20,5	21,4	19,6
Fallzahl	2400	2400	1148	1148	3548	3548	3548

Offensichtlich ist zwischen den ungewichteten und den Ost-West gewichteten Ergebnissen kein allzu großer Unterschied. Eine andere Designgewichtung, die auf ungleichen Auswahlwahrscheinlichkeiten beruht, hat stärkeren Einfluss. Die Auswahlwahrscheinlichkeit des einzelnen Befragten hängt beim ALLBUS von 1992 von der Zahl der Zielpersonen im Haushalt ab, von der so genannten reduzierten Haushaltsgröße.

Bei der Variablen V308 „Sind die Kinder getauft?" waren die Ost-West-Unterschiede wesentlich größer, so dass zwischen ungewichteten und Ost-West gewichteten Ergebnissen auch ohne Designgewichtung auf Grund ungleicher Auswahlwahrscheinlichkeiten der Personen schon große Unterschiede bestanden.

Gewichtung lässt sich mit Schätzfunktionen in Verbindung bringen. Für die Behandlung von Schätzfunktionen und ihren Eigenschaften in der Stichprobentheorie gibt es einen einheitlichen Rahmen.

Ganz allgemein sei eine interessierende Gesamtheit $U = \{1, ..., N\}$ von N Einheiten, z.B. Personen gegeben. Das interessierende Merkmal Y habe die unbekannten Ausprägungen $Y_1, ..., Y_N$, die im Parametervektor $\theta = (Y_1, ..., Y_N)'$ zusammengefasst sind. Der zu schätzende Parameter $T = T(\theta)$ ist eine Funktion des Parametervektors, etwa $T = \bar{Y}$. Die Wahrscheinlichkeit, mit der eine Stichprobe s, also eine Teilmenge von U ausgewählt wird, sei mit $p(s)$ bezeichnet. Um $T(\theta)$ zu schätzen, verwenden wir eine Schätzfunktion $t(s,\theta)$. Diese Funktion darf natürlich nur von den Y_i-Werten abhängen, für die $i \in s$ ist. Eine solche Schätzfunktion heißt linear, wenn $t(s,\theta) = \sum_{i \in s} a_{si} Y_i + b_s$ ist; häufig ist $b_s = 0$. Die Gewichte a_{si} sollten vom Forscher so gewählt werden, dass die Schätzfunktion $t(s,\theta)$ möglichst nahe am zu schätzenden Wert $T(\theta)$ liegt. Ist die Auswahl durchgeführt, dann liegt die Stichprobe fest und man schreibt statt a_{si} oft einfach w_i. Eine weitere Eigenschaft von Schätzfunktionen ist ihre Repräsentativität bezüglich eines oder mehrerer Hilfsmerkmale ξ. Sind X_1,K ,X_N die Ausprägungen des Hilfsmerkmals ξ in der Gesamtheit, so heißt eine Schätzfunktion $t(s,\theta)$ nach Hájek (1981) repräsentativ bezüglich ξ, wenn $t(s,\xi) = T(\xi)$ für $\xi = (X_1, ..., X_N)'$ gilt, wenn also für die Hilfsvariable ξ die Schätzung $t(s,\xi)$ exakt mit dem Populationswert $T(\xi)$ übereinstimmt.

Kommen wir auf das Beispiel mit den Anbauflächen 1987 als den X_i-Werten und den Anbauflächen 1992 als den Y_i-Werten aus *Abbildung 1* zurück. Betrachtet man zur Schätzung von $T(\theta) = \bar{Y}$ den Verhältnisschätzer

$$t_R(s,\theta) = \frac{\bar{y}}{\bar{x}} \cdot \bar{X},$$

wobei $\bar{y}$ und $\bar{x}$ die Stichprobenmittel der y- bzw. der x-Werte sind, und $\bar{X}$ das arithmetische Mittel der x-Werte in der Gesamtheit ist, so ist offensichtlich $t_R(s,\theta)$ repräsentativ bezüglich ξ, da

$$t_R(s,\xi) = \frac{\bar{x}}{\bar{x}} \cdot \bar{X} = \bar{X}$$

gilt. Würde man also den Verhältnisschätzer zur Schätzung der durchschnittlichen Anbaufläche 1987 verwenden, würden die geschätzte und die tatsächliche durchschnittliche Anbaufläche 1987 übereinstimmen. Wie in *Abbildung 1* gezeigt wurde, kann ein repräsentativer Schätzer eine beträchtliche Verbesserung der Schätzgenauigkeit bringen. Der Begriff eines repräsentativen Schätzers ist nicht zu verwechseln mit dem schwammigen Begriff der repräsentativen Stichprobe oder der Repräsentativität von Umfragen, wenn damit ein getreues Abbild der Gesamtheit gemeint ist.

Bei Bevölkerungsumfragen werden häufig komplexe Stichprobenpläne eingesetzt, bei denen Schichtung, Klumpung und Mehrstufigkeit eingehen. Die Primäreinheiten, d.h. die Einheiten, die auf der ersten Stufe ausgewählt werden, sind in der Regel Gemeinden, die mit zur Zahl der Einwohner größenproportionalen Wahrscheinlichkeiten

ausgewählt werden. Die in diesem Zusammenhang für die Summe $T(\theta) = Y_1 + ... + Y_N$ oft verwendete Schätzfunktion ist die Horvitz-Thompson-Schätzfunktion

$$t_{HT}(s,\theta) = \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i \in s} w_i Y_i,$$

wobei π_i die Auswahl- oder Inklusionswahrscheinlichkeit für die i-te Einheit bedeutet. Die Gewichte w_i sind daher gerade die inversen Inklusionswahrscheinlichkeiten. Es ist wohlbekannt, dass der Horvitz-Thompson-Schätzer design-erwartungstreu ist, d.h. es gilt:

$$E_p\left(t_{HT}(s,\theta)\right) = \sum_s p(s) \sum_{i \in s} \frac{Y_i}{\pi_i} = \sum_{i=1}^{N} Y_i = T(\theta).$$

Betrachten wir die ALLBUS-Stichprobe 1994. Die Gemeinden der alten Bundesländer wurden mit einer zur Zahl der darin lebenden Personen (Bevölkerung ab 20 Jahre, Stand 31.12.1991) größenproportionalen Auswahlwahrscheinlichkeit ausgewählt, manche wegen ihrer Größe auch mehrfach. Die Gesamtbevölkerung ab 20 Jahre am 31.12. 1991 in den 8506 Gemeinden belief sich auf 51.067.893. Würde man als Schätzfunktion das 8506-fache des Stichprobenmittels für die 111 ausgewählten sample points in 104 Gemeinden verwenden, käme man auf eine Schätzung von 1.773.460.769 lebenden Personen ab 20 Jahre in Westdeutschland. Der Grund für diese total falsche Schätzung liegt offensichtlich darin, dass die Schätzfunktion nicht zum Auswahlverfahren passt.

Wir wollen diesen Aspekt noch vertiefen und die Zahl der Gemeinden in den einzelnen Bundesländern schätzen (Tabelle 2).

Tabelle 2: Ungewichtete und gewichtete Schätzungen für die Zahl der Gemeinden in den alten Bundesländern

Bundesland	sample points	Statist. JB 1991	ungewichteter Schätzer*	gewichteter Schätzer*
Berlin-West	3	1	230	1
Schleswig-Holstein	5	1131	383	1139
Hamburg	3	1	230	1
Niedersachsen	13	1031	996	791
Bremen	1	2	77	1
Nordrhein-Westfalen	30	396	2299	456
Hessen	10	426	766	378
Rheinland-Pfalz	7	2304	536	1458
Saarland	2	52	153	44
Baden-Württemberg	17	1111	1303	1198
Bayern	20	2051	1533	2013
Summe	111	8506	8506	7480

* ganzzahlig gerundet.

Quelle: ALLBUS 1994.

Der ungewichtete Schätzer für die Zahl der Gemeinden in einem Bundesland wurde wie folgt berechnet:

$$\frac{8506}{111} \times \text{Zahl der Gemeinden im Datensatz je Bundesland.}$$

Gemeinden mit großem Bedeutungsgewicht wurden dabei mehrfach gezählt.

Der Horvitz-Thompson-Schätzer als gewichteter Schätzer für die Zahl der Gemeinden in einem Bundesland wurde wie folgt berechnet:

$$\frac{51067893}{111} \times \sum_{i=1}^{111} z_i,$$

wobei für ein beliebiges Bundesland z_i die Zahl der mindestens 20-Jährigen in der Gemeinde ist, zu der der i-te sample point gehört, falls die Gemeinde überhaupt zu dem entsprechenden Bundesland gehört. Ansonsten ist $z_i = 0$.

Auch hier ist deutlich zu erkennen, dass der ungewichtete Schätzer in der Regel schlechte Werte liefert, etwa, dass in Berlin-West schätzungsweise 230 Gemeinden lägen. Das Profil der gewichteten Schätzungen liegt dagegen wesentlich näher beim tatsächlichen Profil als beim ungewichteten Schätzer. Die einzige Ausnahme bildet Niedersachsen, wo ungewichtet 996 Gemeinden geschätzt werden, gewichtet 791 Gemeinden und in Wahrheit 1031 Gemeinden vorhanden sind. Hier wird ein Problem der Gewichtung offenbar. Es kann im Einzelfall durchaus sein, dass ein ungewichteter Schätzer besser ist als ein gewichteter. Man darf aber daraus nicht den Umkehrschluss ziehen, dass deswegen immer der ungewichtete Schätzer vorzuziehen sei. Dies wird durch obiges Beispiel auch eindrucksvoll widerlegt. Im Unterschied zum ungewichteten Schätzer ist beim gewichteten Schätzer die Summe der geschätzten Gemeinden über ganz Westdeutschland nicht 8506. Bei Kenntnis dieses Wertes könnte ein gewichteter Schätzer einfach dadurch konstruiert werden, dass die Zahlen der letzten Spalte in Tabelle 2 durch 7480 geteilt und mit 8506 multipliziert würden, also eine Anpassungsgewichtung vorgenommen wird, die wir weiter unten noch ausführlich behandeln.

Es soll an dieser Stelle nicht verschwiegen werden, dass die Berechnung eines gewichteten Schätzers $\sum_{i \in s} w_i Y_i$ oder vorzugsweise $\dfrac{\sum_{i \in s} w_i Y_i}{\sum_{i \in s} w_i}$ schwierig oder unmöglich sein kann. Es sind zwar die Bevölkerungszahlen der größeren Gemeinden aus dem statistischen Jahrbuch bekannt. Da aber im Datensatz außer bei einigen großen Gemeinden nicht erkennbar ist, welche Gemeinden ausgewählt wurden, sind diese Gewichte auch nicht berechenbar. Bei Unkenntnis der Gewichte w_i kann aber der eigentlich richtige Schätzer $\sum_{i \in s} w_i Y_i$ nicht berechnet werden. Die Berechnung der letzten Spalte in obiger Tabelle ist nur möglich, weil ZUMA als Auftraggeber die ausgewählten Gemeinden kennt. Aus Datenschutzgründen sind diese aber dem ALLBUS-Nutzer unbekannt.

Ein weiteres Problem kann auftreten, wenn die Inklusionswahrscheinlichkeiten über mehrere Stufen berechnet werden müssen, wie das bei den komplexen Designs von Be-

völkerungsumfragen üblicherweise der Fall ist. Wie groß ist etwa bei einem Random-Route-Verfahren die Inklusionswahrscheinlichkeit für die Auswahl eines Haushalts bei gegebener Gemeinde?

Als Maß für die Güte einer Schätzfunktion wird die Varianz oder bei nicht erwartungstreuen Schätzfunktionen der mittlere quadratische Fehler herangezogen. Die Varianz der Horvitz-Thompson-Schätzfunktion

$$Var_p\left(t_{HT}(s,\theta)\right) = \sum_s p(s)\left(\sum_{i\in s}\frac{Y_i}{\pi_i} - \sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{Y_iY_j}{\pi_i\pi_j}$$

kann im Falle $\pi_{ij} > 0$ für alle $i \neq j$ erwartungstreu durch

$$v_{HT} = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{Y_iY_j}{\pi_i\pi_j}$$

geschätzt werden. Dabei ist π_{ij} die Inklusionswahrscheinlichkeit für die i-te und j-te Einheit mit $\pi_{ii} = \pi_i$. Für ein Auswahlverfahren mit festem Stichprobenumfang gibt es noch die so genannte Yates-Grundy-Schätzfunktion

$$v_{YG} = \sum_{i\in s}\sum_{\substack{j\in s\\ i<j}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2,$$

die den Vorteil hat, positive Varianzschätzwerte zu liefern, falls $\pi_i\,\pi_j > \pi_{ij}$ für alle $i \neq j$ in der Stichprobe gilt. Für viele bekannte Auswahlverfahren ist dies erfüllt, aber z.B. nicht für die systematische Zufallsauswahl. Für komplexe Auswahlverfahren ist es meist nicht möglich, die π_{ij} anzugeben, insbesondere wenn das Auswahlverfahren mehrstufig ist. Die Varianz wird dann durch Resampling-Verfahren geschätzt oder durch Verwenden von Formeln beim Ziehen mit Zurücklegen approximiert. Im Fall der Schichtung der Population $U = U_1 \cup \ldots \cup U_H$ in H Schichten und unabhängiger Auswahl in den einzelnen Schichten lassen sich obige Formeln verallgemeinern.

Als Schätzer wäre

$$t(s,\theta) = \sum_{h=1}^{H} t_{HT}(s_h,\theta_h) = \sum_{h=1}^{H}\sum_{i\in s_h}\frac{Y_{ih}}{\pi_{ih}} = \sum_{i=1}^{H}\sum_{i\in s_h} w_{ih}Y_{ih},$$

mit geschätzter Varianz

$$v = \sum_{h=1}^{H}\sum_{i\in s_h}\sum_{j\in s_h}\frac{\pi_{ijh} - \pi_{ih}\pi_{jh}}{\pi_{ijh}}\frac{Y_{ih}Y_{jh}}{\pi_{ih}\pi_{jh}}$$

zu verwenden, wobei der Index h die Schichtzugehörigkeit kennzeichnet.

Das EU-Projekt DACSEIS (http://www.dacseis.de) hatte die Varianzschätzung bei komplexen Stichprobendesigns zum Gegenstand der Untersuchung und ist eine ausgezeichnete Quelle bei der Suche nach der Einschätzung von Schätzern bei verschiedenar-

tigen Konstellationen. Insbesondere das Recommended Practise Manual enthält wertvolle Empfehlungen.

Die Berechnung der Varianz dient zur Konstruktion von Konfidenzintervallen. Da sich die Varianz der Schätzer bei komplexen Auswahlverfahren beträchtlich von den Schätzern bei uneingeschränkter Zufallsauswahl unterscheiden kann, erhält man in der Regel falsche Konfidenzintervalle, wenn man so tut, als hätte man uneingeschränkt zufällig ausgewählt. Dies wird in der Praxis häufig nicht bedacht. Da die Varianzen von Schätzern bei Mehrstufigkeit und Clusterung meist größer sind als bei uneingeschränkter Zufallsauswahl, sind die Konfidenzintervalle bei Nichtbeachtung des Designs zu klein. Der Designeffekt liefert eine Möglichkeit, die Varianz eines Schätzers unter uneingeschränkter Zufallsauswahl mit Zurücklegen (SRSWR) zu korrigieren. Der Designeffekt *Deff* ist üblicherweise definiert als

$$Deff = \frac{Var_p\left(t(s,\theta)\right)}{Var_{SRSWR}\left(\tilde{t}(s,\theta)\right)}.$$

Das komplexe Auswahlverfahren ist mit p gekennzeichnet, die beiden Schätzer t und $\tilde{t}$ sind häufig identisch. Für hinreichend große Fallzahlen liefert dann

$$t(s,\theta) \pm z_{\alpha/2}\sqrt{Var_{SRSWR}\left(\tilde{t}(s,\theta)\right)}\sqrt{Deff}$$

das richtige Konfidenzintervall. Im Falle einer einfachen geclusterten Stichprobe mit m Einheiten in jedem Cluster ist

$$Deff = 1 + (m - 1) \times \rho.$$

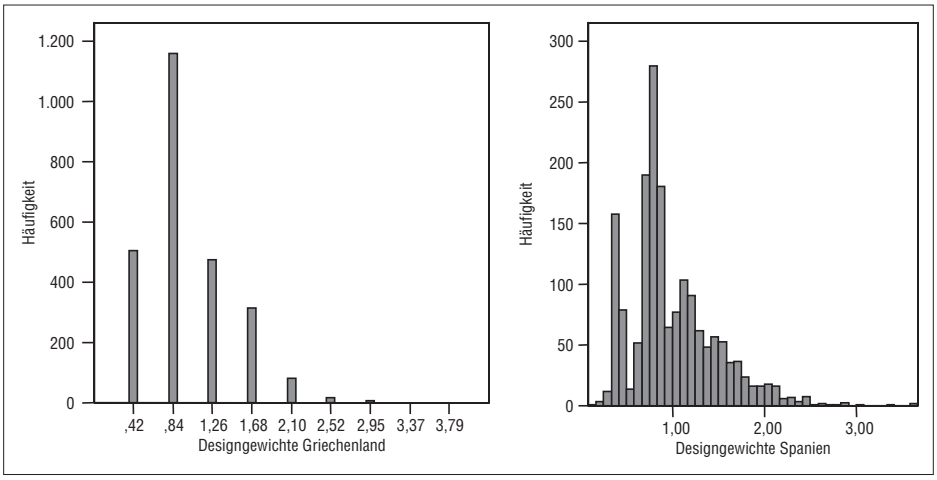
Dabei ist ρ der so genannte Intraklassen-Korrelationskoeffizient (Lohr 1999: 138).

Der Designeffekt spielte eine zentrale Rolle bei den Auswahlverfahren der 22 Länder, die beim European Social Survey (ESS 2002–2003) in Runde 1 teilgenommen haben (siehe http://ess.nsd.uib.no/2003/SamplingReport.pdf). Eine Übersicht über die prognostizierten und geschätzten Designgewichte ist bei Lynn et al. (2004) zu finden. Ein Experten-Panel hat im Vorfeld dafür gesorgt, dass die Stichprobenauswahlen in den einzelnen Ländern methodisch von höchster Qualität waren. Die Daten des ESS sind bei http://ess.nsd.uib.no/ kostenlos herunterladbar und enthalten auch die Designgewichte (Data edition ESS1e05_1spss.zip). Dabei gibt es Länder wie Finnland und Schweden, deren Designgewichte eins sind, und Anteilswerte von Merkmalen in der Gesamtheit daher ungewichtet geschätzt werden können, wenn der Non-response unberücksichtigt bleibt. Die folgende *Abbildung 2* zeigt die Verteilung der Gewichte bei Griechenland und Spanien.

Tabelle 3 liefert einen Überblick über den Anteil der Personen, die in Einpersonenhaushalten leben.

Offensichtlich unterscheiden sich in zehn (kursiv) von 22 Ländern die Ergebnisse in Spalte 2 und 3 von *Tabelle 3* gewaltig. Zu beachten ist noch, dass die Werte aus der Stichprobe sich auf die Gesamtheit der mindestens 15-Jährigen beziehen und daher mit deren Anteil in der Gesamtheit multipliziert werden müssen, um den Anteil der allein lebenden Personen für die Gesamtheit aller Personen in den einzelnen Ländern zu schätzen. Die Frage stellt sich natürlich, ob die gewichteten oder ungewichteten

Abbildung 2: Häufigkeiten der Designgewichte in Griechenland und Spanien



Schätzungen besser sind. Die letzte Spalte in *Tabelle 3* zeigt für die Mitgliedsländer der EU-15 die gesuchten Werte aus dem Jahre 1999. Die Werte für 2003 dürften etwas höher liegen, da ein allgemeiner Trend zu Einpersonenhaushalten zu beobachten ist. Bei Frankreich, Griechenland, Großbritannien, Österreich, Portugal und Spanien liegen demnach die gewichteten Schätzungen wesentlich näher bei den tatsächlichen Werten. Nur bei den Niederlanden liegen die gewichteten und ungewichteten Werte ziemlich gleich weit vom tatsächlichen Wert entfernt. Dies könnte darauf schließen lassen, dass die Einpersonenhaushalte dort einen relativ hohen Non-response aufweisen.

Die Notwendigkeit einer Gewichtung ist also vom Design her begründet, wobei der Non-response noch unberücksichtigt bleibt. In einigen Ländern ergaben sich zunächst extrem hohe Gewichte, die man vermeiden will, da sie zu ineffizienten Schätzern führen können. Dies kann durch Eingrenzung der Spannweite der Gewichte erfolgen, wobei dann allerdings ein Bias in Kauf genommen wird.

Aus dem obigen Beispiel ist auch zu erkennen, wie wichtig es ist, genau zu spezifizieren, für welche Grundgesamtheit Aussagen über Kennzahlen gemacht werden. Da beim ESS die Auswahlgesamtheit nur Personen mit mindestens 15 Jahren enthält, liegt ein Undercoverage-Problem vor, das in diesem Fall dadurch gelöst werden kann, dass mit dem Anteil der mindestens 15-Jährigen an der Gesamtzahl aller Personen nochmals gewichtet wird.

Die Horvitz-Thompson-Schätzfunktion ist die bekannteste Schätzfunktion für eine Gesamtsumme bei ungleichen Auswahlwahrscheinlichkeiten. Der Grund dafür ist ihre Universalität und die Tatsache, dass sie gewisse wünschenswerte statistische Eigenschaften hat. Ihre Verwendung ist aber nicht ohne Einschränkung zu empfehlen. Dies gilt insbesondere, wenn der Stichprobenumfang klein ist oder wichtige Vorinformationen vorliegen, die in die Schätzfunktion eingebaut werden sollten. Im Folgenden soll ein berühmtes Beispiel besprochen werden, das auf einem Disput zwischen Basu (1971)

Tabelle 3: Gewichtete und ungewichtete Anteile der allein lebenden Personen in 22 ESS-Ländern

Land	ESS Stichprobe		Anteil mind. 15 Jahre* (in %)	Anteil der allein lebenden Personen ESS	Anteil der allein lebenden Personen** EU-15
	Ungewichtet (in %)	Gewichtet (in %)		Gewichtet (in %)	1999 (in %)
Belgien	13,9	13,9	82,8	11,5	12
Dänemark	18,0	18,0	81,1	14,6	17
Deutschland	18,9	18,9	85,1	16,1	16
Finnland	22,9	22,9	82,3	18,8	16
Frankreich	23,3	12,6	81,4	10,3	13
Griechenland	18,7	8,0	85,5	6,8	8
Großbritannien	30,3	18,2	81,7	14,9	13
Irland	10,9	10,0	78,9	7,9	7
Israel	12,5	5,9	73,1	4,3	
Italien	9,9	8,7	86,0	7,5	9
Luxemburg	12,7	11,8	80,9	9,5	10
Niederlande	22,9	11,2	81,6	9,1	14
Norwegen	19,4	19,5	80,1	15,6	
Österreich	24,0	11,4	83,9	9,6	12
Polen	8,1	8,1	82,5	6,7	
Portugal	15,8	6,8	83,2	5,7	5
Schweden	21,2	21,2	82,2	17,4	20
Schweiz	27,0	14,8	83,1	12,3	
Slovenien	7,5	7,5	84,7	6,4	
Spanien	14,9	6,1	85,6	5,2	5
Tschechien	25,8	12,0	84,7	10,2	
Ungarn	10,9	10,9	83,8	9,1	

* *Quellen:* U.S. Bureau of the Census, International Data Base; http://www.census.gov/ipc/www/idbagg.html. Table 094. Midyear Population, by Age and Sex / 2003.

Die Zahlen weichen geringfügig ab von denen in http://europa.eu.int/comm/eurostat/newcronos/queen/display.do?screen=detail&language=de&product=YES&root=YES/yearlies/c/ca/caa/caa15632

** Die Familie im Spiegel der amtlichen Statistik. Bundesministerium für Familie, Senioren, Frauen und Jugend. Tabelle 9 Seite 59 und Eurostat: Statistik kurz gefasst. Thema 3 –24/2003 S. 3 Abbildung 5 http://www.europa.eu.int/comm/eurostat/Public/datashop/print-product/DE?catalogue=Eurostat&product=KS-NK-03-024-_ _-N-DE&mode=download

und Godambe beruht und in dem Buch von Brewer (2002) ausführlich behandelt wird.

Ein Zirkusbesitzer plant, seine 50 Elefanten zu verschiffen. Da das Wiegen von Elefanten eine schwierige Angelegenheit ist, wählt der Zirkusbesitzer nur einen Elefanten aus. Aus drei Jahre zurückliegenden Aufzeichnungen geht hervor, dass Sambo das Durchschnittsgewicht der Herde hatte und – wie ihm der Trainer versichert – auch heute noch hat. Er plant daher, eine Schätzung des Gesamtgewichts der Elefanten $Y = Y_1 + ... + Y_{50}$ mittels der Schätzung $50 \cdot Y_{Samba}$ vorzunehmen. Der Zirkusstatistiker ist schockiert über diese bewusste Auswahl, da sie keine erwartungstreue Schätzung von Y zulasse. Zusammen arbeiten die beiden als Kompromiss einen neuen Auswahlplan aus. Der neue Auswahlplan sieht vor, mit Hilfe einer Zufallszahlentabelle Samba mit Wahrscheinlichkeit 99/100 und die anderen 49 Elefanten jeweils mit Wahrscheinlich-

keit 1/4900 auszuwählen. Als Sambo erwartungsgemäß tatsächlich ausgewählt wird, ist der Zirkusbesitzer glücklich. Auf die Frage des Statistikers, wie er denn gedenke, das Gesamtgewicht Y zu schätzen, antwortet er, mit $50 \cdot Y_{Samba}$ natürlich. Oh nein! erwidert der Zirkusstatistiker, das kann unmöglich richtig sein. Er hatte nämlich erst kürzlich einen Artikel in den Annals of Mathematical Statistics gelesen, in dem bewiesen wurde, dass der Horvitz-Thompson-Schätzer der eindeutige hyperzulässige Schätzer in der Klasse aller verallgemeinerten polynomialen unverzerrten Schätzer sei. Sichtlich beeindruckt fragt der Zirkusbesitzer, welchen Wert denn der Horvitz-Thompson-Schätzer im vorliegenden Fall habe. Wegen der Auswahlwahrscheinlichkeit 99/100 sei das $100 \cdot Y_{Samba} / 99$, so die Antwort des Statistikers. Und wie wäre die Schätzung ausgefallen, wenn der große Elefant Jumbo ausgewählt worden wäre, kam die Frage des ungläubigen Besitzers. So wie ich die Horvitz-Thompson-Schätzmethode verstehe, $4900 \cdot Y_{Jumbo}$, antwortet der unglückliche Zirkusstatistiker. Dies ist die Geschichte, wie der Zirkusstatistiker seine Arbeit verlor (und vielleicht Statistikdozent wurde).

Der Grund dafür, dass im vorliegenden Fall der Zirkusbesitzer mit seiner einfachen Methode besser als der Zirkusstatistiker mit seiner theoriebasierten Methode war, liegt zum einen daran, dass der Stichprobenumfang $n = 1$ war und die für eine gute Schätzung vorliegende Vorinformation des Trainers keine Berücksichtigung fand. Hätte der Zirkusstatistiker statt der Horvitz-Thompson-Schätzung den Quotienten $N\sum_{i\in s} w_i Y_i \,/\, \sum_{i\in s} w_i = 50\,Y_{Samba}$ verwendet, wäre er zur selben Schätzung gekommen wie der Zirkusbesitzer. Auch der Verhältnis- und Regressionsschätzer sind nichts anderes als gewichtete Schätzer, deren Verwendung zusätzlich modellbasiert motiviert werden kann. Dass auch sie geeignet sind, bessere Schätzungen zu liefern als die ungewichtete Schätzung, ist weithin bekannt und in *Abbildung 1* sichtbar.

III. Zellgewichtung

Neben der Designgewichtung wird häufig auch die Zellgewichtung verwendet, die auch Redressment oder Anpassungsgewichtung genannt wird. Sie kommt dann in Frage, wenn die Verteilungen externer Variablen, wie Alter, Geschlecht, Bundesland usw. in der Grundgesamtheit verfügbar sind, aber von denen in der Stichprobe differieren. Durch Anpassung an die Verteilung dieser Variablen hofft man, ein Heilmittel für verzerrte Stichproben zu finden. Je nachdem, ob die gemeinsame Verteilung der Anpassungsvariablen oder nur deren Randverteilungen bekannt ist, unterscheidet man zwei Fälle.

1. Die gemeinsame Verteilung von K Variablen ist bekannt. Durch einfache Gewichtung Soll/Ist passt man die Stichprobenverteilung an die Verteilung in der Gesamtheit an. Diese Art der Gewichtung wird auch nachträgliche Schichtung genannt. Eine allgemeine Formel bei K=2 kategorialen Merkmalen lautet:

$$w_{ijk} = \frac{1}{\pi_k}\frac{N_{ij}}{\hat{N}_{ij}} \; ; \; \hat{N}_{ij} = \sum_{k\in s_{ij}} \frac{1}{\pi_k} \; ,$$

wobei π_k die Inklusionswahrscheinlichkeiten erster Ordnung bezeichnen und s_{ij} alle Einheiten der Stichprobe aus *ij*-ter Zelle enthält. Wenn der Ausfallprozess in jeder Zelle zufällig ist, erhält man in der Regel eine gute Schätzung. Problematisch sind Fälle, bei denen etliche Zellen in der Stichprobe unbesetzt sind. Dann muss man nahe beieinander liegende Zellen aggregieren.

2. Wenn nur die Randverteilungen der K Variablen bekannt sind, verwendet man so genannte Raking-Verfahren. Das berühmteste von ihnen beruht auf dem von Deming und Stephan (1941) entwickelten iterative proportional fitting (IPF) Algorithmus. Wir betrachten eine bezüglich Alter und Geschlecht erzeugte 2×2 Tabelle von 400 Personen, wobei Geschlecht die beiden Ausprägungen männlich – weiblich aufweist und das Alter in die zwei Altersklassen jung – alt klassifiziert ist.

Tabelle 4: Häufigkeitstabelle nach Altersklassen und Geschlecht

		Alter		
		alt	jung	Summe
Geschlecht	männlich	100	100	200 (200)
	weiblich	100	100	200 (200)
	Summe	200 (250)	200 (150)	400

Die Werte in Klammern seien die auf 400 normierten Randverteilungen in der Gesamtheit. Wie sind die einzelnen Zellen zu gewichten, damit die Randverteilungen der gewichteten Tabelle mit den Klammerwerten übereinstimmen? Bei Unabhängigkeit von Alter und Geschlecht hätte man *Tabelle 5*.

Tabelle 5: Häufigkeitstabelle nach Altersklassen und Geschlecht bei Unabhängigkeitsannahme

		Alter		
		alt	jung	Summe
Geschlecht	männlich	125	75	200
	weiblich	125	75	200
	Summe	250	150	400

Personen in der Zelle männlich/alt würden demnach mit 1,25 hochgewichtet werden, die in Zelle männlich/jung mit 0,75 heruntergewichtet. Analog sind die anderen Zellen zu interpretieren. Es gibt aber – abgesehen von Ganzzahligkeitsproblemen – unendlich viele Lösungen, die vier Zellen so auszufüllen, dass die gegebenen Randverteilungen genau eingehalten werden.

Tabelle 6: Allgemeine Lösung (λ beliebig)

		Alter		
		alt	jung	Summe
Geschlecht	männlich	$\frac{3s-2a-2b+\lambda}{4}$	$\frac{s-2a+2b-\lambda}{4}$	$s-a$
	weiblich	$\frac{s+2a-2b-\lambda}{4}$	$\frac{-s+2a+2b+\lambda}{4}$	a
	Summe	$s-b$	b	s

Für $\lambda = \frac{s(s-2a-2b)+4ab}{s}$ erhalten wir die Lösung bei Unabhängigkeit. Das Ergebnis des IPF-Algorithmus ergibt sich dadurch, dass das Kreuzprodukt z für die ungewichtete und gewichtete Tabelle identisch ist, also

$$z = \frac{(3s-2a-2b+\lambda)(-s+2a+2b+\lambda)}{(s-2a+2b-\lambda)(s+2a-2b-\lambda)}$$

gilt. Für $z = 1$ ist daher

$$\lambda_{IPF} = \frac{(s-2a)(s-2b)}{s},$$

für $z \neq 1$

$$\lambda_{IPF} = \frac{-s(1+z) \pm 2\sqrt{(a-b)^2 z^2 + 2\left[s(a+b)-a^2-b^2\right]z+(s-a-b)^2}}{1-z}.$$

Die IPF-Lösung hängt also von der Besetzung der Ausgangstabelle ab. Im Beispiel aus *Tabelle 4* ergibt sich wegen $z = 1$ und $s = 2a$ für

$$\lambda_{IPF} = \frac{(s-2a)(s-2b)}{s} = 0.$$

Dies entspricht *Tabelle 5*, dem Unabhängigkeitsfall.

Ein Überblick über die Konstruktion von Gewichten kann in Alexander (1987), Deville et al. (1992, 1993), Gabler (1994), Zaslavsky (1988) und Zieschang (1990) gefunden werden. Ein Vergleich von verschiedenen Lösungen wird in Little und Wu (1991) gegeben.

Im Gegensatz zur Designgewichtung ist die Anpassungsgewichtung nicht designbedingt begründbar. Wird sie nur schematisch verwendet, um Ungereimtheiten in einer verzerrten Stichprobe zu verwischen, ist sie lediglich Kosmetik. Dann ist vor ihr zu warnen (Rothe 1993; Rothe und Wiedenbeck 1993). Der Mangel, der bei schlecht durchgeführten Erhebungen zu Tage tritt, lässt sich auch durch eine Anpassungsgewichtung nicht beheben. Wertvoll kann sie dann sein, wenn die interessierende Variable mit den Anpassungsvariablen in Beziehung steht. Insbesondere bei linearen Be-

ziehungen lassen sich Anpassungsgewichte über ein Regressionsmodell sinnvoll entwickeln.

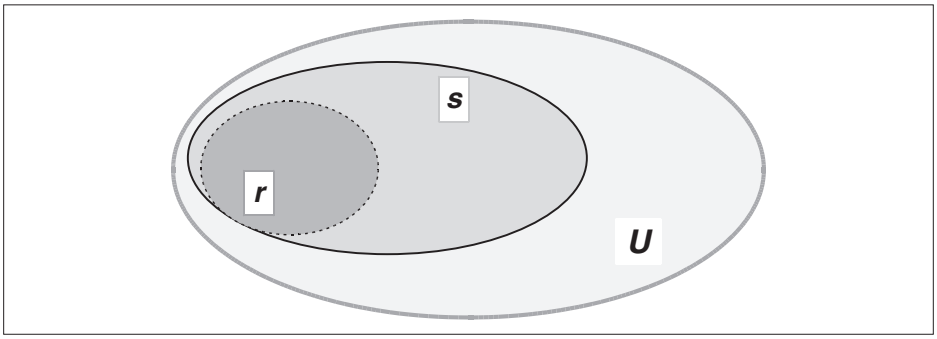
IV. Gewichtung wegen Non-response

Bekanntlich werden Umfragen mehr und mehr durch Non-response beeinträchtigt, wobei seine Höhe in verschiedenen Ländern unterschiedlich groß ist, aber auch innerhalb eines Landes von vielen Faktoren, z.B. der Länge des Fragebogens, der Wahl des Themas, dem Umfragemodus oder dem Interviewer abhängig ist. Näherungsweise 100 Prozent Response gibt es z.B. bei The Living Standards Measurement Study (http://unstats.un.org/unsd/HHsurveys/ch06draft.pdf) in Ländern wie Kirgisische Republik (1998: 99,7 Prozent), Ghana (1998/99: 97,4 Prozent) und Nicaragua (1998–99: 96,3 Prozent), wobei die Berechnung der Response-Rate nicht ganz unproblematisch ist. In Deutschland ist man bei Bevölkerungsumfragen inzwischen gewöhnt, dass die Response-Rate zwischen 50 und 60 Prozent liegt. Umfangreiche Untersuchungen zum Non-response sind bei Schnell (1997) zu finden.

Was kann man tun, wenn Verteilungen von bekannten Merkmalen wie Alter, Geschlecht, Region in der Stichprobe auf Grund von Ausfällen wesentlich anders ist als in der Gesamtheit? Gibt es neben dem Raking bei kategorialen Merkmalen eine andere Möglichkeit, durch Gewichtung dieses Defizit zu korrigieren?

Wir veranschaulichen in der folgenden *Abbildung 3* den Ausfallprozess zunächst im Rahmen eines zweiphasigen Auswahlverfahrens, bei dem zunächst die Stichprobe s und dann als Unterstichprobe die Teilmenge $r \subset s$ ausgewählt wird.

Abbildung 3: Response-Menge r, Stichprobe s, Population U mit $r \subset s \subset U$.



Hätte man die volle Stichprobe s zur Verfügung, wäre aus der Sicht des Statistikers etwa $t(s,\theta) = \sum_{i \in s} d_i Y_i$ ein brauchbarer Schätzer. Da durch den Non-response die y-Werte nur für Personen in der Response-Menge r ermittelt werden können, kann der Schätzer auch nur von diesen abhängen. Man wird daher einen Schätzer $t(r,\theta) = \sum_{i \in r} w_i Y_i$ mit neuen Gewichten w_i wählen, die sich möglichst nahe an den alten Ge-

wichten orientieren und für bestimmte Hilfsvariablen Repräsentativität erzielen. Bezeichnet q_i die Antwortwahrscheinlichkeit für die Einheit i bei gegebener Stichprobe s, so ist ein geeigneter Kandidat

$$w_i = \frac{d_i}{q_i}\left(1 + c_i\left(\sum_{k=1}^{N}\mathbf{x}_k - \sum_{k\in r}\frac{d_k}{q_k}\mathbf{x}_k\right)'\left(\sum_{k\in r}\frac{d_k c_k}{q_k}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{x}_i\right),$$

wobei c_i vom Statistiker festgelegte positive Zahlen sind. Häufig wird $c_i = 1$ gesetzt. Im p-dimensionalen Vektor $\mathbf{x}_k$ sind die Werte der K Hilfsvariablen für die k-te Person zusammengefasst. In der Regel kennt man die Antwortwahrscheinlichkeit q_i für die Einheit i nicht. Oft geht man von homogenen Response-Gruppen aus oder schätzt die q_i durch ein logistisches Regressionsmodell. In der Mehrzahl der Anwendungen wird $q_i = 1$ gesetzt und man erhält die Kalibrierungsgewichte

$$w_i = d_i\left(1 + c_i\left(\sum_{k=1}^{N}\mathbf{x}_k - \sum_{k\in r}d_k\mathbf{x}_k\right)'\left(\sum_{k\in r}d_k c_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{x}_i\right).$$

Der dazugehörige Schätzer ist im Falle $r = s$ identisch mit dem verallgemeinerten Regressionsschätzer, dem so genannten GREG-Schätzer (siehe Särndal et al. 1992; Lundström und Särndal 2002). Einen weiteren Ansatz, der auf dem bedingten Minimaxprinzip beruht, findet man bei Gabler und Häder (1998) mit einer Anwendung auf die ALLBUS-Daten 1996.

Betrachten wir ein einfaches Beispiel, bei dem man die Antwortwahrscheinlichkeiten kennt. In einer Firma werden von den 300 Männern und 1000 Frauen jeweils 100 zufällig ausgewählt. 30 Männer und 50 Frauen antworten davon auf Fragen der Zufriedenheit mit dem Arbeitsplatz. Von den 30 Männern sind 20 zufrieden, bei den 50 Frauen sind es 10. Will man die Zahl aller Mitarbeiter der Firma schätzen, die mit dem Arbeitsplatz zufrieden sind, würde man

$$20 \cdot \frac{1}{\frac{100}{300}\cdot 0{,}3} + 10 \cdot \frac{1}{\frac{100}{1000}\cdot 0{,}5} = 400$$

als Schätzwert berechnen. Der Anteil der mit dem Arbeitsplatz zufriedenen Mitarbeiter beläuft sich daher schätzungsweise auf 31 Prozent. Dabei geht man davon aus, dass alle Männer jeweils mit Wahrscheinlichkeit 0,3 und alle Frauen jeweils mit Wahrscheinlichkeit 0,5 antworten. Weiß man jedoch aus Erfahrung, dass 24 Prozent der Männer und 50 Prozent aller Frauen antworten, hätte man

$$20 \cdot \frac{1}{\frac{100}{300}\cdot 0{,}24} + 10 \cdot \frac{1}{\frac{100}{1000}\cdot 0{,}5} = 250 + 200 = 450$$

als Schätzung. Diese Schätzung hat allerdings einen Nachteil. Hätte man nämlich nach der Zahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter gefragt, hätte man

$$10 \cdot \frac{1}{\frac{100}{300} \cdot 0{,}24} + 40 \cdot \frac{1}{\frac{100}{1000} \cdot 0{,}5} = 125 + 800 = 925$$

erhalten und daher die Zahl aller Mitarbeiter auf 1375 geschätzt. Der geschätzte Anteil der mit der Arbeit zufriedenen Mitarbeiter + der geschätzte Anteil der mit der Arbeit unzufriedenen Mitarbeiter addiert sich nicht zu eins. Diesem Umstand kann man dadurch Rechnung tragen, dass eine Kalibrierung an den 1-Vektor vorgenommen wird. Setzt man für die Kalibrierungsgewichte w_i die Werte $x_i = 1$, $c_i = 1$ für alle i und $d_i = \frac{300}{100} = 3$ für Männer bzw. $d_i = \frac{1000}{100} = 10$ für Frauen, $q_i = 0{,}24$ für Männer bzw. $q_i = 0{,}5$ für Frauen, und daher $w_i = \frac{d_i}{q_i} \cdot \frac{N}{\sum_{Antworter} \frac{d_j}{q_j}} = \frac{130}{11}$ für Männer und $\frac{208}{11}$ für Frauen, so hätte man als Schätzer

$$20 \cdot \frac{130}{11} + 10 \cdot \frac{208}{11} = \frac{4680}{11} = 425{,}45$$

für die Zahl der mit dem Arbeitsplatz zufriedenen Mitarbeiter, d.h. 32,72 Prozent und

$$10 \cdot \frac{130}{11} + 40 \cdot \frac{208}{11} = \frac{9620}{11} = 874{,}55$$

für die Anzahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter, d.h. 67,28 Prozent. Die geschätzte Summe aller Mitarbeiter wäre dann 1300, also die tatsächliche Mitarbeiterzahl.

V. Gewichtung bei modell-basierter Datenanalyse

Bei der Analyse von Daten mittels Regressionsanalyse, Faktorenanalyse oder auch bei der Durchführung von Tests erhebt sich die Frage, ob das Auswahlverfahren als Bestandteil der Gewinnung der Daten Berücksichtigung finden muss oder nicht. Um diese Frage zu beantworten, betrachten wir beispielhaft das klassische Regressionsmodell:

$$y_i = x_i' \beta + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Wenn dieses so genannte Superpopulationsmodell den Datenerzeugungsprozess in der Gesamtheit gut beschreibt, wird

$$B = (X_U' X_U)^{-1} X_U' y_U$$

„nahe“ bei β liegen. Dabei ist X_U die Datenmatrix der x-Werte in der Gesamtheit und y_U der Vektor der y-Werte in der Gesamtheit. B ist als Parameter der Grundgesamtheit

U aufzufassen. Beobachtet man nur eine Stichprobe $s \subset U$, so ist der übliche OLS-Schätzer

$$\hat{\beta}_{OLS} = (X_s^{'} X_s)^{-1} X_s^{'} y_s$$

zwar modellerwartungstreu für β, aber bei Vorliegen der Daten unter Umständen weit weg von dem Wert B. Um B auch näherungsweise designerwartungstreu zu schätzen, ist demnach

$$\hat{\beta}_{WLS} = (X_s^{'} W_s X_s)^{-1} X_s^{'} W_s y_s$$

mit Diagonalmatrix W_s zu verwenden, wobei die Diagonalelemente gerade die inversen Inklusionswahrscheinlichkeiten für die Elemente der Stichprobe sind.

Das Modell spezifiziert daher den zu schätzenden Parameter der Gesamtheit, die Schätzung basiert aber auf dem Stichprobendesign.

Tiefer gehende Überlegungen zu Gewichtung bei Regressionsanalyse werden etwa bei Nathan und Smith (1989) oder Binder und Roberts (2003) behandelt. Den Fall der Schichtung behandeln in diesem Zusammenhang DuMouchel und Duncan (1983) sowie Winship und Radbill (1994). Für die Konstruktion heteroskedastisch konsistenter Schätzer für die Standardfehler verweisen letztere auf die Arbeit von White (1980). Ein ausführliches Beispiel wird bei Korn und Graubard (1999) diskutiert, das auch Lohr (1999: 365ff.) noch einmal aufgreift. Zusammenfassend wird dort gesagt, dass gewichtete und ungewichtete Schätzer unterschiedliche Ergebnisse liefern können, wenn gilt:

– Das Modell ist stark missspezifiziert.
  Im Beispiel Korn und Graubard (1999): Schwangerschaftsdauer und Geburtsgewicht folgen keiner linearen, sondern eher einer quadratischen Regression.
– Eine ausgelassene Variable hat eine starke Korrelation mit den exogenen Variablen und ist hoch korreliert mit den Gewichten.
  Im Beispiel Korn und Graubard (1999): Das Rauchverhalten der Mutter ist aufgeteilt nach Bildung, aber die Hautfarbe der Mutter ist ausgelassen.
– Stichprobenumfänge hängen (indirekt) von der interessierenden Variablen ab.
  Im Beispiel Korn und Graubard (1999): Das Rauchverhalten der Mutter ist klassifiziert mit dem Geburtsgewicht. Säuglinge mit geringem Geburtsgewicht wurden zu häufig ausgewählt.

Nach Chambers und Skinner (2003: 83) spielen bei der Frage der Gewichtung zwei Größen eine wichtige Rolle: 1. Der Bias und die Effizienz: Unter gewissen Umständen kann die Verwendung von Gewichten den Bias verringern, aber auch die Effizienz des Schätzers. 2. Die Frage nach dem Einfluss der Missspezifikation des Modells: Wenn das spezifizierte Modell korrekt ist und das Auswahlverfahren nicht von der interessierenden Variablen, sondern nur von den exogenen Variablen abhängig ist, d.h. wenn die Gewichte – bedingt auf die Werte der exogenen Variablen – keinen Bezug zur Responsevariablen haben, dann erzeugt das Auswahlverfahren keinen Bias in den ungewichteten Schätzern der Regressionskoeffizienten. Wenn aber das Auswahlverfahren von Faktoren abhängig ist, die – bedingt auf die Werte der exogenen Variablen – einen Bezug zur Responsevariablen haben, dann kann das Auswahlverfahren einen Bias im ungewichteten Schätzer induzieren. Der gewichtete Schätzer ist eine einfache Methode, um sich gegen solchen Bias zu schützen.

VI. Software

Geht es bei der Schätzung nur um Punktschätzer für Summen oder Mittelwerte, verwendet man also nur Schätzer der Form $\sum_{i \in s} w_i Y_i$ oder $\sum_{i \in s} w_i Y_i / \sum_{i \in s} w_i$, so kann man die Berechnung bei SPSS mittels des WEIGHT BY-Befehls problemlos durchführen. Die Berechnung der entsprechenden Varianzschätzer und damit der Konfidenzintervalle war bisher allerdings unmöglich. Dies hat sich in der neuen Version 12.0 mit dem Zusatzmodul SPSS Complex Samples geändert. Mit den neuen Prozeduren CSSELECT, CSDESCRIPTIVES und CSTABULATE lassen sich komplexe Stichproben ziehen, gewichtete Punkt- und Varianzschätzer berechnen und auch korrekte Tests etwa bei Kreuztabellen durchführen. Andere statistische Programmpakete wie SAS, Stata oder S-Plus (bzw. R) bieten diese Möglichkeiten schon länger. Durch die Vielfalt der Möglichkeiten, komplexe Stichproben auszuwählen und Schätzungen durchzuführen, sind die Möglichkeiten bei diesen allgemeinen Programmpaketen allerdings meist beschränkt. Bei mehrstufiger Zufallsauswahl ist etwa die Zahl der Stufen begrenzt. Für umfangreichere Einsatzmöglichkeiten sind spezielle Softwareprogramme wie SUDAAN, CLAN, WesVar, Bascula, POULPE auf dem Markt. Ein Vergleich dieser Softwareprodukte und Empfehlungen für deren Einsatz ist im Recommended Practice Manual on Variance Estimation Methods unter http://www.dacseis.de verfügbar.

Literatur

Alexander, Charles, 1987: A Class of Methods for Using Person Controls in Household Weighting. Survey Methodology 2: 183–198.

Basu, Debabrata, 1971: An Essay on the Logical Foundations of Survey Sampling, Part One (with Discussion). S. 203–242 in: *V.P. Godambe* und *D.A. Sprott* (Hg.), Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston.

Bethlehem, Jelke, 1999: Weighting Adjustments for Ignorable Nonresponse. International Conference on Survey Nonresponse. Portland, OR.

Bethlehem, Jelke, 2002: Weighting Nonresponse Adjustments Based on Auxiliary Information. S. 275–288 in: *Robert Groves*, *Don Dillman*, *John Eltinge* und *Roderick Little* (Hg.), Survey Nonresponse. New York: Wiley.

Bethlehem, Jelke, und *Wouter Keller*, 1987: Linear Weighting of Sample Survey Data. Journal of Official Statistics 3: 141–153.

Bihler, Wolf, 2002: Das Stichprobenverfahren der Testerhebungen für einen registergestützten Zensus. Methodenbericht Heft 2. Statistisches Bundesamt: Wiesbaden.

Binder, David, und *Georgia Roberts*, 2003: Design-based and Model-based Methods for Estimating Model Parameters. S. 29–48 in: *Ray Chambers* und *Chris Skinner* (Hg.), Analysis of Survey Data. Chichester: J. Wiley.

Brewer, Ken, 2002: Combined Survey Sampling Inference. Weighing Basu's Elephants. London: Arnold.

Chambers, Ray, und *Chris Skinner*, 2003: Analysis of Survey Data. Chichester: J. Wiley.

Deming, William Edwards, und *Frederick Stephan*, 1941: On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Tables are Known. The Annals of Mathematical Statistics 11: 427–444.

Deville, Jean-Claude, und *Carl-Erik Särndal*, 1992: Calibration Estimators in Survey Sampling. Journal of the American Statistical Association 87: 376–382.

Deville, Jean-Claude, *Carl-Erik Särndal* und *Olivier Sautory*, 1993: Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88: 1013–1020.

DuMouchel, William H., und *Greg J. Duncan,* 1983. Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78: 535–543.

Gabler, Siegfried, und *Sabine Häder*, 1998: A Conditional Minimax Estimator for Treating Nonresponse. ZUMA-Nachrichten Spezial No 4: 334–349.

Gabler, Siegfried, 1992: ALLBUS-Baseline-Studie 1991 und ALLBUS 1992: Ost-West-Gewichtung der Daten. ZUMA-Nachrichten 35: 77–81.

Gabler, Siegfried, *Jürgen H.-P. Hoffmeyer-Zlotnik* und *Dagmar Krebs* (Hg.), 1994: Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Hájek, Jaroslav, 1981: Sampling from a Finite Population. New York: Marcel Dekker.

Kalton, Graham, und *Ismael Flores-Cervantes*, 2003: Weighting Methods. Journal of Official Statistics 19: 81–97.

Kish, Leslie, 1990: Weighting: Why, When, and How. American Statistical Association. 1990 Proceedings of the Survey Research Methods 121–130.

Kish, Leslie, 2003: Weighting for Unequal *Pi*. S. 179–197 in: *Graham Kalton* und *Steven Heeringa* (Hg.), Leslie Kish. Selected Papers. Chichester: J. Wiley.

Korn, Edward, und *Barry Graubard*, 1999: Analysis of Health Surveys. Chichester: J. Wiley.

Lessler, Judith, und *William Kalsbeek*, 1992: Nonsampling Error in Surveys. Chichester: J. Wiley.

Little, Roderick, und *Mei-Miau Wu*, 1991: Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ. Journal of the American Statistical Association 86: 87–95.

Lohr, Sharon L., 1999: Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press.

Lundström, Sixten, und *Carl-Erik Särndal*, 2002: Estimation in the Presence of Nonresponse and Frame Imperfections. Örebro: Statistics Sweden.

Lynn, Peter, Siegfried Gabler, Sabine Häder und *Seppo Laaksonen*, 2004: Methods for Achieving Equivalence of Samples in Cross-National Surveys. ISER Working Papers Number 2004-09. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-09.pdf

Nathan, Gad, und *Fred Smith*, 1989: The Effect of Selection on Regression Analysis. S. 149–163 in: *Chris Skinner*, *Tim Holt* und *Fred Smith* (Hg.), Analysis of Complex Surveys. Chichester: J. Wiley.

Pfeffermann, Danny, 1993: The Role of Sampling Weights when Modeling Survey Data. International Statistical Review 61: 317–337.

Rothe, Günter, 1993: Wie (un)wichtig sind Gewichtungen? Eine Untersuchung am ALLBUS 1986. S. 62–87 in: *Siegfried Gabler, Jürgen H.P. Hoffmeyer-Zlotnik* und *Dagmar Krebs* (Hg.), Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Rothe, Günter, und *Michael Wiedenbeck*, 1993: Stichprobengewichtung: Ist Repräsentativität machbar? S. 46–61 in: *Siegfried Gabler, Jürgen H.P. Hoffmeyer-Zlotnik* und *Dagmar Krebs* (Hg.), Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Särndal, *Carl-Erik*, *Bengt Swensson* und *Jan Wretman*, 1992: Model Assisted Survey Sampling. New York: Springer Verlag.

Schnell, Rainer, 1997: Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Opladen: Leske + Budrich.

White, Halbert, 1980: A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrika 48: 817–838.

Winship, Christopher, und *Larry Radbill*, 1994: Sampling Weights and Regression Analysis. Sociological Methods & Research 23: 230–257.

Zaslavsky, Alan, 1988: Representing Local Area Adjustments by Reweighting of Households. Statistics Canada 14: 265–288.

Zieschang, Kimberly, 1990: Sample Weighting Methods and Estimations of Totals in the Consumer Expenditure Survey. Journal of the American Statistical Association 85: 986–1001.

STICHPROBENZIEHUNG FÜR TELEFONUMFRAGEN IN DEUTSCHLAND

Sabine Häder und Axel Glemser

Zusammenfassung: Über 40 Prozent aller Interviews in der Marktforschung werden in Deutschland gegenwärtig telefonisch durchgeführt. Voraussetzung für eine hohe Qualität der Telefonumfragen ist die adäquate Generierung der Stichproben. In diesem Zusammenhang stellt sich die Frage nach einem geeigneten Auswahlrahmen für Telefonstichproben. Das Telefonbuch kann für die Stichprobenziehung nicht verwendet werden, weil es nicht vollständig ist. Zu viele Haushalte sind nicht eingetragen und unterscheiden sich in wesentlichen Merkmalen von den in das Telefonbuch eingetragenen. Im Folgenden werden Vor- und Nachteile alternativer Auswahlrahmen diskutiert und Designs für die Stichprobenbildung vorgestellt. In den letzten Jahren haben sich in Deutschland das Gabler-Häder-Design im akademischen Bereich und das ADM-Design in der kommerziellen Markt- und Sozialforschung als „state of the art" etabliert. Die fortschreitende Entwicklung im Mobilfunksektor stellt beide Stichprobensysteme vor neue Probleme, für die Lösungen gefunden werden müssen.

I. Einführung

Stichprobenerhebungen spielen im täglichen Leben eine bemerkenswerte Rolle. Speziell im Vorfeld von Wahlen finden z.B. die Prognosen von Meinungsforschungsinstituten in den Diskussionen um die Wahlaussichten der Parteien große Aufmerksamkeit. Aber auch Stimmungsbilder zu aktuellen politischen, kulturellen oder sportlichen Ereignissen werden in den Medien sehr zeitnah präsentiert. Bei den meisten dieser Erhebungen handelt es sich um Telefonbefragungen. Mit dieser Erhebungsmethode ist es den Umfrageinstituten möglich, innerhalb relativ kurzer Zeit große Datenmengen zu erzeugen. Dabei sind aber die Verfügbarkeit eines vollständigen Auswahlrahmens sowie die adäquate Stichprobenziehung und Schätzung entscheidend für die Qualität der Ergebnisse. Im Folgenden soll deshalb erklärt werden, wie gegenwärtig in Deutschland qualitativ hochwertige Telefonstichproben gebildet werden.

Nach einem Überblick über die Forschungsgeschichte der Telefonbefragungen in Deutschland werden die Vorteile derartiger Erhebungen diskutiert. Es folgt eine Beschreibung des deutschen Telefon-Nummerierungssystems. Damit wird die Grundlage für das Verständnis der sich anschließenden Darstellung verschiedener Verfahren für die Bildung von Zufallsstichproben gelegt. Abschließend wird ein spezielles Problem bei der Telefonstichprobengenerierung besprochen, das in letzter Zeit zunehmend an Bedeutung gewinnt. Der wachsende Anteil von Haushalten, die nur noch über ein Mobiltelefon erreichbar sind und auf einen herkömmlichen Festnetzanschluss verzichten, stellt die etablierten Stichprobensysteme vor neue Herausforderungen. Die damit

verbundenen Probleme werden dargestellt und die Schwierigkeiten möglicher Lösungswege diskutiert.

II. Geschichtliches

In den USA setzten sich Telefonbefragungen schon Ende der 1970er Jahre durch, als die Ausstattung der Haushalte mit Telefon die 90-Prozent-Marke überschritten hatte. In Westdeutschland erreichte die Telefondichte erst ein Jahrzehnt später dieses Niveau. Ende der 1980er Jahre konnte dann jedoch davon ausgegangen werden, dass zwischen Haushalten mit und ohne Telefon keine wesentlichen Unterschiede mehr bestehen. Zum Beispiel wurden einkommensschwache Haushalte durch spezielle Sozialtarife unterstützt, so dass diese nicht systematisch aus der Auswahlgesamtheit heraus fielen. Zugleich waren etwa 97 Prozent der Haushalte mit Telefonanschluss auch im Telefonbuch verzeichnet, da eine Pflicht zum Eintrag bestand. Nur in seltenen Fällen waren Ausnahmegenehmigungen zu erhalten. Diese Situation führte zu einem Anstieg des Anteils der Telefoninterviews an den Interviews, die insgesamt in der Markt- und Sozialforschung durchgeführt wurden. Bei TNS Infratest[1] erhöhte sich z.B. der Anteil der Telefoninterviews von 14 Prozent im Jahr 1983 auf 38 Prozent im Jahr 1989 (vgl. www.adm-ev.de).

Die deutsche Wiedervereinigung stoppte diesen Trend jedoch erst einmal. Da die Telefondichte in Ostdeutschland nach dem Mauerfall 1989 nur 17 Prozent betrug, waren gesamtdeutsche Telefonbefragungen nicht möglich, zumal sich Telefonhaushalte und Haushalte ohne Telefon in wesentlichen soziodemographischen Merkmalen systematisch unterschieden. Die Deutsche Post/Deutsche Telekom unternahm in der Folge große Anstrengungen, um die technischen Voraussetzungen für den Anschluss aller Haushalte an das Telefonnetz zu schaffen. Ende der 1990er Jahre war das gelungen, Ostdeutschland verfügt gegenwärtig über eines der modernsten Telefonnetze Europas. Folgerichtig stieg der Anteil der Telefoninterviews wieder an:

Tabelle 1: Quantitative Interviews der Mitgliedsinstitute des ADM nach Befragungsart (in Prozent)

	1990	1991	1995	1996	1997	2002	2003
Persönliche Interviews	65	60	60	45	44	33	28
Telefoninterviews	22	30	30	44	40	41	43
Schriftliche Interviews	13	10	10	11	16	21	19
Online-Interviews	–	–	–	–	–	5	10

Quelle: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM) e.V.

In der Zwischenzeit war allerdings ein anderes Problem aufgetreten, das dem Siegeszug der Telefonbefragungen in Deutschland Einhalt gebot. Im Jahr 1992 war die Pflicht zum Eintrag in das Telefonbuch aufgehoben worden. Insbesondere in Ostdeutschland

1 TNS: Abkürzung für den Marktforschungskonzern Taylor Nelson Sofres (London), dem Infratest seit 2003 angehört.

ließen sich viele der neu angeschlossenen Haushalte nicht registrieren. In Westdeutschland wurde diese Option häufig nach Umzügen und damit verbundenen Anschlussänderungen gewählt. Inhaber anonymer Anschlüsse waren z.B. jünger, lebten häufiger in Städten, hatten andere Parteipräferenzen und waren formal höher gebildet als Personen, die sich in das Telefonbuch hatten eintragen lassen (Häder 1996: 52ff.). 1997 wurde ein Anteil Nicht-Eingetragener von 10,6 Prozent in den alten und 18,7 Prozent in den neuen Bundesländern ermittelt (Follmer und Smid 1998: 48). Inzwischen dürfte der Anteil der „Nonpubs", d.h. der Anteil der nicht eingetragenen Festnetz-Telefonhaushalte an den Haushalten mit Festnetztelefon über 25 Prozent liegen. In urbanen Regionen sind die Anteile noch höher. Damit stellt sich das Problem eines adäquaten Auswahlrahmens für die Stichprobenziehung. Konnte für telefonische Befragungen bis Ende der 1980er Jahre das Telefonbuch als hinreichend vollständiger und aktueller Auswahlrahmen verwendet werden, ist dies seit einiger Zeit nicht mehr möglich. Dies gilt insbesondere auch deshalb, weil der Verdacht besteht, dass sich Telefonanschlussinhaber zum Teil noch in die gedruckten Telefonbücher eintragen lassen, aber der Veröffentlichung ihres Eintrages auf CD-ROMs, die die Grundlage für die Stichprobenziehung bilden (Deutschmann und Häder 2000: 77ff.), nicht zustimmen.

III. Vorteile von Telefonbefragungen

Dass Telefonumfragen inzwischen zum am häufigsten eingesetzten Befragungsmodus geworden sind, ist auf eine Reihe von allgemeinen Vorteilen zurückzuführen, die telefonische Befragungen im Vergleich zu den herkömmlichen persönlich-mündlichen oder schriftlichen Befragungen aufweisen:

Keine Klumpung. Insbesondere bei bundesweiten Umfragen, aber auch bei Erhebungen in einzelnen Bundesländern oder kleineren Regionen kann die Stichprobe so gezogen werden, dass sie keine Klumpung aufweist. Dies ist ein Vorteil zum Beispiel gegenüber den Einwohnermeldeamtsstichproben für persönlich-mündliche oder schriftliche Befragungen.

Legt man für die Berechnung des Design-Effektes auf Grund von Clusterung das folgende Modell (Kish 1987, siehe auch Gabler, Häder und Lahiri 1999) zugrunde,

$$\mathrm{deff}_c = 1 + (\bar{b} - 1)\rho$$

mit

$\bar{b}$ mittlere Zahl der Befragten pro Cluster;
ρ Intraklassenkorrelationskoeffizient,

ergibt sich, dass die Varianz des Punktschätzers gegenüber einer ungeklumpten Stichprobe in Abhängigkeit von der Homogenität in den Clustern und der mittleren Zahl von Befragten pro Cluster um den Faktor deff_c erhöht ist (vgl. Gabler und Häder 2000: 85f.). Das bedeutet, dass bei ungeklumpten Stichproben die Konfidenzintervalle um die Punktschätzer u.U. erheblich schmaler sind als bei geklumpten Stichproben gleichen Umfangs.

Viele Kontaktversuche. Das häufige Kontaktieren des ausgewählten Haushaltes bzw. der ausgewählten Person ist eine Voraussetzung für die Erhöhung der Response-Rate bei einer Befragung. Bei persönlich-mündlichen Befragungen ist jeder Kontaktversuch am Wohnort des zu Befragenden sehr kostspielig. Im Gegensatz dazu können Kontaktversuche via Telefon zu unterschiedlichen Tageszeiten (üblich im ADM: ab 17:00 Uhr und nicht später als 21:00 Uhr) und an unterschiedlichen Wochentagen vorgenommen werden. Computergestützte Terminverwaltungssysteme unterstützen die zeitoptimale Wiedervorlage der Nummern.

Anonyme Interviewsituation. Die Interviewsituation im Telefoninterview ist dadurch gekennzeichnet, dass kein visueller Kontakt zwischen Interviewer und Befragungsperson besteht. Das kann bei sensiblen bzw. heiklen Fragen dazu führen, dass ehrlicher und weniger sozial erwünscht geantwortet wird.

CATI (computer aided telephone interview). Die mittlerweile eingesetzte Hard- und Software für Telefonumfragen erlaubt eine Vielzahl von Funktionen, die die Durchführung von Umfragen erleichtern, z.B. komplexe Filterführungen, Konsistenzchecks und die Verwaltung der Nummern mit Hilfe von Sample-Management-Systemen. Dadurch, dass die Antworten direkt bei der Befragung auf maschinenlesbare Datenträger eingegeben werden, liegen die Daten relativ schnell nach Ende der Feldzeit vor.

Gute Interviewerkontrolle. Telefonbefragungen finden im Allgemeinen in Telefonstudios statt, in denen ein Supervisor zum einen für Rückfragen zur Verfügung steht, zum anderen die ordnungsgemäße Durchführung der Interviews kontrolliert. Eine derartige Kontrolle ist bei keinem anderen Befragungsmodus möglich.

Relativ preiswert. Verglichen mit persönlich-mündlichen Befragungen ist der Preis pro Interview bei telefonischen Befragungen niedriger, da Kostenbestandteile wie z.B. die Wegekosten entfallen.

Diese genannten Vorteile haben zu einer großen Akzeptanz telefonischer Befragungen unter Umfrageforschern geführt.

IV. Nummerierungssystem

Um zu verstehen, wie schwierig sich die Stichprobenziehung für Telefonumfragen unter den Bedingungen in Deutschland gestaltet, ist es notwendig, sich kurz mit dem Nummerierungssystem im Festnetzbereich zu beschäftigen.

Es existieren 5.200 Ortsnetzbereiche, die Deutschland vollständig abdecken. Die Grenzen der Ortsnetzbereiche sind allerdings nicht mit Gemeindegrenzen oder anderen administrativen Grenzen kongruent. Dies erschwert z.B. eine Schichtung nach Bundesländern, wenn diese eigentlich sinnvoll wäre, weil die zu erhebenden Variablen eine regionale Abhängigkeit vermuten lassen. Beispielsweise deckt der Ortsnetzbereich 0621 die Gemeinden Mannheim, Ilvesheim, Edingen-Neckarhausen und Ludwigshafen am Rhein ab. Die ersten drei Gemeinden liegen im Bundesland Baden-Württemberg, die letzte in Rheinland-Pfalz. Außerdem werden viele Gemeinden von mehreren Ortsnetzbereichen geschnitten. So gibt es z.B. in der Gemeinde Neustadt an der Weinstraße Rufnummern mit den Ortsnetzkennzahlen 06321 und 06327.

Jedem Ortsnetz ist eine Kennzahl (Ortsnetzkennzahl) zugeordnet. Diese beginnen nach einer führenden Null mit einer Ziffer zwischen 2 und 9. Die Ortsnetzkennzahlen haben eine Länge von zwei bis fünf Stellen (ohne die führende Null). Sie erlauben einen Rückschluss auf die geographische Lage des Anschlusses (vgl. www.regtp.de); die Länge der Ortsnetzkennzahlen zusätzlich auf die Größe des Ortsnetzbereichs.

Die Rufnummern hatten bislang zwischen drei und neun Stellen. Aus der Länge der Rufnummern konnte man Hinweise auf das Alter und den Urbanisierungsgrad der Region des Telefonanschlusses gewinnen: Je älter die Nummer und je ländlicher die Region, desto kürzer war die Nummer. Die Länge von Vorwahl und Rufnummer variierte von 7 bis 11 Stellen.

Da in einigen Ortsnetzbereichen allerdings inzwischen eine Rufnummernknappheit eingetreten ist, werden dort seit Mai 2003 elfstellige Rufnummern vergeben. Dies gilt u.a. für eine Vielzahl ostdeutscher sehr kleiner Ortsnetzbereiche mit fünf Stellen (z.B. 033762 Zeuthen). Künftig wird die Beziehung zwischen der Länge der Rufnummer und dem Urbanisierungsgrad demnach nicht mehr ohne weiteres erkennbar sein.

Innerhalb eines Ortsnetzbereiches variiert die Länge der Telefonnummern also. Das hat zur Folge, dass bestimmte Rufnummern – technisch bedingt – nicht existieren können. Ist z.B. im Ortsnetzbereich 06321 die Rufnummer 3370 vergeben, kann es keine längere Nummer mit den gleichen Anfangsziffern geben, also z.B. sind die Nummern 3370x, 3370xx usw. ausgeschlossen. Das ist ein Grund für die großen Lücken, die in der Aufeinanderfolge der Rufnummern innerhalb eines Ortsnetzes bestehen.

An eine Standardisierung der Länge der Rufnummern, wie sie beispielsweise in der Schweiz oder in Großbritannien vorgenommen wurde, wird lt. Auskunft der Regulierungsbehörde für Telekommunikation und Post (RegTP) derzeit nicht gedacht (vgl. Hofmann 1998: 39).

Abbildung 1: Blockstruktur in einem Ortsnetzbereich

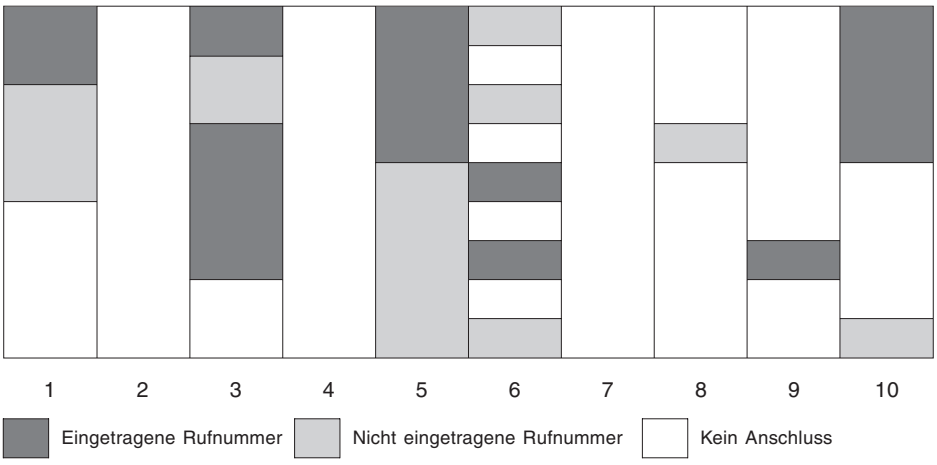


Abbildung 1 veranschaulicht die Anordnung der Rufnummern innerhalb eines Ortsnetzbereiches. Stellt man sich die Rufnummern statt auf einem Zahlenstrahl in Blöcken gleicher Länge angeordnet vor, ergibt sich folgendes Bild: In einigen Blöcken gibt

es eingetragene und nicht eingetragene Rufnummern sowie „Lücken", d.h. (noch) nicht angeschlossene Rufnummern (Blöcke 1; 3; 6; 10). Im Block 5 sind bereits alle Rufnummern vergeben, und zwar zu gleichen Teilen mit eingetragenen und nicht eingetragenen Rufnummern. Darüber hinaus gibt es die erwähnten leeren Blöcke (Blöcke 2; 4; 7) und Blöcke, die erst sehr gering besetzt sind. In Block 8 gibt es lediglich eine nicht eingetragene Nummer, in Block 9 eine eingetragene Nummer.

Die Kenntnis dieser Anordnung der Rufnummern in den Ortsnetzbereichen ist eine Voraussetzung für das Verständnis der in Abschnitt V vorgestellten Stichprobendesigns.

V. Stichprobendesigns

Zu Beginn jeder Stichprobenerhebung – also auch der für eine telefonische Umfrage – ist zunächst eine exakte Abgrenzung der Grundgesamtheit notwendig. Das bedeutet, dass eine sachliche, räumliche und zeitliche Definition der Menge von Elementen (Zielpersonen) vorgenommen werden muss, für die die aus der Stichprobe gewonnenen Aussagen gültig sein sollen (Grundgesamtheit oder Population).

Danach ist der Umfang der Grundgesamtheit abzuschätzen, um vorhersagen zu können, wie viele Einheiten (z.B. Haushalte) voraussichtlich kontaktiert werden müssen, um überhaupt eine bestimmte Anzahl von Zielpersonen zu erreichen. Sollte z.B. eine bundesweite Befragung zur Europawahl durchgeführt werden, lässt sich der Umfang der Grundgesamtheit aus Angaben des Statistischen Bundesamtes relativ einfach abschätzen. Zur Europawahl 1999 waren 60,6 Mio. Bürger wahlberechtigt, der in der Studie interessierende Anteil an der Gesamtbevölkerung ist also „groß". Außer in der kleinen Zahl von Haushalten, in denen ausschließlich nicht wahlberechtigte Ausländer oder Personen unter 18 Jahren wohnen, wird man in jedem Privathaushalt potenziell wenigstens eine Zielperson ermitteln können.

Wenn die interessierenden Segmente in der Gesamtbevölkerung allerdings relativ klein sind, ist die Erhebung mit einem wesentlich höheren zeitlichen und finanziellen Aufwand verbunden, denn die zur Grundgesamtheit gehörenden Haushalte (in denen wenigstens eine Zielperson lebt) müssen erst durch ein Screening ermittelt werden. So sollten beispielsweise vor einigen Jahren in einer Telefonbefragung Haushalte, in denen Kinder im Alter von maximal acht Jahren leben, befragt werden. Der entsprechende Anteil von Haushalten mit mindestens einem Kind in diesem Alter betrug zu dieser Zeit in Westdeutschland etwa 13 Prozent, in Ostdeutschland ca. 11 Prozent. Es war also von einer längeren Feldzeit und höheren Kosten auszugehen, als sie für eine allgemeine Bevölkerungsbefragung üblich sind. Dies muss bei der Planung einer solchen Studie unbedingt berücksichtigt werden.

Nachdem die Grundgesamtheit definiert ist, auf die sich die Untersuchung beziehen soll, stellt sich die entscheidende Frage, ob ein geeigneter Auswahlrahmen existiert, d.h. ob es eine akzeptable Aufstellung mit allen Elementen der Grundgesamtheit gibt. Wenn dies der Fall ist, erleichtert das die Stichprobenziehung erheblich; wenn nicht, müssen Umwege in Kauf genommen werden.

Ideal wäre eine Liste mit folgenden Eigenschaften:
- Auf der Liste sind alle Zielpersonen mit ihren Telefonnummern einmal erfasst, d.h. es fehlen keine Personen und es sind keine Personen mehrfach aufgelistet.
- Die Liste ist hinreichend aktuell, d.h. es kann davon ausgegangen werden, dass die Zielpersonen tatsächlich (noch) über die angegebenen Rufnummern erreichbar sind.
- Die Liste ist für die Erhebung zugänglich und ihre Nutzung nicht zu teuer.
- Die Liste enthält Zusatzinformationen für eine mögliche Schichtung, d.h. Merkmale, die im Zusammenhang mit Untersuchungsmerkmalen stehen dürften.

Wie schon erwähnt, war das Telefonbuch in Deutschland bis 1992 ein akzeptabler Auswahlrahmen für telefonische Bevölkerungsbefragungen (in Westdeutschland). In den letzten Jahren ist aber der Anteil sogenannter „Nonpubs" stark angestiegen. Da sich Eingetragene und Nicht-Eingetragene in wesentlichen sozio-demographischen Merkmalen unterscheiden, würde die Ziehung aus dem Telefonbuch in der Verteilung dieser und davon abhängiger Merkmale zu verzerrten Stichproben führen. Deshalb ist das Telefonbuch mittlerweile für die Stichprobenziehung bei Befragungen der allgemeinen Bevölkerung der Bundesrepublik oder auch einzelner Gemeinden nicht mehr geeignet. Dies bedeutet aber nicht, dass es überhaupt keine Verwendung bei der Stichprobenplanung mehr finden kann. Für die Konstruktion von Migrantenstichproben (vgl. Humpert und Schneiderheinze 2000) oder die Auswahl von Arbeitsstätten oder Institutionen ist es weiterhin wertvoll.

Mitunter existieren Listen, die Grundgesamtheiten in dem Sinn „akzeptabel" abbilden, dass der Unterschied zwischen der Grundgesamtheit und der sich aus der Liste ergebenden Auswahlgesamtheit vernachlässigt werden kann. Dies kann z.B. für Mitglieder bestimmter Vereinigungen, Organisationen oder Institutionen sowie Kunden von Firmen der Fall sein. Derartige Listen können selbstverständlich als Grundlage für die Stichprobenziehung genutzt werden.

Weitaus häufiger tritt der Fall ein, dass für die Grundgesamtheit einer geplanten Studie keine Liste vorliegt, auf der alle potenziellen Zielpersonen mit Namen und Telefonnummern verzeichnet sind. Hier ergibt sich die Frage, mit welchen Methoden trotzdem Stichproben generiert werden können, die (möglichst) allen Telefonnummern von zur Grundgesamtheit gehörenden Einheiten eine positive Auswahlchance geben.

1. Random Digit Dialling

Um sowohl eingetragene als auch nicht eingetragene Telefonnummern in die Stichprobe einzuschließen, wäre es zunächst denkbar, das aus den USA bekannte Random Digit Dialling – die einfache Zufallsziffernanwahl – anzuwenden. Dazu müsste für alle 5.200 (oder auch nur für interessierende ausgewählte) Ortnetzbereiche in Deutschland jeweils die Spannweite zwischen der niedrigsten und der höchsten eingetragenen Telefonnummer bestimmt werden. Diese würde die Menge der als Telefonnummern denkbaren Ziffernfolgen umfassen, aus denen die Stichprobe zu ziehen wäre. Dazu ein Beispiel: Für den Ortsnetzbereich „06321" ist die niedrigste Nummer eines privaten Haushalts 2001, die höchste 6790152. Insgesamt sind für diesen Ortsnetzbereich gegenwärtig etwa 21.000 Nummern im Telefonbuch veröffentlicht. Wenn man nun eine

Stichprobe aus der bestehenden Spannweite zöge, dann würde man nur in 0,3 Prozent der Fälle auf eine eingetragene Nummer treffen. Selbst wenn ein Anteil von nicht eingetragenen Nummern von 25 Prozent (gegenwärtiger Bundesdurchschnitt) unterstellt wird, läge die Trefferquote („Hitrate") unter einem Prozent. Damit dürfte klar sein: Dieser Ansatz ist in Deutschland auf Grund des komplizierten Nummernsystems nicht praktikabel, da er entschieden zu zeit- und kostenintensiv ist. Zudem existiert ein weiteres Problem: Zwar werden die anzuwählenden Ziffernfolgen mit gleichen Wahrscheinlichkeiten ausgewählt. Bei der praktischen Umsetzung haben aber – bedingt durch das deutsche Nummerierungssystem mit seinen unterschiedlichen Rufnummernlängen – nicht alle Haushalte die gleiche Auswahlwahrscheinlichkeit: Die Inhaber kürzerer Nummern haben eine höhere Chance, in die Stichprobe zu gelangen, als die Telefonhaushalte mit längeren Nummern.

2. Randomize Last Digits

Eine andere, ebenfalls in den USA entwickelte Lösung des Problems der Einbeziehung von Nummern auch nicht eingetragener Telefonteilnehmer ist das sogenannte „Randomize Last Digits". Hierbei werden Nummern aus dem Telefonbuch gezogen und anschließend ihre letzten beiden Stellen durch zufällig erzeugte Ziffern ersetzt. Bei diesem Vorgehen sind allerdings die Inklusionswahrscheinlichkeiten für die anzuwählenden Ziffernfolgen nicht gleich. Sie hängen davon ab, wie viele weitere Nummern im gleichen 100er Block eingetragen sind. Ein 100er Block ist definiert als der Stamm einer Telefonnummer, der nach dem Abschneiden der beiden letzten Stellen verbleibt. Er umfasst die Menge aller unterschiedlichen Ziffernfolgen, die sich durch die zufällige Ersetzung dieser letzten beiden Ziffern generieren lassen – also genau 100. So liegt z.B. die Telefonnummer 5129815 im Block 51298xx, der alle Ziffernfolgen von 5129800 bis 5129899 beinhaltet. Um die ungleichen Inklusionswahrscheinlichkeiten beim Randomize Last Digits wieder auszugleichen, müssten die in den Interviews erhaltenen Datensätze im nachhinein jeweils mit der Inversen der Eintragsdichte, d.h. der Inversen der Zahl eingetragener Nummern in diesem Block, gewichtet werden, was aufwändig wäre. So würde ein Datensatz, der aus einem Interview mit einem Teilnehmer stammt, in dessen Block noch 40 weitere eingetragene Nummern stehen, mit dem Faktor 1/40 multipliziert werden. Ein anderer Teilnehmer, in dessen Block nur zwei eingetragene Nummern vorkommen, erhielte dagegen den Faktor 0,5.

3. Gabler-Häder-Design

Am Zentrum für Umfragen, Methoden und Analysen (ZUMA) wurde Mitte der 1990er Jahre von Gabler und Häder ein Ansatz entwickelt, der eingetragenen und nicht eingetragenen Telefonnummern die gleiche Chance gibt, in die Stichprobe zu gelangen (vgl. Häder und Gabler 1997). Dieser soll im Folgenden kurz vorgestellt werden:

– Für jeden Ortsnetzbereich ist es möglich, genau die 100er Blöcke festzustellen, in denen wenigstens eine eingetragene Telefonnummer vorkommt. In *Abbildung 1* sind das die Blöcke 1, 3, 5, 6, 9 und 10.
– Wie bereits erwähnt, bestehen zwischen diesen 100er Blöcken häufig große Lücken. Diese sind der wesentliche Grund für die nur geringe Effizienz bei der Anwendung von reinen Zufallsziffernanwahlen. Wir gehen davon aus, dass in diesen Lücken keine nicht eingetragenen Nummern vorkommen. Eine Situation wie im Block 8 der *Abbildung 1* sollte in der Realität nur äußerst selten vorkommen.
– Für jeden der mit mindestens einer eingetragenen Nummer besetzten 100er Blöcke werden anschließend alle möglichen Ziffernfolgen generiert (bei K belegten Blöcken also insgesamt K × 100 Ziffernfolgen, im Beispiel in *Abbildung 1* wären es 6 × 100 = 600 Ziffernfolgen).

Aus dieser Menge wird schließlich (z.B. uneingeschränkt zufällig oder geschichtet) eine vorher bestimmte Zahl von Ziffernfolgen gezogen.

Der Auswahlrahmen, den ZUMA für die akademische Sozialforschung bereitstellt, hat sich über die Zeit wie folgt entwickelt:

Tabelle 2: Entwicklung der Zahl der Blöcke, der Zahl der eingetragenen Nummern sowie der Eintragsdichte

Datum	Zahl der Blöcke	Eingetragene Nummern	Eintragsdichte
April 1999	647.636	30.654.324	47,33
April 2000	721.146	30.970.423	42,95
Oktober 2000	757.092	30.635.916	40,47
April 2001	805.447	30.560.982	37,94
Oktober 2001	844.745	30.341.964	35,92
Juli 2002	904.748	29.937.636	33,09
Januar 2003	946.616	29.833.453	31,52
Juli 2003	975.656	29.542.697	30,28
Januar 2004	1.015.461	30.452.423	29,99
Juli 2004	1.037.791	28.863.646	27,81

Quelle: Gabler und Schürle (2002), eigene Berechnungen.

Bei nur unwesentlich verringerter Zahl der eingetragenen Rufnummern hat sich die Zahl der Blöcke erheblich vergrößert. Diese sinkende Eintragsdichte mag mit der beginnenden Nutzung von freien Rufnummernblöcken durch Telekom-Konkurrenten zusammenhängen. Da die Erhöhung der Zahl der Blöcke nicht durch eine im gleichen Maße ansteigende Zahl nicht eingetragener Nummern begleitet wird, verringert sich die Trefferquote (Quotient aus Zahl der Nummern, die einen Anschluss repräsentieren, zur Zahl der insgesamt angewählten Ziffernfolgen) beim Gabler-Häder-Design. Dies ist gleichbedeutend mit einer sinkenden Effizienz, da mehr Anwahlversuche pro Anschluss notwendig werden. Für größere Umfrageinstitute, die über Autodialer verfügen, ist das kein großes Problem, wohl aber für kleine Telefonlabors insbesondere an Universitäten, die sich diese Technik nicht leisten können.

4. ADM-Design

Die „Arbeitsgemeinschaft Telefonstichproben“ im ADM (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.) befasst sich mit methodischen und praktischen Fragen, die bei der Anlage und Durchführung von telefonischen Befragungen in der Institutsmarktforschung auftreten. Ziel ist die Optimierung von Methoden und die Festlegung bzw. Kontrolle der gemeinsamen Standards. Zu den Aufgaben des Gremiums gehört auch die Erstellung und Aktualisierung der ADM-Telefonauswahlgrundlage und der damit verbundenen Basisdateien.[2]

Das ADM-Telefonmastersample entspricht im Hinblick auf den Randomisierungsalgorithmus eingetragener Nummern dem Gabler-Häder-Design (s.o.). Das heißt, auch beim ADM-Design erfolgt die Entwicklung des Auswahlrahmens (Heckel 2002) nicht als Teilschritt der Stichprobenziehung bei einem singulären Projekt, sondern bereits im vorhinein. Um von einer ADM-Stichprobe sprechen zu können, müssen zusätzlich zur Ziehung der Stichprobe aus der ADM-Auswahlgrundlage weitere Bedingungen erfüllt sein, denn das ADM-Design zeichnet sich darüber hinaus durch ein spezielles Stichprobenmodell aus (von der Heyde 2002). Eine ADM-Telefonstichprobe ist als mehrfach stratifizierte Zufallsauswahl konzipiert, denn durch Schichtung kann die Präzision der Stichprobe verbessert werden, indem die Streuung im Vergleich zur reinen Zufallsauswahl reduziert wird.

ADM-Telefonstichproben lassen sich durch zwei Auswahlebenen charakterisieren. Zum einen sind sie als stratifizierte Haushaltsstichproben auf Flächenbasis konzipiert, zum andern kann innerhalb des gewählten Haushalts im Zuge der Stichprobenrealisation zufällig eine Person ausgewählt werden. Bei TNS Infratest wird die Zielperson mit dem Schwedenschlüssel (Kish-Methode) ermittelt (vgl. hierzu *Abschnitt VI*).

Zur Bildung einer haushaltsproportionalen Flächenstichprobe sind zwei Bedingungen zu erfüllen. Zum einen müssen möglichst kleinräumig Angaben über die Anzahl der Haushalte verfügbar sein und zweitens müssen die Telefonnummern mit diesem Flächenraster kompatibel gemacht, also einer räumlichen Verortung unterzogen werden. Die erste Bedingung ist notwendige Voraussetzung zur Berechnung der Schichtungsmatrix und der Allokation. Die zweite Bedingung ermöglicht es, die berechneten Anzahlen zu ziehender Rufnummern in den jeweils betreffenden Einheiten des Flächensystems per Zufall auszuwählen.

Haushaltsangaben werden für das ADM-System über die sogenannte Gemeindedatei verfügbar gemacht. Dieser Datenbestand, der jährlich aktualisiert wird, verzeichnet für alle rechtlich selbständigen Städte und Gemeinden anhand der amtlichen Gemeindekennziffer (GKZ)

– die Zahl der Haushalte insgesamt,
– die Zahl der Haushalte mit deutscher Bezugsperson,
– die Anzahl der Wohnbevölkerung,
– die Anzahl deutscher Bevölkerung,

2 Im ADM sind derzeit 47 Institute zusammengeschlossen. 14 davon sind Mitglieder in der Arbeitsgemeinschaft Face-to-Face-Strichproben, 22 Institute sind in der Arbeitsgemeinschaft Telefonstichproben.

– BIK-Gemeindetypen sowie weitere Angaben.[3]

Mit der Gemeindedatei liegt damit ein äußerst flexibles Instrumentarium zur Schichtung und Allokation vor. Kleinste räumliche Einheit ist die GKZ, die daher auch als Verbindungsglied zu den Telefonnummern fungiert.

Damit die Auswahl der Rufnummern (technisch) gelingt, müssen diese den Schichtzellen über ihren räumlichen Bezug zugeordnet werden können – d.h. es wird für jede Rufnummer die GKZ benötigt. In den verfügbaren Telefonverzeichnissen gibt es zunächst jedoch keine GKZ, sondern zu den Telefonnummern nur postalische Ortsangaben. Dies bedeutet, dass zu den Einträgen mittels Postleitzahl und Ort die eindeutige GKZ zugeordnet werden muss. Den generierten Rufnummern fehlen naturgemäß solche Angaben.

Daher wird bei der Erstellung der Auswahlgrundlage in einem ersten Schritt für jeden 100er-Block separat die Verteilung der GKZs der eingetragenen Rufnummern ermittelt. Im zweiten Schritt wird dieses Verteilungsmuster auf die generierten Nummern im 100er-Block übernommen, indem jeder künstlich erzeugten Rufnummer bis zu drei verschiedene GKZs aus der Menge der GKZs der Einträge zugewiesen werden. Zudem wird zur jeweiligen GKZ die Wahrscheinlichkeit (W1, W2, W3) aus der Verteilung der Einträge vermerkt (siehe *Abbildung 2*).

Dieses Datenmaterial bildet die Grundlage für die eindeutige Verortung der Rufnummern, die – als dritter Schritt – im Rahmen der hausinternen Implementierung des Mastersamples bei den Instituten geschehen muss. Bei TNS Infratest wird hierbei für jede generierte Rufnummer zufallsproportional eine GKZ der ausgewiesenen Auswahlmenge gezogen. Dies bedeutet im Ergebnis, dass pro 100er-Block die GKZ-Verteilung der erzeugten Telefonnummern der GKZ-Verteilung der Einträge entspricht.[4]

Das Vorgehen bei einer Stichprobenziehung lässt sich in vier Teilschritte untergliedern: Auf Basis dieses Gemeindedatenbestandes ist erstens das Befragungsgebiet zu definieren (äußere Grenzen), zweitens die Stratifikation vorzunehmen (innere Differenzierung) und drittens die Allokationsrechnung durchzuführen. In der Regel wird die Allokation mit der Angabe „Haushalte insgesamt" durchgeführt. Ergebnis der Allokationsrechnung ist neben dem Nettosoll – also der Anzahl zu realisierender Interviews pro Schichtzelle – auch die Zahl der benötigten Telefonnummern pro Schichtzelle. Im vierten Schritt sind diese als Stichprobenbrutto auszuwählen.

---

3 Da es keine amtliche Haushaltsfortschreibung auf Ebene der Gemeinden gibt, wird hier die Haushaltsschätzung von BIK (Kurzform für das Marktforschungsunternehmen BIK Aschpurwis + Behrens GmbH, Hamburg, zum Haushaltsschätzmodell siehe: http://www.bik-gmbh.de/texte/haushalt2002.pdf (06.02.2004) verwendet. Es handelt sich hierbei um ein Gemeindeklassifizierungsmodell, das neben der Ortsgröße auch die räumliche Verflechtung in die Betrachtung mit einbezieht. Der Ansatz basiert auf den grundlegenden Arbeiten von Boustedt und wurde für Deutschland von BIK fortgeschrieben und weiterentwickelt; siehe: http://www.bik-gmbh.de/produkte/data/regionen.html sowie
http://www.bik-gmbh.de/texte/BIK-Regionen2000.pdf (06.02.2004).

4 Für 13 Großstädte Deutschlands liegt zudem eine Stadtteilkennung für die Rufnummern vor: Die Vergabe des Stadtteilkennzeichens für die generierten Rufnummern erfolgt analog der GKZ-Verortung. Alternative Verfahrensweisen der eindeutigen Ortszuweisung der generierten Rufnummern sind möglich. Dieser Realisierungsschritt obliegt den einzelnen Beziehern der ADM-Auswahlgrundlage.

Abbildung 2: Verortung generierter Telefonnummern

Basis:

Block: 089/8411300 ... 089/8411399
Einträge: 40 Nummern
davon mit GKZ: 09730000 (20 St.); 09751123 (12 St.); 09751456 (8 St.)

Datensatz für eine generierte Nummer:

Telnr.	GKZ1	W1	GKZ2	W2	GKZ3	W3
089/8411302	09730000	0,50	09751123	0,30	09751456	0,20

Die äußere Abgrenzung des Befragungsgebietes einer konkreten Studie sollte mittels amtlicher Gebietseinheiten definierbar sein. Bundesländer, Nielsengebiete,[5] Regierungsbezirke und Landkreise sind hier also unproblematische Einheiten. Auch kleinräumigere Gebiete wie bspw. listenweise vorgegebene Städte und Gemeinden sind selektierbar. Ballungsräume (z.B. Berlin und „Speckgürtel") sind über die BIK-Konzeption operationalisierbar und liegen vor. Wesentlich mühevoller sind Befragungsgebiete festzulegen, die vom Auftraggeber mit Hilfe von Postleitzahlenlisten vorgegeben werden. Da die postalischen Gebietsraster in Deutschland nicht deckungsgleich mit der amtlichen Gebietssystematik sind, muss in einem solchen Fall vorab die Postleitzahlenliste in eine Gemeindeliste umgewandelt werden.

Für die Stratifikation der Haushaltsstichprobe gilt im wesentlichen das Gleiche wie für die äußere Abgrenzung: Die amtlichen Gebietseinheiten bilden die Basis zur inneren Flächendifferenzierung. Darüber hinaus werden die BIK-Typen zur Stratifikation herangezogen. Damit lässt sich bspw. für eine bundesweite Befragung eine mehrdimensionale Schichtungsmatrix aufspannen (Bundesland × Regierungsbezirk × BIK-Typen). Die Wahl des Schichtmodells ist im wesentlichen von zwei Aspekten abhängig: Müssen in der Stichprobe Teilgebiete abgebildet werden, so sind diese auch in der Schichtung wiederzugeben. Die Anzahl der Schichtungszellen sollte im Hinblick auf die Fallzahl so gewählt werden, dass möglichst wenig Zellen durch die Allokation auf Null berechnet werden.

Die Allokationsrechnung wird mit dem Cox-Verfahren durchgeführt (vgl. Cox 1987). Dabei wird zunächst die Haushaltsverteilung der Matrix in Rändern und Zellen auf die Anzahl der gewünschten Rufnummern als Eckwert heruntergebrochen. Der ganzzahlige Anteil des berechneten Wertes stellt die jeweilige Mindestbesetzung dar. Die verbleibenden Nachkommaanteile werden in einem iterativen Prozess solange in Rändern und Zellen gerundet, bis die Summen der gerundeten Zellenverteilung den gerundeten Randverteilungen entsprechen.

Das ermittelte Schichtsoll wird per Zufallsauswahl auf die jeweils schichtzugehörigen Gemeinden verteilt, wobei die Anzahl der Haushalte als Bedeutungsgewicht in diese Ziehung einfließt. Im Ergebnis liegt so für jede Gemeinde die Zahl zu ziehender Rufnummern vor, die anschließend per einfacher Zufallsauswahl aus dem Mastersam-

5 Zusammenfassung der Bundesländer zu 7 Gruppen, die in Bezug auf die Bevölkerung in etwa gleich gewichtig sind. Der Begriff geht zurück auf das Marktforschungsunternehmen A.C. Nielsen.

ple gezogen werden kann. Bei dieser Ziehung werden gesperrte Rufnummern, die etwa für ein anderes Projekt bereits ausgewählt worden waren, nicht einbezogen.

Die Bruttostichprobe liegt somit als reine Haushaltsstichprobe vor. Die zweite Auswahlstufe – die Personenstichprobe – wird, wie erwähnt, im Rahmen der Stichprobenrealisierung zu Beginn des Interviews umgesetzt.

In den Fragenkatalog des Interviews müssen darüber hinaus unabhängig vom Gegenstand und Thema der Studie drei designbedingte Abfragen eingebunden werden, die sich als Konsequenz aus dieser Stichprobenanlage ergeben:

1. Der Wohnort des Befragten sollte abgefragt werden, denn für einen Teil der generierten Rufnummern ist im Rahmen der Auswahl die falsche GKZ zugeordnet worden. Wichtig bei der Aktualisierung dieser Information ist, dass sich hierdurch auch eine Änderung in der Schichtzuordnung des Interviews ergeben kann. Solche „Zellenwechsler" sind in der Steuerungsmatrix für die Realisierung der Stichproben zu berücksichtigen.
2. Im Rahmen der Zielpersonenauswahl ist bereits die Größe der Auswahlgesamtheit innerhalb des Haushalts bekannt; i.d.R. ist das die Anzahl der Personen ab 14 Jahren im Haushalt. Dieser Wert muss in den Befragungsdaten gespeichert werden, da dies eine gewichtungsrelevante Größe ist. Die Auswahlwahrscheinlichkeit der Personen hängt von diesem Wert ab und daher bildet sein Kehrwert ein Designgewicht.
3. Das zweite Designgewicht, das im Interview ermittelt werden muss, ist die Anzahl der Festnetztelefonnummern, unter denen der Haushalt zu erreichen ist. Dieser Faktor führt auf Haushaltsebene zu unterschiedlichen Auswahlwahrscheinlichkeiten, die ex post per Gewichtung zu korrigieren sind. Bei der Abfrage gilt es allerdings zu beachten, dass in die Berechnung nur solche Rufnummern einbezogen werden dürfen, die tatsächlich zu einem Telefonapparat schalten. Nummern, die exklusiv für Faxgeräte oder Modemverbindungen genutzt werden, sind hier nicht mitzuzählen, da sie im Rahmen der Stichprobenrealisation erkannt und automatisch ausgesteuert werden.

Das ADM-Design zeichnet sich also durch mehr aus, als lediglich das Ziehen von Stichproben aus der ADM-Auswahlgrundlage. Es wird der Ansatz einer stratifizierten Haushaltsstichprobe auf Flächenbasis verfolgt, der im Zuge der Realisierung der Interviews seine Komplettierung als Personenstichprobe erfährt. Dieses Vorgehen erfordert, dass die erhobenen Befragungsdaten per Gewichtung justiert werden, damit die Parameter der Grundgesamtheit korrekt geschätzt werden können.[6]

5. Auswahl der Zielperson im Haushalt

Handelt es sich bei der Befragung um die Erfassung personenbezogener Merkmale, muss nach der Ziehung der Ziffernfolgen und deren Anwahl die Zielperson im Haushalt bestimmt werden.

Die Phase der Kontaktaufnahme und insbesondere die Auswahl der zu Interviewenden ist entscheidend für die weitere Teilnahme- bzw. Vermittlungsbereitschaft der Per-

6 Siehe als praktisches Beispiel für eine Studienanlage Glemser (2002).

son, die den Anruf entgegennimmt. Während dieses Prozesses treten erfahrungsgemäß sehr häufig Verweigerungen auf, da entweder das Anliegen der Auswahlprozedur nicht verstanden wird oder keine Auskunftswilligkeit über die Struktur des Haushaltes besteht (vgl. Frey et al. 1990; Forsman und Berg 1992; Hagan und Meier Collier 1983).

Deshalb wäre es aus forschungsökonomischer Sicht am günstigsten, nach erfolgter Kontaktaufnahme möglichst wenige Informationen über demographische Merkmale aller Haushaltsmitglieder vom Gesprächspartner abzufordern. Dem widerspricht jedoch die statistische Forderung, dass jede im Haushalt lebende Person eine Chance haben muss, als Respondent ausgewählt zu werden (sofern sie zur Grundgesamtheit gehört).

In der Literatur werden insbesondere die folgenden Auswahlprozeduren diskutiert: a) Kish-Methode, b) Birthday-Methode.

Zu a) Die Kish-Methode (Kish 1949) gilt als methodisch sauberstes Verfahren zur Ermittlung der Zielperson im Haushalt. „The most rigorous approach requires a full listing of all adults in the household by age and sex. A random selection of one person from such a listing provides an equal non-zero probability of selection to all adult household members" (Hagan und Meier Collier 1983: 548). Die Kish-Methode erfordert also aufwändige Angaben über alle im Haushalt lebenden Personen und ist deshalb nach oben genannten forschungsökonomischen Gesichtspunkten nachteilig: „Die Kish-Technik der Trennung von Befragtenauswahl und Befragung führt auf Grund des sehr differenzierten Auswahlverfahrens leicht zu einem Abbruch der Haushaltskontakte; nur sehr erfahrene und sorgfältig arbeitende Interviewer werden diese Auswahl korrekt durchführen können; verglichen mit ähnlichen Verfahren sind die Kosten sehr hoch" (Frey et al. 1990: 76).

Um denkbaren Abbrüchen bei der Zielpersonenauswahl entgegenzuwirken, kann bei der Umsetzung dieses Verfahrens für CATI-Erhebungen auch eine modifizierte und weniger zeitintensive Variante der Kish-Methode angewandt werden. Dabei wird nicht die komplette Zielpersonenstruktur nach Alter und Geschlecht des Haushalts erhoben, sondern es wird zunächst nur die Anzahl der zur Grundgesamtheit gehörenden Personen im Haushalt abgefragt. Anschließend wird in den Haushalten mit mehreren Zielpersonen durch uneingeschränkte Zufallsauswahl eine Person gewählt. Der Interviewer bittet die Kontaktperson dann beispielsweise um Fortführung der Befragung mit der zweitältesten Person im Haushalt (die zur Grundgesamtheit gehört).

Mit dieser Vereinfachung der Kish-Methode ist auf der einen Seite zwar die Möglichkeit zur formalen Kontrolle der Richtigkeit der Zielpersonenauswahl im Vergleich zum ursprünglichen Konzept eingeschränkt. Andererseits aber ist in der praktischen CATI-Situation die Kontrollierbarkeit eines möglichen bzw. notwendigen Personenwechsels für den Interviewer naturgemäß nicht so groß wie in der Face-to-Face-Feldarbeit. Die Kontaktphase verliert durch diese Maßnahme auf jeden Fall an Komplexität für Interviewer wie Interviewte.

Zu b) Birthday-Methode: Eine Methode zur Auswahl der Zielperson im Haushalt, die in den letzten Jahren verstärkt eingesetzt wird, ist die Birthday-Methode. Bei diesem Auswahlmechanismus wird die Kontaktperson im Haushalt um Auskunft darüber gebeten, wer von den zur Grundgesamtheit gehörenden Personen zuletzt Geburtstag hatte. In einer allgemeinen Bevölkerungsbefragung könnte die entsprechende Frage z.B. lau-

ten: „Wer von den Personen, die in ihrem Haushalt leben und 18 Jahre oder älter ist, hatte als letzter Geburtstag?“ Mit der genannten Person würde dann das Interview geführt bzw. zunächst ein Termin für das Interview vereinbart. Alternativ ist auch die Frage nach der zur Grundgesamtheit gehörenden Person, die als nächste Geburtstag hat, möglich.

Mit dieser Methode wird eine Randomisierung der Auswahl der Zielperson angestrebt. Tatsächlich kann davon ausgegangen werden, dass die Geburtstage in einem Haushalt zufällig über das Jahr gestreut sind. Bei Feldzeiten für Befragungen gibt es allerdings durchaus Präferenzen (z.B. Meidung von Ferienzeiten). Damit ist die gleiche Wahrscheinlichkeit der Auswahl innerhalb des Haushaltes aus praktischen Gründen nicht gegeben. „The ‚last birthday‘ procedure is not random because the respondent is determined when the telephone number is selected and the field date is determined, but the procedure does increase the diversity of the respondents’ characteristics“ (Kennedy 1993).

Außerdem haben Experimente ergeben, dass in 20 bis 25 Prozent der kontaktierten Haushalte die „falsche“ Person interviewt worden ist, d.h. nicht diejenige, die als letzte Geburtstag hatte, da die Angaben der Kontaktperson im Haushalt fehlerhaft waren (vgl. Lavrakas 2000: 2).

Eine generelle Empfehlung für eine der beiden Auswahlmechanismen zu geben, wäre methodisch nicht hinreichend abgesichert. Dafür sind die in Experimenten erzielten Ergebnisse zu Methodenvergleichen zu widersprüchlich. „The results of these studies are somewhat mixed, and it is not clearly demonstrated that alternatives to enumeration methods provide uniformly higher response rates“ (Brick und Collins 1997: 4).

6. Ausschöpfungsquote

Persönlich-mündliche Befragungen galten lange Zeit als „Königsweg“ in der empirischen Sozialforschung. Dazu trugen Vorteile dieses Befragungsmodus wie die Visualisierbarkeit von Skalen (Vorlage optischer Hilfsmittel wie Listen und Karten) und die im Vergleich zu Telefonumfragen höheren Ausschöpfungsquoten bei (Wüst 1998: 10). Mittlerweile hat sich das Antwortverhalten zwischen beiden Befragungsarten jedoch angeglichen, d.h. es schwankt in Abhängigkeit von der Länge, dem Thema und anderen Merkmalen der Erhebung gleich stark von Umfrage zu Umfrage. Bei der allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) bewegen sich die Ausschöpfungsraten seit Jahren zwischen 50 und 60 Prozent. Ausschöpfungsquoten dieser Größenordnung können durchaus auch bei Telefonumfragen ausgewiesen werden:

– Politbarometer Januar 1995, Erhebungsinstitut: Forschungsgruppe Wahlen, n=1023, Ausschöpfungsquote: 53,9 Prozent (Stögbauer 2000: 93);
– DNW-Nachwahlstudie 1998, Erhebungsinstitut forsa: n=2019, Ausschöpfungsquote: 51,32 Prozent (http://hypatia.ss.uci.edu/democ//archive/german98.pdf);
– Mediennutzungs-Umfrage 2000, Erhebungsinstitut: Sozialwissenschaftliches Umfragezentrum der Universität Duisburg: n=6111, Ausschöpfungsquote: 36,0 Prozent (Deutschmann und Häder 2002: 70).

Natürlich ist die Ausschöpfungsquote nur ein Hinweis auf die Qualität einer Umfrage. So kann eine Erhebung mit einer Ausschöpfungsquote von 60 Prozent stärker verzerrt sein als eine mit einer Ausschöpfungsquote von 50 Prozent, wenn die Ausfälle systematisch sind, d.h. bestimmte Bevölkerungsgruppen betreffen. Insofern müssen sich an jede empirische Erhebung Analysen zur Qualität der Daten anschließen. A priori-Aussagen zur Abbildungsgüte lassen sich weder für Telefonumfragen noch für persönlich-mündliche Befragungen treffen.

VI. Mobilfunknutzung: Achillesferse der Festnetzstichproben?

Die gängigen Verfahren zur Zufallsauswahl von Telefonnummern stützen sich sämtlich auf Telefonbucheinträge in den Ortsnetzen und damit auf den sogenannten Festnetzbereich. Vernachlässigt werden dabei die Rufnummern der Mobilfunknetze. Es stellt sich angesichts der rasanten Entwicklung in diesem Sektor seit Ende der 1990er Jahre die Frage, inwieweit mit Hilfe von Festnetzstichproben bevölkerungsrepräsentative Daten telefonisch erhoben werden können.

Die Mobilfunkbetreiber zeichnen das Bild einer überaus dynamischen Wachstumsbranche. Nach ihren Angaben gab es im Jahr 2003 rund 62 Mio. vergebene Rufnummern in Deutschland.[7] Auch das Zahlenmaterial der Regulierungsbehörde für Telekommunikation und Post liefert vergleichbare Angaben: Es wurden zur Jahresmitte 2003 rund 63,5 Mio. Mobilfunknutzer ausgewiesen. Der deutlichste Zuwachs der Mobilfunknutzung erfolgte zwischen 1999 und 2000 mit einer Verdoppelung der Nutzerzahlen von 23,5 Mio. auf 48 Mio., wobei der bei weitem größte Teil dieses Zuwachsvolumens dem Prepaid-Geschäft zuzuschreiben ist: Die Zahl der Vertragskunden wuchs während dieses Geschäftsjahres von rund 18 Mio. auf mehr als 22 Mio. an, die Zahl der Prepaid-Kartennutzer von 5,5 Mio. auf rund 26 Mio.[8]

Auf Basis dieser Angaben erreicht heute die Penetrationsrate in der Bevölkerung einen Wert von etwa 77 Prozent. Zwei Aspekte lassen diesen Wert jedoch als zu hoch angesetzt erscheinen. Zum einen lässt es sich nicht feststellen, ob im Bereich der Prepaid-Kunden Mehrfachzählungen und nicht mehr genutzte Anschlüsse ausgeklammert werden. Zum anderen werden Nummern, die für die exklusive geschäftliche Nutzung verwandt werden, in den Kundenstatistiken nicht separat ausgewiesen.

Das Wachstum der Mobilfunknutzung zeigt sich auch in den Befragungsdaten des Face-to-Face-Omnibus von TNS Infratest – wenngleich auf einem niedrigeren Niveau. Mit der Hochrechnung dieser Befragungsdaten zum Jahresende 2003 werden 59,7 Mio. Mobilfunknutzer geschätzt. Auf Ebene der Haushalte – was im Sinne einer Fest-

7 T-Mobile (13.01.2004): http://www.t-mobile-international.com/CDA/pr_-_tmo_1__halbjahr_03,215,newsid-2162,de.html?w=1024&h=600; Vodafone (13.01.2004): http://www.vodafone.de/unternehmen/ueber_vodafone/13680.html; O2 (13.01.2004): http://de.o2.com/de/intro/welcome/unternehmen0/company.idxx.21.ver.2.os.1.nnie.2.html; EPlus (13.01.2004): http://www.eplus-unternehmen.de/corporate/1/1_1/1_1_r3.asp

8 Siehe: Tätigkeitsbericht 2002/2003 der Regulierungsbehörde für Telekommunikation und Post. Bonn Dezember 2003, S. 29ff.; http://www.regtp.de/imperia/md/content/aktuelles/Taetigkeitsbericht_2002_2003.pdf (13.01.2004)

netzstichprobe zunächst der bessere Anknüpfungspunkt ist – wird ein Anteilswert von rund 67 Prozent der Haushalte gemessen, die über Mobilfunk verfügen. 1999 lag dieser Wert noch bei 23 Prozent. Im Jahr 2000 wurden rund 52 Prozent, in 2001 56 Prozent und in 2002 61 Prozent Mobilfunkhaushalte gemessen. Im Rahmen dieses Wachstumsprozesses nahm auch der Anteil der Haushalte, die über zwei und mehr Geräte verfügen, ständig zu.[9]

Angesichts dieser Entwicklungen drängen sich im Hinblick auf Telefonstichproben einige Fragen auf:

– Inwieweit führt die verstärkte Mobilfunknutzung zur Substitution des Festnetzes?
– Gibt es Haushalte, die nicht mehr über ein herkömmliches Festnetztelefon verfügen und nur noch Mobilfunk nutzen?
– Wie groß ist der Anteil solcher „reinen" Mobilfunkhaushalte?
– Wie lassen sich diese Haushalte charakterisieren?

Zur Beantwortung der aufgeworfenen Fragen wird hier wiederum auf das Datenmaterial des Face-to-Face-Omnibus von TNS Infratest zurückgegriffen. Im Jahr 2003 wurde, wie in *Tabelle 3* dargestellt, der Anteil der Haushalte, die nur über Mobilfunk erreichbar sind, mit 5,5 Prozent ausgewiesen. Seit dem Jahr 2000 stieg dieser Anteil nur sehr geringfügig an, hingegen betrug dieser Wert 1999 erst 1,6 Prozent! Im gleichen Zeitraum sanken sowohl der Anteil der Haushalte, die über Festnetz erreichbar sind, sowie der Anteil der Haushalte, die nicht über einen Telefonanschluss verfügen. Im Gesamtbild ist auch hier die deutlichste Veränderungsrate zwischen 1999 und 2000 festzustellen.

Tabelle 3: Haushalte nach Telefonanschluss

	Erhebungsjahr				
Angaben in %	1999	2000	2001	2002	2003
Festnetzanschluss	95,2	93,0	92,8	92,9	92,9
nur Mobilfunk	1,6	4,1	5,0	5,4	5,5
Telefonhaushalte insgesamt	96,8	97,1	97,8	98,3	98,4
kein Telefonbesitz	3,2	2,9	2,2	1,7	1,6

Quelle: TNS Infratest Face-to-Face-Omnibus, n=30000 (p.a.) haushaltsgewichtete Angaben.

Die Nutzung von Mobiltelefonen stellt also nicht zwangsläufig ein Problem bei der Bildung von Festnetztelefonstichproben dar. Sie wird dann problematisch, wenn für den Einsatz eines Handys auf einen herkömmlichen Festnetzanschluss verzichtet wird. Angesichts der Entwicklung in den Jahren 1999 bis 2003 ist in gewissem Umfang eine derartige Substitution festzustellen. Mit der Hochrechnung für Befragungsdaten 2003

9 Der Face-to-Face-Omnibus von TNS Infratest basiert auf einer ADM-Standard-Random-Stichprobe und wird per CAPI erhoben. Die Fallzahl beträgt ca. 30000 Interviews p.a. Siehe zum Erhebungsdesign Behrens und Löffler (1999). Die Neukonzeption des ADM-Face-to-Face-Stichprobensystems wurde zum Ende des Jahres 2003 abgeschlossen. Bis Ende 2004 wurde das System in den Häusern der Mitgliedsinstitute implementiert. Der Anteil der Mobilfunknutzung dürfte mit diesem methodischen Instrumentarium tendenziell unterschätzt werden, wenn unterstellt wird, dass die Mobilfunknutzung in den mobilen Schichten der Gesellschaft stark verbreitet ist, die auch in einer Face-to-Face-Stichprobe unterrepräsentiert sind.

werden rund 3,6 Mio. Personen geschätzt, die nicht mehr über das Festnetz telefonisch zu erreichen sind. Darüber hinaus ist jedoch ebenfalls zu konstatieren, dass mit der Zunahme der Mobilfunknutzung auch sinkende Anteile der Haushalte einhergehen, die über kein Telefon verfügen. Auch diesbezüglich scheinen gewisse Austauschprozesse stattzufinden.[10]

Die Anteile der reinen Mobilfunkhaushalte wie auch der Haushalte ohne Telefonanschluss weisen in Ostdeutschland überdurchschnittliche, in Westdeutschland dagegen unterdurchschnittliche Werte auf. Das Ausmaß dieses Musters ist 2003 wesentlich ausgeprägter als dies noch 1999 der Fall war (siehe *Tabelle 4*). Spitzenwerte nehmen 2003 Berlin, Brandenburg, Mecklenburg-Vorpommern und Sachsen-Anhalt mit Anteilswerten von jeweils rund 10 Prozent ein. In Nordrhein-Westfalen, Baden-Württemberg und Bayern liegen beide Kategorien unter dem Durchschnitt.

Tabelle 4: Haushalte nach Telefonanschluss und Region

	1999			2003		
Angaben in %	Gesamt	West	Ost	Gesamt	West	Ost
Festnetzanschluss	95,2	96,0	92,3	92,9	93,8	89,1
nur Mobilfunk	1,6	1,3	2,8	5,5	4,8	8,6
Telefonhaushalte insgesamt	96,8	97,3	95,1	98,4	98,6	97,7
Kein Telefonbesitz	3,2	2,7	4,9	1,6	1,4	2,3

Quelle: TNS Infratest Face-to-Face-Omnibus, n=30000 (2003) haushaltsgewichtete Angaben, West inkl. Westberlin, Ost inkl. Ostberlin.

Eine Auswertung der Telefonanschlussarten nach demographischen Angaben liefert einige Hinweise darauf, welche Personen verstärkt nur Mobilfunk nutzen und somit in den festnetzbasierten Stichproben unterrepräsentiert sind. Wie in *Tabelle 5* dargestellt wird, handelt es sich vor allem um Personen im Alter zwischen 20 und 29 Jahren (zum Teil auch bis 39 Jahren), die verstärkt nur mit dem Handy telefonieren. Demgegenüber zeichnen sich Personen über 50 Jahre durch eine geringe Affinität zum Mobilfunk aus. Zudem sind die exklusiven Mobilfunknutzer eher Männer als Frauen.

In der Gesamtschau der dargestellten Befunde mit weiteren Auswertungen aus dem Face-to-Face-Omnibus von TNS Infratest ergibt sich ein relativ klares Bild, welche Personengruppen über das Festnetz schlechter erreichbar sind. Bei Mobilfunknutzern, die über keinen Festnetzanschluss verfügen, handelt es sich besonders häufig um junge Männer, die in Berlin oder Ostdeutschland wohnen, ein geringes Einkommen haben und überwiegend in Einpersonenhaushalten leben. Es ist wahrscheinlich, dass es sich vorrangig um Auszubildende, Studenten und junge Berufsstarter handelt.

Insgesamt ist festzuhalten, dass die Erreichbarkeit der Haushalte im Festnetz weitaus besser ist, als die Erfolgsgeschichte der Mobilfunknutzung zunächst nahe legen würde. Die Nutzung von Mobiltelefonen führt nicht zwangsläufig zur Aufgabe des Festnetzanschlusses. Die erdrutschartige Zunahme in den Jahren 1999–2000 scheint

10 Die neueste Einkommens- und Verbrauchsstichprobe des Statistischen Bundesamtes gibt rund 4 Prozent Haushalte an, die nur noch über Mobilfunk erreichbar sind und keinen Festnetzanschluss mehr haben (siehe: http://www.destatis.de/presse/deutsch/pm2004/p2210024.htm).

Tabelle 5: Telefonnutzung und Demographie

	Geschlecht		Alter					
Angaben in %	M	F	14–19	20–29	30–39	40–49	50–59	60++
Deutschland insg.	48,0	52,0	7,8	12,3	18,7	17,6	13,8	29,7
Festnetzanschluss	47,5	52,5	7,7	11,2	18,4	17,7	14,1	30,8
nur Mobilfunk	58,2	41,8	9,5	37,7	25,8	15,5	7,4	4,1
Telefonhaushalte insg.	48,0	52,0	7,8	12,3	18,8	17,6	13,8	29,6
Kein Telefonbesitz	47,4	52,6	7,7	11,5	17,2	14,6	12,0	37,4

Quelle: TNS Infratest Face-to-Face-Omnibus, n=30000 (2003), personengewichtete Angaben.

sich momentan nicht weiter zu verfestigen. Möglicherweise wird eine solche Entwicklung auch durch die Zuwächse der privaten Internetanschlüsse gebremst, die derzeit immer noch auf einem Festnetzanschluss basieren, von welchem dann eben auch telefoniert werden kann.

Ein Aspekt der momentanen Situation ist indes besonders kritisch: Die Mobilfunknutzung ist kein gleichmäßig verteiltes Phänomen. Hierbei gibt es eine regionale und eine strukturelle Komponente. Besonders bei Schwerpunktstudien – sei es z.B. bei einer Untersuchung in einzelnen Regionen Ostdeutschlands oder z.B. bei einer Studie, die sich ausschließlich an Jugendliche wendet – gilt es dies zu berücksichtigen.

VII. Probleme bei der Bildung von Mobilfunkstichproben

Wenn zukünftig die Nutzung von Mobilfunktelefonen in größerem Maße als gegenwärtig auch zur Aufgabe des Festnetzanschlusses führen sollte, dann müssen die Stichprobenstatistiker gerüstet sein, auch diese Bevölkerungsgruppe mit Telefonstichproben erreichen zu können. Zwar existiert bislang noch keine praktikable – d.h. mit vertretbarem Aufwand realisierbare und somit auch finanzierbare – Lösung für die Konstruktion eines Auswahlrahmens für Mobilfunkstichproben. Ausgehend von der Verfügbarkeit des benötigten Datenmaterials sind aber prinzipiell vier Ansatzpunkte denkbar:

1. Die Netzbetreibergesellschaften wären theoretisch in der Lage, mit ihren Kundenstammdaten eine mögliche Ausgangsbasis zu liefern. Aber auf Grund der gesetzlichen Bestimmungen sind diese personenbezogenen Daten natürlich nicht öffentlich verfügbar und stehen nur den Netzbetreibern selbst zur Verfügung, wenn bspw. die Kundschaft befragt werden soll. Neben dem Datenschutzargument spricht noch ein weiterer Aspekt gegen diesen Ansatz. Es ist unklar, in welchem Umfang den Providern Kundendaten im Bereich der Prepaid-Telefonie vorliegen.

2. In den öffentlich zugänglichen elektronischen Telefonbüchern finden sich auch Einträge mit Mobilfunknummern. Hieraus listenweise zu selektieren, erscheint jedoch in vielfacher Hinsicht problematisch. Die Anzahl der Einträge ist im Vergleich zu den publizierten Teilnehmerzahlen der Mobilfunkgesellschaften äußerst gering (vergleiche *Tabelle 6*). Darüber hinaus fällt auf, dass es sich bei Mobilfunkeinträgen in vielen Fällen auch um Zweitnennungen von Personen handelt, die zu ihrer „normalen" Nummer im Ortsnetz einen weiteren Eintrag mit der Mobilfunknummer veranlasst haben. Eine

solche Auswahlliste kann unmöglich für alle Mobilfunkteilnehmer ein valides Abbild liefern. Auf Grund des geringen Umfangs dieser Liste und wegen nicht auszuschließenden Ungleichmäßigkeiten beim Eintragsverhalten ist mit vollkommen unkontrollierbaren Verzerrungen zu rechnen.

3. Zunächst vielversprechender erscheint die Idee, im Rahmen eines Accesspanels befragungsbereite Personen mit Mobilfunktelefonen zu gewinnen. Mit einem solchen Vorgehen wird die Problematik jedoch nur verlagert. Zu prüfen ist hier nicht nur die Zusammensetzung des Panels als solches, sondern vielmehr sein Zustandekommen und die Rekrutierungsmechanismen. Es erscheint bei genauerer Betrachtung unrealistisch, gerade solche Personen für die Teilnahme in einem Panel gewinnen zu können, die sich entweder nicht eintragen lassen und/oder keinen Festnetzanschluss haben.

4. Schließlich ist zur Bildung einer Auswahlgrundlage auch an die Generierung von Rufnummern zu denken. Dies wäre prinzipiell möglich entweder als rein zufällig generierte Zahl mit entsprechender Ziffernlänge oder als Randomize-Last-Digits-Verfahren auf Basis der im Telefonbuch verzeichneten Einträge.[11] Auf der einen Seite haben solche Generierungsmodelle gegenüber den ersten drei Ansätzen den klaren Vorteil, nicht a priori die Auswahlchancen bestimmter Gruppen auf Null zu setzen – das Vorhandensein eines Eintrags (Ansatz 1 und 2) oder die Erfolgschancen der Rekrutierung (Ansatz 3) spielen hier zunächst keine Rolle. Bei genauerer Betrachtung zeigt sich allerdings, dass die blockbezogene Generierung mit dem Gabler-Häder-Verfahren für den Mobilfunksektor problematisch, weil im Vergleich zur Anzahl der Netzteilnehmer der erzeugte Auswahlrahmen wesentlich zu klein ist. Es gibt also eine ganze Reihe von Anschlüssen, für die kein Nummernblock erzeugt wird, weil keine einzige Rufnummer aus dem betreffenden Block eingetragen ist. Auf der anderen Seite hat aber die Generierung gegenüber den Varianten 1 bis 3 den deutlichen Nachteil, dass mit einer ganz erheblichen Zahl von Fehlkontakten zu rechnen ist, denn die Dichte gültiger Telefonanschlüsse in den Nummernblöcken ist in den Mobilfunknetzen bisweilen so niedrig, dass eine forschungsökonomische Studienabwicklung kaum möglich ist. Für den ADM-Mobilfunkpool ergibt sich im Durchschnitt ein geschätzter Anteil gültiger Rufnummern von knapp unter 50 Prozent (vgl. *Tabelle 6*). Würde man demgegenüber eine Stichprobe auf Basis einer freien Generierung – also ohne Berücksichtigung von Einträgen – ziehen, läge der Anteil gültiger Rufnummern noch deutlich weiter unterhalb der 50-Prozentmarke, wenn man den strukturellen Befunden aus *Tabelle 6* unterstellt, dass generierte Rufnummern immer mehr ungültige Nummern aufweisen als eingetragene Nummern.

Welcher Weg zur Konstruktion einer Auswahlgesamtheit zukünftig eingeschlagen werden wird, hängt nicht nur von der weiteren genauen Beobachtung des Mobilfunksektors und von Analysen zum Eintragsverhalten der Mobilfunknutzer ab, sondern setzt darüber hinaus auch eine methodisch-konzeptionelle Auseinandersetzung voraus, wie ein sinnvolles Stichprobendesign aussehen kann.

---

11 Im Rahmen der jährlichen Aktualisierung der ADM-Telefonstichproben wird ein Pool mit Mobilfunknummern miterzeugt. Die Generierung wurde entsprechend dem Gabler-Häder-Verfahren durchgeführt. Einige Strukturmerkmale dieses Nummernraumes werden – im Vergleich zu den Teilnehmerzahlen der Netzbetreibergesellschaften bzw. der RegTP – in *Tabelle 6* dargestellt.

Tabelle 6: Generierung von Mobilfunknummern

	T-Mobile		Vodafone		E-Plus		O_2		Alle
Teilnehmerzahl (in Mio.)[1]	26,04		24,02		7,96		5,18		63,20
eingetragene Rufnummern (in Mio.)[2]	1,01		0,77		0,19		0,30		2,27
generierte Rufnummern (in Mio.)[2]	12,37		10,73		6,73		3,00		32,83
Frame-Größe (in Mio.)[2]	13,38		11,50		6,92		3,30		35,10
Frame-Abdeckung (in %)	51,4		47,9		86,9		63,7		55,5
Eintragsdichte (in %; Basis Frame)	7,5		6,7		2,7		9,1		6,5
Eintragsdichte (in %; Basis Teilnehmer)	3,9		3,2		2,4		5,8		3,6
	gelis-tet	gene-riert	gelis-tet	gene-riert	gelis-tet	gene-riert	gelis-tet	gene-riert	∅
Anteil gültiger Rufnummern (in %)[3]	67,7	48,0	76,7	63,5	54,5	29,2	50,8	34,0	49,2

1 *Quelle:* RegTP Tätigkeitsbericht 2003.

2 ADM Mobilfunkpool (2003), Zählungen TNS Infratest.

3 Für jeden Provider wurden 1.000 generierte und 1.000 eingetragene Rufnummern per Zufallsauswahl aus dem ADM-Mobilfunkpool gezogen und im Rahmen der Feldarbeit bei TNS Infratest qualifiziert. Die Anteilswerte sind als Momentaufnahme nur eingeschränkt aussagekräftig.

Es wäre denkbar, die konventionellen Festnetzstichproben mit Mobilfunknummern zu ergänzen: Die Schwierigkeit dabei ist, eben auch solche Personen im richtigen Ausmaß mit Mobilfunknummern in der Auswahlgesamtheit zu haben, die via Festnetz nicht zu erreichen sind. Ein solches kombiniertes Stichprobenmodell erfordert jedoch auch die Klärung, wie die Mobilfunkinterviews (als Personenstichprobe!) mit den Festnetzinterviews (als Haushaltsstichprobe!) in einem Gesamtdatensatz verrechnet werden können. Insbesondere müsste das Problem der Korrektur unterschiedlicher Auswahlchancen gelöst werden.

Alternativ könnte eine Telefonstichprobe auch als reine Mobilfunkstichprobe ohne Festnetznummern anlegt werden. Dann wäre natürlich – spiegelbildlich zur heutigen Situation – zu prüfen, ob damit wiederum nicht auch wesentliche Gruppen ausgeschlossen werden. Es bedarf hier keiner weiteren Ausführungen, dass dies bislang kein erfolgversprechender Weg ist.

Eine weitere Schwierigkeit ergibt sich daraus, dass bisher die Auswahl von Telefonnummern im Rahmen des ADM als stratifizierte Stichprobe umgesetzt wird. Regionale Kriterien und auch Ortsgrößenklassen dienen als Schichtungsvariablen. Dies setzt vorab eine räumliche Verortung der Rufnummern voraus, was für den Mobilfunksektor bisher nicht ohne weiteres machbar ist. Darüber hinaus stellt aber gerade die Mobilität – also die Loslösung des Anschlusses von einem fest definierten geographischen Punkt – das Konzept der Flächenstichprobe in Frage.

Ferner entfällt die Möglichkeit, Studien mit regionalem Fokus durchzuführen, weil entsprechende Filterkriterien nicht vorhanden sind. Regionale Schwerpunktstudien sind also mit enormem Screeningaufwand und vielen Fehlkontakten verbunden – praktisch also kaum mehr durchführbar.

Weiterhin wird bei ADM-Festnetzstichproben aus den Einträgen anhand von Schlüsselbegriffen ein Kennzeichen gesetzt, ob es sich um einen gewerblichen Eintrag handelt. Diese Rufnummern werden bei der Ziehung ausgeschlossen. Im Mobilfunkbereich ist dies praktisch kaum möglich – eine Mobilfunkstichprobe wird daher immer

einen erheblichen Teil rein geschäftlicher Rufnummern umfassen, die dann in der Feldarbeit zu Fehlkontakten führen.

Stehen diese konzeptionellen Fragestellungen in engem Zusammenhang mit der oben diskutierten Frage nach der Verfügbarkeit des benötigten Datenmaterials, so gibt es neben diesen beiden Diskussionslinien noch eine Reihe technischer und praktischer Probleme, die im Rahmen der Feldarbeit mit Mobilfunknummern auftauchen und bislang nicht befriedigend gelöst sind:

Die geringe Eintragsdichte führt zu astronomischen Bruttoansätzen, so dass forschungsökonomisch nicht sinnvoll gearbeitet werden kann (vgl. *Tabelle 6*).

Die Autodialer müssen speziell konfiguriert werden, da je nach Netzbetreiber unterschiedliche Signale zurückgemeldet werden. Hier erschwert die Rufnummernportabilität[12] die korrekte Zuordnung der Codes. Ferner kann der Autodialer Voicemailboxen nur bedingt erkennen und legt diese z.T. den Interviewern vor. Angesichts hoher Anteile von inaktiven Voiceboxen im Prepaid-Bereich ist dies ein weiteres Hindernis für effektive Feldarbeit. Sofern bei solchen Voiceboxen nur die Standardansage des Providers geschaltet ist, kann nicht einmal zweifelsfrei entschieden werden, ob dies eine gültige Rufnummer ist, die später wieder vorgelegt werden sollte, weil sie nur temporär nicht erreicht werden konnte, oder ob es sich um einen nicht genutzten Anschluss und somit um einen endgültigen Ausfall handelt.

Sicherlich müssen auch mehr Erfahrungen gesammelt werden, wie sich möglichst viele erfolgversprechende Kontakte tatsächlich in Interviews ummünzen lassen. Mehrere Situationen sind denkbar, in denen ein technisch-erfolgreicher Anruf eines Instituts auf dem Handy denkbar ungünstig für die Zielperson ist und somit zur Verweigerung führt.

Der Vollständigkeit halber muss abschließend konstatiert werden, dass auch das Studienbudget stärker belastet wird, denn die Kosten für die Telefonie steigen mit dem Wechsel in ein Mobilfunknetz deutlich an.

Die Beschäftigung mit dem Thema Mobilfunk wird immer mehr zur wichtigen Aufgabe der Stichprobenstatistiker – auf theoretischer und praktischer Seite. Von „Entwarnung" kann noch keine Rede sein. Zum einen müssen die bisherigen Beobachtungen zur Abdeckungsqualität der Festnetzstichproben weiter fortgeschrieben werden. Zum anderen müssen zukünftig mehr Methodentests zur Mobilfunktelefonie durchgeführt werden. Es gilt dabei sowohl die technischen Fragen zu lösen als auch verschiedene Auswahlmethoden zu prüfen.

12 Unter Rufnummernportabilität wird die Mitnahme der ursprünglichen zugeteilten Mobilfunknummer einschließlich der Netzkennzahl beim Wechsel des Providers verstanden. Das bedeutet, dass sich unter einer Nummer, die mit 0171 beginnt, nicht unbedingt eine Nummer im Netz von T-Mobile verbirgt, sondern es könnte sich genauso gut um einen Anschluss im Vodafone-Netz handeln. Über das Ausmaß der Rufnummernportabilität gibt es nur wenige Erkenntnisse.

Literatur

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (ADM) e.V.: http://www.adm-ev.de.

Behrens, Kurt, und *Ute Löffler*, 1999: Aufbau des ADM-Stichprobensystems. S. 69–72. in: ADM und AG.MA (Hg.), Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis. Opladen: Leske + Budrich.

Brick, Michael, und *Mary Collins*, 1997: An Experiment in Random-Digit-Dial Screening. National Household Education Survey.

Cox, Lawrence H., 1987: A Constructive Procedure for Unbiased Controlled Rounding. Journal of the American Statistical Association 82: 520–524.

Deutschmann, Mark, und *Sabine Häder*, 2000: Nicht-Eingetragene in CATI-Surveys. S. 68–84 in: *Siegfried Gabler* und *Sabine Häder* (Hg.), Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann Verlag.

Follmer, Robert, und *Menno Smid*, 1998: Nichteingetragene Telefonnummern: Ergebnisse eines Methodentests. S. 43–57 in: *Siegfried Gabler, Sabine Häder* und *Jürgen H. P. Hoffmeyer-Zlotnik* (Hg.), Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag.

Forsman, Gösta, und *S. Berg*, 1992: Telephone Interviewing and Data Quality, an Overview and Empirical Study. Linköping University.

Frey, James H., Gerhard Kunz und *Günther Lüschen,* 1990: Telefonumfragen in der Sozialforschung. Opladen: Westdeutscher Verlag.

Gabler, Siegfried, und *Sabine Häder,* 1997: Überlegungen zu einem Stichprobendesign für Telefonumfragen in Deutschland. ZUMA-Nachrichten 41: 7–18.

Gabler, Siegfried, Sabine Häder und *Partha Lahiri,* 1999: A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering. Survey methodology 1(25): 105–106.

Gabler, Siegfried, und *Sabine Häder,* 2000: Über Design-Effekte. S. 73–97 in: *Peter Ph. Mohler* und *Paul Lüttinger* (Hg.), Querschnitt. Festschrift für Max Kaase. Mannheim: ZUMA.

Gabler, Siegfried, und *Josef Schürle,* 2002: Zur Stabilität des Gabler-Häder-Auswahlrahmens. S. 59–67 in: *Siegfried Gabler* und *Sabine Häder* (Hg.), Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann Verlag.

Glemser, Axel, 2002: ADM Telefonstichproben in der Praxis. S. 46–58 in: *Siegfried Gabler* und *Sabine Häder* (Hg.), Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann Verlag.

Hagan, Dan E., und *Charlotte Meier Collier,* 1983: Must Respondent Selection Procedures for Telephone Surveys be Invasive? Public Opinion Quarterly 47(183): 547–556.

Häder, Sabine, 1996: Wer sind die „Nonpubs"? Zum Problem anonymer Anschlüsse bei der Stichprobenziehung für Telefonumfragen. ZUMA-Nachrichten 39: 45–68.

Heckel, Christiane, 2002: Erstellung der ADM-Telefonauswahlgrundlage. S. 11–31 in: *Siegfried Gabler* und *Sabine Häder* (Hg.), Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann Verlag.

Heyde, Christian von der, 2002: Das ADM-Telefonstichproben-Modell. S. 32–45 in: *Siegfried Gabler* und *Sabine Häder* (Hg.), Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster: Waxmann Verlag.

Hofmann, Heinrich, 1998: Grundsatzfragen der Rufnummernverwaltung. S. 30–42 in: *Siegfried Gabler, Sabine Häder* und *Jürgen H. P. Hoffmeyer-Zlotnik* (Hg.), Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag.

Humpert, Andreas, und *Klaus Schneiderheinze,* 2000: Stichprobenziehung für telefonische Zuwandererumfragen – Einsatzmöglichkeiten der Namenforschung (Onomastik). ZUMA-Nachrichten 47.

Kennedy, John. M., 1993: A Comparison of Telephone Survey Respondent Selection Procedures. Paper presented at the annual AAPOR meeting. http://www.Indiana.edu/~csr/aapor93.html

Kish, Leslie, 1949: A Procedure for Objective Respondent Selection Within the Household. JASA 44: 380–387.

Kish, Leslie, 1987: Weighting in Deft^2. Survey Statistician 17.1.

Lavrakas, Paul J., 2000: A Further Investigation of the Last Birthday Respondent Selection Method and within-unit Coverage Error. Paper submitted to AAPOR 2000. http://www.csr.ohio-state.edu/scholarship/lastbirthday.pdf.

Stögbauer, Andrea, 2000: Ausschöpfungsprobleme telefonischer Umfragen. Eine Zwischenbilanz praktischer gesamtdeutscher Erfahrung. S. 91–104 in: *Volker Hüfken* (Hg.), Methoden in Telefonumfragen. Opladen: Westdeutscher Verlag.

Wüst, Andreas M., 1998: Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften als Telefonumfrage. ZUMA-Arbeitsbericht 98/04.

PROBABILISTISCHE SELEKTIONSMODELLE*

Ulrich Pötter

Zusammenfassung: Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Befragte geben manchmal keine Auskunft oder verweigern das Interview. Der Anteil der vollständigen Antworten ist selten größer als 40 Prozent. Was weiß man über die anderen 60 Prozent? Kann man problemlos die Angaben der 40 Prozent so behandeln als hätte man Angaben von allen ausgewählten Befragten? Probabilistische Selektionsmodelle sind ein Hilfsmittel, über Zusammenhänge zwischen erhaltenen und nicht erhaltenen Angaben nachzudenken. Ihre Anwendungsmöglichkeiten beschränken sich nicht auf die Analyse fehlender Angaben in Umfragen. Sie werden auch für die Untersuchung gruppierter, zensierter oder sonstiger unvollständiger Daten sowie für die Modellierung von Prozessen mit Selbstselektionen und die Evaluation von Maßnahmen benutzt. Nach einem kurzen Überblick über verschiedene Anwendungen wird zunächst diskutiert, welche Annahmen Selektionsmodellen zugrunde liegen. Es zeigt sich, dass die Verwendung probabilistischer Modelle auf Voraussetzungen verweist, die über empirisch zugängliche Annahmen über soziale Sachverhalte weit hinausgehen. Der Anteil modellimmanenter Spekulation kann allerdings verringert werden, wenn auch partielle Angaben der Befragten mit einbezogen werden. Die Grundlagen von Selektionsmodellen werden daher nicht nur für fehlende, sondern auch für gruppierte und vergröberte Angaben formuliert. Abschließend werden verschiedene Möglichkeiten der Sensitivitätsanalyse für probabilistische Selektionsmodelle dargestellt und die Beschränkungen enger Formulierungen demonstriert. Fehlende und gruppierte Einkommensangaben im ALLBUS 1996 dienen zur Illustration.

I. Einleitung

Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Schließlich steht es jedem Befragten frei, die Zumutung eines Interviews zurückzuweisen oder auf einzelne Fragen nicht zu antworten. Zudem mag ein Befragter nicht willens oder in der Lage sein, Auskünfte in der gewünschten Präzision zu geben. In Umfragen ist der Anteil vollständiger Angaben zu einer bestimmten Frage selten höher als 40 Prozent. Dennoch sollen Umfrageergebnisse Aufschluss über soziale Sachverhalte geben. Wenn dies nachvollziehbar gelingen soll, muss ein Zusammenhang zwischen den erhaltenen Angaben und den potentiellen Angaben aller ausgewählten Befragten hergestellt werden. Denn allein aufgrund der vollständigen Angaben soll es möglich sein, Aussagen über statistische Sachverhalte zu treffen, die sich auf alle Befragten beziehen. Wenn aber über 60 Prozent der Befragten nichts oder wenig bekannt ist, dann sind zur Rechtfertigung solcher Aussagen offenbar starke Annahmen erforderlich. Probabilisti-

* Ich danke Andreas Behr, Eva Berns, Andreas Diekmann, Götz Rohwer und einem anonymen Gutachter für kritische Kommentare und hilfreiche Diskussionen.

sche Selektionsmodelle sind ein Hilfsmittel, solche Annahmen explizit zu formulieren und über ihre Konsequenzen nachzudenken.

Das Grundschema probabilistischer Selektionsmodelle lässt sich leicht angeben: Man möchte Aussagen über Aspekte der Verteilung einer statistischen Variablen Y machen. Nur wurden die Werte dieser Variablen nicht vollständig beobachtet. Dagegen ist die Verteilung einer Variablen Y^* bekannt, von der angenommen wird, sie stände in einer Beziehung zu der interessierenden Variablen Y. Zum Beispiel mag ein Befragter auf die Frage nach dem Einkommen mit einer genauen Angabe, mit einer ungefähren Angabe etwa in Form eines Intervalls, oder gar nicht antworten. Der Wert der Variablen Y^* ist dann entweder der Wert von Y, oder ein Intervall, in welches Y fällt, oder das Intervall $(0, \infty)$, das „keine Angabe" repräsentiert. Wäre bekannt, unter welchen Umständen jemand bei gegebenem Wert von Y eine mehr oder weniger genaue Auskunft Y^* gibt, dann könnte aus der Verteilung von Y^* und den Umständen auf Aspekte der Verteilung von Y geschlossen werden. Wird für die Antwortmöglichkeiten Y^* bei gegebenem Y eine bedingte Wahrscheinlichkeit angegeben, dann soll dieses probabilistische Modell im Folgenden ein allgemeines Selektionsmodell für (Y, Y^*) heißen.

In der Literatur zur Stichprobentheorie sind probabilistische Selektionsmodelle nur sehr zurückhaltend diskutiert worden. So schrieb Tore Dalenius, damals Präsident der International Association of Survey Statisticians: „I take a dim view of the usefulness of these endeavors on two grounds. (1) First, it appears utterly unrealistic to postulate ‚response probabilities' which are independent of the varying circumstances under which an effort is made to elicit a response. ... (2) ... it seems unavoidable to introduce assumptions of unknown validity about probabilities. In summary, I am inclined to reject approaches to the nonresponse problem which involve ‚response probabilities'" (Dalenius in Madow und Olkin 1983, Band 3: 412).

Und Mohler et al. verweisen auf die großen Schwankungen in den Ausschöpfungsraten von Stichproben, „die sich nicht mehr auf statistischen Zufall zurückführen lassen" (Mohler et al. 2003: 11). Daher hat man sich in dieser Tradition darauf beschränkt, Bereiche möglicher Schlussfolgerungen auszuweisen, die sich allein auf die Angaben der Befragten und Konsistenzannahmen stützen.

Dagegen sind sowohl in der mathematischen Statistik wie auch in der Biometrie und Ökonometrie seit etwa 30 Jahren eine Fülle von probabilistischen Selektionsmodellen entwickelt worden. Einerseits ist geklärt worden, unter welchen Modellvorstellungen relativ einfache Verfahren des Umgangs mit fehlenden Daten gerechtfertigt werden können. In solchen Modellen können die Einzelheiten des Zustandekommens unvollständiger Daten weitgehend ignoriert werden. Einen guten Überblick darüber geben die Bücher von Schafer (1997) und Little und Rubin (2002). Andererseits sind in der ökonometrischen Tradition hauptsächlich nicht ignorierbare Selektionsmodelle und entsprechende Schätzverfahren vorgeschlagen worden. Insbesondere die frühen Arbeiten von Heckman (1976, 1979) haben einen kaum zu überschätzenden Einfluss auf viele Bereiche der Sozialwissenschaften ausgeübt. Neuere Überblicke geben Nicoletti (2002) und Vella (1998). Obwohl die Abhängigkeit dieser Modelle von einer unübersichtlichen Mischung von Annahmen über Verteilungen, funktionale Formen von Regressionen, latente Variablen und Ausschlussrestriktionen früh kritisiert wurde (z.B. in Wainer

1986, 1989), haben sie sich in einigen Bereichen der Sozialwissenschaften als dominante analytische Methode durchgesetzt.

Dieser Artikel gibt einen Überblick über neuere Entwicklungen in der statistischen Literatur. Insbesondere werden die Form von Annahmen, die Einbeziehung unvollständiger Angaben und die Verwendung von Sensitivitätsanalysen diskutiert. In Abschnitt II wird zunächst ein kurzer Überblick über häufig verwandte probabilistische Selektionsmodelle und ihre Anwendungen in den Sozialwissenschaften gegeben. Anschließend wird am Beispiel der Einkommensangaben im ALLBUS 1996 die Form der Annahmen diskutiert, die in probabilistische Modelle einfließen (Abschnitt III). Die Annahmen verweisen nicht nur auf potentiell empirisch zugängliche soziale Sachverhalte, sondern enthalten immer auch modellimmanente Anteile, die nicht reifiziert werden sollten. Das mag die Zurückhaltung gegenüber solchen Modellen in der Stichprobentheorie rechtfertigen. Allerdings kann der Anteil modellimmanenter Spekulation verringert werden, wenn auch gruppierte, zensierte und andere partielle Angaben in die Analyse einbezogen werden. Abschnitt V beschreibt die Vorgehensweise für den Fall ignorierbarer Selektionsmodelle. Im Abschnitt VI wird die Ignorierbarkeit von Selektionsmodellen auch für gruppierte und vergröberte Angaben in einem wahrscheinlichkeitstheoretischen Rahmen definiert. Anschließend werden zwei Schätzverfahren vorgestellt, die die Einbeziehung von Kovariablen in Selektionsmodelle erlauben. Einige nicht ignorierbare Modelle werden im folgenden Abschnitt VIII kurz vorgestellt. Abschließend werden Techniken der Sensitivitätsanalyse dargestellt (Abschnitt IX). Sie erlauben eine Abschätzung der Abhängigkeit von Schlussfolgerungen von einigen zentralen modellimmanenten Annahmen und sind daher ein wesentliches Hilfsmittel für die Beurteilung von Selektionsmodellen.

Wenn möglichst alle stochastischen Annahmen eines Selektionsmodells systematisch variiert werden, so zeigt sich, dass der Bereich möglicher Schlussfolgerungen probabilistischer Modelle sehr groß werden kann. Zudem deckt er sich häufig mit den Bereichen, die im Rahmen der klassischen Stichprobentheorie entwickelt wurden. Globale Sensitivitätsanalysen probabilistischer Selektionsmodelle führen daher zu Abschätzungen, die mit denen der Stichprobentheorie vergleichbar sind. Die Formulierung verschiedener Annahmen in probabilistischen Selektionsmodellen ermöglicht eine Diskussion über den Zusammenhang zwischen erhaltenen und nicht erhaltenen Angaben, die stichprobentheoretische Überlegungen ergänzen können.

II. Selektionsmodelle in den Sozialwissenschaften

Umfragedaten bilden eine wesentliche empirische Grundlage aller Sozialwissenschaften. Aber schon das Verfahren, mit dem Befragte ausgewählt werden, ist häufig mit dem Hinweis hinterfragt worden, die Auswahl sei selektiv gewesen. Ein oft und gern zitiertes Beispiel ist der spektakuläre Misserfolg der Wahlvorhersage der Zeitschrift Literary Digest für die Präsidentenwahl 1936 in den USA. Das Literary Digest hatte 60 Prozent der Stimmen für den Republikaner Landon vorhergesagt, aber Roosevelt gewann die Wahl mit 62 Prozent. Das Literary Digest hatte eine Stichprobe von Telefon- und Autobesitzern befragt. Eine klassische Erklärung des Fehlschlags besagt, Telefon- bzw.

Autobesitz sei damals ein Anzeichen von Reichtum gewesen und reichere Personen hätten eher republikanisch gestimmt. Das Selektionsargument bezieht sich auf den gewählten Rahmen der Stichprobe, die Basis für die Auswahl von Befragten. Aber eine einfache Überlegung zeigt, dass diese Selektion nicht allein für den Misserfolg verantwortlich sein kann. 1936 besaßen ca. 40 Prozent der Haushalte ein Telefon. Hätten die Telefon- und Autobesitzer in der Tat zu 60 Prozent für Landon gestimmt, dann hätten von allen Haushalten, die weder Auto noch Telefon besaßen, über 75 Prozent für Roosevelt stimmen müssen. Betrachtet man die abgegebenen Wählerstimmen, so hätte der entsprechende Anteil sogar größer als 90 Prozent sein müssen (Bryson 1976). Das Verhältnis der Odds für Roosevelt in den beiden Gruppen der Telefonbesitzer und derjenigen ohne Telefon müsste also mehr als 1:20 betragen. Das Literary Digest hatte 10 Millionen Fragebogen verschickt, aber nur 2,3 Millionen zurückerhalten. Es liegt nahe zu vermuten, dass Landon-Anhänger eher als Roosevelt-Anhänger auf die Umfrage geantwortet haben. Wird angenommen, die 10 Millionen Befragten hätten tatsächlich zu 62 Prozent für Roosevelt gestimmt, muss das Verhältnis der Odds für Roosevelt in den beiden Gruppen der Antwortenden und der nicht Antwortenden nur 1:3 betragen, um zu der Diskrepanz zwischen Vorhersage und Wahlergebnis zu führen. Ein Verhältnis der Odds von 1:3 ist eine weit realistischere Größenordnung als die 1:20, die für die These einer Selektion zwischen Telefonbesitzern und Nichtbesitzern angenommen werden müsste.

Daten aus Nachbefragungen haben dann auch die Bedeutung der Unterschiede im Antwortverhalten der Landon- bzw. Rooseveltanhänger bestätigt (Squire 1988; Cahalan 1989). Squire (1988: 132) schließt: „The analysis here should also call attention to the other potential problem with any survey: nonresponse bias. ... Consumers of public opinion surveys, as well as practitioners, must be reminded of this potential problem in order to avoid a future disaster like the Literary Digest poll of 1936."

Brysons Überlegungen über den Misserfolg der Wahlvorhersage des Literary Digest verweisen zwar auf die möglicherweise gravierenden Folgen unvollständiger Angaben in Umfragen, benutzen aber keine probabilistischen Modelle etwa über das Antwortverhalten von Landon- bzw. Rooseveltanhängern. Seit der Mitte der 1970er Jahre sind Verfahren entwickelt worden, die auf der Basis probabilistischer Modelle für Teilnahme- und Antwortentscheidungen der Befragten versuchen, die Folgen unvollständiger Angaben abzuschätzen. Eine Variante, die auf Arbeiten von Heckman (1976, 1979, 1990) zurückgeht, ist von Engelhardt (1999) vorgestellt worden. Sie untersucht die Einkommensangaben in der Berliner Altersstudie (BASE), einer nach Alter und Geschlecht geschichteten Zufallsstichprobe von Berlinern und Berlinerinnen über 70 Jahren, die auf der Basis des Einwohnermelderegisters gezogen wurde (Mayer und Baltes 1996). An der Erstbefragung haben 928 Personen teilgenommen, das entspricht ca. 49 Prozent der Ausgangsstichprobe. Engelhardt verwendet neben dem Einkommen die Variablen Geschlecht, Alter, Familienstand, Schul- und Berufsausbildung, Wohnform, Interviewform sowie einen Demenzindikator. Vollständige Angaben zu diesen Variablen liegen für 842 Personen vor, zusätzliche Angaben zum Einkommen nur für 716 oder 77 Prozent der Personen, die an der Befragung teilgenommen haben. Engelhardt (1999: 716f.) unterstellt eine Wahrscheinlichkeit für die Antwort jeder Person auf die Einkommensfrage, die über einen Probit-Link linear von allen Variablen (bis auf den

Familienstand) abhängt. Außerdem nimmt sie an, das logarithmierte Einkommen aller Personen habe eine lineare, homoskedastische Regression auf alle Variablen (bis auf Alter, Demenzindikator und Interviewform) und folge einer bedingten Normalverteilung. Unter diesen Annahmen kann aus der bedingten Verteilung der beobachteten Einkommen auf die bedingte Verteilung der Einkommen aller Personen geschlossen werden.

Engelhardt vergleicht die entsprechenden Ergebnisse mit einer linearen Regression, die nur die vollständigen Angaben berücksichtigt. Sie schließt, „die Heckman-Korrektur [bietet] aber auch Möglichkeiten, die in der explorativen Analyse liegen. Wenn – wie im Beispiel – die selektionskorrigierte Regressionsanalyse zu demselben Resultat kommt wie die unkorrigierte Schätzung, erhöht dies das Vertrauen in die OLS-Regression“ (1999: 721). Sie betont aber die Abhängigkeit der Ergebnisse von den unterstellten Annahmen (1999: 713f.) und zeigt, dass zumindest diejenigen Annahmen, die sich überprüfen lassen, wohl nicht gelten (1999: 719). Um tatsächlich ein erhöhtes „Vertrauen in die unkorrigierte Schätzung“ zu haben, müsste an Stelle eines einzigen alternativen Selektionsmodell, das zudem auf zweifelhaften Annahmen beruht, mehrere Selektionsmodelle verglichen werden. Heckmans Modell bietet aber keinen systematischen Ansatzpunkt, Modellannahmen zu variieren bzw. die Auswirkungen verletzter Annahmen quantitativ abzuschätzen. Die Möglichkeit, mit Hilfe von probabilistischen Selektionsmodellen über die Folgen unvollständiger Angaben nachzudenken, kann im Rahmen dieses Modells nur begrenzt genutzt werden.

In anderen Bereichen, etwa der historischen Demographie, werden dagegen manchmal probabilistische Selektionsmodelle eingesetzt, um über die Aussagekraft von Angaben zu spekulieren. So existieren oft nur unvollständige Angaben über die Lebensdauer von Menschen. Z.B. gibt es zu Geburts- bzw. Taufangaben in Kirchenregistern in vielen Fällen keine Angaben über das Todesdatum. Allerdings gibt es manchmal weitere Ereignisse wie Heiraten oder Kindergeburten, die aufgezeichnet wurden. Es folgt, dass die Person mindestens das Alter bei diesem Ereignis erreicht hat. Ist Y das Lebensalter einer Person, dann ist Y^* entweder der Wert von Y, falls ein Todesdatum registriert wurde, oder aber das Intervall $(T_{\max}, \infty)$, wobei $T_{\max}$ der Zeitpunkt des letzten Ereignisses ist, das registriert wurde. Wenn angenommen wird, diese Angaben sagten nichts anderes als dass jemand älter als $T_{\max}$ geworden ist, dann können Verfahren der Ereignisanalyse eingesetzt werden. Dagegen kann eingewandt werden, der Grund des Fehlens eines Todesdatums sei i.d.R. die Abwanderung der Personen. Dann wäre das Datum der Abwanderung eine untere Grenze für das erreichte Lebensalter und $T_{\max}$ wäre immer kleiner als dieses Zensurereignis. Im Ergebnis würden Verfahren der Ereignisanalyse die Risikomengen unterschätzen und damit Sterberaten überschätzen. Selbst wenn es keine Angaben über Abwanderungen gibt, kann mithilfe eines probabilistischen Modells für die Zwischenereignisse sowie die Abwanderungszeiten über die Verteilung der Lebensdauern nachgedacht werden (Gill 1997; Jonker 2003).

In den Sozialwissenschaften sind Selektionsmodelle eher selten im Zusammenhang mit unvollständigen Angaben in Umfragen oder Registerdaten behandelt worden. Stattdessen dominieren Anwendungen, die sich auf Größen beziehen, die ihre Bedeutung nur im Rahmen eines vorab definierten Modells gewinnen. So untersuchen Diekmann und Wyder (2002) Reputationseffekte bei Internetauktionen, wobei sie auch die erzielten Preise der Auktionen heranziehen. Sie versuchen dabei auch diejenigen Auktionen

einzubeziehen, für die gar kein Gebot abgegeben wurde und für die daher auch kein erzielter Preis existiert. Sie argumentieren: „Die Regressionsschätzung basiert allerdings nur auf der Stichprobe der 99 erfolgreichen Auktionen, da nur für diese ein Verkaufspreis vorliegt. Nun könnte es sich hierbei um ein selektives Sample handeln. ... Die Zwei-Stufen-Schätzmethode von Heckman ist eine Alternative, um einen eventuellen Stichprobenauswahlfehler zu kontrollieren" (2002: 687).

Ein „Stichprobenauswahlfehler" könnte aber nur vorliegen, wenn auch den Auktionen ohne Gebote ein „Preis" zukäme. Diekmann und Wyder unterstellen wohl eine Größe, die mit „Zahlungsbereitschaft" umschrieben werden kann. Ein Gebot wird abgegeben, falls die Zahlungsbereitschaft eines potentiellen Auktionsteilnehmers größer als das Mindestgebot der Auktion ist. Also fehlen Angaben über die Zahlungsbereitschaft, wenn der Startpreis der Auktion höher als diese Zahlungsbereitschaft ist. Daher könnte das Problem ähnlich wie fehlende Angaben in Umfragen oder Registern behandelt werden. Zwar bleibt unklar, was „Zahlungsbereitschaft" unabhängig von einem konkreten Gebot in einer gegebenen Auktion bedeuten könnte, sogar, welcher Gruppe von Personen diese Größe zugeschrieben werden soll. Aber selbst wenn dem Konzept eine gewisse Plausibilität zugestanden wird, dann greift der Versuch, ein einfaches Selektionsmodell für fehlende Angaben zu verwenden, zu kurz. Denn zum einen ist bei Auktionen ohne Gebote die Zahlungsbereitschaft kleiner als der Startpreis, so dass die Zahlungsbereitschaft nicht vollständig unbekannt ist. Zum anderen ist der erzielte Preis auch nicht gleich der Zahlungsbereitschaft des Höchstbietenden, sondern (bei mehr als einem Gebot) gleich der Zahlungsbereitschaft des Bieters mit dem zweit höchsten Gebot.[1] Ein Selektionsmodell für „Zahlungsbereitschaft" müsste beide Aspekte, den der zusätzlichen Information aus den Mindestgeboten und den der erzielten Preise als zweit höchste Zahlungsbereitschaft berücksichtigen. Die Konzentration auf eine Schätzmethode behindert aber oft die Formulierung probabilistischer Modelle für Konzepte, die sich auf fehlende, abgeschnittene oder zensierte und nach Größe selektierte Beobachtungen stützen.

In den Sozialwissenschaften werden Selektionsmodelle auch zur Evaluation sozialpolitischer Maßnahmen und zur Kausalanalyse herangezogen. Die Idee besteht darin, zunächst eine Variable Y festzulegen, die den Erfolg einer Maßnahme oder die Wirkung einer Ursache darstellen soll. Unterstellt wird dann die gleichzeitige Existenz der Variablen (Y_0, Y_1), wobei Y_0 den Wert von Y annimmt, der sich ergeben hätte, wenn jemand nicht an der Maßnahme teilgenommen hätte, Y_1 den bei Teilnahme. Der Erfolg der Maßnahme ließe sich dann etwa durch $Y_1 - Y_0$ ausdrücken. Natürlich kann eine Person nur entweder an einer Maßnahme teilnehmen oder nicht teilnehmen. Y_0 und Y_1 können also nicht gleichzeitig beobachtet werden. Ist $\mathbb{Y}$ der Wertebereich der Variablen Y und $\mathbb{Y}^*$ der Wertebereich der beobachteten Variablen Y^*, dann ist

$$\mathbb{Y}^* = (\mathbb{Y} \times \{\mathbb{Y}\}) \cup (\{\mathbb{Y}\} \times \mathbb{Y})$$

1 Genauer: Bietet Person A zunächst maximal 20 Euro und später B 10 Euro, so beträgt der Auktionspreis 10 Euro. Der Auktionspreis gibt also die Zahlungsbereitschaft von B wieder. Bietet andererseits zunächst B 10 Euro, A später 20 Euro, so ist der Auktionspreis 10 Euro plus Mindesterhöhung der Auktion.

Entweder wird der Wert von Y_0 beobachtet, nicht aber der von Y_1, für den nur $y_1 \in \mathbb{Y}$ bekannt ist. Die Beobachtung ist also $(y_0, \mathbb{Y})$. Oder der Wert von Y_1 wird beobachtet, nicht aber der von Y_0. Dann kann die Beobachtung durch $(\mathbb{Y}, y_1)$ angegeben werden. In dieser Formulierung entspricht das Evaluationsproblem einem Problem unvollständiger Angaben. Wird ein passendes probabilistisches Selektionsmodell unterstellt, dann ist die gemeinsame Verteilung von (Y_0, Y_1) identifiziert. Andersherum ist aber klar, dass (Y_0, Y_1) außerhalb dieses Selektionsmodells gar nicht definiert ist. Die Abhängigkeit von willkürlich gesetzten Modellannahmen und die kontrafaktische Formulierung des Evaluationsproblems sind oft kritisiert worden (z.B. Dawid 2000). Die kontrafaktische Formulierung des Problems ist fragwürdig, weil sie einen wohldefinierten Wert für das Ergebnis von Ereignissen voraussetzt, die gar nicht stattgefunden haben. Und im Unterschied zu fehlenden Angaben in Umfragen oder Registerdaten kann die Abhängigkeit der Ergebnisse von den Annahmen des Selektionsmodells nie empirisch ergänzt oder kritisiert werden. Fehlen Angaben in Kirchenregistern, so können Angaben aus Lehns- und Pachtregistern, Handwerksrollen, Sippenbüchern und Gerichtsakten herangezogen werden. Interessiert die Verteilung von Einkommen, dann können Personen, die einmal die Auskunft verweigert haben, nochmals befragt werden. Zudem geben Steuerstatistiken, Sozialversicherungsmeldungen und Lohnstatistiken weitere Auskunft. Aber was in parallelen Welten geschehen würde, in denen alles bis auf die Teilnahme an bestimmten Maßnahmen gleich wäre, entzieht sich jedem Versuch empirischer Überprüfung. Die Betonung der formalen Äquivalenz zwischen beiden Situationen verwischt oft die inhaltlichen Unterschiede.

Rubin führt zwei weitere Unterscheidungen an: „The formal equivalence of these problems ... is highly useful conceptually ..., but it is, I believe, not helpful to muddle the distinction when trying to generate sound, practical statistical advice. The reasons for this conclusion are that (a) the estimands (the things we want to estimate) are fundamentally different for these situations, and (b) the processes that create the missing data are typically very different, both by investigators' design and by nature's devices" (Rubin in Wainer 1992: 183f.). Rubins Punkt (b) verweist zurück auf den Status unvollständiger Daten in den beiden Fällen. Verweigert jemand die Auskunft im Interview, so liegt das in seinem oder ihrem Ermessen, hängt aber nicht von den Interessen und Vorstellungen des Forschers ab. Dagegen sind Daten im kontrafaktischen Modell der Evaluation ebenso wie Überlegungen zur „Zahlungsbereitschaft" nur unvollständig aufgrund der Modellvorstellungen des Forschers. Erzielte Preise in Auktionen sind ebenso wie Ergebnisse von Arbeitsmarktmaßnahmen zumindest prinzipiell beobachtbar. „Zahlungsbereitschaft" im Rahmen von Auktionen oder der Erfolg einer Maßnahme (definiert als $Y_1 - Y_0$) ist dagegen nie beobachtbar. Auch wenn diese Unterscheidung eine graduelle ist, so muss doch genau angegeben werden, welche Aussagen getroffen werden sollen. Rubins Punkt (a) soll im Folgenden für unvollständige Angaben in Umfragen diskutiert werden.

III. Beispiel: Einkommensangaben im ALLBUS

Das Haushaltsnettoeinkommen ist von zentraler Bedeutung in vielen Bereichen der Sozialforschung, etwa der Armutsforschung und der Haushaltstheorie. Dennoch ist selbst über das mittlere Hausaltsnettoeinkommen empirisch wenig bekannt. Das Statistische Jahrbuch 2001 weist auf der Basis der Einkommens- und Verbrauchsstichprobe (EVS) 1998 einen monatlichen Mittelwert von 5115 DM aus, 5346 DM in Westdeutschland, 4059 DM in Ostdeutschland (Statistisches Bundesamt 2001: 566ff.). Dagegen weist der Datenreport 1999 auf der Basis des SOEP für das Jahr 1996 ein mittleres Haushaltsnettoeinkommen von 1978 DM (2061 DM West, 1644 DM Ost) aus (Statistisches Bundesamt 2000: 584). Die beiden Datensätze (EVS und SOEP) unterscheiden sich zwar deutlich: Die EVS ist eine Quotenstichprobe mit über 60000 beteiligten Haushalten, das SOEP ist eine weit kleinere Zufallsstichprobe.[2] Aber wegen der großen Abweichungen wäre es sicherlich wünschenswert, unabhängigen Aufschluss über die Verteilung des Haushaltseinkommens in der BRD zu erhalten. Von den großen regelmäßigen sozialwissenschaftlichen Umfragen enthält auch der ALLBUS eine Frage nach dem Haushaltsnettoeinkommen sowie nach dem persönlichen Einkommen. Der ALLBUS 1996 wurde als Melderegisterstichprobe durchgeführt. Grundgesamtheit waren Personen ab 18 Jahren in Privathaushalten (einschließlich deutsch sprechender Ausländer) in West- und Ostdeutschland. Dabei wurden 3518 Interviews realisiert, 2402 davon in den alten und 1116 in den neuen Bundesländern. Personen in den neuen Bundesländern sind also überproportional befragt worden. Ihr Anteil in der Nettostichprobe beträgt 31,7 Prozent, der Bevölkerungsanteil betrug ca. 19 Prozent. Die berichtete Ausschöpfungsquote betrug 54,2 Prozent, d.h. nur 54,2 Prozent der angestrebten Interviews wurden realisiert (Koch 2002: 33).[3] Von den 3518 Befragten, die einem Interview zustimmten, antworteten gerade einmal 1772 oder 50,4 Prozent auf die Frage nach dem monatlichen Haushaltsnettoeinkommen. Weitere 906 Personen oder 25,8 Prozent machten Angaben in gruppierter Form. Insgesamt hat man also nur von etwa 41 Prozent der ausgewählten Personen eine valide Antwort auf die Frage nach dem Haushaltseinkommen erhalten, darunter 14 Prozent in gruppierter Form. Darüber hinaus gibt es aber noch 293 Personen, die zwar Angaben zu ihrem persönlichen Einkommen machten, aber keine Angaben zum Haushaltseinkommen. Diese Angaben könnten als untere Grenzen für das Haushaltseinkommen benutzt werden. Die Fallzahlen sind in der folgenden Tabelle zusammengestellt.

2 Der Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung (Bundesregierung 2001) enthält einen hilfreichen Vergleich verschiedener verfügbarer Datenquellen. Fehlende und unvollständige Einkommensangaben im SOEP und ihre Behandlung sind im Datenreport nicht angegeben. Die Arbeiten von Frick und Grapka (2003), Riphahn und Serfling (2002) und Schräpler (2004) geben einen Überblick.

3 Untersuchungen über den möglicherweise selektiven Ausfall von geplanten Interviews im ALLBUS, dem so genannten Unit-Nonresponse, sind von Koch (1997) und Schneekloth und Leven (2003) vorgelegt worden.

s_1	Bruttostichprobe	$\lvert s_1 \rvert = n_1 = 6491$
s_2	Nettostichprobe	$\lvert s_2 \rvert = n_2 = 3518$
s_3	Haushaltseinkommen angegeben	$\lvert s_3 \rvert = n_3 = 2678$
s_4	genaue Angaben	$\lvert s_4 \rvert = n_4 = 1772$
s'_3	Einkommen in [1,9999]	$\lvert s'_3 \rvert = n'_3 = 6491$
s'_4	genaue Angaben in [1,9999]	$\lvert s'_4 \rvert = n'_4 = 6491$
s_5	keine Angabe Haushaltseinkommen, aber persönliches Einkommen angegeben	$\lvert s_5 \rvert = n_5 = 293$
s_6	keine Angabe Haushaltseinkommen, aber genaues persönliches Einkommen	$\lvert s_6 \rvert = n_6 = 213$

Abbildung 1: Kern-Dichte-Schätzung der genauen Angaben zum Haushaltseinkommen in DM, eingeschränkt auf das Intervall [1,9999]. Als Kern wurde eine Normalverteilung mit Standardabweichung 166 benutzt. Der Dichte-Schätzer ist etwas unterglättet.

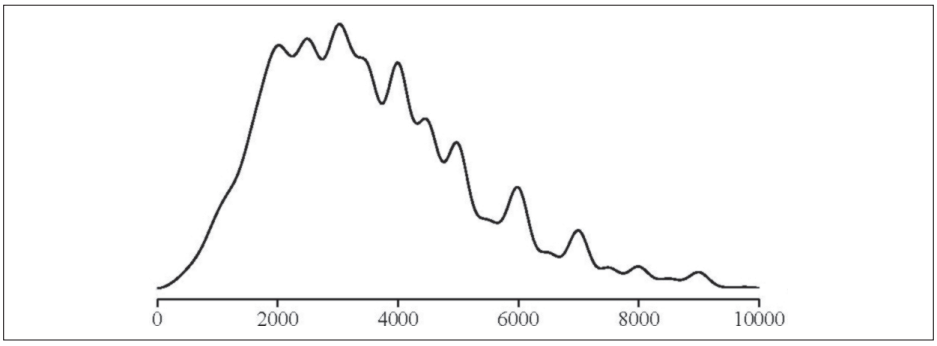


Abbildung 1 zeigt die Dichte der genauen Angaben zum Haushaltsnettoeinkommen. Diese Angaben sind insbesondere in den höheren Einkommensbereichen zu einem großen Teil auf volle 1000 DM Beträge gerundet. So sind von den 122 Angaben von mehr als 7000 DM 120 in vollen 100 DM Beträgen und 107 in vollen 500 DM Beträgen angegeben, dagegen 85 Angaben in vollen 1000 DM Beträgen. Auf der anderen Seite sind von den 802 Angaben von 3000 DM oder weniger nur 277 in vollen 500 DM Beträgen angegeben. Die Rundungsregeln der Befragten sind unbekannt, hängen aber offenbar von der Einkommenshöhe ab. Auch die „genauen" Einkommensangaben können nur als grobe Näherung des tatsächlichen verfügbaren Haushaltseinkommens betrachtet werden. Die Rundung in den Einkommensangaben führt zu einer weiteren Unsicherheit bei der Analyse der Daten, die im Folgenden aber nicht systematisch untersucht wird.

Folgt man dem üblichen Verfahren und ignoriert fehlende und gruppierte Angaben, so ergibt sich ein (ungewichtetes) mittleres Haushaltsnettoeinkommen von 3676 DM (3909 DM West, 3171 DM Ost). Beschränkt man sich bei der Berechnung des mittleren Haushaltseinkommens auf Einkommen unter 10000 DM und benutzt nur die genauen Angaben, also die Teilstichprobe s'_4, so ergibt sich ein Mittelwert von

3560.[4] Wird das Stichprobendesign ignoriert und eine einfache Zufallsauswahl unterstellt, so kann ein 95 Prozent-Konfidenzintervall konstruiert werden: (3480, 3640). Dabei werden die unterschiedlichen Auswahlsätze für Ost- und Westdeutschland unterschlagen, obwohl sich die Mittelwerte in Ostdeutschland (3158 DM) und Westdeutschland (3749 DM) deutlich unterscheiden und ostdeutsche Personen deutlich überrepräsentiert sind. Das geht aber gar nicht anders. Denn weder können die Gewichte für die Teilstichprobe s_2 verwandt werden: dort beträgt der Anteil an Personen in Ostdeutschland 31,7 Prozent, während er in der Teilstichprobe s'_4 32,0 Prozent beträgt.[5] Noch können Gewichte benutzt werden, die sich aus dem Anteil in der Teilstichprobe s'_4 ergeben, denn diese würden sich von Stichprobe zu Stichprobe ändern und erlaubten keine statistischen Aussagen. Man erhält also einen Mittelwert, der wegen der sehr unterschiedlichen Auswahlsätze in Ost- und Westdeutschland wohl nichts über das durchschnittliche Haushaltseinkommen in Deutschland sagt, und ein Konfidenzintervall, dessen Überdeckungseigenschaften schlicht unbekannt sind.[6]

Soll der Rahmen der klassischen Stichprobentheorie nicht vollständig verlassen werden, dann muss die gesamte (realisierte) Stichprobe einschließlich der gruppierten Angaben betrachtet werden. Dies erfordert Verfahren zur Behandlung unvollständiger Angaben. Das einfachste Verfahren hält an der Idee fest, die Angaben der Befragten als fixes Datum zu behandeln. Unvollständige Angaben werden als Bereiche aufgefasst, in denen der tatsächliche Wert liegt. Ist die Auskunft eines Befragten, das Einkommen liege zwischen 2000 DM und 2500 DM, dann wird unterstellt, das exakte Einkommen sei eine der Zahlen 2000, 2001, 2002, ..., 2499. Bei einer Statistik wie dem Mittelwert wird die Berechnung für jeden der möglichen Werte in diesem Bereich durchgeführt. Das Ergebnis ist ein Bereich von Mittelwerten. Handelt es sich um zusammenhängende Intervalle, dann reicht es, die Statistik für die Extremwerte der Antwort-

4 Die Beschränkung des betrachteten Einkommenbereichs ist sowohl für nicht-probabilistische wie für probabilistische Modelle notwendig. Für nicht-probabilistische Überlegungen ist das unmittelbar einsichtig, weil Haushalten ohne Einkommensangaben jedes beliebige Einkommen zukommen könnte. Aber auch in probabilistischen Modellen sind Einschränkungen notwendig. Denn ohne solche Einschränkungen ist der Erwartungswert nicht einmal ein stetiges Funktional bezüglich der Kolmogorov-Metrik $d(F,G) := \sup_y |F(y) - G(y)|$ zwischen Verteilungsfunktionen F und G (Lehmann 1999: 391). Ohne Einschränkungen existieren also keine gleichmäßig konsistenten Tests, der Bootstrap funktioniert nicht etc. Wird das Verhalten von Schätzern nicht einfach unter einem als „wahr" angenommenen Modell untersucht, sondern werden alle Modelle zugelassen, die mit den gegebenen Beobachtungen statistisch verträglich sind, dann ist der Bereich möglicher Erwartungswerte unendlich groß, selbst wenn die Existenz endlicher vierter Momente unterstellt wird (Davies 1995: 205). Im Zusammenhang mit Selektionsmodellen ist das Problem von Robins und Ritov (1997) untersucht worden (vgl. Abschnitt VII).

5 In der Teilstichprobe s'_3 ergibt sich sogar ein Anteil von 34,1 Prozent, ein „signifikanter" Unterschied zu 31,7 Prozent.

6 Zur Berechnung eines Gesamtmittelwerts kann auch einfach der bekannte Bevölkerungsanteil in Ost- und Westdeutschland zur Kombination der Ergebnisse in Ost- und Westdeutschland benutzt werden. Dann werden Auswahlsätze einfach ignoriert. Für komplexere Fragen eignet sich eine solche naive Randanpassung allerdings nicht. Zusammenhänge zwischen Stichprobengewichten und Unit-nonresponse werden von Kalton (2002), Kalton und Flores-Cervantes (2003) diskutiert. Little und Vartivarian (2003) kritisieren einige klassische Verfahren.

intervalle auszuwerten.[7] Das Verfahren kann nur dann sinnvoll angewandt werden, wenn die Bereiche unvollständiger Daten beschränkt sind. Daher können nur Einkommensangaben etwa unter 10000 DM behandelt werden. Betrachtet man zunächst s'_3, so ergibt sich für die möglichen Mittelwerte das Intervall (3589,3785). Das Intervall der möglichen Mittelwerte ist deutlich länger als das naive Konfidenzintervall des letzten Absatzes. Um aber die Auswahlwahrscheinlichkeiten des ursprünglichen Stichprobendesigns verwenden zu können, muss zumindest die Teilstichprobe s_2 betrachtet werden, also zusätzlich die 885 Befragten, die gar keine Angaben machten.[8] Dann ergibt sich ein ungewichtetes Intervall der möglichen Mittelwerte von (2687,5348). Wird gar die Bruttostichprobe s_1 betrachtet, ergibt sich das Intervall (1457,7479). Für diese Intervalle könnten nun „korrekt" gewichtete Versionen und Konfidenzintervalle ausgerechnet werden. Nur sind die Intervalle selbst schon viel zu groß, um von praktischem Interesse zu sein.

IV. Stichproben und probabilistische Auswahlmodelle

Da beide Ansätze selten weiterhelfen, wurde versucht, die Fragestellung umzuformulieren. Der hierbei zumeist eingeschlagene Weg opfert einen wesentlichen Ausgangspunkt der Stichprobentheorie, der von den Berichten der Befragten als fixem Datum ausgeht. Stattdessen werden die Angaben der Befragten als Realisationen von Zufallsvariablen im Sinne der Wahrscheinlichkeitstheorie aufgefasst. Für probabilistische Modelle existieren bereits Methoden zur Analyse unvollständiger Daten. Zudem können in diesem Rahmen auch Modelle der Entstehung unvollständiger Angaben entwickelt werden.

Die Durchführung eines solchen Ansatzes ist konzeptionell weit schwieriger und konsequenzenreicher als oft angenommen wird. Einen Teil des Weges gehen Superpopulationsmodelle: Sie unterstellen, dass die interessierenden Größen in einer Gesamtheit U durch einen Zufallsprozess zustande gekommen seien, der sich durch eine Wahrscheinlichkeitsverteilung F_θ beschreiben lässt. Etwa: Das Einkommen der Bevölkerung der BRD wird als Realisation von $N := |U|$ unabhängigen und identisch lognormalverteilten Zufallsvariablen erzeugt. Das ist offenbar keine realistische Annahme über das Zustandekommen von Einkommen. Die Metaphorik des „als ob durch einen Zufallsprozess zustandegekommen" erlaubt aber relativ kompakte Beschreibungen von empirischen Verteilungen durch die Parameter θ sowie einen Anschluss an die Stichprobentheorie, denn die realisierten Werte der Zufallsvariablen werden für die Stichprobenziehung als fix angenommen. Der Superpopulationsansatz hält also auf der Ebene der Stichprobenziehung an der Idee der Angaben der Befragten als fixem Datum fest. Allerdings wird das ursprüngliche Problem, Aussagen über die Verteilung eines Merkmals in der Gesamtheit U zu gewinnen, durch ein anderes ersetzt: Aussagen über

7 Die Grundidee ist recht alt (Cochran 1977: Kap. 13.2). Aber schon bei Statistiken wie der Varianz ergeben sich Probleme, effiziente Algorithmen für die Berechnung der Intervalle zu finden (Fishman und Rubin 1998; Rohwer und Pötter 2001: Kap. 19; Ferson et al. 2002). Manski (2003) gibt einen guten Überblick über neuere Ergebnisse. Für Kreuztabellen werden neuere Verfahren von Dobra und Fienberg (2000) beschrieben.

8 Ich rechne die $n_3 - n'_3 = 45$ Angaben außerhalb von [1,9999] dazu.

θ zu gewinnen. Diese Parameter sind nur durch die Beziehung auf die unterstellte Modellklasse $\{F_\theta \mid \theta \in \Theta\}$ definiert. Ihnen entspricht kein Wert, der sich allein aus der Beobachtung der Werte der Variablen in der Gesamtheit U gewinnen ließe. Somit wird der realistische Ansatz der klassischen Stichprobentheorie unterlaufen.

Der Superpopulationsansatz geht aber noch nicht weit genug. Wenn man sich für das Antwortverhalten von Befragten interessiert, so müsste im Superpopulationsansatz angenommen werden, dieses Verhalten sei bereits vor jeder Befragung festgelegt, und zwar ganz unabhängig davon, ob jemand tatsächlich befragt wurde oder nicht. Ob also jemand auf die Frage nach dem Haushaltseinkommen gar nicht antwortet oder nur in gruppierter Form, wäre vor jeder Befragung schon entschieden. Denn der Superpopulationsansatz unterstellt fixe Werte (Realisationen des Zufallsprozesses) in der Gesamtheit U zum Zeitpunkt der Stichprobenziehung. Im nächsten Schritt werden wie in der klassischen Theorie Stichproben aus diesen fixen Werten gezogen. Stichprobenfunktionen wie Mittelwerte hängen auf der Stufe der Stichprobenziehung allein davon ab, wer aus der Gesamtheit U in die Stichprobe gelangt. Dies ermöglicht den Anschluss an Ergebnisse der klassischen Theorie, hat aber zur Folge, dass das Antwortverhalten aller Mitglieder der Gesamtheit vor jeder Stichprobenziehung festgelegt sein muss.

Soll nicht nur die Tatsache unvollständiger Daten konstatiert, sondern auch ihr Zustandekommen reflektierbar gemacht werden, dann wird es notwendig, auch Variablen wie Interviewform, Merkmale der Interviewer und vieles mehr zu betrachten. Die Annahme, all diese Variablen seien vor jeder Stichprobenziehung für alle Personen der Gesamtheit festgelegt, ist nicht nur fatalistisch und völlig unrealistisch, sondern würde auch die Spezifikation eines Stichprobendesigns wegen der notwendigen Details praktisch unmöglich machen.

Soll an probabilistischen Auswahlmodellen festgehalten werden, dann muss schließlich ganz auf Elemente der Stichprobentheorie verzichtet werden. Sowohl die interessierenden Größen wie das Haushaltseinkommen, die Stichprobenziehung und das Antwortverhalten der Befragten werden in einem einzigen probabilistischen Modell beschrieben. Ist $(\Omega, \mathbb{B}, \lambda)$ ein hinreichend großer Wahrscheinlichkeitsraum, mit dem alle diese Variablen beschrieben werden können, dann lässt sich etwa das Haushaltseinkommen als Funktion von $u \in U$ und $\omega \in \Omega$ auffassen:

$$Y : U \times \Omega \to \{1,2,3,...\} =: \mathbb{Y}$$

$Y(u,\omega)$ ist also das Haushaltseinkommen, das einer Person $u \in U$ bei Realisierung von $\omega \in \Omega$ zukommt. Eine Person u hat in Abhängigkeit von ω verschiedene Einkommen. Aber es gibt ein $\omega_0 \in \Omega$, für das $(Y(u,\omega_0),\ u \in U)$ den Haushaltseinkommen $(y(u), u \in U)$ in der BRD entspricht.

Der Zusammenhang mit Aussagen über Durchschnitte der $Y(.,\omega)$ über alle $u \in U$ wird hergestellt, indem allen u gleiche Wahrscheinlichkeitsverteilungen zugeschrieben werden und die Unabhängigkeit von $Y(u,.)$ und $Y(u',.)$ für verschieden u, u' angenommen wird.

Tatsächlich in der Stichprobe beobachtet wird aber nur eine mengenwertige Variable mit dem Merkmalsraum

$$\mathbb{Y}^* := \{\{y\} \mid y \in \mathbb{Y}\} \cup \{[1,400), [400, 800), ..., [15000, \infty)\} \cup \{\mathbb{Y}\}$$

der genaue oder gruppierte Angaben bzw. gar keine Angabe darstellt. Es sei nun

$$S\colon \Omega \rightarrow P(U)\setminus\{\varnothing\}$$

eine Stichprobe, wobei $P(U)$ die Potenzmenge von U bezeichnet und S bzgl. $(\Omega, \mathbb{B})$ messbar sein soll. Für die Menge der befragten Personen $u \in S(\omega)$ kann eine neue Variable mit dem Wertebereich $\mathbb{Y}^*$ konstruiert werden, die das „angegebene Haushaltseinkommen" repräsentiert. Um das Problem zu umgehen, allen Personen unabhängig von der Befragung ein Antwortverhalten zuzuschreiben, wird zu $\mathbb{Y}^*$ noch ein Symbol „*" hinzugefügt. Dann kann die Definition auf alle Personen $u \in U$ ausgedehnt werden und es ergibt sich

$$Y^*: U \times \Omega \rightarrow \{\{y\} \mid y \in \mathbb{Y}\} \cup \{[1{,}400), ..., [15000, \infty)\} \cup \{\mathbb{Y}\} \cup \{*\}$$

wobei $Y^*(u,\omega) = *$ für $u \notin S(\omega)$ gesetzt wird. Die Abbildung Y^* repräsentiert das „in der Stichprobe s angegebene Haushaltseinkommen". Mit dieser Konstruktion ist man nicht gezwungen, über die Antworten nicht befragter Personen zu spekulieren. Sowohl für eine gegebene Stichprobe s, also eingeschränkt auf die Menge $\{\omega \mid S(\omega) = s\}$, als auch auf ganz Ω sind $Y(u,.)$ und $Y^*(u,.)$ Zufallsvariablen im Sinn der Wahrscheinlichkeitstheorie. Der Zusammenhang zwischen $Y(u,.)$ und $Y^*(u,.)$ lässt sich durch probabilistische Modelle darstellen. Sie beziehen sich zunächst auf eine Person u. Es muss zusätzlich angenommen werden, die Zufallsvariablen $(Y(u,.),\ u \in U)$ bzw. $(Y^*(u,.),\ u \in U)$ seien stochastisch unabhängig und identisch verteilt. In diesem Rahmen können nun Konsequenzen unvollständiger Angaben für statistische Aussagen abgeschätzt werden.

Die etwas aufwendige Notation ist notwendig, um Verwechslungen zwischen Durchschnitten über die Gesamtheit U und Verteilungen, Erwartungswerten etc. bezüglich des Wahrscheinlichkeitsraums $(\Omega, \mathbb{B}, \lambda)$ zu vermeiden. In der Literatur erscheint die Verwechslung häufig nach einem nicht kenntlich gemachten Übergang von stichprobentheoretischen zu probabilistischen Argumenten. So schreibt z.B. P. Holland: „A probability will mean nothing more nor less than a proportion of units in U. The expected value of a variable is merely its average value over all of U" (Holland 1986: 945). Später verwendet er aber die stochastische Unabhängigkeit zwischen Variablen (1986: 948f.), ohne zu bemerken, dass stochastisch unabhängige Variablen auf endlichen Räumen U nur selten existieren. Eine ähnliche Verwechslung findet sich noch bei Vytlacil (2002: 332).

Durchschnitte über U und Durchschnitte über den Wahrscheinlichkeitsraum $(\Omega, \mathbb{B}, \lambda)$ führen nicht nur zu unterschiedlichen numerischen Ergebnissen, sie sind nicht einmal konzeptionell verbunden.[9] Zwar garantieren asymptotische Aussagen wie starke

9 Eine weitere Konsequenz probabilistischer Ansätze betrifft Designvariablen wie die Ost/West-Differenzierung im ALLBUS. Wird eine solche Variable als Konstante bzw. als degenerierte Zufallsvariable aufgefasst, kann sie wie in der klassischen Stichprobentheorie verwandt werden. Insbesondere können Gewichtungsverfahren benutzt werden, um Designaspekte zu berücksichtigen. Werden Designvariablen dagegen als Zufallsvariablen wie alle anderen behandelt, dann hängen diese Variablen von allen anderen Variablen eines Modells ab und Gewichtungsverfahren verlieren ihre Gültigkeit. Die beiden Ansätze führen z.B. bei Regressionsmodellen zu unterschiedlichen Ergebnissen.

Gesetze oder Ergodensätze, dass die beiden Durchschnitte im Grenzwert ($|U| \to \infty$ oder $\mathbb{E}(|S|)/|U| \to c \notin \{0, 1\}$, $|U| \to \infty$ etc.) gleich sind. Dies sind aber probabilistische Aussagen, die sich auf die Modellebene beziehen, also ein probabilistisches Modell auf $(\Omega, \mathbb{B}, \lambda)$ voraussetzen. Es sind mathematische Konstruktionen, die keine Aussage über empirische Verhältnisse wie Einkommensverteilungen begründen können. Und Grenzwertüberlegungen führen eine zusätzliche Abstraktionsebene ein, die über probabilistische Formulierungen von Antworten auf Fragen nach dem Einkommen hinausgehen.

Le Cam und Yang schreiben hierzu: „It must be pointed out that the asymptotics of the ‚standard i.i.d. case' are of little relevance to practical use of statistics, in spite of their widespread study and use. The reason for this is very simple. One hardly ever encounters fixed families $\{p_\theta \mid \theta \in \Theta\}$ with a number of observations that will tend to infinity. There are not that many particles in the visible universe! The use of such considerations is an abuse of confidence that has been foisted upon unsuspecting students and practitioners owing to the fact that we, as a group, possess limited analytical abilities and, perforce, have to limit ourselves to simple problems. ... The use of asymptotics ‚as $n \to \infty$' for the standard i.i.d. case seems to be based on an entirely unwarranted act of faith" (Le Cam und Yang 1990: 99f.). Selbst wenn asymptotische Argumente als relevant angesehen werden, so wird man konstatieren müssen, dass unterschiedliche Modelle für F zu Ergebnissen führen können, die offenbar nichts über den Durchschnitt von Werten aller $u \in U$ sagen.

Die Abwendung von stichprobentheoretischen Konzepten zugunsten probabilistischer Modelle erfordert zudem eine Klärung der Annahmen, die in probabilistischen Modellen verwandt werden. Insbesondere die Annahme unabhängiger und identisch verteilter Zufallsvariablen ist keine Annahme, die verändert oder aufgegeben werden könnte, ohne den Rahmen des Modells zu sprengen. Sie verweist auf keine gesellschaftlichen Sachverhalte, ebenso wenig wie die Leinwand eines Gemäldes auf Eigenschaften der dargestellten Dinge verweist. Entsprechend gibt es auch keine empirischen Anhaltspunkte, aufgrund derer sich die Annahme zurückweisen ließe. Die Annahme ist weder wahr noch falsch, sondern ein Ausgangspunkt für alle probabilistischen Modelle. Wird in einem nächsten Modellierungsschritt eine Modellklasse, z.B. $\{F_\theta \mid \theta \in \Omega\}$ vorgeschlagen, so wird immer schon die Unabhängigkeit und identische Verteilung der so beschriebenen Zufallsvariablen unterstellt. Auch eine Modellklasse kann daher weder wahr noch falsch sein. Aber die Wahl einer Modellklasse kann sich bei einem Vergleich von Realisierungen der Zufallsvariablen mit empirischen Verteilungen als unangemessen erweisen. Eine solche Kritik von Modellvorschlägen, so notwendig und hilfreich sie ist, führt allerdings selbst unter idealisierten Bedingungen nicht zu der eindeutigen Wahl eines probabilistischen Modells. Der spekulative Spielraum, den probabilistische Modelle immer bieten, kann gerade bei der Behandlung von Daten mit unvollständigen Angaben produktiv genutzt werden. Denn dabei muss immer überlegt werden, was der Fall gewesen sein könnte. Probabilistische Modelle bieten einen Rahmen, eine Vielzahl alternativer Möglichkeiten einfach zu benennen und gegeneinander abzuwägen.

Am einfachsten wäre es, wenn eine Angabe $Y^*(u,\omega)$ nur die offensichtliche Information $Y(u,\omega) \in Y^*(u,\omega)$ enthalten würde.[10] Dann bräuchte man sich bei Schätzungen keine Gedanken über den Zusammenhang von $Y(u,.)$ und $Y^*(u,.)$ zu machen. Der Ansatz sei an zwei Beispielen demonstriert: Zunächst sei die Verteilung F_θ von $Y(u,.)$ durch einen endlich-dimensionalen Vektor $\theta \in \mathbb{R}^k$ parametrisiert, etwa $Y(u,.) =_d \mathrm{N}(\mu,\sigma^2)$, also normalverteilt mit Erwartungswert μ und Varianz σ^2, $\theta = (\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Der Beitrag einer Beobachtung $y(u)$ zur Likelihood $L(\theta; y(u), u \in s)$ ist dann $\phi(y(u); \mu,\sigma^2)$, wobei $\phi(.; \mu,\sigma^2)$ die Dichte der Normalverteilung mit Erwartungswert μ und Varianz σ^2 ist. Sind nun $\{Y(u,.), u \in s\}$ stochastisch unabhängig für gegebenes s, dann ist die Likelihood das Produkt der Dichten

$$L\Big((\mu,\sigma^2); y(u), u \in s\Big) = \prod_{u \in s} \phi\Big(y(u);\ \mu,\sigma^2\Big)$$

wobei unterstellt wird, der Stichprobenplan habe nichts mit den Variablen $Y(u,.)$ zu tun. Werden alle Informationen über das Zustandekommen einer Realisation $y^*(u)$ von $Y^*(u,.)$ vernachlässigt, wird also nur $y(u) \in y^*(u)$ berücksichtigt, und sind die $Y^*(u,.)$ weiterhin stochastisch unabhängig, dann wird die Likelihood zu[11]

$$L\big(\theta; y^*(u), u \in s\big) = \prod_{u \in s} \int_{v \in y^*(u)} dF_\theta(v)$$

Im Fall der Angaben zum Haushaltseinkommen im ALLBUS ergibt sich

$$\begin{aligned} &L\big(\theta; y^*(u), u \in s_2\big) \\ &= \prod_{u \in s_4} \phi\big(y(u);\theta\big) \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} \phi(v;\theta)\, dv \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} \phi(v;\theta)\, dv \end{aligned}$$

Der erste Faktor repräsentiert den Beitrag der genauen Beobachtungen, der zweite den der gruppierten Angaben, und der letzte gibt den Beitrag der Verweigerungen (inklusive keine Angabe/weiß nicht) wieder. Der letzte Term ist konstant 1 und damit unabhängig von den Parametern, so dass man sich auf die ersten beiden Terme konzentrieren kann. Maximiert man diese Likelihoodfunktion, ergibt sich für die Nettostichprobe s_2 und ohne Berücksichtigung der unterschiedlichen Auswahlsätze für Ost- und Westdeutschland $\hat{\mu} = 3807$ und $\hat{\sigma} = 2047$ sowie ein modellbasiertes 95 Prozent-Konfidenzintervall für μ von (3729, 3885). Der naive Ansatz, der die $n_3 - n_4 = 906$ gruppierten Angaben ganz unberücksichtigt lässt und nur die Daten der Stichprobe s_4 benutzt, er-

10 Im Folgenden wird unterstellt, dass die Befragten nicht „lügen", also immer $Y(u,\omega) \in Y^*(u,\omega)$ gilt. Letzteres war, ohne einen probabilistischen Rahmen, bereits bei der Betrachtung intervallwertiger Statistiken unterstellt worden.

11 Die Formulierung ist bei absolut stetigen Verteilungen wie der Normalverteilung offenbar nicht korrekt. Denn falls es mindestens eine exakte Beobachtung gibt, wird der entsprechende Term 0. Es gibt verschiedene Vorschläge, wie auch absolut stetige Verteilungen in dieser allgemeinen Form behandelt werden können, vgl. Jacobsen und Keiding (1995), Gill et al. (1997) und Nielsen (2000).

gibt einen Mittelwert von 3676 mit einem (ungewichteten) 95 Prozent-Konfidenzintervall (3582, 3770). Der Wert 3676 liegt 131 DM unter dem Wert, der sich unter Berücksichtigung der gruppierten Angaben innerhalb des Normalverteilungsmodells ergibt, sogar außerhalb des Konfidenzintervalls (3729, 3885).

Anstelle einer Normalverteilung kann auch unterstellt werden, $\log(Y(u,.))$ sei normalverteilt mit den Parametern (μ,σ^2). Die Maximierung der entsprechenden Likelihood führt zu $\hat{\mu} = 8{,}1113$ und $\hat{\sigma} = 0{,}5371$. Da $\mathbb{E}_\theta(Y) = \exp(\mu + \sigma^2/2)$ für log-normalverteilte Zufallsvariablen Y ist, ergibt sich als Schätzung des Erwartungswerts 3849 mit dem (modellbasierten) Konfidenzintervall (3765, 3932), berechnet mit der Deltamethode (Lehmann 1999: 85ff.). Der Erwartungswert unter diesem Modell ist um 173 DM größer als der naive Durchschnitt. Wählt man schließlich die log-logistische Verteilung mit $F_\theta(y) = \theta y/(1 + \theta y)$, dann ergibt sich $\hat{\theta} = 2{,}9879 \times 10^{-4}$. Der Erwartungswert unter diesem Modell ist allerdings ∞. Man würde auch einen Erwartungswert von ∞ erhalten, wenn die Dichte etwa der log-normalen Verteilung rechts von einem beliebig großen Wert y_0 durch die entsprechende Dichte der log-logistischen Verteilung ersetzt würde. Aber eine solche Veränderung in der Wahl der Modellklasse lässt sich empirisch nicht beurteilen, weil y_0 immer größer als alle Beobachtungen gewählt werden kann. Der Mittelwert von Merkmalen einer endlichen Menge U ist dagegen sicher immer endlich. Hier zeigt sich der bereits angedeutete Perspektivenwechsel: An Stelle eines Durchschnitts über alle $u \in U$ interessiert der Durchschnitt über die $\omega \in \Omega$, durch die der Parameter θ erst seine Bedeutung erhält. Es gibt aber zwischen dem Mittelwert einer endlichen Menge U und dem Erwartungswert einer Zufallsvariablen $Y(u_0,.)$ keine notwendigen Beziehungen.

Es könnte scheinen, das Problem entstehe durch eine zu enge Wahl der Modellklasse und könne durch nichtparametrische Verfahren gelöst werden. Wird „nur" angenommen, die $(Y(u,.),\ u \in s)$ seien unabhängig und identisch verteilt, dann ist die nichtparametrische Likelihood für die Verteilung F bei Daten $\{Y^*(u,.) \mid u \in s\}$

$$L\big(F; y^*(u), u \in s\big) = \prod_{u \in s} \int_{v \in y^*(u)} dF(v)$$

Im Fall von s_2 ergibt sich

$$\begin{aligned} &L\big(F; y^*(u), u \in s\big) \\ &= \prod_{u \in s_4} F\big(y(u)\big) - F\big(y(u)_-\big) \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} dF(v) \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} dF(v) \end{aligned}$$

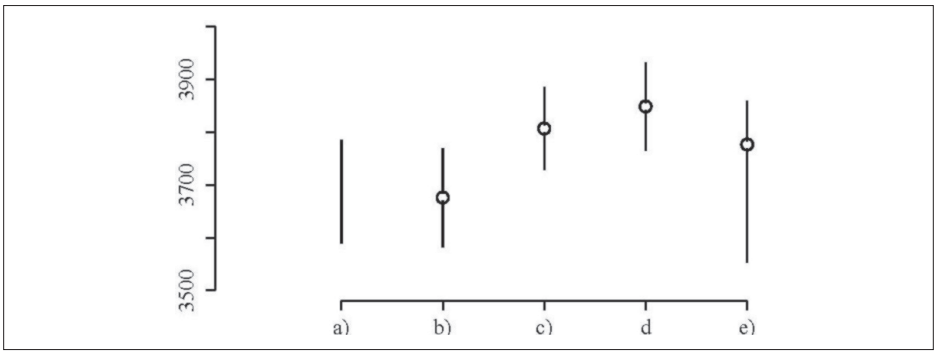
wobei $F(y) - F(y_-)$ die Sprunghöhe der Funktion F an der Stelle y ist. Werden nur diskrete Verteilungen F betrachtet, dann existiert häufig ein Maximum der Likelihoodfunktion, etwa $\hat{F}$. $\hat{F}$ wird nichtparametrischer Maximum-Likelihood-Schätzer (NPMLE) der Verteilungsfunktion F genannt. Als Schätzer des Erwartungswerts kann $\hat{\mu} := \int v\, d\hat{F}(v)$ verwandt werden.

Im Fall der Haushaltseinkommen ergibt sich $\hat{\mu} = 3777$. Ein 95 Prozent-Konfidenzintervall, basierend auf 10000 Bootstrap-Replikationen, ist (3553, 3859).[12] Auf den

12 Die Verteilung der geschätzten Mittelwerte für die Replikationen ist multimodal. Daher ergibt

ersten Blick könnte es scheinen, als ergäbe sich immer ein endlicher Schätzwert $\hat{\mu}$, der nun nicht mehr von der Wahl einer Modellklasse abhinge. Das ist bei unvollständigen Angaben aber nicht der Fall. Der NPMLE ist nicht eindeutig definiert, wenn es gruppierte Beobachtungen gibt, aber keine genauen Beobachtungen in dieses Intervall fallen. Dann kann F auf dem Intervall beliebig definiert werden, ohne den Wert der Likelihoodfunktion zu ändern. Fällt insbesondere keine genaue Beobachtung in die größte Einkommensklasse „Einkommen größer als 15000 DM", während das Intervall wenigstens einmal genannt wird, dann kann die zugehörige Masse in $\hat{F}$ beliebig gegen ∞ verschoben werden. Das geschieht in den Bootstrap-Replikationen so oft, dass die obere Schranke des Konfidenzintervalls ehrlicherweise durch ∞ ersetzt werden müsste. In der Tat ist in der Berechnung des „Konfidenzintervalls" der replizierte Datensatz einfach die Angabe „Einkommen größer als 15000 DM" durch den Wert 22500 ersetzt worden, falls keine genauen Beobachtungen in das Intervall fielen. Jeder andere Wert > 15000 wäre aber genauso möglich. Auch unter nichtparametrischen Modellen ergibt sich ein beliebig großer Unterschied zwischen dem ursprünglich interessierenden Durchschnitt von Werten einer Gesamtheit U und dem Erwartungswert $\int v\, dF(v) = \int Y(u_0,\omega)\, d\lambda(\omega)$, einem Durchschnitt über $\omega \in \Omega$.

Abbildung 2: Geschätztes mittleres Haushaltseinkommen und 95 Prozent Konfidenzintervalle. a) Nichtparametrische Schranken auf 3/3, b) Normalverteilung ohne gruppierte Angaben, c) Normalverteilung unter Einschluss gruppierter Angaben, d) Lognormalverteilung unter Einschluss gruppierter Angaben, e) Nichtparametrischer Mittelwert unter Einschluss gruppierter Angaben



VI. Ignorierbare Ausfälle: MAR, CAR und all das

Der Perspektivenwechsel führt zu einer Abkehr von dem Versuch, probabilistische Aussagen über den Zusammenhang eines Mittelwerts in einer Gesamtheit U mit Mittelwerten über Stichproben zu formulieren. Das mag gerechtfertigt sein, wenn stattdessen

sich ein sehr asymmetrisches Konfidenzintervall, wenn wie hier die Perzentile der Verteilung zur Konstruktion benutzt werden.

die Verwendung probabilistischer Modelle zum Verständnis von Effekten unvollständiger Angaben beiträgt, zumal über deren Zustandekommen nur spekuliert werden kann. Zunächst muss geklärt werden, unter welchen Bedingungen die Verwendung von $L(\theta; y^*(u), u \in s)$ im vorigen Abschnitt begründet werden kann. Dies kann nicht immer der Fall sein, selbst wenn für alle Personen $u \in U$ identische Beziehungen zwischen $Y(u,.)$ und $Y^*(u,.)$ unterstellt werden.

Sei z.B. $\mathbb{Y} = \{1000, 1500, 2000\}$ und $\mathbb{Y}^* = \{\{1000, 1500\}, \{1000\}, \{1500\}, \{2000\}\}$. Ist nun $y(u) = 1000$, dann kann u entweder {1000} berichten, also den genauen Betrag nennen, oder aber mit {1000, 1500} antworten. Entsprechendes gilt für $y(u) = 1500$. Ist nun $Y(u,.)$ auf {1000, 1500, 2000} gleichverteilt und berichten alle Personen $y^*(u) = \{1000, 1500\}$ falls $y(u) = 1000$, sonst aber immer den genauen Betrag, dann ergibt sich die folgende Verteilung auf $\mathbb{Y}^*$:

$$\Pr(Y^*(u,.) = \{1000, 1500\}) = 1/3$$
$$\Pr(Y^*(u,.) = \{1000\}) = 0$$
$$\Pr(Y^*(u,.) = \{1500\}) = 1/3$$
$$\Pr(Y^*(u,.) = \{2000\}) = 1/3$$

Werden die Beobachtungen wie im letzten Abschnitt behandelt, so wird $\Pr(Y^*(u,.) = \{1000, 1500\}) = \Pr(Y(u,.) = 1000) + \Pr(Y(u,.) = 1500)$ gesetzt. Ist die Verteilung von $Y^*(u,.)$ bekannt, so ergibt sich als Verteilung von $Y(u,.)$: $\Pr(Y(u,.) = 1000) = 0$, $\Pr(Y(u,.) = 1500) = 2/3$, $\Pr(Y(u,.) = 2000) = 1/3$. Für den Zusammenhang zwischen Y und Y^* wird $\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1000) = 1/2 = \Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1500)$ angenommen. Die bedingten Wahrscheinlichkeiten von $\{Y^*(u,.) = \{1000, 1500\}\}$ sind für beide möglichen Bedingungen $\{Y(u,.) = 1000\}$ und $\{Y(u,.) = 1500\}$ gleich. Kombiniert man die unterstellte Verteilung von $Y(u,.)$ mit den beiden konditionalen Verteilungen, dann ergibt sich in der Tat die Verteilung von $Y^*(u,.)$. Die Interpretation wäre: Ist $Y(u,.) = 1500$, dann antwortet u mit Wahrscheinlichkeit 1/2 entweder {1500} oder {1000, 1500}. Die entscheidende Annahme ist offenbar die über die bedingten Verteilungen $\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1000)$ und $\Pr(Y^*(u,.) = \{1000, 1500\} \mid Y(u,.) = 1500)$. Die bedingten Verteilungen beschreiben den Zusammenhang zwischen $Y(u,.)$ und $Y^*(u,.)$.

Sei nun allgemein $Y(u,.)$ eine Zufallsvariable mit Werten in der endlichen Menge y und $Y^*(u,.)$ eine Zufallsvariable mit Werten in $\mathbb{Y}^* \subseteq P(\mathbb{Y}) \setminus \{\emptyset\}$ auf dem gemeinsamen Raum $(\Omega, \mathbb{B}, \lambda)$. Dann soll $Y^*(u,.)$ zufällige Vergröberung (CAR, coarsened at random) von $Y(u,.)$ heißen, wenn eine der folgenden äquivalenten Bedingungen für alle $y^* \in \mathbb{Y}^*$ und für alle $y \in y^*$ erfüllt ist:

$$\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) \text{ ist konstant auf } y \in y^* \quad (1)$$
$$\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) = \Pr(Y^*(u,.) = y^* \mid Y(u,.) \in y^*) \quad (2)$$
$$\{Y^*(u,.) = y^*\} \perp \{Y(u,.) = y\} \mid \{Y(u,.) \in y^*\} \quad (3)$$
$$\Pr(Y(u,.) = y \mid Y^*(u,.) = y^*) = \Pr(Y(u,.) = y \mid Y(u,.) \in y^*) \quad (4)$$

Dabei bedeutet $A \perp B|C$ die bedingte stochastische Unabhängigkeit der Ereignisse A und B gegeben C. $\{\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y) \mid y^* \in \mathbb{Y}^*, y \in y^*\}$ soll im Folgenden Selektionsmodell heißen.

Die Bedingungen (1) und (2) beschreiben die Situation ausgehend vom Wert $y(u)$ der zugrunde liegenden Variablen $Y(u,.)$. Bei gegebenem $y(u)$ müssen nur noch die Antwortmöglichkeiten $Y^*(u,.)$ in Erwägung gezogen werden. Ist etwa das tatsächliche Haushaltseinkommen eines Befragten 1234 DM, dann kann er im ALLBUS entweder 1234 oder das Intervall [1000, 1250) angeben, oder er antwortet gar nicht. Aus diesen Antwortalternativen wählt u mit den bedingten Wahrscheinlichkeiten $\Pr(Y^*(u,.) = y^* | Y(u,.) = 1234)$. Ist dagegen $y(u) = 1123$, so erzwingt (1) eine Auswahl zwischen den drei Antwortmöglichkeiten 1123, [1000,1250) und $[1,...,\infty)$ mit den gleichen Wahrscheinlichkeiten wie im Fall $y(u) = 1234$. Für alle möglichen Einkommen im Bereich [1000, 1250) entscheidet sich u nach den gleichen Wahrscheinlichkeiten zwischen einer genauen Angabe, der gruppierten Angabe oder gar keiner Angabe. Für Werte in einem anderen Gruppierungsintervall kann sich eine andere Aufteilung zwischen den Antwortmöglichkeiten „genau" und „gruppiert" ergeben. Die Bedingung (2) ist nur eine Umformulierung dieser Beschreibung. Denn (2) verlangt, dass die Entscheidung, kein Haushaltseinkommen anzugeben, unabhängig von dem tatsächlichen Einkommen getroffen wird, während die Entscheidung zwischen einer gruppierten oder genauen Angabe innerhalb eines Gruppierungsintervalls nicht von der tatsächlichen Höhe des Einkommens abhängt. Wenn bekannt ist, dass das tatsächliche Einkommen $Y(u,.) \in$ [1000, 1250) ist, dann verlangt (3) die Unabhängigkeit des Ereignisses „gruppierte Angabe" von den tatsächlichen Einkommen innerhalb des Intervalls. Ist $y^* = \mathbb{Y} = [1,2,...,\infty)$, dann ist die Bedingung $Y(u,.) \in [1,2,...,\infty)$ immer erfüllt und (3) verlangt die (unbedingte) stochastische Unabhängigkeit von $\{Y^*(u,.) = [1,2,...,\infty)\}$ und $Y(u,.)$. Ist dagegen $Y(u,.) \in y^* = \{y\}$, dann gilt $Y(u,.) = y$, und die Bedingung (3) ist automatisch erfüllt.

Die Bedingung (4) ist besonders hilfreich, weil sie nicht von den zugrunde liegenden Werten $Y(u,.)$, sondern von den beobachteten Angaben $Y^*(u,.)$ ausgeht. Ist etwa $y^*(u) = [1000, 1250)$, dann verlangt (4), dass die Verteilung der tatsächlichen Werte $Y(u,.)$ sich nicht von der bedingten Verteilung der $Y(u,.)$ unterscheidet, wenn bekannt ist, dass $Y(u,.)$ in dem Intervall [1000,1250) liegt. Genau dies ist in der Konstruktion der Likelihoodfunktionen im letzten Abschnitt verwandt worden.

Schreibt man $\Pr_\theta(Y(u,.) = y)$, um die Abhängigkeit der Verteilung von einem Parameter $\theta \in \Theta$ anzugeben, und entsprechend $\Pr_\gamma(Y^*(u,.) = y^* | Y(u,.) = y)$ mit $\gamma \in \Gamma$ für das Selektionsmodell, dann kann die Verteilung der Beobachtungen $Y^*(u,.)$ wie folgt aufgespalten werden:

$$
\begin{aligned}
&\Pr_{\theta,\gamma}(Y^*(u,.) = y^*) \\
&= \sum_{y \in y^*} \Pr_{\theta,\gamma}\left(Y^*(u,.) = y^*, Y(u,.) = y\right) \\
&= \sum_{y \in y^*} \Pr_\theta\left(Y(u,.) = y\right) \Pr_\gamma\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) = y\right) \\
&= \Pr_\gamma\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) = y,\, y \in y^*\right) \sum_{y \in y^*} \Pr_\theta\left(Y(u,.) = y\right) \\
&= \Pr_\theta\left(Y(u,.) \in y^*\right) \Pr_\gamma\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) \in y^*\right)
\end{aligned}
$$

Die dritte Gleichung folgt aus der CAR-Bedingung (1), die letzte Gleichung aus (2). Sind die Parameter θ und γ variationsunabhängig, gibt es also zu jedem Element $(\theta, \gamma) \in \Theta \times \Gamma$ eine Wahrscheinlichkeitsverteilung $\Pr_{\theta,\gamma}$, dann kann bei Likelihoodbetrachtungen für θ der Selektionsteil $\Pr_\gamma(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ vernachlässigt werden. Es reicht,

$$L\big(\theta; y^*(u), u \in s\big) = \prod_{u \in s} \Pr_\theta\big(Y(u,.) \in y^*\big)$$

zu maximieren. Antworten die Befragten entweder mit einer genauen Angabe oder gar nicht, dann ist $y^* = \{\{y\} \mid y \in \mathbb{Y}\} \cup \{\mathbb{Y}\}$. In diesem Fall impliziert CAR die Möglichkeit, sich nur auf die genauen Angaben beschränken zu können. Die CAR-Bedingung ist in diesem Zusammenhang auch MAR (missing at random) genannt worden.

Es kann gezeigt werden, dass es zu jeder vorgelegten Verteilung auf $\mathbb{Y}^*$ immer eine Verteilung auf $\mathbb{Y}$ und ein Selektionsmodell gibt, das die CAR-Bedingung erfüllt (Gill et al. 1997: 262; Heitjan 1994; Heitjan und Rubin 1991; Grünwald und Halpern 2003). Ist insbesondere $\{y\} \in \mathbb{Y}^*$ für alle $y \in \mathbb{Y}$ und $\Pr(Y^*(u,.) = \{y\}) > 0$, dann ist die Verteilung von $Y(u,.)$ durch die CAR-Bedingung sogar eindeutig bestimmt. Ist die Verteilung von $Y^*(u,.)$ bekannt, dann kann immer ein CAR-Modell unterstellt werden. Mit anderen Worten: Keine noch so große Menge an Daten erlaubt es, zwischen einem ignorierbaren Selektionsmodell, das die CAR-Bedingung erfüllt, und nicht ignorierbaren Modellen (wie am Anfang des Abschnitts) zu unterscheiden.[13] Selektionsmodelle sind nicht identifizierbar: Wird ein beliebiges Selektionsmodell vorgeschlagen, so kann immer ein CAR-Modell angegeben werden, der ebenso gut zu den Daten passt.

Wenn von einem CAR-Modell ausgegangen wird, dann braucht, basierend auf der Likelihoodtheorie, kein Selektionsmodell angegeben zu werden. Manchmal erscheint es aber sinnvoll, sich selbst in der CAR-Situation ein Bild des Selektionsprozesses zu machen. Wird ein (semi-)parametrisches Modell für den Selektionsprozess gewählt, dann hat dies empirische Konsequenzen, kann sich also als falsch erweisen. Denn unter der CAR-Bedingung sind auch die $\Pr(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ eindeutig bestimmt, falls nur $\Pr(Y^*(u,.) = y^*) > 0$ ist. Die Annahme einer Klasse $\Pr_\gamma(Y^*(u,.) = y^* \mid Y(u,.) \in y^*)$ von Selektionsmodellen kann unter der CAR-Bedingung zumindest potentiell aus empirischen Gründen zurückgewiesen werden, jedenfalls dann, wenn neben exakten und vollständig fehlenden Angaben auch partielle Angaben zur Verfügung stehen und modelliert werden.

Zusammenfassend lässt sich sagen, dass erstens probabilistische Überlegungen zu einer nicht trivialen Charakterisierung von Bedingungen führen, unter denen klassische Methoden für unvollständige Beobachtungen korrekt sind: die CAR-Bedingungen. Zweitens zeigt sich, dass Selektionsmodelle empirisch nicht identifiziert sind: Es kann immer ein CAR-Modell konstruiert werden, das die Daten exakt reproduziert, ganz unabhängig davon, wie die Daten „tatsächlich" entstanden sind. Die CAR-Annahme und damit die Verwendung klassischer Likelihood-Methoden lässt sich empirisch nicht hinterfragen. Drittens ist es selbst unter der CAR-Bedingung möglich, einige (semi-)

13 Das Ergebnis gilt nicht nur für endliche Mengen $\mathbb{Y}$, sondern im Wesentlichen auch in allgemeinen Räumen. Allerdings wird dann die Formulierung sehr aufwendig (Gill et al. 1997: 273ff.).

parametrische Selektionsmodelle empirisch zurückzuweisen, wenn neben exakten und vollständig fehlenden Angaben auch gruppierte oder andere partielle Angaben vorliegen und entsprechend modelliert werden. Man darf aber bei all dem Fortschritt nicht vergessen, dass probabilistische Modelle untersucht werden. Die Modelle können nicht umstandslos mit dem realen Verhalten von Befragten gleichgesetzt werden. Denn die Modelle unterstellen u.a., alle Befragten würden ihr Antwortverhalten nach einer Wahrscheinlichkeitsverteilung auswürfeln, die zudem für alle gleich wäre.

VII. Ignorierbare Ausfälle: Konditionale CAR-Modelle

In vielen Fällen gibt es neben den Angaben ($y^*(u)$, $u \in s$) zu den interessierenden Größen ($y(u)$, $u \in U$) weitere Informationen. Dabei kann es sich um Designvariablen handeln, deren Werte zumindest für die intendierte Stichprobe bekannt sind, um Angaben über die Kontaktaufnahme oder um Angaben des Befragten aus anderen Teilen des Interviews. Im ALLBUS gibt es z.B. Angaben für alle Befragten aus s_2 zu Geschlecht, Alter, Haushaltsgröße, Staatsangehörigkeit und Befragungsgebiet (Ost/West). Nun kann die globale Annahme einer CAR-Bedingung unrealistisch erscheinen. Gleichwohl könnte die CAR-Bedingung getrennt für alle Teilmengen gelten, die durch die zusätzlichen Angaben $X(u,.)$ definiert werden.

Wenn die zusätzlichen Angaben in einem Vektor X zusammengefasst werden, dann kann ein entsprechender Zufallsvektor konstruiert werden: $X : U \times \Omega \rightarrow \mathbb{X}$. Die CAR-Bedingungen können konditional auf die Werte dieser Kovariablen formuliert werden:

$$\Pr\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) = y, X(u,.) = x\right) \text{ ist konstant auf } y \in y^* \tag{5}$$

$$\begin{aligned}&\Pr\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) = y, X(u,.) = x\right)\\ &= \Pr\left(Y^*(u,.) = y^* \,\middle|\, Y(u,.) \in y^*, X(u,.) = x\right)\end{aligned} \tag{6}$$

$$\left\{Y^*(u,.) = y^*\right\} \perp \left\{Y(u,.) = y\right\} \;\middle|\; \left\{Y(u,.) \in y^*\right\}, X(u,.) \tag{7}$$

$$\begin{aligned}&\Pr\left(Y(u,.) = y \,\middle|\, Y^*(u,.) = y^*, X(u.,) = x\right)\\ &= \Pr\left(Y(u,.) = y \,\middle|\, Y(u,.) \in y^*, X(u,.) = x\right)\end{aligned} \tag{8}$$

In vielen Texten wird suggeriert, die CAR-Annahme werde plausibler, wenn nur genügend „Informationen“ in Form von Kovariablen in das Selektionsmodell einbezogen werden, wenn also ein möglichst großer Vektor $X(u,.)$ gewählt wird. So schreiben Little und Rubin: „... we believe that in situations where good covariate information is available and included in the analysis, the missing at random (MAR) assumption may often be a reasonable approximation to reality, thus obviating the need for a sensitivity analysis to model nonignorable nonresponse“ (Little und Rubin in Scharfstein et al. 1999: 1130).

Die Argumentation beruht auf einer fehlerhaften Gleichsetzung von „Information“ mit bedingten Verteilungen. Während im umgangssprachlichen Gebrauch des Wortes

eine „bessere Information“ immer zu einem besseren Verständnis einer Situation oder eines Ereignisses beiträgt, gilt dies nicht für die Einbeziehung zusätzlicher Kovariabler in den Bedingungen (5) – (8).[14]

Eine Idee, die schon von Pearson Anfang des 20. Jahrhunderts formuliert wurde, geht davon aus, dass Y eine lineare Regression auf den Vektor X besitzt, $\mathbb{E}(Y|X) = X\beta$ (Lawley 1943). Die lineare Beziehung soll dabei sowohl für die Teilmenge der vollständigen Angaben als auch für die Gesamtheit U mit dem jeweils gleichen β gelten. An Stelle der Identität der Erwartungswerte in allen Teilstichproben wird also nur die Identität der Regressionsfunktion sowie eine konstante bedingte Varianz gefordert. Haben $Y|X$ und $Y|X, I[R = 1]$ die gleiche Verteilung, dann ist das Modell sicherlich CAR. Wird nur die Gleichheit der Erwartungswerte und Varianzen gefordert, so ergibt sich ein etwas allgemeineres Selektionsmodell.

Im Pearson-Lawley Modell kann zunächst β auf der Teilstichprobe mit vollständigen Angaben geschätzt werden; in einem zweiten Schritt aufgrund von $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$ ein Schätzer des Erwartungswerts von Y durch

$$\hat{\mu} = \int \hat{\mathbb{E}}(Y \mid X = x)\, d\hat{F}_X(x) = \frac{1}{|s|} \sum_{u \in s} x(u)\hat{\beta}$$

konstruiert werden. Ist die Regression auch homoskedastisch, dann lassen sich Schätzer für die Varianzen angeben. Im Fall des ALLBUS ergibt sich $\hat{\mu} = 3763$ mit dem 95 Prozent Konfidenzintervall (3660, 3866), wenn als Kovariablen das Alter, Geschlecht, Haushaltsgröße und Staatsangehörigkeit benutzt werden.[15] Der naive Mittelwert, der nur vollständige Angaben berücksichtigt, ist um 90 DM kleiner und befindet sich am unteren Rand des Konfidenzintervalls.

Im Rahmen des Pearson-Lawley Ansatzes können auch die gruppierten Angaben aus $s_3 \setminus s_4$ berücksichtigt werden, indem die Methoden des letzten Abschnitts auf die Residuen $y(u) - x(u)\beta$ angewandt werden. Außerdem kann an Stelle der linearen Regression ein beliebiges (parametrisches oder semiparametrisches) Regressionsmodell benutzt werden. Ein Nachteil der Methode ist aber ihre Abhängigkeit von Annahmen bezogen auf die Regressionsgleichung. Diese ist in der Regel nicht direkt von Interesse. Im Fall des ALLBUS soll eine Aussage über das mittlere Haushaltseinkommen getroffen werden, nicht aber über einen Regressionszusammenhang. Zudem sind die Kova-

14 Sind z.B. U, V normalverteilte unabhängige Zufallsvariablen mit Erwartungswert 0 und Varianz 1, dann sind $Y_1 := U + V$ und $Y_2 := U - V$ unabhängig und normalverteilt, also $Y_1 \perp Y_2$. Dagegen ist die bedingte Kovarianz, wenn die „Information“ $U = u$ gegeben ist: $\mathrm{Cov}(Y_1, Y_2 \mid U = u) = \mathbb{E}(Y_1 Y_2 \mid U = u) - \mathbb{E}(Y_1 \mid U = u)\mathbb{E}((Y_2 \mid U = u) = \mathbb{E}((u + V)(u - V)) - u^2 = \mathrm{Var}(V) - u^2 = 1 - u^2$. $Y_1 \perp Y_2 \mid U = u$ gilt also nur, falls $U = 0$ ist, ein Ereignis mit Wahrscheinlichkeit 0. Die Einführung der zusätzlichen „Information“ $U = u$ führt von unabhängigen Variablen Y_1 und Y_2 zu korrelierten bedingten Variablen. Die Bedingung (7) kann durch die Einbeziehung weiterer Kovariabler also auch verletzt werden. Weitere Probleme bei der Interpretation von „Information“ bei bedingten Modellen diskutieren Dubra und Echenique (2004).

15 Die 6 Beobachtungen ohne Altersangabe wurden ausgeschlossen, die Schätzung von β erfolgte auf s_4 ohne diese Beobachtungen. Die Haushaltsgröße bezieht sich auf die Anzahl der Personen im Haushalt, einschließlich der Kinder. Die Angabe wurde gruppiert, indem Haushalten mit mehr als 4 Personen der Wert 5 zugeordnet wurde. Bei der Staatsangehörigkeit wird nur unterschieden, ob jemand einen ausländischen Pass hat oder nicht.

riablen X, die für alle Befragten zur Verfügung stehen, hauptsächlich durch das Stichprobendesign und die Fragebogenkonstruktion bestimmt, nicht durch inhaltliche Überlegungen. Daher wird versucht, den Regressionszusammenhang $\mathbb{E}(Y|X)$ möglichst allgemein, d.h. ohne parametrische Annahmen, zu modellieren. Das Haushaltseinkommen ist sicherlich keine lineare Funktion des Alters. Theoretisch gibt es keinen sinnvollen Zusammenhang zwischen dem Alter eines Befragten und dem Haushaltseinkommen, es sei denn, es handelt sich um Ein-Personen-Haushalte. Der lineare Term für das Alter sollte daher flexibler, etwa durch Spline-Funktionen dargestellt werden. Zudem sollten möglichst alle Interaktionen zwischen den Kovariablen berücksichtigt werden. Wenn aber keine parametrischen Annahmen oder wenigstens Annahmen über die Glätte der Regressionsbeziehung getroffen werden können, dann gibt es innerhalb des Ansatzes nicht einmal ein Schätzverfahren, das gleichmäßig konsistent ist (Robins und Ritov 1997: 294f.). Selbst wenn diese theoretischen Schwierigkeiten ignoriert werden, bleibt das praktische Problem, ein relativ stabiles und gleichzeitig allgemein akzeptierbares Regressionsmodell für Y gegeben X zu formulieren.

Eine Möglichkeit, zumindest einige der theoretischen Probleme zu umgehen, ergibt sich aus einem Rückgriff auf eine Idee der Stichprobentheorie und verwendet gewichtete Schätzgleichungen. Wird nur zwischen vollständigen und fehlenden Angaben unterschieden, dann kann ein probabilistisches Modell für das Fehlen einer Angabe in Abhängigkeit von den Kovariablen X formuliert werden. Wird $R(u,.) = 1$ gesetzt, falls $Y(u,.)$ beobachtet wurde, $R(u,.) = 2$ sonst, dann ist ein Modell für das Fehlen von Angaben über $Y(u,.)$ etwa durch $\pi(u,x) := \Pr(R(u,.) = 1 \mid X(u,.) = x) = \Pr(R(u,.) = 1 \mid X(u,.) = x,\ Y(u,.) = y)$ bestimmt. Die Variable $\pi(u,X)$ wird häufig „Propensity Score" genannt. Ignorierbarkeit des Selektionsmodells besteht gerade in der Unabhängigkeit des Propensity Scores $\pi(u,x)$ von y, also in der Annahme $R(u,.) \perp Y(u,.) \mid X(u,.)$. Ist die Auswahlwahrscheinlichkeit für alle Kovariablenwerte größer als eine positive Schranke, $\pi(u,X) > \sigma > 0$ für alle $X(u,.)$, dann gilt

$$
\begin{aligned}
&\mathbb{E}\left(\frac{I\left[R(u,.)=1\right]Y(u,.)}{\pi(u,X)}\right)\\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{I\left[R(u,.)=1\right]Y(u,.)}{\pi(u,X)} \;\middle|\; Y(u,.)=y, X(u,.)=x\right)\right)\\
&= \mathbb{E}\left(\frac{Y(u,.)}{\pi(u,X)}\mathbb{E}\left(I\left[R(u,.)=1\right] \;\middle|\; Y(u,.)=y, X(u,.)=x\right)\right)\\
&= \mathbb{E}\left(\frac{\pi(u,X)Y(u,.)}{\pi(u,X)}\right) = \mathbb{E}\left(Y(u,.)\right)
\end{aligned}
$$

Daher ist

$$
\hat{\mu} = \sum_{u\in s}\frac{I\left[R(u,.)=1\right]Y(u,.)}{\pi(u,X)} \Bigg/ \sum_{u\in s}\frac{I\left[R(u,.)=1\right]}{\pi(u,X)}
$$

ein erwartungstreuer Schätzer des Erwartungswerts der $Y(u,.)$, und zwar ganz unabhängig von der Form des bedingten Erwartungswerts $\mathbb{E}(Y \mid X)$. Allerdings muss ein Schät-

zer für $\pi(u,X)$ angegeben werden. Wird z.B. ein Logit-Modell $\hat{\pi}(u, x(u)) = \exp(x(u)\hat{\beta})/1+\exp(x(u)\hat{\beta})$ verwendet, dann ergibt sich für das mittlere Haushaltseinkommen $\hat{\mu}$ = 3765 mit einem Konfidenzintervall von (3599, 3931).

Das deutlich größere Konfidenzintervall des gewichteten Schätzers im Vergleich zum Pearson-Lawley-Schätzer ist eine Konsequenz sowohl der abgeschwächten probabilistischen Annahmen als auch der nur unvollständigen Nutzung der Verteilung der $X(u,.)$ sowie der gemeinsamen Verteilung von $Y(u,.)$ und $X(u,.)$ auf $\{u \in s \mid R(u,.) = 1\}$ für die Schätzung. Ein Teil der Information kann durch die Addition eines weiteren, von Funktionen der $X(u,.)$ und $I[R(u,.) =1]\,Y(u,.)$ abhängigen Terms zur Schätzgleichung zurückerhalten werden.[16] Allerdings können im Rahmen gewichteter Schätzer partielle Angaben nur sehr rudimentär Berücksichtigung finden. Die Schätzgleichung beruht zunächst nur auf vollständigen Angaben, partielle Angaben werden im Gegensatz zum Pearson-Lawley Ansatz oder zu Likelihood-Methoden nur über weitere additive Terme in die Schätzgleichung eingeführt.

VIII. Nicht ignorierbare Ausfälle

Bei der Diskussion von Selektionsmodellen wird überlegt werden müssen, was geschieht, wenn sie nicht die CAR-Bedingung erfüllen. In diesem Fall hängt $\Pr(Y^*(u,.) = y^* \mid Y(u,.) = y)$ aufgrund von (1) von $y \in y^*$ ab, und diese bedingten Wahrscheinlichkeiten könnten modelliert werden. Andersherum kann von (4) ausgegangen und entsprechend $\Pr(Y(u,.) = y \mid Y^*(u,.) = y^*)$ modelliert werden. Eine Reihe anderer Möglichkeiten der Konstruktion nicht ignorierbarer Selektionsmodelle ist denkbar. Das wohl bekannteste Modell ist von Heckman (1976) vorgeschlagen und zumeist für den Fall vollständig fehlender Werte verwandt worden: $y^* = \{\{y\} \mid y \in \mathbb{Y}\} \cup \{\mathbb{Y}\}$. Sei $R(u,.)$ eine Variable, die das Antwortverhalten von u bei der Frage nach dem Haushaltseinkommen darstellt, also $R(u,.) := 1$, wenn $Y(u,.)$ genau angegeben wird, $R(u,.) := 2$ sonst. Wenn nun die Existenz einer Variablen $R^*(u,.)$ mit $R(u,.) = 1 \Leftrightarrow R^*(u,.) \geq 0$ und ein Zusammenhang zwischen $Y(u,.)$ und $R^*(u,.)$ postuliert wird, dann ergibt sich für den Erwartungswert der Stichprobe mit vollständigen Antworten $\mathbb{E}(Y(u,.) \mid R^*(u,.) \geq 0)$. Dies muss nicht mit dem unkonditionalen Erwartungswert $\mathbb{E}(Y(u,.))$ übereinstimmen, wenn $R^*(u,.)$ und $Y(u,.)$ stochastisch abhängig sind. Andererseits erfordert die CAR-Bedingung bei fehlenden Werten die Gleichheit von $\mathbb{E}(Y(u,.))$ und dem bedingten Erwartungswert von $Y(u,.)$ gegeben $R(u,.) = 1$. Das Modell ist also sicher nicht CAR, falls ein stochastischer Zusammenhang zwischen $R^*(u,.)$ und $Y(u,.)$ unterstellt wird.

16 Verschiedene Varianten der Einbeziehung dieser Information sind sowohl von Qin et al. (2002) als auch von Robins und seinen Mitarbeitern untersucht worden. Theoretisch können „optimale“ erweiterte Schätzfunktionen angegeben werden, die die (asymptotische) Varianz von $\hat{\mu}$ minimieren. Die optimale Wahl einer Schätzfunktion hängt von der Spezifikation eines Modells der bedingten Verteilung von $Y(u,.)$ gegeben $X(u,.)$ bzw. gegeben $X(u,.)$, $R(u,.) = 1$ ab (Rotnizky und Robins 1997; Scharfstein und Irizarry 2003; van der Laan und Robins 2003). Beide Verteilungen sind aber nicht von direktem Interesse und ihre nichtparametrische Schätzung, die zusätzliche probabilistische Annahmen vermeidet, ist bei den üblichen Stichprobengrößen wie im ALLBUS sehr instabil.

Jede gemeinsame Verteilung von $Y(u,.)$ und $R^*(u,.)$ erzeugt ein Selektionsmodell, das nur dann ignorierbar ist, wenn $Y(u,.) \perp R^*(u,.)$ gilt. Denn der Ausdruck

$$\Pr\left(Y^*(u,.) = \mathbb{Y} \mid Y(u,.) = y\right) = \Pr\left(R(u,.) = 2 \mid Y(u,.) = y\right)$$
$$= \int_{-\infty}^{0} f\left(v \mid Y(u,.) = y\right) dv$$

müsste aufgrund von (1) konstant in y sein, wenn es sich um ein CAR-Modell handelte. Es ist aber von vornherein klar, dass über ein solches nicht ignorierbares Selektionsmodell empirisch wenig zu sagen sein wird, da weder $R^*(u,.)$ noch $Y(u,.)$ tatsächlich beobachtet werden.

Sind $Y(u,.)$ und $R^*(u,.)$ gemeinsam normalverteilt mit $\mathbb{E}(Y(u,.)) := \mu$, $\mathbb{E}(R^*(u,.)) := \mu^*$, $\mathrm{Var}(R^*(u,.)) := 1$, $\mathrm{Var}(Y(u,.)) := \sigma^2$ und $\mathrm{Corr}(Y(u,.), R^*(u,.)) := \rho$, dann ist die bedingte Dichte von $Y(u,.)$ gegeben $R(u,.) = 1$

$$\phi\left(y \mid R^*(u,.) \geq 0\right)$$
$$= \frac{1}{\Pr\left(R(u,.) = 1\right)} \frac{1}{\sigma} \phi\left((y-\mu)/\sigma\right) \Phi\left(\frac{1}{\sqrt{1-\rho^2}} \left(\mu^* + \frac{\rho}{\sigma}(y-\mu)\right)\right)$$

wobei ϕ bzw. Φ hier die Dichte bzw. Verteilungsfunktion der standardisierten Normalverteilung bezeichnen (Copas und Li 1997: 59f.). Der letzte Term ist die bedingte Wahrscheinlichkeit für eine vollständige Angabe bei gegebenem Wert von $Y(u,.)$

$$\Pr\left(R = 1 \mid Y(u,.) = y\right) = \Phi\left(\frac{1}{\sqrt{1-\rho^2}} \left(\mu^* + \frac{\rho}{\sigma}(y-\mu)\right)\right)$$

Diese bedingte Wahrscheinlichkeit entspricht der bedingten Wahrscheinlichkeit in (1). Die CAR-Bedingung gilt genau dann, wenn der Koeffizient von y 0 ist, wenn also $\rho = 0$ ist.[17]

Da $R(u,.)$ für alle $u \in s$ bekannt ist, kann wegen $\Phi(\mu^*) = 1 - \Phi(0: \mu^*) = \Pr(R^*(u,.) \geq 0) = \Pr(R(u,.) = 1)$ ein Schätzer von μ^* durch $\Phi^{-1}(\mathrm{P}(R = 1))$ konstruiert werden, wobei $\mathrm{P}(R = 1)$ den Anteil vollständiger Beobachtungen in der Stichprobe angibt. Dagegen müssen μ, σ und ρ auf der Basis der vollständigen Beobachtungen geschätzt werden. Insbesondere der Parameter ρ, der die Abweichung von einem CAR-Modell darstellt, lässt sich nur aufgrund der Abweichung der Verteilung der vollständigen Beobachtungen von einer Normalverteilung identifizieren. Wenn im Fall des Haushaltseinkommens die genauen Angaben zur Berechnung der Parameter μ, σ, ρ verwandt werden, so erhält man die völlig unplausiblen Werte $\hat{\mu} = 1312$, $\hat{\sigma} = 3119$, $\hat{\rho} = 0{,}987$. Der geschätzte Erwartungswert liegt selbst außerhalb des konsistenten Mittel-

17 In der ökonometrischen Literatur wird Heckmans Modell oft als ein Beispiel für eine „selection on unobservables" angeführt (Nicoletti 2002: 3). Die Bezeichnung soll wohl auf die Abhängigkeit des Modells von der unbeobachteten Variablen R^* hinweisen. Die Bezeichnung ist irreführend, denn wie die letzte Gleichung zeigt, hängt die Selektion nicht von unbeobachtbaren Größen ab, sondern von den Werten von $Y(u,.)$, etwa dem Haushaltseinkommen. Letzteres ist zwar nicht für alle Befragten $u \in s$ bekannt, aber es ist sicher nicht „unbeobachtbar".

wertintervalls auf der Bruttostichprobe s_1 (vgl. Abschnitt III.). Nach diesem Ergebnis hätte ein erheblicher Teil der Population ein stark negatives Haushaltseinkommen. Zudem wäre die „Antwortwortbereitschaft" $R^*(u,.)$ fast perfekt mit dem Einkommen korreliert. Benutzt man dagegen die logarithmierten Einkommen, ergibt sich als Erwartungswert 3849 DM und $\hat{\rho} = -0{,}845$, also ein stark negativer Zusammenhang zwischen Einkommen und „Antwortbereitschaft". Etwas stabilere Ergebnisse wird man nur erhoffen können, wenn Kovariablen mit großem Effekt auf $R(u,.)$ angegeben werden können, die keinen Einfluss auf $Y(u,.)$ haben. Aber selbst dann wird die Identifikation des Modells im Wesentlichen durch Linearitätsannahmen im Modell für $R^*(u,.)$ ermöglicht. Denn da es zu gegebenen Daten immer ein (konditionales) CAR-Modell gibt, kann der Koeffizient von y in der bedingten Wahrscheinlichkeit $\Pr(R(u,.) = 1 \mid Y(u,.) = y, X(u,.) = x)$ immer als 0 angenommen werden, wenn nur der Einfluss der übrigen Kovariablen $X(u,.)$ allgemein modelliert wird. Die Identifikation des Koeffizienten von y und damit der Korrelation ρ erfordert eine parametrische Einschränkung der Wirkung der Kovariablen. Aber die Auswirkungen etwa von Linearitätsannahmen in $\Pr(R(u,.) = 1 \mid Y(u,.) = y, X(u,.) = x)$ sind ebenso wenig wie die Auswirkungen willkürlicher Verteilungsannahmen einfach zu überblicken.

Immerhin ergibt das Heckmansche Modell einen ersten Ansatzpunkt zur Modellierung nicht ignorierbarer Selektionsprozesse. Zudem können auch partielle Angaben relativ leicht einbezogen werden. In der Tat kann das Modell so erweitert werden, dass viele häufig auftretende Probleme mit unvollständigen Angaben in diesem Rahmen formuliert werden können (z.B. Crouchley und Ganjali 2002). Weiterhin wurden Verteilungsannahmen abgeschwächt (Das et al. 2003) und asymptotische Analysen verfeinert (Rotnitzky et al. 2000). Ausgehend von Heckmanschen Selektionsmodellen können die Auswirkungen „lokaler" Abweichungen von den Annahmen approximiert werden (Copas und Li 1997).

IX. Sensitivität

In der sozialwissenschaftlichen Praxis reicht es zumeist nicht, sich für ein Selektionsmodell zu entscheiden und sodann Ergebnisse nur unter dieser Modellannahme zu präsentieren. Wird mit einer CAR-Annahme begonnen, so muss dennoch Rechenschaft über die Konsequenzen der Annahme abgelegt werden. Das ist umso wichtiger, als die CAR-Annahme dazu verführen könnte, nicht über Selektionsprozesse nachzudenken. Zwar können „lokale" Abweichungen von der CAR-Annahme allgemein beschrieben werden. Damit ist es auch möglich, die Konsequenzen der CAR-Annahme für statistische Aussagen zu approximieren (Copas und Eguchi 2001). Aber probabilistische CAR-Modelle können kaum als realistische Modelle für Stichprobenausfälle angesehen werden. Daher wird eine lokale Approximation, so nützlich sie theoretisch ist, zumeist nicht befriedigen können. Die Approximation läuft Gefahr, die Konsequenzen von Annahmen über den Selektionsprozess nur in den engen Grenzen probabilistisch ähnlicher Modelle zu untersuchen. Rosenbaums Sensitivitätsanalyse (2002: Chap. 4) knüpft an Überlegungen über die Variable $R^*(u,.)$ an, wie sie auch in Heckmans Modell verwandt wird. Rosenbaums Methode geht von einer größeren Menge von Selektionsmo-

dellen aus und erlaubt damit eine umfassendere Abschätzung der Effekte von nicht ignorierbaren Selektionsprozessen. Die Methode verbleibt aber in der unterstellten Modellwelt und ist zudem bisher nur für spezielle Statistiken formuliert worden.

Robins und Mitarbeiter haben vorgeschlagen, von den gewichteten Schätzfunktionen

$$\sum_{u \in s} \frac{I\left[R(u,.)=1\right]}{\pi(u,X)}\left(Y(u,.)-\mu\right)$$

auszugehen und den Propensity Score auch als Funktion von $Y(u,.)$ zu modellieren (Rotnitzky und Robins 1997; Robins 1997; Rotnitzky et al. 1998; Scharfstein et al. 1999; Scharfstein und Irizarry 2003). Da

$$\mathbb{E}\left(1-\frac{I\left[R(u,.)=1\right]}{\pi(u,X,Y)}\right)$$
$$=1-\mathbb{E}\left(\frac{1}{\pi(u,X,Y)}\mathbb{E}\left(I\left[R(u,.)=1\right] \mid X(u,.),Y(u,.)\right)\right)=0$$

ist, kann die Schätzfunktion erweitert werden:

$$\sum_{u \in s} \frac{I\left[R(u,.)=1\right]}{\pi(u,X,Y)}\left(Y(u,.)-\mu\right)+\left(1-\frac{I\left[R(u,.)=1\right]}{\pi(u,X,Y)}\right)\phi(X;\mu)$$

Wird etwa logit $\pi(u, x, y) = x(u)\beta + \alpha y(u)$ gesetzt und

$$\phi(X;\mu,\alpha)=\frac{\mathbb{E}\left(Y(u,.)\exp\left(\alpha Y(u,.)\right) \mid R(u,.)=1, X(u,.)\right)}{\mathbb{E}\left(\exp\left(\alpha Y(u,.)\right) \mid R(u,.)=1, X(u,.)\right)}-\mu$$

gewählt, dann ergibt sich ein doppelt robuster Schätzer, der auch dann konsistent ist, wenn die Auswahlgleichung fehlspezifiziert ist. Dabei kann der Koeffizient α von $Y(u,.)$ in der Auswahlgleichung $\pi(u,X,Y)$ nur aufgrund der Beobachtungen mit vollständigen Angaben zu $Y(u,.)$ geschätzt werden, wenn also $R(u,.) = 1$ ist. Da außerdem zu jedem Selektionsmodell auch ein (konditionales) CAR-Modell gebildet werden kann, wenn nur $\pi(u,X)$ flexibel genug gewählt wird, ist α auch in sehr großen Datensätzen kaum stabil schätzbar. Robins und Mitarbeiter haben daher vorgeschlagen, eine Reihe von Werten des Koeffizienten α fest zu wählen und die Auswirkungen auf die Schätzung von μ zu notieren. Wie ihre Simulationen und Beispiele aber zeigen, sind auch diese Schätzungen sehr instabil. Zudem erscheint noch unklar, wie partielle Angaben in diesem Ansatz direkt berücksichtigt werden können (Robins 1997; Robins und Gill 1997).

Manskis Abschätzungen der Folgen von Selektivität (Manski 1993; Manski und Horowitz 2000; Manski 2003; Zaffalon 2002; Manski und Tamer 2002) erfolgen ähnlich wie die Abschätzungen in Abschnitt III und verzichten auf probabilistische Annahmen über den Selektionsprozess. Sie führen nur dann zu relativ engen Intervallen, wenn $Y(u,.)$ beschränkt ist. Im Fall des Haushaltseinkommens ergeben sich daher Intervalle für die möglichen Werte der Durchschnitte wie schon im Abschnitt III. Diese

Intervalle sind auch die theoretischen Grenzen der Methode von Robins et al., wenn ihr Koeffizient α alle Werte zwischen $-\infty$ und ∞ durchläuft (Scharfstein et al. 1999: 1108).

Raghhunathan hat in der Diskussion der Arbeit von Manski und Horowitz (2000: 86) die Weite der Intervalle beklagt: „I am afraid that I agree with Cochran (1977) that such an approach is so conservative as to be of little value in most practical settings for inferential purposes." Manski und Horowitz antworteten: „The width of the bounds reflects the information available from the data per se about the population parameters of interest. The width also indicates the relative importance of the data and untestable assumptions in determining the values of point estimates. ... Readers should be told that the point estimates are sensitive to untestable assumptions and that different assumptions could produce widely different results" (Manski und Horowitz 2000: 87f.).

X. Diskussion

Soziologen haben bisher die Möglichkeiten und vor allem die Grenzen von Selektionsmodellen selten zur Kenntnis genommen. Zwar gab es immer wieder Arbeiten in verbreiteten Zeitschriften, die das Thema aufgegriffen haben (Berk und Ray 1982; Stolzenberg und Relles 1997; Winship und Mare 1992). Aber in empirischen Arbeiten wird das Problem oft vollständig ignoriert, mit einem kurzen Verweis auf MAR-Modelle abgetan oder mit einem Heckman Modell erledigt. Bei Ausschöpfungsraten von 50 Prozent in Umfragen wird man es sich nicht auf Dauer leisten können, Selektionsmodelle, ihre Grundlagen und ihre Konsequenzen zu ignorieren.

Nun kann über die Entstehung fehlender Angaben immer nur spekuliert werden. Man ist auf Informationen einschlägiger Untersuchungen über Antwortverhalten angewiesen. Existieren aber gruppierte, zensierte oder fehlklassifizierte Berichte der Befragten, dann müssen diese Angaben ernst genommen und in die Berechnung von Statistiken einbezogen werden. Es werden Verfahren benötigt, die auch diese partiellen Angaben berücksichtigen. Selektionsmodelle erlauben es, über die Voraussetzungen und Konsequenzen solcher Verfahren nachzudenken. Zudem sind in diesem Fall Spekulationen über Selektionsprozesse empirische Grenzen gesetzt und Selektionsmodelle können auch ohne Rückgriff auf zusätzliche Informationen beurteilt werden. Dazu können Methoden der Sensitivitätsanalyse einen wesentlichen Beitrag leisten. In dieser Rolle werden Selektionsmodelle auch für die empirische Sozialforschung an Bedeutung gewinnen.

Literatur

ALLBUS, 1998: Codebuch des kumulierten ALLBUS 1980–96, ZA-Nummer 1795, Release 98.01. Köln: Zentralarchiv für Empirische Sozialforschung.

Berk, Richard A., und *Subash C. Ray*, 1982: Selection Bias in Sociological Data. Social Science Research 11: 352–398.

Bryson, Maurice C., 1976: The Literary Digest Poll: Making of a Statistical Myth. American Statistician 30: 184–185.

Bundesregierung, 2001: Lebenslagen in Deutschland. Daten und Fakten. Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung. Bundestagsdrucksachen Nr. 14/5990.

Cahalan, Don, 1989: The Digest Poll Rides Again. Public Opinion Quarterly 53: 129–133.

Cochran, William G., 1977: Sampling Techniques, 3. Aufl. New York: Wiley.

Copas, John B., und *H. G. Li*, 1997: Inference for Non-random Samples (with Discussion). Journal of the Royal Statistical Society, B, 59: 55–95.

Copas, John B., und *Shinto Eguchi*, 2001: Local Sensitivity for Selectivity Bias. Journal of the Royal Statistical Society, B, 63: 871–895.

Crouchley, Rob, und *Mojtaba Ganjali*, 2002: The Common Structure of Several Models for Non-ignorable Dropout. Statistical Modeling 2: 39–62.

Das, Mitali, *Whitney K. Newey* und *Francis Vella*, 2003: Nonparametric Estimation of Sample Selection Models. Review of Economic Studies 70: 33–58.

Davies, Laurie, 1995: Data Features. Statistica Neerlandica 49: 185–245.

Dawid, A. Philip, 2000: Causal Inference without Counterfactuals (with Discussion). Journal of the American Statistical Association 95: 407–448.

Diekmann, Andreas, und *David Wyder*, 2002: Vertrauen und Reputationseffekte bei Internet-Auktionen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 674–693.

Dobra, Adrian, und *Stephen E. Fienberg*, 2000: Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs. Proceedings of the National Academy of Sciences 97: 11885–11892.

Dubra, Juan, und *Federico Echenique*, 2004: Information is not about Measurability. Mathematical Social Sciences 47: 177–185.

Engelhardt, Henriette, 1999: Lineare Regression mit Selektion: Möglichkeiten und Grenzen der Heckman-Korrektur. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51: 706–723.

Ferson, Scott, *Lev Ginzburg*, *Vladik Kreinovich*, *Luc Longpré* und *Monica Aviles*, 2002: Computing Variance for Interval Data is NP-hard. ACM SIGACT News 33: 108–118.

Fishman, George S., und *David S. Rubin*, 1998: Best- and Worst-case Variances when Bounds are Available for the Distribution Function. Computational Statistics & Data Analysis 29: 35–53.

Frick, Joachim R., und *Markus M. Grabk*a, 2003: Missing Data in the German SOEP: Incidence, Imputation and its Impact on the Income Distribution. Discussion Papers 376. Berlin: DIW.

Gill, Richard D., 1997: Nonparametric Estimation under Censoring and Passive Registration. Satistica Neerlandica 51: 35–54.

Gill, Richard D., *Mark J. van der Laan* und *James M. Robins*, 1997: Coarsening at Random: Characterizations, Conjectures, Counter-examples. S. 255–294 in: *D.Y. Lin* und *Thomas R. Fleming* (Hg.), 1997: Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis. Berlin: Springer.

Grünwald, Peter D., und *Joseph Y. Halpern*, 2003: Updating Probabilities. Journal of Artificial Intelligence Research 19: 243–278.

Heitjan, Daniel F., 1994: Ignorability in General Incomplete-Data Models. Biometrika 81: 701–708.

Heitjan, Daniel F., und *Donald B. Rubin*, 1991: Ignorability and Coarse Data. The Annals of Statistics 19: 2244–2253.

Heckman, James J., 1976: The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models. Annals of Economic and Social Measurement 5: 475–492.

Heckman, James J., 1979: Sample Selection Bias as a Specification Error. Econometrica 47: 153–161.

Heckman, James J., 1990: Varieties of Selection Bias. American Economic Review 80: 313–318.

Holland, Paul W., 1986: Statistics and Causal Inference (with Discussion). Journal of the American Statistical Association 81: 945–970.

Jacobsen, Martin, und *Niels Keiding*, 1995: Coarsening at Random in General Sample Spaces and Random Censoring in Continuous Time. The Annals of Statistics 23: 774–786.

Jonker, Marianne A., 2003: Estimation of Life Expectancy in the Middle Ages. Journal of the Royal Statistical Society, A, 166: 105–117.

Kalton, Graham, 2002: Models in the Practice of Survey Sampling (revisited) (with Discussion). Journal of Official Statistics 18: 129–161.

Kalton, Graham, und *Ismael Flores-Cervantes*, 2003: Weighting Methods. Journal of Official Statistics 19: 81–97.

Koch, Achim, 1997: Teilnahmeverhalten beim ALLBUS 1994. Soziodemographische Determinanten von Erreichbarkeit, Befragungsfähigkeit und Kooperationsbereitschaft. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 98–122.

Koch, Achim, 2002: 20 Jahre Feldarbeit im ALLBUS: Ein Blick in die Blackbox. ZUMA Nachrichten 51: 9–37.

Lawley, D.N., 1943: A Note on Karl Pearson's Selection Formulae. Proceedings of the Royal Society of Edinburgh A, 62, 28–30.

Le Cam, Lucien, und *Grace Lo Yang*, 1990: Asymptotics in Statistics. Some Basic Concepts. Berlin: Springer.

Lehmann, Erich Leo, 1999: Elements of Large-Sample Theory. Berlin: Springer.

Little, Roderick J., und *Donald B. Rubin*, 2002: Statistical Analysis with Missing Data. 2. Aufl. New York: Wiley.

Little, Roderick J., und *Sonya Vartivarian*, 2003: On Weighting the Rates in Non-response Weights. Statistics in Medicine 22: 1589–1599.

Madow, William G., und *Ingram Olkin*, 1983: Incomplete Data in Sample Surveys. 3 Bände. New York: Academic Press.

Manski, Charles F., 1993: The Selection Problem in Econometrics and Statistics. S. 73–84 in *G.S. Maddala*, *C.R. Rao* und *H.D. Vinod* (Hg.), 1993: Handbook of Statistics, Vol. 11. Amsterdam: Elsevier.

Manski, Charles F., 2003: Partial Identifiability of Probability Distributions. Berlin: Springer.

Manski, Charles F., und *Joel L. Horowitz*, 2000: Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data (with Discussion). Journal of the American Statistical Association 95: 77–88.

Manski, Charles F., und *Elie Tamer*, 2002: Inference on Regressions with Interval Data on a Regressor or Outcome. Econometrica 70: 519–546.

Mayer, Karl Ulrich, und *Paul B. Baltes*, 1996: Die Berliner Altersstudie. Berlin: Akademie-Verlag.

Mohler, Peter, *Achim Koch* und *Siegfried Gabler*, 2003: Alles Zufall oder? Ein Diskussionsbeitrag zur Qualität von face to face Umfragen in Deutschland. ZUMA Nachrichten 53: 10–15.

Nicoletti, Cheti, 2002: Correcting for Sample Selection Bias: Alternative Estimators Compared. http://www.iser.essex.ac.uk/activities/seminars/Monday_Afternoons/archive/papers/poverty3.pdf

Nielsen, Søren Feodor, 2000: Relative Coarsening at Random. Statistica Neerlandica 54: 79–99.

Qin, Jing, *Denis Leung* und *Jun Shao*, 2002: Estimation with Survey Data under Nonignorable Nonresponse or Informative Sampling. Journal of the American Statistical Association 97: 193–200.

Riphahn, Regina T., und *Oliver Serfling*, 2002: Item Non-response on Income and Wealth Questions. Forschungsinstitut zur Zukunft der Arbeit. Discussion Paper No. 573.

Robins, James M., 1997: Non-response Models for the Analysis of Non-monotone Non-ignorable Missing Data. Statistics in Medicine 16: 21–37.

Robins, James M., und *Richard D. Gill*, 1997: Non-response Models for the Analysis of Non-monotone Ignorable Missing Data. Statistics in Medicine 16: 39–56.

Robins, James M., und *Ya'acov Ritov*, 1997: Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models. Statistics in Medicine 16: 285–319.

Rohwer, Götz, und *Ulrich Pötter*, 2001: Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Rosenbaum, Paul R., 2002: Observational Studies. 2. Aufl. Berlin: Springer.

Rotnitzky, Andrea, und *James M. Robins*, 1997: Analysis of Semi-parametric Regression Models with Non-ignorable Non-response. Statistics in Medicine 16: 81–102.

Rotnitzky, Andrea, *James M. Robins* und *Daniel O. Scharfstein*, 1998: Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. Journal of the American Statistical Association 93: 1321–1339.

Rotnitzky, Andrea, *David R. Cox*, *Matteo Bottai* und *James M. Robins*, 2000: Likelihood-based Inference with Singular Information Matrix. Bernoulli 6: 243–284.

Schafer, Joseph L., 1997: Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Scharfstein, Daniel O., *Andrea Rotnitzky* und *James M. Robins*, 1999: Adjusting for Nonignorable Drop-out Using Semiparametric Nonresponse Models (with Discussion). Journal of the American Statistical Association 94: 1096–1146.

Scharfstein, Daniel O., und *Rafael A. Irizarry*, 2003: Generalized Additive Selection Models for the Analysis of Studies with Potentially Nonignorable Missing Outcome Data. Biometrics 59: 601–613.

Schneekloth, Ulrich, und *Ingo Leven*, 2003: Woran bemisst sich eine „gute" allgemeine Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. ZUMA Nachrichten 53: 16–57.

Schräpler, Jörg-Peter, 2004: Respondent Behavior in Panel Studies – A Case Study for Income-Nonresponse by Means of the SOEP. Erscheint in: Sociological Methods & Research.

Squire, Peverill, 1988: Why the 1936 Literary Digest Poll Failed. Public Opinion Quarterly 52: 125–133.

Statistisches Bundesamt, 2000: Datenreport 1999. Zahlen und Fakten über die Bundesrepublik Deutschland. Bundeszentrale für politische Bildung, Schriftenreihe Band 365, Bonn.

Statistisches Bundesamt, 2001: Statistisches Jahrbuch. Wiesbaden.

Stolzenberg, Ross M., und *Daniel A. Relles*, 1997: Tools for Intuition about Sample Selection Bias and its Correction. American Sociological Review 62: 494–507.

van der Laan, Mark J., und *James M. Robins*, 2003: Unified Methods for Censored Longitudinal Data and Causality. Berlin: Springer.

Vella, Francis, 1998: Estimating Models with Sample Selection Bias: A Survey. The Journal of Human Resources 33: 127–169.

Vytlacil, Edward, 2002: Independence, Monotonicity, and Latent Index Models: An Equivalence Result. Econometrica 70: 331–341.

Wainer, Howard, 1986: Drawing Inferences from Self-selected Samples. Mahwah: Lawrence Erlbaum.

Wainer, Howard, 1989, 1992: Eeelworms, Bullet Holes, and Geraldine Ferraro: Some Problems with Statistical Adjustment and some Solutions (with Discussion). Journal of Educational Statistics 14: 121–140. Wieder abgedruckt S. 129–207 in: *Juliett Popper Shaffer* (Hg.), 1992: The Role of Models in Nonexperimental Social Science: Two Debates. Washington: American Educational Research Association and American Statistical Association.

Winship, Christopher, und *Robert D. Mare*, 1992: Models for Sample Selection Bias. Annual Review of Sociology 18: 327–350.

Zaffalon, Marco, 2002: Exact Credal Treatment of Missing Data. Journal of Statistical Planning and Inference 105: 105–122.

KOGNITIONSPSYCHOLOGIE UND UMFRAGEFORSCHUNG: ALTERSABHÄNGIGE KONTEXTEFFEKTE*

Norbert Schwarz und Bärbel Knäuper

Zusammenfassung: Die Kooperation von Kognitionspsychologen und Umfrageforschern hat zu einem guten Verständnis der kognitiven und kommunikativen Prozesse geführt, die Kontexteffekten in Umfragen zugrunde liegen. Auf dieser Basis können Vorhersagen über *differentielle* Kontexteffekte abgeleitet werden, wie wir am Beispiel *altersabhängiger* Kontexteffekte zeigen. Generell nehmen Fragereihenfolgeeffekte mit dem Alter ab, aber Antwortreihenfolgeeffekte mit dem Alter zu. Beide Effekte sind auf altersabhängige Minderungen der Gedächtnisleistung zurückzuführen. Ebenso nehmen Kontexteffekte bei Verhaltensberichten mit dem Alter zu, es sei denn das Verhalten ist von hoher persönlicher Relevanz. Die Größe der beobachteten Einflüsse kann ausreichen, die ordinale Platzierung von Altersgruppen umzukehren. Jeder beobachtete Altersunterschied in Einstellungs- und Verhaltensberichten kann a) reale Unterschiede, b) differentielle methodische Einflüsse oder c) eine unbekannte Mischung beider widerspiegeln.

I. Einleitung

Umfragedaten sind nur so sinnvoll wie die Antworten, die Befragte geben. Seit den Anfangstagen der Umfrageforschung ist bekannt, dass diese Antworten vielfach kontextabhängig sind (Cantril 1944; Payne 1951). Die zugrunde liegenden kognitiven und kommunikativen Prozesse sind Gegenstand eines aktiven Forschungsprogramms an der Schnittstelle von kognitiver (Sozial-)Psychologie und Umfragemethodologie (zum Überblick siehe Schwarz 1999; Sirken et al. 1999; Sudman et al. 1996; Tourangeau et al. 2000). In den zwei Jahrzehnten seit den beiden Konferenzen, die diese Entwicklung im Herbst 1983 (siehe Beiträge in Jabine et al. 1984) und Sommer 1984 (siehe Beiträge in Hippler et al. 1987) initiierten, erlebte die Kooperation von Umfragemethodologen und Psychologen eine schnelle Institutionalisierung. Zahlreiche Umfragezentren an Universitäten und Regierungseinrichtungen, insbesondere in den USA, unterhalten „cognitive laboratories" und Kognitionspsychologie fand Eingang in die Curricula amerikanischer Graduiertenstudiengänge der Umfrageforschung. Auf der theoretischen Seite sind die Bedingungen von Kontexteffekten in der Einstellungsmessung ebenso bekannt (zum Überblick siehe Schwarz und Bohner 2001; Sudman et al. 1996) wie die Variablen, die die Validität von Verhaltensberichten beeinflussen (zum Überblick siehe

* Die berichteten Untersuchungen wurden gefördert durch das National Institute on Aging (AG14111-01 an N. Schwarz, D. C. Park, und B. Knäuper). Anschrift: Norbert Schwarz, Institute for Social Research, University of Michigan, Ann Arbor, MI 418106-1248, USA, nschwarz @umich.edu; Bärbel Knäuper, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, Canada H3A 1B1, barbel.knauper@mcgill.ca.

Schwarz und Oyserman 2001). Auf der praktischen Seite haben insbesondere kognitive Pretestmethoden (zum Überblick siehe die Beiträge in Schwarz und Sudman 1996), die eine kostengünstige Explorierung von Fragebogenproblemen erlauben, einen festen Platz in der Vorbereitung größerer Studien gefunden.

Das zunehmende theoretische Verständnis der kognitiven und kommunikativen Prozesse bei Befragungen erlaubte es zugleich, theoretische Fortschritte auf anderen Gebieten psychologischer Forschung für die Umfragemethodologie fruchtbar zu machen. So haben z.B. Theorien und Befunde der kognitiven Alternsforschung (zum Überblick siehe die Beiträge in Park und Schwarz 2000) wesentliche Implikationen für das Antwortverhalten älterer Befragter. Ebenso erlauben Befunde der kulturvergleichenden Psychologie (zum Überblick siehe Fiske et al. 1998) eine theoretische und empirische Analyse von kulturellen Unterschieden im Befragungsverhalten, die allerdings beim derzeitigen Forschungsstand auf Unterschiede zwischen kollektivistischen Kulturen Asiens und individualistischen Kulturen westlicher Industrieländer begrenzt ist (für erste Befunde siehe Schwarz 2003). In beiden Fällen wurde deutlich, dass psychologische Unterschiede im Befragungsprozess in alterspezifischen oder kulturspezifischen Kontexteffekten resultieren können, die direkte Vergleiche zwischen Altersgruppen oder Ländern erschweren. Der vorliegende Beitrag zeigt dies am Beispiel altersabhängiger Kontexteffekte.

Aus der Sicht der empirischen Sozialforschung beinhalten altersabhängige Kontexteffekte, dass beobachtete Unterschiede in den Einstellungs- und Verhaltensberichten verschiedener Altersgruppen entweder a) einen realen Unterschied in Meinung und Verhalten, b) einen Unterschied im Antwortprozess oder c) eine unbekannte Mischung beider Einflüsse widerspiegeln können. Wie die ausgewählten Beispiele zeigen werden, ist die Größe altersabhängiger Kontexteffekte oft hinreichend, die ordinale Platzierung von Kohorten umzukehren: Je nach Reihenfolge der Fragen oder Antwortvorgaben können z.B. entweder jüngere oder ältere Befragte als die „konservativere" Altersgruppe erscheinen.

Dieser Beitrag gibt einen Überblick über relevante Befunde. Einer kurzen Einführung in kognitive Veränderungen im Alter folgt eine Diskussion des Einflusses der Fragenreihenfolge und Antwortreihenfolge. In der Regel nimmt der Einfluss der Fragenreihenfolge mit dem Alter ab, aber der Einfluss der Antwortreihenfolge mit dem Alter zu. Beide Effekte sind auf eine abnehmende Kapazität des Arbeitsgedächtnisses zurückzuführen. Aus dem gleichen Grund müssen ältere Befragte bei Verhaltensberichten stärker auf Schätzstrategien zurückgreifen, was zu stärkeren Kontexteffekten bei Verhaltensberichten führt, es sei denn, das jeweilige Verhalten ist für ältere Befragte von besonderer Relevanz.

II. Kognitive Veränderungen im Alter

Umfragen konfrontieren Befragte mit komplexen kognitiven Aufgaben (Strack 1994; Strack und Martin 1987; Tourangeau 1984). Sie müssen zunächst die intendierte Bedeutung der Frage verstehen, was eine Berücksichtigung des jeweiligen Konversationskontextes und häufig einen Rückgriff auf frühere Fragen und Antworten erfordert. Da-

nach müssen sie relevante Information aus dem Gedächtnis abrufen und ein Urteil bilden, das für die Beantwortung der Frage relevant ist. Dieses Urteil müssen Befragte sodann in das vom Forscher vorgegebene Antwortformat einpassen, das zu diesem Zweck noch erinnert werden muss. Wenn die Frage sensitive Themen betrifft, mögen Befragte außerdem ihre Antwort aus Gründen sozialer Erwünschtheit „editieren", was weitere kognitive Operationen erfordert. Insbesondere unter dem Zeitdruck von Telefoninterviews, wo die typische Frage-Antwortsequenz meist unter einer Minute liegt und keine externen Gedächtnishilfen verfügbar sind, stellen diese Aufgaben hohe Anforderungen an die Kapazität des Arbeitsgedächtnisses. Diese Kapazität nimmt mit zunehmendem Alter ab (zum Überblick siehe Park 2000). Da zudem kognitive Operationen mit zunehmendem Alter mehr Zeit beanspruchen (Salthouse Babcock und Shaw 1991), ist es für ältere Befragte oft schwierig, relevante Operationen durchzuführen, bevor die benötigte Information nicht mehr im Arbeitsgedächtnis verfügbar ist. Die Abnahme der Arbeitsgedächtniskapazität beginnt im frühen Erwachsenenalter und ist linear, wobei einige Forschungsergebnisse einen schnelleren Abbau ab etwa 65 Jahren nahe legen (Schaie 1996).

Altersabhängige Veränderungen des Arbeitsgedächtnisses legen spezifische Hypothesen zum Auftreten altersabhängiger Kontexteffekte nahe. Wie weiter unten für drei verschiedene Arten von Kontexteffekten aufgezeigt wird, wurde für diese Hypothesen empirische Unterstützung gefunden (für eine umfassendere Diskussion von Altersunterschieden in Befragungen siehe die Beiträge in Schwarz et al. 1999).

III. Der Einfluss der Fragenreihenfolge

Fragereihenfolgeeffekte spiegeln wider, dass Personen zur Urteilsbildung selten auf alle potentiell relevanten Informationen zurückgreifen. Vielmehr brechen sie die Informationssuche ab, sobald „genug" Information erinnert wurde, um mit hinreichender subjektiver Sicherheit ein Urteil zu bilden (Bodenhausen und Wyer 1987). Das Urteil beruht daher auf der Information, die zum Befragungszeitpunkt am ehesten in den Sinn kommt. Häufig ist dies Information, die gerade zur Beantwortung einer früheren Frage aus dem Gedächtnis abgerufen wurde. Dies resultiert in ausgeprägten Einflüssen der Fragereihenfolge. So korreliert z.B. die Häufigkeit von Rendezvous nur $r = -.12$ mit der Lebenszufriedenheit von Studenten, wenn die Frage zur Lebenszufriedenheit der Frage zu Rendezvous vorausgeht, aber $r = +.66$ wenn beide Fragen in umgekehrter Reihenfolge gestellt werden (Strack et al. 1988; siehe auch Schwarz et al. 1991). Da die allgemeine Lebenszufriedenheit auf der Basis einer Vielzahl unterschiedlicher Aspekte beurteilt werden kann, haben jene Aspekte den größten Einfluss, die zum Urteilszeitpunkt am ehesten in den Sinn kommen.

Die Richtung des Einflusses ist dabei eine Funktion der Informationsnutzung. Information, die in die Repräsentation des Urteilsobjekts (im obigen Beispiel „Mein Leben") eingeht, resultiert in Assimilationseffekten. Information, die in die Repräsentation des Standards eingeht, an dem das Urteilsobjekt bewertet wird, resultiert in Kontrasteffekten (Schwarz und Bless 1992). Sowohl die Verfügbarkeit als auch die Nutzung

von Information sind z.T. eine Funktion des Forschungsinstruments (siehe Sudman et al. 1996, Kapitel 5 für eine Diskussion der relevanten Fragebogenvariablen).

Das in der Umfrageforschung wohl bekannteste Beispiel für einen Kontrasteffekt ist das nachfolgend beschriebene Experiment von Schuman und Presser (1981). In einer Telefonumfrage mit einer Zufallsstichprobe (N = 777) von Einwohnern der US-Metropole Detroit stellten die Autoren im August 1979 zwei Fragen zur Einstellung gegenüber legalen Abtreibungen. Die beiden Fragen beinhalten unterschiedlich starke Legitimierungen des Abtreibungswunsches.

Frage A: „Do you think it should be possible for a pregnant woman to obtain a legal abortion *if there is a strong chance of serious defect in the baby?*" (hohe Legitimität)

Frage B: „Do you think it should be possible for a pregnant woman to obtain a legal abortion *if she is married and does not want any more children?*" (geringe Legitimität)

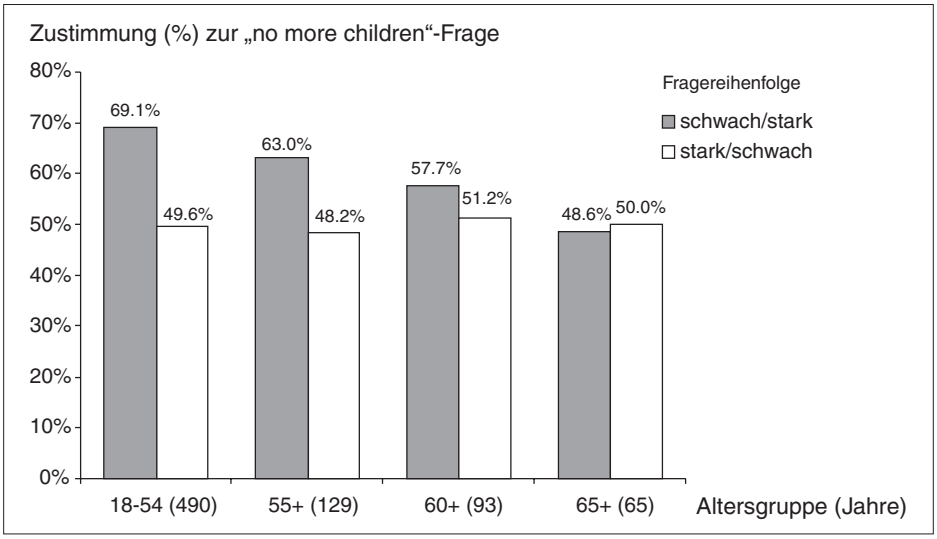
Wie zu erwarten, findet Frage A mehr Zustimmung als Frage B. Wichtiger ist jedoch, dass die Zustimmung zu Frage B von der Reihenfolge abhängig ist: In zahlreichen Experimenten in verschiedenen Ländern nimmt die Zustimmung zu Frage B ab, wenn Frage A zuerst gestellt wird (siehe Schuman 1992). Relativ zum drohenden „birth defect", den Frage A in den Sinn ruft, erscheint der simple Wunsch, keine weiteren Kinder haben zu wollen als weniger legitim.

Eine Sekundäranalyse der Daten von Schuman und Presser (1981) zeigt jedoch, dass dieser Kontrasteffekt nur bei jüngeren Personen auftritt (Knäuper 1999a; Knäuper et al. 2002). *Abbildung 1* veranschaulicht diesen Befund. Für Personen unter 55 Jahren ist die Zustimmung zu Frage B um 19,5 Prozent geringer wenn sie Frage A folgt. Dieser Reihenfolgeeffekt nimmt mit zunehmendem Alter ab und tritt für Befragte über 65 Jahre nicht mehr auf.

Wie *Abbildung 1* zeigt, legen unterschiedliche Fragenreihenfolgen unterschiedliche Schlüsse über Altersunterschiede in der Einstellung zu Abtreibung nahe. Wenn die „no more children"-Frage zuerst gestellt wird, würden wir folgern, dass Ältere legale Abtreibung weniger befürworten als Jüngere. Geht dieser Frage hingegen die „birth defect"-Frage voraus, würden wir folgern, dass Einstellungen gegenüber Abtreibung vollkommen altersunabhängig sind. Wie dieses Beispiel zeigt, können altersabhängige Kontexteffekte zu unterschiedlichen substantiellen Schlussfolgerungen führen – und ob Ältere oder Jüngere die „konservativere" Einstellung berichten, mag lediglich eine Funktion der Fragereihenfolge sein.

Nachfolgende Laborexperimente zeigten, dass die beobachteten Altersunterschiede auf Unterschiede in der Gedächtnisleistung zurückzuführen sind (Knäuper et al. 2003). An diesen Experimenten nahmen N = 82 jüngere Personen (19 bis 42 Jahre, M = 29.45) und N = 165 gesunde ältere Personen (65 bis 100 Jahre, M = 75.34) teil. Die jüngeren Teilnehmer wurden anhand von Zeitungsanzeigen rekrutiert, während die älteren Personen über den Versuchspersonen-Pool des Center for Applied Research in Cognitive Aging an der University of Michigan und Zeitungsanzeigen rekrutiert wurden. Für die statistischen Analysen gruppierten wir die 165 Personen auf der Basis ihrer Performanz in einem Gedächtnistest, der die Fähigkeit erfasst, Information für kurze Zeit im Gedächtnis zu behalten und gleichzeitig relevante Operationen mit der Information auszuführen („reading span"-Test, Salthouse und Babcock 1991). Wie erwar-

Abbildung 1: Sekundäranalyse des Experiments von Schumann und Presser (1981): Zustimmung (in Prozent) zur „no more children"-Frage als Funktion von Fragereihenfolge und Alter (Stichprobengröße in Klammern)



Quelle: Nach Tabelle 3 aus „Age differences in question and response order effects" von Bärbel Knäuper, 1999a (S. 355).

tet zeigten Ältere mit guter Gedächtnisleistung Fragereihenfolgeeffekte, während dies bei Älteren mit schlechter Gedächtnisleistung nicht der Fall war. Bei einer Replikation des Experiments von Schumann und Presser ergab sich für jüngere Personen ein Reihenfolgeeffekt von 25,4 Prozent ($t(59.94) = 2.98$, $p < .004$). Ältere Personen mit guter Gedächtnisleistung tendierten ebenso dazu, von der Reihenfolge beeinflusst zu sein und zeigten einen Reihenfolgeeffekt von 18,6 Prozent ($t(80.57) = 1.75$, $p < .08$). Für Ältere mit geringer Gedächtnisleistung wurde hingegen nur ein insignifikanter Unterschied von 5,6 Prozent gefunden ($t(78) = 0.51$, *ns*). Experimente mit anderen Fragen replizierten dieses Muster.

Befunde dieser Art zeigen, dass a) Fragereihenfolgeeffekte mit dem Alter abnehmen und b) diese Veränderung auf eine verminderte Leistung des Arbeitsgedächtnisses zurückzuführen ist. Diese verminderte Erinnerung an frühere Frageninhalte und die zu ihrer Beantwortung abgerufene Information entspricht funktional dem, was Umfrageforscher oft durch die Einführung von Füllfragen zu erreichen hoffen. Wir vermuten, dass die beobachtete altersabhängige Abnahme von Fragereihenfolgeeffekten auf Fragen beschränkt ist, die emotional wenig involvierend sind, was für die meisten Fragen in standardisierten Umfragen zutrifft. Emotional involvierende Gedankeninhalte könnten hingegen ältere Befragte länger beeinflussen als jüngere Befragte, da Ältere es generell schwieriger finden, emotionale Gedanken zu kontrollieren (vgl. Park 2000). Relevante Daten zu dieser Hypothese sind derzeit nicht verfügbar.

Der beobachtete Zusammenhang zwischen Gedächtnisleistung und der Abnahme von Fragereihenfolgeeffekten spricht zugleich gegen die Annahme, dass Ältere weniger

von vorausgegangenen Fragen beeinflusst werden, da sie „stärkere“ oder besser „kristallisierte“ Einstellungen haben (vgl. z.B. Sears 1986). In der Tat zeigt sich, dass andere Kontexteffekte mit dem Alter zunehmen, konsistent mit der vermuteten Gedächtnisabhängigkeit von altersspezifischen Kontexteffekten. Der nächste Abschnitt präsentiert relevante Befunde.

IV. Der Einfluss der Antwortreihenfolge

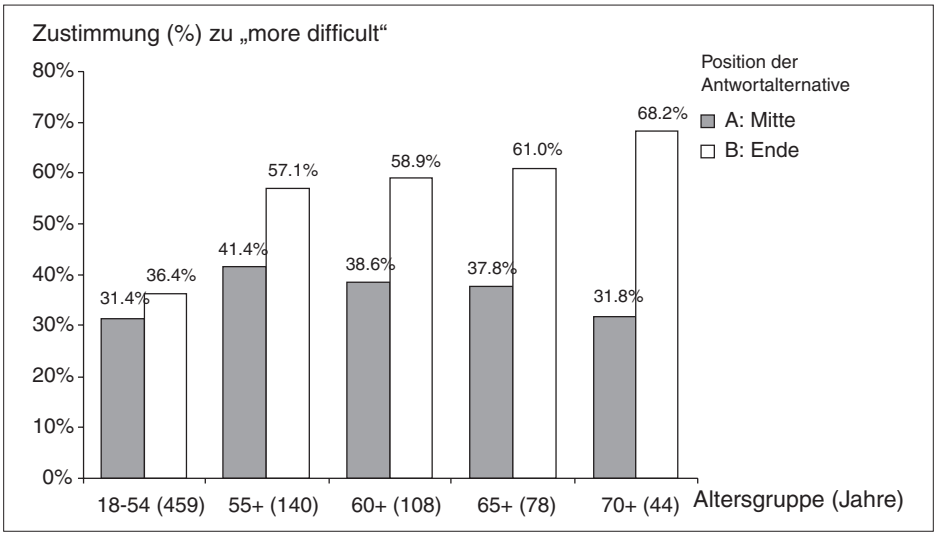
Bei der Beantwortung von Fragen mit mehreren Antwortalternativen sollen Befragte idealerweise a) die Frage und alle Antwortalternativen im Gedächtnis behalten, b) die Implikationen jeder Antwortalternative evaluieren und c) die zutreffendste Alternative auswählen. Die Realität weicht jedoch häufig von dieser Hoffnung ab, wie das Auftreten von Antwortreihenfolgeeffekten zeigt. Solche Effekte treten auf, wenn eine Frage mehrere Antwortalternativen anbietet, für die ein Befragter potentiell unterstützende Argumente generieren kann. Die meisten Befragten können z.B. leicht einige Gründe generieren, warum Scheidung erleichtert werden sollte („Eine schlechte Ehe ist eine Belastung für die Partner“; „Streitende Eltern sind schlecht für die Kinder“), aber sie könnten ebenso leicht einige Gründe gegen Scheidungserleichterung finden („Man soll nicht so schnell aufgeben“; „Eine intakte Familie ist gut für die Kinder“). Unter solchen Bedingungen hängt die Antwort auf die Frage „Sollte eine Scheidung leichter oder schwerer zu erhalten sein?“ davon ab, welche Antwortalternative (leichter, schwerer) Befragte zuerst erwägen. Beginnen sie mit „leichter“, fallen ihnen ebenso Argumente für Zustimmung ein als wenn sie mit „schwerer“ beginnen – und sobald solche Argumente verfügbar sind, wird die jeweils andere Alternative mit geringerer Wahrscheinlichkeit elaboriert (für ein detailliertes Prozessmodell siehe Schwarz et al. 1992; Sudman et al. 1996, Kapitel 6). Über welche Antwortalternative Befragte zuerst nachdenken ist eine Funktion des Erhebungsmodus (Krosnick und Alwin 1987). Bei schriftlicher Vorlage werden die Alternativen in der Reihenfolge der Präsentation überdacht, was in „primacy“-Effekten resultiert (d.h. die gleiche Antwortalternative wird eher gewählt, wenn sie am Anfang statt am Ende der Liste steht). Werden die Antwortalternativen jedoch vorgelesen, beginnen Befragte mit der letzten Alternative, die noch „im Ohr“ ist, was in „recency“-Effekten resultiert (d.h. die gleiche Antwortalternative wird eher gewählt wenn sie zuletzt statt zuerst vorgelesen wird).

Die Fähigkeit, nach dem Nachdenken über die zuerst elaborierte Alternative noch über die anderen Alternativen nachzudenken und die jeweiligen Argumente im Kopf zu behalten und gegeneinander abzuwägen, nimmt mit zunehmendem Abbau des Arbeitsgedächtnisses ab. Dies ist insbesondere der Fall, wenn die Alternativen, wie in Telefoninterviews, ohne visuelle Unterstützung präsentiert werden. Es ist daher zu erwarten, dass ältere Befragte in Telefoninterviews stärkere „recency“-Effekte zeigen als jüngere Befragte. Dieser Altersunterschied sollte jedoch geringer ausfallen, wenn die Alternativen schriftlich präsentiert werden, da dies die Gedächtnisbelastung reduziert. Empirisch ist dies der Fall.

Eine Meta-Analyse von Telefonumfragen (Knäuper 1999b) ergab, dass „recency“-Effekte mit dem Alter zunehmen, wie ein Beispiel zeigen soll. Schuman und Presser

(1981) stellten die folgende Frage im September 1979 in einer Telefonumfrage in einer Zufallsstichprobe von Einwohnern der US-Metropole Detroit (N = 466): *„Should divorce in this country be easier to obtain, more difficult to obtain, or stay as it is now?"* Abhängig von der jeweiligen experimentellen Bedingung wurde die Antwortalternative „more difficult to obtain" entweder als zweite oder als letzte Antwortalternative vorgelesen. Insgesamt stimmten die Befragten der Antwortalternative „more difficult" eher zu, wenn sie als letztes dargeboten wurde, ein so genannter Recency-Effekt. Sekundäranalysen zeigen jedoch ausgeprägte Altersunterschiede. Wie *Abbildung 2* zeigt, steigt der Recency-Effekt mit zunehmendem Alter an. Für jüngere Personen (bis 54 Jahre) zeigt sich ein nicht signifikanter Unterschied von 5 Prozent (χ^2 = 2.63, df = 2, *ns*). Dieser Unterschied nimmt mit dem Alter zu und erreicht 36,3 Prozent für Personen über 70 Jahre (χ^2 = 10.98, df = 2, p < .007). Hervorzuheben ist wiederum, dass man zu unterschiedlichen Schlussfolgerungen über Alters- oder Generationsunterschiede kommt, je nachdem in welcher Reihenfolge die Antwortalternativen dargeboten werden: Während die Einstellungen gegenüber Scheidung mit zunehmendem Alter in der einen Reihenfolgebedingung deutlich konservativer zu werden scheinen, zeigt sich kein zuverlässiger Altersunterschied unter der anderen Reihenfolgebedingung.

Abbildung 2: Scheidungsexperiment: Zustimmung (in Prozent) zu „more difficult" als Funktion der Position der Antwortalternative und Alter (Stichprobengröße in Klammern)

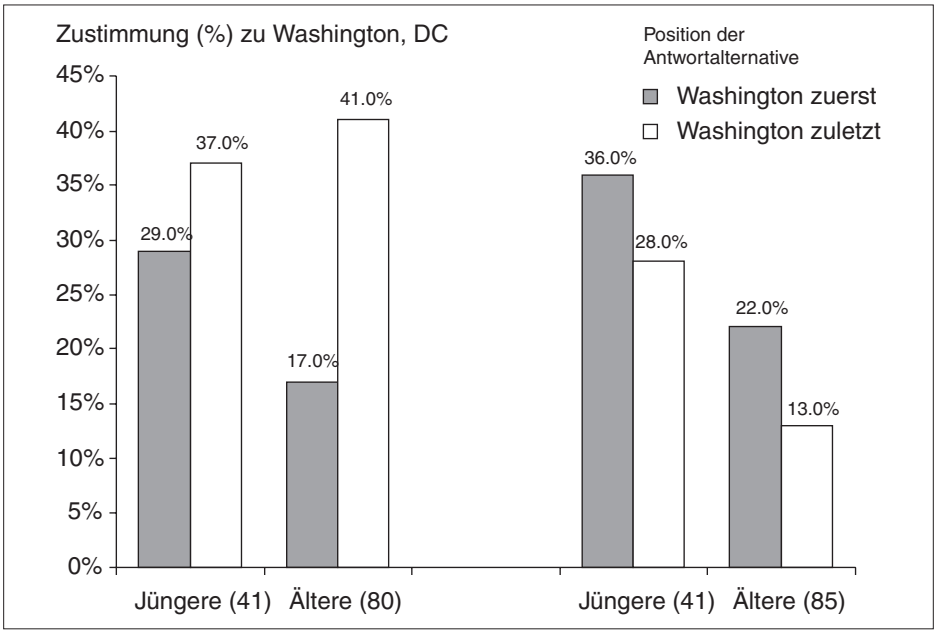


Quelle: Nach Tabelle 6 aus „The Impact of Age and Education on Response Order Effects in Attitude Measurement", von B. Knäuper (1999b), *Public Opinion Quarterly.*

Wie theoretisch zu erwarten, treten altersabhängige Unterschiede in Antwortreihenfolgeeffekten primär bei auditiver Präsentation der Antwortalternativen auf. In einem weiteren Laborexperiment mit der oben bereits beschriebenen Stichprobe fragten wir die jüngeren (20–40 Jahre, N = 82) und älteren (65–100 Jahre, N = 165) Befragten zum Beispiel „Which of the following four cities do you find most attractive?" Die

Antwortalternative „Washington, D.C." wurde entweder zuerst oder zuletzt präsentiert. Unter auditiver Präsentation erhielten wir „recency"-Effekte, d.h. Washington, D.C., wurde eher gewählt, wenn es zuletzt statt zuerst vorgelesen wurde. Die Größe dieses Effekts betrug 8 Prozent für jüngere Befragte und 24 Prozent für ältere Befragte (siehe *Abbildung 3*), was die Befunde der Meta-Analyse (Knäuper 1999b) repliziert. Wurden die Fragen hingegen schriftlich präsentiert, traten die erwarteten „primacy"-Effekte auf: Washington, D.C., wurde nun eher gewählt, wenn es am Anfang statt am Ende der Liste präsentiert wurde. Da die schriftliche Präsentation die Gedächtnisbelastung reduziert, erwarteten wir unter dieser Bedingung keinen Altersunterschied. Empirisch war dies mit einem „primacy"-Effekt von 8 Prozent für jüngere und 9 Prozent für ältere Befragte der Fall.

Abbildung 3: Antwortreihenfolgeeffekte als Funktion von Präsentationsmodus (auditiv versus schriftlich) und Alter (Stichprobengröße in Klammern)



Hervorzuheben ist, dass diese Daten wiederum differentielle Altersunterschiede in Stadtpräferenzen nahe legen, die zudem vom Modus der Datenerhebung abhängen (siehe *Abbildung 3*). Welche Folgerungen wir über Stadtpräferenzen oder über Altersunterschiede in Stadtpräferenzen ziehen würden, ist somit weitgehend eine Funktion methodischer Variablen. Hervorzuheben ist außerdem der hohe Einfluss des Erhebungsmodus auf die berichteten Präferenzen älterer Befragter: Ihre Wahl von Washington, D.C., reicht von 13 Prozent (schriftlich/zuletzt) bis 41 Prozent (auditiv/zuletzt), selbst unter konstanter Reihenfolge der Antwortalternativen. Befunde aus Umfragen mit versus ohne visuelle Hilfen sind daher für ältere Befragte nur bedingt vergleichbar,

selbst wenn die Fragen und die Reihenfolge der Antwortvorgaben identisch sind (für einen Überblick zu „mode effects“ siehe Schwarz et al. 1991).

V. Verhaltensberichte

In vielen Umfragen werden Befragte gebeten, über die Häufigkeit verschiedener Verhaltensweisen, von Einkäufen bis zu Gesundheitsbeschwerden, zu berichten. In der Regel hoffen die Forscher, dass Befragte möglichst alle relevanten Episoden erinnern, die Episoden korrekt datieren und sodann zählen, wie viele Episoden in die relevante Referenzperiode fallen (Bradburn et al. 1979). Zahlreiche Untersuchungen zeigen jedoch, dass diese „Erinnern und Zählen“-Strategie von Befragten selten verfolgt wird. Häufig wiederkehrendes Alltagsverhalten ist im Gedächtnis nicht in getrennten Episoden repräsentiert. Vielmehr verschwimmen einzelne Episoden in einer generischen Repräsentation des Verhaltenstypus, die eine getrennte Identifikation einzelner Episoden nur begrenzt erlaubt (Bradburn et al. 1987; Linton 1982; Schwarz 1990; Strube 1987). Die Gedächtnisrepräsentation von Alltagsverhalten entspricht daher eher der Form allgemeinen Wissens, ohne spezifische Indikatoren von Ort und Zeit, als der Form episodischer Gedächtnisinhalte. Während experimentelle Untersuchungen gezeigt haben, dass die Erinnerung an spezifische Episoden durch geeignete Suchreize verbessert werden kann (zum Überblick siehe Belli 1998; Schwarz und Oyserman 2001, und die Beiträge in Schwarz und Sudman 1994), sind der praktischen Nutzung dieser Möglichkeit bei heterogenen Stichproben enge Grenzen gesetzt. Das Modell des „Erinnerns und Zählens“ ist daher nur für selten gezeigtes Verhalten von hoher persönlicher Wichtigkeit zutreffend. Für häufig auftretendes Alltagsverhalten sind Befragte hingegen auf Schätzstrategien angewiesen.

Eine dieser Schätzstrategien besteht in der Nutzung von Information, die der Fragebogen selbst bereitstellt. Dies ist insbesondere der Fall, wenn Befragten Häufigkeitsskalen vorgelegt werden, die spezifische Werte anbieten. Schwarz et al. (1985) baten Befragte zum Beispiel, ihren täglichen Fernsehkonsum auf einer der beiden folgenden Skalen zu berichten:

Niedrige Häufigkeitsvorgaben	*Hohe Häufigkeitsvorgaben*
bis zu ½ Stunde	bis zu 2½ Stunden
½ bis 1 Stunde	2½ bis 3 Stunden
1 bis 1½ Stunden	3 bis 3½ Stunden
1½ bis 2 Stunden	3½ bis 4 Stunden
2 bis 2½ Stunden	4 bis 4½ Stunden
mehr als 2½ Stunden	mehr als 4½ Stunden

Sie fanden, dass 37,5 Prozent der Befragten einen Fernsehkonsum von mehr als 2½ Stunden pro Tag berichteten, wenn ihnen die hohe Skala vorgelegt wurde, während dies nur für 16,2 Prozent der Befragten der Fall war, wenn ihnen die niedrige Skala vorgelegt wurde. Dieser Einfluss von Häufigkeitsskalen wurde für viele Lebensbereiche belegt, von der Häufigkeit körperlicher Beschwerden (Schwarz und Scheuring 1992) bis zum Kaufverhalten (Menon et al. 1995) und der Zahl sexueller Kontakte (Tourangeau und Smith 1996; siehe Schwarz 1996, Kapitel 5, zum Überblick).

Der Effekt tritt auf, weil Befragte davon ausgehen, dass der Forscher eine sinnvolle Skala konstruiert hat. Sie nehmen an, dass die Skala das Wissen des Forschers über das untersuchte Phänomen widerspiegelt und daher „typische" Häufigkeiten im Mittelbereich der Skala repräsentiert sind, während die Extreme der Skala den Extremen der Verteilung entsprechen. Auf der Basis dieser Annahme ziehen sie die Skala als Bezugsrahmen zur Schätzung ihrer eigenen Verhaltenshäufigkeiten heran, was in den beobachteten Skaleneinflüssen resultiert. Diese Skaleneinflüsse nehmen ab, je besser das Verhalten im Gedächtnis repräsentiert ist – also unter Bedingungen, unter denen Befragte nicht auf eine Schätzstrategie zurückgreifen müssen (Menon et al. 1995; Ji et al. 2000).

Tabelle 1: Einfluss der Höhe der Antwortvorgaben auf Verhaltensberichte als Funktion von Frageinhalt und Alter

	Häufigkeitsvorgaben		
	Niedrig	Hoch	Unterschied
Alltägliches Verhalten			
Fleischkonsum			
Jüngere	24%	43%	19%
Ältere	19%	63%	44%
Kauf von Geburtstagsgeschenken			
Jüngere	42%	49%	7%
Ältere	46%	61%	15%
Krankheitssymptome			
Kopfschmerzen			
Jüngere	37%	56%	19%
Ältere	11%	10%	1%
Sodbrennen			
Jüngere	14%	33%	19%
Ältere	24%	31%	7%

Anmerkung: Jüngere (20–49 Jahre, $N = 143$) und ältere Befragte (60–100 Jahre, $N = 233$). Angaben beziehen sich auf Fleischkonsum von 10mal pro Monat oder öfter; Kauf von Geburtstagsgeschenken 5mal pro Jahr oder öfter; Kopfschmerzen und Sodbrennen zweimal pro Monat oder öfter. Adaptiert von Knäuper, Schwarz und Park (im Druck). Reprinted by permission.

Aus der Beobachtung, dass der Einfluss von Häufigkeitsvorgaben umso größer ist, je weniger gut das Verhalten im Gedächtnis repräsentiert ist, lässt sich die Vorhersage ableiten, dass der Einfluss von Antwortalternativen typischerweise für ältere Befragte stärker sein sollte als für jüngere Befragte. Die verfügbaren Daten unterstützen diese Vorhersage. *Tabelle 1* zeigt die Daten eines Laborexperiments mit $N = 233$ älteren Personen (60 bis 100 Jahre, $M = 76.21$) und $N = 143$ jüngeren Personen (20 bis 49 Jahre, mittleres Alter 35.29). Die älteren Befragten wurden wiederum über Zeitungsanzeigen oder den Befragten-Pool der University of Michigan rekrutiert. Die jüngeren Befragten wurden alle über Zeitungsanzeigen rekrutiert. 52,8 Prozent der älteren und 52,4 Prozent der jüngeren Gruppe waren Frauen. Wie *Tabelle 1* zeigt, fanden Knäuper, Schwarz und Park (2004), dass die älteren Befragten in ihren Berichten stärker als jüngere Personen von der Höhe der Häufigkeitsalternativen beeinflusst wurden, wenn sie über alltägliches Verhalten berichten sollten (z.B. das Kaufen eines Geburtstagsgeschenks, $t(372) = 2.19$, $p < .03$, oder die Häufigkeit von Fleischkonsum, $t(181) = 3.83$,

$p < .001$). Andererseits nimmt die Wichtigkeit mancher Lebensbereiche mit dem Alter zu. So schenken ältere Personen z.B. ihren Gesundheitsproblemen wesentlich mehr Aufmerksamkeit als Jüngere (siehe Borchert et al. 1999). Dies lässt erwarten, dass Jüngere bei Fragen nach der Häufigkeit körperlicher Beschwerden eher auf Schätzstrategien angewiesen sind als Ältere, da ihnen ihre allgemein bessere Gedächtnisleistung wenig hilft, wenn sie ihre körperlichen Symptome wenig beachten. Wie *Tabelle 1* zeigt, war dies der Fall. Die Symptomberichte älterer Befragter wurden von der Skala weniger beeinflusst als die Berichte jüngerer Befragter (Kopfschmerzen: $t(372) = 2.97$, $p < .003$; Sodbrennen: $t(372) = 3.01$, $p < .003$).

Hervorzuheben ist, dass wir wiederum unterschiedliche Schlussfolgerungen über Altersunterschiede in Verhaltenshäufigkeiten ziehen würden, je nachdem welche Antwortskala in der Umfrage verwendet wurde. Beispielsweise würden wir folgern, dass Altersunterschiede im Fleischkonsum nur gering sind, wenn die Skala niedrige Häufigkeiten vorgab, aber durchaus groß, wenn die Skala hohe Häufigkeiten vorgab. Um solche systematischen Einflüsse von Häufigkeitsskalen zu vermeiden, ist es empfehlenswert, Häufigkeitsberichte in einem offenen Antwortformat zu erfassen („Wie viele Stunden pro Tag schauen Sie fern? ___ *Stunden pro Tag.*"). Während dieses Format keineswegs die Validität der Antworten garantiert, vermeidet es zumindest *systematische* Verzerrungen.

VI. Abschließende Bemerkungen

Wie dieser Überblick zeigt, stellen *altersabhängige* Kontexteffekte eine besondere Herausforderung für die empirische Sozialforschung dar: Beobachtete Altersunterschiede in den berichteten Einstellungen oder Verhaltensweisen können entweder auf a) realen Unterschieden in Meinung oder Verhalten, b) Unterschieden im Antwortprozess oder c) einer unbekannten Mischung beider Faktoren beruhen. Ohne experimentelle Variationen ist es nicht möglich, die relativen Einflüsse dieser Faktoren zu isolieren. Da die Größe altersabhängiger Kontexteffekte oft ausreicht, die ordinale Platzierung von Altersgruppen auf der Einstellungs- oder Verhaltensdimension umzukehren, ist dieses Problem von großer praktischer Relevanz.

In ähnlicher Weise dokumentieren erste Untersuchungen zum Einfluss kultureller Unterschiede in Kognition (z.B. Ji et al. 2000) und Kommunikation (z.B. Haberstroh et al. 2002) auf Einstellungs- und Verhaltensberichte ausgeprägte *kulturabhängige* Kontexteffekte. Die Größe dieser Einflüsse war wiederum hinreichend, die ordinale Platzierung der untersuchten Länder (China versus Deutschland oder USA) auf der jeweiligen Meinungs- oder Verhaltensdimension umzukehren.

In allgemeineren Begriffen zeigen diese Untersuchungen zu alters- und kulturabhängigen Kontexteffekten, dass psychologische Variablen zu Fehlschlüssen über soziale Gruppen führen können, wenn sie a) eng mit soziodemographischen Variablen (wie Alter oder Herkunftsland) korreliert sind und b) die kognitiven und kommunikativen Prozesse beeinflussen, die Befragungen zugrunde liegen. Praktikable Lösungen für diese Probleme zu finden ist derzeit eine der größten Herausforderungen für Forschung an der Schnittstelle von Psychologie und Umfragemethodologie.

Literatur

Belli, Robert, 1998: The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. Memory 6: 383–406.

Bodenhausen, Galen V., und *Robert S. Wyer*, 1987: Social Cognition and Social Reality: Information Acquisition and Use in the Laboratory and the Real World. S. 6–41 in: *Hans-Jürgen Hippler, Norbert Schwarz* und *Seymour Sudman* (Hg.), Social Information Processing and Survey Methodology. New York: Springer Verlag.

Bradburn, Norman, Seymour Sudman et al., 1979: Improving Interviewing Methods and Questionnaire Design. San Francisco: Jossey-Bass.

Cantril, Hadley (Hg.), 1944: Gauging Public Opinion. Princeton: Princeton University Press.

Fiske, Alan P., Shinobu Kitayama, Hazel R. Markus und *Richard E. Nisbett*, 1998: The Cultural Matrix of Social Psychology. S. 915–981 in: *Daniel T. Gilbert* und *Susan T. Fiske* (Hg.), The Handbook of Social Psychology (Vol. 2, 4th edition). Boston, MA: McGraw-Hill.

Haberstroh, Susanne, Daphna Oyserman, Norbert Schwarz, Ulrich Kühnen und *Li-Jun Ji*, 2002: Is the Interdependent Self More Sensitive to Question Context than the Independent Self? Self-construal and the Observation of Conversational Norms. Journal of Experimental Social Psychology 38: 323–329.

Hippler, Hans-Jürgen, Norbert Schwarz und *Seymour Sudman* (Hg.), 1987: Social Information Processing and Survey Methodology. New York: Springer Verlag.

Jabine, Thomas B., Miron L. Straf, Judith M. Tanur und *Roger Tourangeau* (Hg.), 1984: Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines. Washington, DC: National Academy Press.

Ji, Li-Jun, Norbert Schwarz und *Richard E. Nisbett*, 2000: Culture, Autobiographical Memory, and Behavioral Frequency Reports: Measurement Issues in Cross-cultural Studies. Personality and Social Psychology Bulletin 26: 586–594.

Knäuper, Bärbel, 1999a: Age Differences in Question and Response Order Effects. S. 341–363 in: *Norbert Schwarz, Denise C. Park, Bärbel Knäuper* und *Seymour Sudman* (Hg.), Cognition, Aging, and Self-reports. Philadelphia, PA: Psychology Press.

Knäuper, Bärbel, 1999b: The Impact of Age and Education on Response Order Effects in Attitude Measurement. Public Opinion Quarterly 63: 347–370.

Knäuper, Bärbel, Norbert Schwarz und *Denise C. Park*, 2002: Selbstberichte im Alter. S. 75–98 in: *Andreas Motel-Klingebiel* und *Uwe Kelle* (Hg.), Perspektiven einer empirischen Alternssoziologie. Opladen: Leske + Budrich.

Knäuper, Bärbel, Norbert Schwarz und *Denise C. Park,* 2004: Frequency Reports across Age Groups. Journal of Official Statistics 20(1): 91–96.

Knäuper, Bärbel, Norbert Schwarz, Denise C. Park und *Andreas Fritsch,* 2003: The Perils of Interpreting Cohort Differences in Attitude Reports: Question Order Effects Decrease with Age. Manuskript, McGill University.

Krosnick, Jon A., und *Duane F. Alwin*, 1987: An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement. Public Opinion Quarterly 51: 201–219.

Linton, Marigold, 1982: Transformations of Memory in Everyday Life. S. 77–91 in: *Ulrich Neisser* (Hg.), Memory Observed: Remembering in Natural Contexts. San Francisco: Freeman.

Menon, Geeta, Priya Raghubir und *Norbert Schwarz*, 1995: Behavioral Frequency Judgments: An Accessibility-diagnosticity Framework. Journal of Consumer Research 22: 212–228.

Park, Denise C., 2000: The Basic Mechanisms Accounting for Age-related Decline in Cognitive Function. S. 3–22 in: *Denise C. Park* und *Norbert Schwarz* (Hg.), Cognitive Aging: A Primer. Philadelphia, PA: Psychology Press.

Park, Denise C., und *Norbert Schwarz* (Hg.), 2000: Cognitive Aging: A Primer. Philadelphia, PA: Psychology Press.

Payne, Stanley L., 1951: The Art of Asking Questions. Princeton: Princeton University Press.

Salthouse, Timothy A., und *Renee L. Babcock*, 1991: Decomposing Adult Age Differences in Working Memory. Developmental Psychology 27: 763–776.

Salthouse, Timothy A., Renee L. Babcock und *Raymond J. Shaw,* 1991: Effects of Adult Age on Structural and Operational Capacities in Working Memory. Psychology and Aging 6: 118–127.

Schaie, K. Warner, 1996: Intellectual Development in Adulthood. S. 266–280 in: *James E. Birren* und *K. Warner Schaie* (Hg.), Handbook of the Psychology of Aging. San Diego: Academic Press.

Schuman, Howard, 1992: Context Effects: State of the Past/State of the Art. S. 5–20 in: *Norbert Schwarz* und *Seymour Sudman* (Hg.), Context Effects in Social and Psychological Research. New York: Springer Verlag.

Schuman, Howard, und *Stanley Presser*, 1981: Questions and Answers in Attitude Surveys. New York: Academic Press.

Schwarz, Norbert, 1990: Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction. S. 98–119 in: *Clyde Hendrick* und *Margaret S. Clark* (Hg.), Research Methods in Personality and Social Psychology (Review of Personality and Social Psychology, Vol. 11). Beverly Hills, CA: Sage.

Schwarz, Norbert, 1996: Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation. Hillsdale, NJ: Erlbaum.

Schwarz, Norbert, 1999: Self-reports: How the Questions Shape the Answers. American Psychologist 54: 93–105.

Schwarz, Norbert, 2003: Culture-sensitive Context Effects: A Challenge for Cross-cultural Surveys. S. 93–100 in: *Janet A. Harkness, Fons J. R. van de Vijver* und *Peter Ph. Mohler* (Hg.), Cross-cultural Survey Methods. New York: Wiley.

Schwarz, Norbert, und *Herbert Bless*, 1992: Constructing Reality and its Alternatives: Assimilation and Contrast Effects in Social Judgment. S. 217–245 in: *Leonard L. Martin* und *Abraham Tesser* (Hg.), The Construction of Social Judgment. Hillsdale: Erlbaum.

Schwarz, Norbert, und *Gerd Bohner*, 2001: The Construction of Attitudes. S. 436–457 in: *Abraham Tesser* und *Norbert Schwarz* (Hg.), Blackwell Handbook of Social Psychology: Intraindividual Processes (Vol. 1). Oxford, UK: Blackwell.

Schwarz, Norbert, Hans-Jürgen Hippler, Brigitte Deutsch und *Fritz Strack*, 1985: Response Categories: Effects on Behavioral Reports and Comparative Judgments. Public Opinion Quarterly 49: 388–395.

Schwarz, Norbert, Hans-Jürgen Hippler und *Elisabeth Noelle-Neumann* 1992: A Cognitive Model of Response Order Effects in Survey Measurement. S. 187–201 in: *Norbert Schwarz* und *Seymour Sudman* (Hg.), Context Effects in Social and Psychological Research. New York: Springer Verlag.

Schwarz, Norbert, und *Daphna Oyserman*, 2001: Asking Questions about Behavior: Cognition, Communication, and Questionnaire Construction. American Journal of Evaluation 22: 127–160.

Schwarz, Norbert, Denise C. Park, Bärbel Knäuper und *Seymour Sudman* (Hg.), 1999: Cognition, Aging, and Self-reports. Philadelphia, PA: Psychology Press.

Schwarz, Norbert, und *Bettina Scheuring*, 1992: Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Anwortvorgaben des Fragebogens lernen. Zeitschrift für Klinische Psychologie 22: 197–208.

Schwarz, Norbert, Fritz Strack, Hans-Jürgen Hippler und *George Bishop*, 1991: The Impact of Administration Mode on Response Effects in Survey Measurement. Applied Cognitive Psychology 5, 193–212.

Schwarz, Norbert, Fritz Strack und *Hans-Peter Mai,* 1991: Assimilation and Contrast Effects in part-whole Question Sequences: A Conversational Logic Analysis. Public Opinion Quarterly 55 (1): 3–23.

Schwarz, Norbert, und *Seymour Sudman* (Hg.), 1992: Context Effects in Social and Psychological Research. New York: Springer Verlag.

Schwarz, Norbert, und *Seymour Sudman* (Hg.), 1994: Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer Verlag.

Schwarz, Norbert, und *Seymour Sudman* (Hg.), 1996: Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Sears, David O., 1986: College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. Journal of Personality and Social Psychology, 51 (3): 515–530.

Sirken, Monroe, Douglas Hermann, Susan Schechter, Norbert Schwarz, Judith Tanur und *Roger Tourangeau* (Hg.), 1999: Cognition and Survey Research. New York: Wiley.

Strack, Fritz, 1994: Zur Psychologie der standardisierten Befragung. Berlin: Springer Verlag.

Strack, Fritz, und *Leonard L. Martin*, 1987: Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. S. 123–148 in: *Hans-Jürgen Hippler, Norbert Schwarz* und *Seymour Sudman* (Hg.), Social Information Processing and Survey Methodology. New York: Springer Verlag.

Strack, Fritz, Leonard L. Martin und *Norbert Schwarz*, 1988: Priming and Communication: The Social Determinants of Information Use in Judgments of Life-satisfaction. European Journal of Social Psychology 18: 429–442.

Strube, Gerhard, 1987: Answering Survey Questions: The Role of Memory. S. 86–101 in: *Hans-Jürgen Hippler, Norbert Schwarz* und *Seymour Sudman* (Hg.), Social Information Processing and Survey Methodology. New York: Springer Verlag.

Sudman, Seymour, Norman Bradburn und *Norbert Schwarz*, 1996: Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco, CA: Jossey-Bass.

Tourangeau, Roger, 1984: Cognitive Science and Survey Methods: A Cognitive Perspective. S. 73–100 in: *Thomas B. Jabine, Miron L. Straf, Judith M. Tanur* und *Roger Tourangeau* (Hg.), Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines. Washington, DC: National Academy Press.

Tourangeau, Roger., Lance J. Rips und *Kenneth Rasinski,* 2000: The Psychology of Survey Response. Cambridge: Cambridge University Press.

Tourangeau, Roger, und *Smith, Tom W.,* 1996: Asking Sensitive Questions. The Impact of Data Collection, Mode, Question Format, and Question Context. Public Opinion Quarterly 60, 275–304.

ONLINE-BEFRAGUNG

Probleme und Chancen verschiedener Arten von Online-Erhebungen*

Mick P. Couper und Elisabeth Coutts

Zusammenfassung: In den letzten fünf Jahren hat die Anzahl der durchgeführten Online-Befragungen weiter zugenommen. Gleichzeitig ist die Zahl der Veröffentlichungen zu methodischen Problemen und Verfahren der Web-Befragungen gestiegen. Aufbauend auf der Arbeit von Couper (2000) werden hier zunächst die Hauptfehlerquellen von Befragungen im Allgemeinen und deren spezifische Ausprägung bei Online-Surveys im Besonderen dargestellt. Hierzu gehören Probleme wie Unterschiede in der Internet-Nutzung zwischen sozialen Gruppen, des Nonresponse und Fehler aufgrund der Gestaltung der Erhebungsinstrumente. Anschließend werden die Arten von Online-Befragungen typisiert, wobei zunächst zwischen nicht-zufallsgesteuerten und zufallsgesteuerten Methoden unterschieden wird. Abschließend werden Möglichkeiten und Grenzen webbasierter Surveys vor dem Hintergrund neuer Forschungsergebnisse diskutiert.

I. Einleitung

Mit der zunehmenden Ausbreitung des Internets wächst auch die Anzahl Online-Umfragemethoden sowohl im Bereich der Markt- und Meinungsforschung als auch der akademischen Forschung. Online-Umfragemethoden umfassen u.a. verschiedene Spielarten von E-Mail-Befragungen, bei denen der Fragebogen per E-Mail verschickt wird, sowie Web-Befragungen, bei denen ein Fragebogen online in Form eines Web-Formulars auszufüllen ist. Während sich einige Meinungsforscher und Wissenschaftler durch die schnelle Ausbreitung von Umfragen im Internet oder im World Wide Web (WWW) dazu veranlasst sehen, das Ende der traditionellen Erhebungsmethoden zu prophezeien, stehen andere der zukünftigen Rolle von Web-Befragungen eher skeptisch gegenüber. Unabhängig von den persönlichen Meinungen über die Zukunft der Online-Befragung ist jedoch klar, dass die Untersuchung von methodischen Aspekten internetbasierter Umfragen fester Bestandteil weiterer Forschungsarbeiten sein sollte.

Der Forschungsbedarf gründet unter anderem darin, dass es zwar immer einfacher wird, Web-Befragungen durchzuführen, sich die Gewährleistung einer hohen Qualität der Umfragen aber auch gerade dadurch zunehmend schwierig gestaltet. So können

* Wir danken Daniel Heller, Frauke Kreuter, Reto Meyer, Wojtek Przepiorka, Claudia Queißer, Hélène Venningen und einem anonymen Gutachter für hilfreiche Kommentare, und Nicole Schneider für Unterstützung bei der Vorbereitung des Manuskripts. Ben Jann gilt unserer besonderer Dank für die gründliche und kritische Durchsicht des Manuskripts. Wir bedanken uns auch bei ZUMA, wo Mick Couper während der Vorbereitung des Manuskripts Gastwissenschaftler war. Teile des Artikels basieren auf einer Publikation von Couper (2000).

Umfragedaten für eine breite Masse von Individuen kostengünstig erhoben werden. Und nicht nur Forscher haben Zugang zu dieser ungeahnt großen Anzahl potenzieller Befragter. Die relativ geringen Kosten ermöglichen es nahezu jeder Person, die über einen Internetanschluss verfügt, eine eigene Web-Befragung durchzuführen. Gerade dadurch entsteht aber die Gefahr, dass es immer schwieriger wird, „gute" und „schlechte" Befragungen voneinander zu trennen. Zudem hängt der Wert von Online-Umfragen wie auch bei anderen Ansätzen von der Teilnahmebereitschaft der anvisierten Befragungspersonen ab. So könnte die Methode der Online-Befragung ihre eigenen Anwendungsvoraussetzungen untergraben, indem die Internet-Nutzer, bedingt durch die günstige Durchführung, mit derart vielen Umfragen überhäuft werden, dass sie sich der Teilnahme gänzlich verschließen, oder ihre Entscheidung über die Teilnahme nur noch selektiv aufgrund des Inhalts, des Themas, des Unterhaltungswerts oder anderer Merkmale der Befragung fällen. Qualitativ hochwertige, auf einem guten Forschungsdesign basierende Web-Befragungen könnten somit in der Masse vieler anderer Versuche untergehen, Befragungsdaten im Internet zu sammeln. Dieser Effekt war bereits im Bereich der Telefonumfragen zu beobachten, wo die starke Verbreitung von Telemarketing die Verwendung der Methode zu wissenschaftlichen Zwecken bedroht.

Zu beachten ist allerdings, dass die Beurteilung der Qualität einer Umfrage (ob Web-Survey oder Umfrage eines anderen Typs) im Hinblick auf den intendierten Gebrauch der gewonnenen Daten, also im Hinblick auf die Aussagen, die auf Basis der Daten getroffen werden sollen, erfolgen sollte. Die „Qualität" einer Umfrage lässt sich also als die Tauglichkeit der erfassten Daten für den beabsichtigten Verwendungszweck konzipieren („fitness for use", Juran und Gryna 1980). Generell bezeichnet O'Muircheartaigh (1997: 1) fehlerbehaftete Arbeit oder Arbeit mangelnder Qualität als „work purporting to do what it does not do" und erläutert dies bezüglich der Qualität von Umfragen wie folgt: „Broadly speaking, every survey has an objective, an outcome, and a description of that outcome. Error (quality failures) will be found in the mismatches among these elements." Die Qualität von Umfragen sollte also auch hinsichtlich Eigenschaften wie zum Beispiel der Messgenauigkeit, der Kosten oder der termingerechten Beendigung der Arbeiten evaluiert werden. Weiterhin sollte die Qualität eines bestimmten Ansatzes in Anbetracht alternativer Forschungsdesigns beurteilt werden, die ebenfalls zur Untersuchung der spezifischen Fragestellung geeignet wären. Schließlich sollte die Qualität eines Ansatzes nicht auf der Basis einer einzigen Anwendung bewertet werden.

Das Ziel des vorliegenden Beitrags ist, die derzeit am stärksten verbreiteten Typen von Web-Befragungen hinsichtlich gängiger Qualitätskriterien und Fehlerquellen zu evaluieren. Nachfolgend werden zunächst die Hauptfehlerquellen bei Surveys und deren spezifische Ausprägung bei Web-Befragungen diskutiert. Zu den Hauptfehlerquellen zählen abdeckungsbezogene Fehler, Stichprobenfehler, Fehler durch Nonresponse und Messfehler (vgl. z.B. Groves 1989). Anschließend werden die unterschiedlichen Arten von Online-Befragungen typisiert und auf ihre Anfälligkeit bezüglich der besprochenen Fehlerquellen untersucht.

II. Abdeckungsbezogener Fehler und Stichprobenfehler

Zwei der wichtigsten Fehler, die im Rahmen einer Stichprobenziehung vorkommen können, sind abdeckungsbezogene Fehler („coverage errors") und Stichprobenfehler („sampling errors").

Abdeckungsbezogene Fehler resultieren aus einer unvollständigen Berücksichtigung der Zielpopulation durch die Auswahlgesamtheit. Die Zielpopulation (oder angestrebte Grundgesamtheit) kann als die Menge der Individuen betrachtet werden, die man untersuchen möchte, bzw. als die Population, auf die man Rückschlüsse zu ziehen beabsichtigt (z.B. die erwachsene Bevölkerung eines Landes). Die Auswahlgesamtheit hingegen („sampling frame") wird von Groves (1989: 82) als die Menge der Personen definiert, für die vor der Stichprobenziehung eine Auflistung vorgenommen werden kann. Beispiele für Auswahlgesamtheiten sind etwa alle privaten Telefonanschlüsse (für eine Telefonbefragung) oder alle persönlichen E-Mail-Adressen der Mitglieder einer Organisation (für eine Online-Mitgliederbefragung). Stichprobenfehler andererseits entstehen aufgrund der Tatsache, dass nicht alle Mitglieder der Auswahlgesamtheit in eine jeweilige Stichprobe aufgenommen werden können. Falls das Auswahlverfahren wiederholt würde, erhielte man eine leicht unterschiedliche Menge von ausgewählten Personen. Man beachte, dass sich das Konzept des Stichprobenfehlers explizit auf Zufallsstichproben bezieht, das heißt auf Stichproben, bei denen jedem Mitglied der Auswahlgesamtheit eine bekannte Auswahlwahrscheinlichkeit größer als Null zugeordnet werden kann. Während sich der abdeckungsbezogene Fehler also auf das Problem bezieht, dass ein Teil der Zielpopulation nicht durch die Auswahlgesamtheit erfasst wird (im Fall von Online-Befragungen wären dies z.B. Personen ohne Zugang zum Internet), tritt der Stichprobenfehler bei der Ziehung einer Stichprobe aus dieser Auswahlgesamtheit auf.

Web-Befragungen sind besonders anfällig für abdeckungsbezogene Fehler, da der Befragungsmodus die Auswahlpopulation einschränkt und zudem die Auswahlgesamtheit für Web-Befragungen oftmals gar nicht eindeutig festgelegt werden kann. Die meisten Fragestellungen beziehen sich auf eine Zielpopulation, die über die Population der Internetnutzer hinausgeht, es sei denn, die Zielpopulation wird im Voraus explizit auf die zum Zeitpunkt der Untersuchung aktiven Internetnutzer eingeschränkt. So stellen abdeckungsbezogene Fehler derzeit wohl die Hauptursache für die mangelnde Verallgemeinerbarkeit von Ergebnissen aus Online-Befragungen dar. Wie Groves et al. (2004) feststellen, hängt der abdeckungsbezogene Fehler von zwei Faktoren ab. Einerseits ist der Abdeckungsgrad von Bedeutung, also der Anteil der Zielpopulation, der durch die Auswahlgesamtheit abgedeckt wird. Andererseits hängt der abdeckungsbezogene Fehler auch von der Größe der Differenzen bezüglich der Untersuchungsvariablen ab, die zwischen den in der Auswahlgesamtheit enthaltenen Personen und den sonstigen Mitgliedern der Zielpopulation bestehen.

Neben der mangelnden Abdeckung der Zielpopulation besteht ein weiteres Problem darin, die Auswahlgesamtheit für Web-Befragungen überhaupt erst zu konstruieren. Selbst wenn die Abdeckung vollständig wäre, wenn also zum Beispiel jede Person in einem bestimmten Land Zugang zum Internet hätte, wären die Schwierigkeiten erheblich, eine Auflistung der Auswahlgesamtheit zum Zweck der Ziehung einer Zufallsstichprobe zu erstellen. Es gibt kein Internet-Verzeichnis, das einem Telefonverzeichnis

oder einem Registerauszug ähnlich wäre, und eine zufällige Generierung von E-Mail-Adressen wie bei telefonischen Random-Digit-Dialing-Befragungen (RDD-Befragung, dabei werden einige Ziffern bekannter Telefonnummern durch zufällig generierte Zahlen ersetzt) lässt sich kaum realisieren. Auch das Problem, eine angemessene Auswahlgesamtheit zu bilden, gefährdet somit die Gültigkeit der Schlussfolgerungen aus Web-Befragungen.

Wenden wir uns zunächst dem Problem des Abdeckungsgrads („coverage") zu, also dem Anteil der Zielpopulation, der potenziell über das Internet erreicht werden kann. Während die Verfechter von Online-Befragungen als Grund für ihren Optimismus auf die erstaunliche Wachstumsrate des Internets Ende der 1990er und Anfang der 2000er Jahre verweisen, so bleibt die zukünftige Entwicklung doch ungewiss. Zudem weichen die Schätzungen für die Abdeckungsdichte von Internetzugängen stark voneinander ab, wobei es zugegebenermaßen schwierig ist, eine Größe zu messen, die sich so schnell verändert. Zum Teil beruhen die Abweichungen auf Definitionsproblemen („Zugang" versus „Nutzung", Schätzungen auf Haushalts- oder Personenebene, Zugang zu Hause oder im Büro usw.), sie sind aber auch durch die unterschiedliche Qualität der Befragungen bedingt, die zur Ermittlung der Abdeckungsdichte des Internets durchgeführt werden.

Die offiziellen Statistiken der Europäischen Union deuten auf sehr unterschiedliche Abdeckungsgrade in den Mitgliedsländern hin. Insgesamt scheint die Internetanschlussdichte in Europa aber etwas niedriger zu sein als in den USA.[1] 42 Prozent der Haushalte in den 25 derzeitigen EU-Mitgliedsstaaten hatten im Jahr 2004 Zugang zum Internet (Eurostat 2005). Die niedrigsten Zugangsraten finden sich in den neuen Mitgliedsstaaten wie Litauen (12 Prozent), Ungarn (14 Prozent) oder der Tschechischen Republik (19 Prozent) sowie in den südeuropäischen Ländern wie Griechenland (17 Prozent), Portugal (26 Prozent) und Italien (34 Prozent). Die höchsten Zugangsraten weisen die skandinavischen Länder auf, zum Beispiel Norwegen (60 Prozent), Dänemark (69 Prozent) oder Island (81 Prozent).

Die offiziellen Statistiken der deutschsprachigen Länder weisen einen Verbreitungsgrad auf, der dem europäischen Durchschnitt entspricht oder sogar etwas darüber liegt. Diesen Zahlen zufolge nahm die Internetanschlussdichte in den letzten zwei Jahren zudem beträchtlich zu. Das Statistische Bundesamt Deutschland (2005) gibt an, dass Anfang 2004 rund 50 Prozent der 16- bis 74-Jährigen das Internet regelmäßig von zu Hause aus oder im Büro benutzten.[2] Zudem hatten 2004 60 Prozent der deutschen Haushalte mit mindestens einem Haushaltsmitglied unter 75 Jahren Zugang zum Internet, verglichen mit 46 Prozent im Jahr 2002. Statistik Austria (2005) berichtet für österreichische Haushalte etwas niedrigere Zahlen: Unabhängig von der Alterszusam-

1 Die wahrscheinlich genauesten Messungen der Anzahl Internetzugänge und ihrer Nutzer in den USA stammen aus dem Current Population Survey (CPS). Den CPS-Statistiken aus dem Jahr 2003 zufolge verfügt mehr als die Hälfte der US-Haushalte (55 Prozent) über einen Internetzugang (National Telecommunications and Information Administration 2004).

2 Die Ergebnisse für Deutschland wurden mittels einer Quoten-Stichprobe gewonnen, je nach Bundesland mit einer postalischen oder einer Telefonbefragung (Statistisches Bundesamt 2005: 44–45). Für weitere Schätzungen der Internetanschlussdichte in Deutschland siehe auch den (N)Onliner Atlas von TNS Infratest und Initiative D^{21} bei www.nonliner-atlas.de. Die Daten für die Schweiz und Österreich stammen aus Telefonbefragungen mit Zufallsstichprobe.

mensetzung hatten rund 40 Prozent der österreichischen Haushalte Anfang 2004 zu Hause einen Internetzugang, verglichen mit 31 Prozent im Jahr 2002. Die offiziellen Schweizer Daten beziehen sich auf Personen und nicht auf Haushalte. Gemäß dem Schweizerischen Bundesamt für Statistik (2005a) hatten Ende 2004 etwa zwei Drittel der Schweizer Bevölkerung zu Hause einen Internetanschluss, während 50 Prozent an ihrem Arbeitsplatz oder in der Schule auf das Internet zugreifen konnten. Zudem gaben 68 Prozent der Befragten an, das Internet in den vergangenen sechs Monaten genutzt zu haben, 41 Prozent sogar „täglich" oder „fast täglich".

Diese Zahlen belegen zwar die rasche Zunahme der Internetnutzung in den letzten Jahren, aber Zugang und Nutzung sind nicht bei allen Bevölkerungsanteilen gleich. Ob Online-Befragungen in Zukunft repräsentativ für die Gesamtbevölkerung sein können, hängt vom weiteren Anstieg der Internetnutzung ab. Die beiden wichtigsten Fragen sind hierbei, wie schnell die Entwicklung voranschreitet und wie weit dieser Zuwachs gehen wird. Die Wachstumskurve wird sich unweigerlich abflachen und irgendwann einen stabilen Zustand erreichen. Wann dies der Fall sein wird und auf welchem Niveau ist jedoch noch ungewiss.

Zum heutigen Zeitpunkt ist der Zugang zum Internet und dessen Nutzung in Europa, den USA und anderen Ländern noch keineswegs umfassend, und das dürfte auch noch einige Zeit so bleiben. Die Tatsache, dass ein großer Teil der Bevölkerung derzeit über das Internet nicht erreicht wird, ist eine ernsthafte Herausforderung für Versuche, Zufallsstichproben aus der Gesamtbevölkerung für Web-Befragungen zu entwickeln. Die Frage ist dabei nicht nur, wie groß der Anteil der Bevölkerung ist, der mit einer bestimmten Methode nicht erreicht werden kann. Das schwerwiegendere Problem, das durch die geringe Abdeckungsrate entsteht, liegt in dem Ausmaß, in dem sich Personen, die potenziell mit der Internet-Technologie erreicht werden können, von der Gesamtbevölkerung in Bezug auf die Untersuchungsvariablen unterscheiden.

Die demographischen Unterschiede zwischen Personen mit Web-Zugang und Personen ohne Zugang sind bereits gut dokumentiert. Die Daten der statistischen Ämter im deutschen Sprachraum deuten auf eine so genannte digitale Kluft („digital divide") hin. Markante Unterschiede finden sich u.a. in Bezug auf folgende Variablen:

- *Einkommen.* In Deutschland hatten 2004 87 Prozent der Haushalte mit einem monatlichen Nettoeinkommen von EUR 2.600 oder mehr Zugang zum Internet. Die Quote für Haushalte mit einem monatlichen Nettoeinkommen von EUR 1.300 oder weniger betrug jedoch lediglich 34 Prozent (Statistisches Bundesamt 2005: 28). In der Schweiz benutzten 2004 81 Prozent der Personen mit einem monatlichen Einkommen von SFR 10.000 oder mehr das Internet mehrmals pro Woche. Für Personen mit einem monatlichen Einkommen von SFR 4.000 oder weniger betrug die Quote 25 Prozent (2005a).
- *Bildung.* In der Schweiz verwendeten 2004 81 Prozent der Hochschulabsolventen das Internet mehrmals pro Woche, im Vergleich zu 52 Prozent der Absolventen einer Lehre oder Berufsschule und 31 Prozent der Personen mit obligatorischer Schule als höchstem Abschluss. In Österreich hatten 2003 82 Prozent der Hochschulabsolventen das Internet innerhalb des letzten Jahres genutzt, im Vergleich zu 24 Prozent der Personen mit lediglich Pflichtschulabschluss und 21 Prozent der Personen ohne Pflichtschulabschluss (Statistik Austria 2005: Tabelle 5.33).
- *Geschlecht.* In Deutschland nutzten im ersten Quartal 2004 63 Prozent der Männer und 53 Prozent der Frauen das Internet. In der Schweiz lag der Anteil der Internetnutzer unter den Männern noch deutlicher über dem Anteil bei den Frauen: 63 gegenüber 46 Prozent. Ein ähnliches

Muster findet sich in den Daten für Österreich: 60 Prozent der Männer und 48 Prozent der Frauen gaben 2003 an, das Internet in den vergangenen 12 Monaten genutzt zu haben.

– *Alter.* In Deutschland lebten 2004 85 Prozent der 10- bis 24-Jährigen in einem Haushalt mit Internetzugang, im Vergleich zu 32 Prozent der über 54-Jährigen. In der Schweiz benutzten 78 Prozent der 20- bis 29-Jährigen das Internet regelmäßig, gegenüber 31 Prozent der über 49-Jährigen. In Österreich variierte der Prozentsatz derjenigen, die das Internet in den letzten sechs Monaten genutzt hatten, zwischen 81 Prozent bei den 16- bis 24-Jährigen und 10 Prozent bei den 65- bis 74-Jährigen.

Selbst wenn die demographischen Charakteristika der Web-Nutzer mit denen der allgemeinen Bevölkerung übereinstimmen würden, wäre das Problem noch nicht gelöst. Die eigentliche Frage ist, wie ähnlich sich die beiden Populationen in Bezug auf die interessierenden Untersuchungsvariablen wie Einstellungen, Verhaltensweisen, Kauf- oder Wahlabsichten sind. Die hierfür relevanten Forschungsergebnisse basieren auf Studien, bei denen parallele Untersuchungen mit verschiedenen Befragungstechniken und Stichproben durchgeführt wurden. So berichtet Faas (2004) von einer Studie, bei der die Online-Befragten mit einer Zufallsstichprobe von face-to-face interviewten Personen verglichen wurden. Im Vergleich zu den face-to-face-Befragten äußerten die Online-Befragten häufiger die Absicht, Bündnis 90/Die Grünen oder die FDP zu wählen und weniger häufig die CDU. Auf ähnliche Weise verglich eine britische Studie (Sparrow und Curtice 2004) drei Gruppen von Befragten aus einer RDD-Telefonbefragung: 1. alle Befragten, 2. die Befragten, die angaben, einen Internetzugang zu haben und 3. die Befragten, die angaben, einen Internetzugang zu haben, die auch einwilligten, an einer Reihe weiterer Befragungen teilzunehmen. Gemäß Sparrow und Curtice waren Vollzeit arbeitende Männer im Vergleich zur Kontrollgruppe in beiden Internet-Gruppen überrepräsentiert, ältere Befragte waren dagegen unterrepräsentiert. Aber noch deutlicher gestalteten sich die Unterschiede in einer Reihe von Einstellungsfragen. In einigen Fällen lagen die Abweichungen gegenüber der Gesamtheit der telefonisch Befragten bei über 15 Prozentpunkten. Sparrow und Curtice (2004: 40) ziehen daraus die folgende Schlussfolgerung: „[That] the differences found have no discernable pattern to them ... certainly makes it difficult to tell when an internet poll is going to provide a reasonable estimate of public opinion in general and when not." Diese Anmerkungen entsprechen bereits früher geäußerten Bedenken von Flemming und Sonner (1999) hinsichtlich des Mangels an vorhersagbaren Mustern in Bezug auf die Unterschiede bei Antworten aus einer Web- und einer RDD-Stichprobe.[3]

Zusammenfassend deuten diese Ergebnisse darauf hin, dass sich die „Internet-Bevölkerung" (falls sich eine solche überhaupt bestimmen lässt) in vielen Aspekten von der allgemeinen Bevölkerung unterscheidet. Obwohl das Web sich in einem Stadium massiven Wachstums und ständiger Veränderung befindet, dürften diese Unterschiede noch für geraume Zeit bestehen bleiben. Die Herausforderung für Forscher, die sich mit Web-Befragungen befassen, besteht darin, kreative Möglichkeiten zu finden, mit solchen Befragungstechniken einen möglichst großen Anteil der gewünschten Zielpopulationen zu erfassen. Die Alternative dazu ist, Verallgemeinerungen auf begrenzte Po-

3 Siehe aber Schonlau, Zapert, Simon et al. (2003) für eine Diskussion des Einflusses der Fragecharakteristika (z.B. Sachbezogenheit, Anzahl der Antwortkategorien usw.) auf die Abweichungen der Antworten zwischen Web- und anderen Befragungen.

pulationen einzuschränken (z.B. Internetnutzer, die einer bestimmten Sprache mächtig sind).

Um schließlich nochmals auf die Stichprobenfehler zurückzukommen: Vielen Web-Surveys liegt die falsche Annahme zugrunde, dass umfangreiche Stichproben automatisch zu zuverlässigeren Ergebnissen führen oder dass die Stichprobengröße (oder genauer gesagt die Anzahl der Befragten) die einzige Determinante des Stichprobenfehlers sei. Inferenzstatistische Aussagen sind jedoch nur bei Zufallsstichproben möglich. Bei andersartigen Stichproben sind alle Bemühungen zur Verallgemeinerung auf eine Population, zur Berechnung des Stichprobenfehlers oder eines Konfidenzintervalls irreführend. Hier muss man zwischen wissenschaftlichen Befragungen, die Rückschlüsse auf eine Population erlauben, und schlichter Datenerfassung, bei der es vor allem auf die Zahl der Befragten, statt auf die Repräsentativität ankommt, unterscheiden. Man sollte diese beiden Ansätze nicht miteinander verwechseln.

III. Nonresponse-Fehler

Einige Arten von Web-Befragungen haben – wie im Folgenden zu zeigen sein wird – Lösungen für das Problem des Abdeckungsgrads gefunden. Aber selbst wenn die angestrebte Grundgesamtheit einen hohen Abdeckungsgrad aufweist und man für diese Population eine Auswahlgesamtheit identifizieren kann, beeinträchtigt Nonresponse auch bei Online-Befragungen die Brauchbarkeit der Daten. Ein Nonresponse-Fehler liegt vor, wenn nicht alle Befragten in der Stichprobe gewillt oder in der Lage sind, die Fragen zu beantworten bzw. an der Befragung teilzunehmen. Wie auch der abdeckungsbezogene Fehler ist der Nonresponse-Fehler nicht nur abhängig von der Teilnahmequote an sich, sondern auch von der Größe des Unterschieds zwischen den Befragten und den Nicht-Befragten bezüglich der Untersuchungsvariablen (Groves und Couper 1998).

Wenn die Auswahlgesamtheit nicht identifiziert werden kann, ist das Ausmaß an Nonresponse schwierig zu bestimmen. Wenn man z.B. in einem Web-Portal eine offene Einladung ausspricht, an einer Befragung teilzunehmen, ist die Anzahl der Personen, die zur Teilnahme in Frage kommen, normalerweise nicht bekannt. Man kann bei einer solchen Befragung daher auch keine Nonresponsequote berechnen. Der Nonresponse-Fehler lässt sich nur in solchen Fällen nachvollziehbar messen oder bewerten, bei denen die Auswahlgesamtheit und die Auswahlwahrscheinlichkeit bekannt sind, das heißt bei Umfragen mit einer Zufallsstichprobe. Üblicherweise geschieht die Auswahl bei solchen Studien über eine Liste von E-Mail-Adressen (z.B. von Studierenden, Mitarbeitenden oder Organisationsmitgliedern) oder mittels der Rekrutierung über einen anderen Befragungsmodus (z.B. über Telefoninterviews).

Um Vorstellungen über das Nonresponse-Problem in Online-Befragungen gewinnen zu können, müssen wir uns also vorrangig auf Befragungen stützen, die zufallsgesteuerte Stichproben verwenden. Studien über Nonresponse im Internet vergleichen üblicherweise die Rücklaufquote bei einer Web-Befragung mit dem Rücklauf bei einer parallel durchgeführten postalischen Befragung. Die bisherigen Ergebnisse solcher Studien sind uneinheitlich, deuten aber eher darauf hin, dass die Teilnahmequoten bei

Web-Befragungen im Allgemeinen niedriger sind als bei postalischen Befragungen (z.B. Kwak und Radler 2002; Batinic 2003; Hayslett und Wildemuth 2004). Es gibt jedoch auch Gegenbeispiele, bei denen die Teilnahmequote im Web-Modus höher war als bei der postalischen Befragung (z.B. Cobanoglu, Warde und Moreo 2001; Kiernan et al. 2005; McCabe et al. im Druck), ungefähr gleich (z.B. Forsman und Varedian 2002) oder nur leicht niedriger (Kaplowitz, Hadlock und Levine 2004; Reips und Franek 2004). Eine neuere Metaanalyse (Lozar Manfreda et al. 2005) veröffentlichter als auch unveröffentlichter Studien kommt zum Ergebnis, dass sich die Teilnahmequoten bei Web-Befragungen nicht signifikant von den Teilnahmequoten bei anderen Befragungsmodi (in erster Linie postalischen Befragungen) unterscheiden. Eine mögliche Erklärung für den letzthin beobachteten kleiner werdenden Unterschied zwischen den Antwortquoten von Web-Umfragen und postalischen Befragungen ist, dass E-Mail-Datenbanken vollständiger geworden sind. Für einige Populationen, z.B. Studierende, sind die E-Mail-Adressen womöglich sogar aktueller als die Informationen zur Postanschrift.

Trotz der relativen Verbesserung der Teilnahmequoten in den letzten Jahren gibt es einige Nonresponse-Ursachen, die speziell bei Online-Umfragen wirksam sind. Einer der Vorzüge des Internets ist die Möglichkeit, multimediale Elemente in die Befragungen einzubinden. Aber die Verwendung solcher Komponenten (die aktuelle Browser, Plug-ins, Cookies usw. erfordern) kann einen Teil der Zielpopulation auch davon abhalten, den Fragebogen auszufüllen. So beobachteten Schwarz und Reips (2001) höhere Abbruchraten bei einer Befragung mit einem Web-Formular, das Java-Software auf dem Computer des Benutzers erforderte, als bei einer Befragung mit einem Formular ohne diese Java-Elemente. Weiterhin erhielten Dillman, Tortora, Conradt und Bowker (1998) sowie Lozar Manfreda, Batagelj und Vehovar (2002) höhere Nonresponse-Raten für Web-Befragungen, in denen rechenintensive Grafik-Elemente verwendet wurden. Whitcomb und Porter (2004) berichten ein ähnliches Ergebnis für eine E-Mail-Befragung.

Langsame Modems, unzuverlässige Verbindungen usw. können die zeitlichen Kosten für das Ausfüllen von Online-Fragebögen erhöhen und die Teilnahmequote deswegen weiter beeinträchtigen. Bei einigen Internet-Anbietern schlagen zudem die Verbindungskosten zu Buche, was die Teilnahmebereitschaft weiter beeinträchtigen kann. In Deutschland verfügten 2004 nur 30 Prozent der Haushalte mit einem privaten Internetzugang über eine Breitbandverbindung, der Anteil in Österreich lag 2003 bei 27 Prozent (Statistik Austria 2005: Tabelle 5.32; Statistisches Bundesamt 2005). Fast alle übrigen Haushalte nutzten eine Einwahlverbindung für den Internetzugang. In solchen Fällen entstehen durch die Teilnahme an einer Web-Befragung in der Regel monetäre Kosten für die Befragten.

Darüber hinaus sind nicht alle Benutzer gleichermaßen mit E-Mail oder mit dem Web vertraut. Gewillte Teilnehmer wissen womöglich nicht, wie sie einen Online-Fragebogen aufrufen und ausfüllen müssen (vgl. Dillman und Bowker 2001: 165–169, für eine Übersicht zu möglichen Stolpersteinen für weniger erfahrene Internetnutzer). Selbst bei technisch eher versierten und regelmäßigen Nutzern des Webs scheint die Wahrscheinlichkeit, an einer Online-Befragung teilzunehmen, mit fortschreitender technischer Expertise und der Intensität der Internetnutzung zuzunehmen (Zhang 2000; Kwak und Radler 2002; Lukawetz 2002; Carini, Hayek, Kuh et al. 2003).

Bedenken hinsichtlich der Sicherheit einer Internet-Verbindung können eine weitere Ursache für Nonresponse bei Online-Befragungen sein. Ein häufig vermuteter Vorteil von E-Mail-, Web- oder anderen selbstadministrierten Befragungen sind geringe Verzerrungen aufgrund sozialer Erwünschtheit. Aber Bedenken hinsichtlich der Sicherheit des Internet, das heißt, wenn die Befragten befürchten, sie können persönlich identifiziert werden, können diesem Vorteil entgegen wirken. Dies führt womöglich zu einer höheren Nonresponse-Rate oder zu weniger aufrichtigen Antworten auf Fragen zu sensiblen Themen (vgl. z.B. Reips und Franek 2004; Heerwegh et al. 2005; Joinson und Reips im Druck). Weiterhin bewahren einige Unternehmen Kopien der E-Mails ihrer Mitarbeiter auf, was potenzielle Befragte davon abhalten könnte, vom Arbeitsplatz aus an einer Befragung teilzunehmen.

Ein nützlicher Aspekt von Web-Befragungen für die Erforschung dieser und anderer Einflussfaktoren auf Nonresponse-Raten ist die Möglichkeit, zwischen verschiedenen Ursachen von Nonresponse zu unterscheiden. Im Unterschied zu einer postalischen Befragung ist es möglich zu bestimmen, ob die Zielpersonen das Befragungsinstrument überhaupt aufgerufen haben, oder ob eine Befragung während des Ausfüllens abgebrochen wurde, und an welchem Punkt dies geschah (Bosnjak, Tuten und Bandilla 2001; Bosnjak, Tuten und Wittman 2005). Faktoren, die die Nonresponsequote an der einen Stelle verringern (z.B. die Anzahl der Teilnehmer, die den Fragebogen gar nicht aufrufen), können die Nonresponsequote an anderer Stelle auch erhöhen (z.B. Nonresponse bei Einzelfragen, vgl. etwa die „high hurdle technique" von Reips 2002). Die Untersuchung von Möglichkeiten zur Steigerung der Rücklaufquote zieht deswegen nicht nur die Rücklaufquote insgesamt als Erfolgskriterium heran, sondern befasst sich zum Beispiel auch mit dem Drop-Out-Anteil (Anteil Befragte, die die Befragung abbrechen).

Im Unterschied zu anderen Arten von Befragungen weisen Web-Befragungen am Anfang des Antwortprozesses eine zusätzliche Hürde auf, da ein potenzieller Teilnehmer zuerst die Website aufsuchen muss, auf der die Befragung zu finden ist. Ein Großteil der Nonresponse bei Web-Befragungen tritt bereits an dieser Stelle auf, das heißt, bei den meisten Ausfällen handelt es sich um Personen, die überhaupt nicht auf das Befragungsinstrument zugegriffen haben (Lozar Manfreda et al. 2002). Motivationstechniken, die sich bei postalischen Befragungen bewährt haben, reichen eventuell nicht aus, um die Befragten dazu zu bewegen, in einem zusätzlichen Schritt den Fragebogen online aufzurufen. Außerdem haben viele dieser Techniken (z.B. schriftliche Vorankündigung, persönliche Unterschrift, professioneller Briefkopf, monetäre Anreize usw., vgl. Dillman 2000) vielleicht gar kein funktionales Äquivalent bei vollständig elektronischen Web-Befragungen.

Zum Beispiel zählen Vorankündigungen und Erinnerungsschreiben seit langem zur bewährten Praxis bei postalischen Befragungen. Ihr Effekt bei Web- und E-Mail-Befragungen scheint aber deutlich schwächer zu sein. Der vermutlich steigernde Einfluss einer Vorankündigung auf die Rücklaufquote konnte weder bei E-Mail-Befragungen (Sheehan 2001) noch bei Web-Befragungen (Lozar Manfreda et al. 2005) festgestellt werden. Es scheint auch keine deutlichen Belege für die positive Wirkung von Erinnerungsschreiben auf die Rücklaufquote zu geben (Lozar Manfreda et al. 2005), eher sogar Anzeichen dafür, dass mehr als ein Erinnerungskontakt kontraproduktiv sein könn-

te (Cook, Heath und Thompson 2000). Einzelne Studien belegen zwar einen positiven Effekt von Erinnerungskontakten auf die Rücklaufquote, deuten aber auch darauf hin, dass Mehrfachkontakte bei einer Web-Befragung weniger wirksam waren als bei einer parallel durchgeführten postalischen Befragung (Zhang 2000; Kwak und Radler 2002).

Materielle Anreize („incentives") sind ebenfalls Standard-Motivationstechniken, die bei Web-Befragungen nicht auf dieselbe Weise angewandt werden können wie bei traditionellen Befragungsarten. In einer Metaanalyse fasst Göritz (2004) die wichtigsten Probleme bezüglich der Verwendung von Anreizen bei Online-Umfragen zusammen. Zum Beispiel zeigen die Erfahrungen mit postalischen Befragungen, dass Anreize im Voraus geleistet werden müssen, um wirksam zu sein. Es ist zwar durchaus möglich, im Rahmen einer Web-Befragung Anreize zu schaffen, indem man im Voraus eine Belohnung vergibt (entweder online – mit elektronischen Gutscheinen oder einer elektronischen Überweisung – oder offline, falls die Postanschrift des Teilnehmers bekannt ist), es bleibt jedoch unklar, ob eine solche Auszahlung dieselben Auswirkungen auf die Teilnahme hat wie bei einer postalischen Befragung.

Drei allgemeine Trends ergeben sich aus den Forschungsergebnissen zu den Auswirkungen von Anreizen auf die Rücklaufquoten in Web-Befragungen: Erstens ist der Effekt – wenn überhaupt vorhanden – nicht groß, zumindest was die derzeitigen Formen von solchen Anreizen betrifft (Lozar Manfreda et al. 2005; Göritz im Druck). Zweitens spielen Anreize eine größere Rolle für die Verringerung der Drop-Out-Rate als für die Erhöhung der Teilnahmequote (Göritz im Druck). Drittens scheinen Befragte eine sichere, direkt zugängliche Bezahlung, die in zeitlicher Nähe zur Teilnahme an der Befragung erfolgt, zu bevorzugen (Birnholtz et al. 2004; Tuten, Galesic und Bosnjak 2004). Dieser Aspekt erscheint besonders wichtig, weil eine beliebte Methode, den Rücklauf von Web-Umfragen zu erhöhen, darin besteht, den Befragten die Teilnahme an einer Verlosung anzubieten. Manche Befragte ziehen aber eine sichere Entlohnung einem Los vor (z.B. O'Neil, Penrod und Bornstein 2003; Göritz eingereicht).

All diese Unklarheiten beim Einsatz von Motivationstechniken sind womöglich nur temporäre Probleme, die sich mit zunehmenden Kenntnissen über die Umsetzung von Internet-Befragungen lösen. Während sich zudem auch das Problem mangelnder Abdeckung mit der Zeit reduziert, wird wahrscheinlich vor allem das Nonresponse-Problem in den Vordergrund rücken. Darüber hinaus könnten Web-Befragungen von Anfang an mit einem Nachteil gegenüber anderen Datenerfassungsarten behaftet sein: Web-Befragungsmethoden verfügen nicht über den zeitlichen Puffer, der es zum Beispiel ermöglichte, telefonische Befragungen zu einer ausgereiften Methode zu entwickeln, noch ehe das Telefonmarketing in größerem Umfang aufgekommen ist. Im Web scheinen Direktvermarkter und Spammer der Befragungsbranche bereits weit voraus zu sein.

Zusammenfassend ist festzuhalten, dass die Bestimmung der Zielpersonen mithilfe von Zufallsstichproben auf der Basis von Listen oder ähnlichen Verfahren eine Messung der Nonresponse-Rate ermöglicht, die Entwicklung von Strategien zur Verringerung von Ausfällen fördert und zusätzliche Variablen für die Anpassungen der Umfrage-Ergebnisse unter Berücksichtigung von Nonresponse liefert. Es besteht jedoch noch erheblicher Forschungsbedarf, um das Entscheidungsverhalten der Zielpersonen, an ei-

ner Befragung teilzunehmen, besser verstehen zu können und Strategien zur Minimierung der potenziellen Verzerrungen durch Nonresponse zu entwickeln.

IV. Messfehler

Die Probleme der Stichprobenausfälle (und Unterthemen wie Auswahlverfahren, der Abdeckungsgrad und Nonresponse-Fehler) haben die Diskussion um die Qualität von Web-Befragungen in den 1990er und Anfang der 2000er Jahre beherrscht. Zunehmend gerät jedoch nun auch das Problem der Messfehler ins Blickfeld. Vereinfacht gesagt versteht man unter einem Messfehler die Abweichung der Antwort eines Befragten vom „wahren" Wert. Die Ursachen von Messfehlern bei selbstadministrierten Befragungen können entweder auf Seiten der Befragten liegen (fehlende Motivation, Verständnisprobleme, absichtliche Fälschungen der Angaben etc.) oder auf Seiten des Instruments (schlechte Wortwahl, mangelhafte Gestaltung, technische Probleme etc.). Um Messfehler zu minimieren, muss das Befragungsinstrument einfach zu verstehen und auszufüllen sein, es muss die Motivation der Befragten aufrechterhalten und den Befragten eine vertrauliche Behandlung ihrer Angaben glaubwürdig zusichern (für aktuelle Zusammenfassungen der Forschungsergebnisse in diesem Gebiet siehe Tourangeau, Rips und Rasinski 2000; Biemer et al. 2004).

Obwohl ein wachsendes Interesse an der Gestaltung von Web-Fragebögen besteht (vgl. z.B. Dillman 2000), soll auf dieses Thema in diesem Beitrag nicht ausführlich eingegangen werden. Allgemein lässt sich allerdings festhalten, dass noch offen ist, welche Design-Praktiken sich von traditionellen Befragungstechniken auf das Web übertragen lassen. Die Gestaltung könnte bei Web-Befragungen eine wesentlich wichtigere Rolle spielen als bei anderen Techniken, da einerseits mehr Design-Werkzeuge zur Verfügung stehen (Farbe, Ton, Animation etc.), sich aber andererseits das Erscheinungsbild des Fragebogens auf verschiedenen Computern unter Umständen unterscheidet. So ist es derzeit zum Beispiel möglich, dass sorgfältig entworfene Formulare gar nicht in der beabsichtigten Form übertragen werden. Im Unterschied zur Papierform kann das Erscheinungsbild eines Fragebogens zudem je nach Zielperson aufgrund unterschiedlicher Browser-Einstellungen, Variationen der Hardware etc. variieren. Dillman und Bowker (2001) weisen darauf hin, dass diese Diskrepanzen eine erhebliche Quelle von Messfehlern darstellen können.

Eine weitere Quelle von Messfehlern ist typisch für Panel- oder Längsschnittbefragungen (wie sie bei Online-Befragungen häufig vorkommen): die sog. Panel-Konditionierung (oder der „time-in-sample bias"). Panel-Konditionierung tritt durch die regelmäßige Befragung derselben Panelmitglieder auf (vgl. Kalton und Citro 1993). Aufgrund der Erfahrung, die diese Teilnehmer in früheren Befragungen gesammelt haben, können ihre Antworten zunehmend systematisch von den Antworten von Personen abweichen, die an der gleichen Befragung zum ersten Mal teilnehmen. Bedingt durch die Schwierigkeiten bei der Erstellung einer geeigneten Auswahlgesamtheit, behelfen sich viele Befragungsunternehmen mit Online-Panels (auch Access-Panels genannt). Solche Panels beruhen auf jeglicher Art von Rekrutierung, und die rekrutierten Personen oder Haushalte erklären sich bereit, an mehreren Online-Befragungen teilzunehmen. Die

Verbreitung dieser Strategie heißt, dass Panel-Effekte, die die Antworten der Befragten beeinträchtigen können, bei vielen Web-Surveys von Bedeutung sind. Schon die bloße Teilnahme an einem laufenden Panel an sich kann das Verhalten und die Einstellung der Befragten verändern, auch wenn der Befragungsinhalt im Verlauf der Zeit variiert wird. Es werden zwar bereits statistische Korrekturverfahren eingesetzt, um Panel-Effekten entgegenzuwirken, die Natur und das Ausmaß dieser Effekte müssen aber noch genauer untersucht werden.

Die bisherigen Ausführungen sollen nicht den Anschein erwecken, dass hochwertige Web-Befragungen gar nicht möglich seien. Im Gegenteil, Web-Befragungen bieten der Forschung enorme Möglichkeiten für preiswerte selbstadministrierte Befragungen unter Verwendung einer großen Auswahl an Stimulus-Materialien (Ton, Abbildungen, Video etc.), die bislang nicht verfügbar oder für den breiten Einsatz in Befragungen mit Interviewern zu teuer waren. Der Einsatz von computergestützten Methoden erlaubt die Einbeziehung von Design-Merkmalen, die sich bei Befragungen in Papierform nur schwer implementieren lassen (z.B. individuelle Anpassung von Wortwahl, Reaktionszeitmessung, Bearbeitung in Echtzeit etc.). Das Potenzial von Web-Befragungen ist also in der Tat sehr groß. Nur sind durch das Aufkommen des Web als Werkzeug zur Erfassung von Befragungsdaten die herkömmlichen Überlegungen bezüglich Repräsentativität und Replizierbarkeit von Forschungsergebnissen nicht obsolet geworden.

V. Arten von Online-Befragungen

Nach den Ausführungen zu den möglichen Fehlerquellen in Online-Umfragen sollen nun die wichtigsten Arten von heutzutage gebräuchlichen Web-Befragungen diskutiert werden. Es gibt viele Varianten von Web-Befragungen, die sich aber wie folgt typisieren (zusammengefasst in *Tabelle 1*) und in Bezug auf ihren beabsichtigten Zweck sowie im Kontext der unterschiedlichen Fehlerquellen beurteilen lassen.

Tabelle 1: Arten von Online-Befragungen

nicht-zufallsgesteuerte Ansätze	*zufallsgesteuerte Ansätze*
1. Web-Befragungen zu Unterhaltungszwecken	4. Intercept-Befragungen
2. Befragungen mit uneingeschränkt selbstrekrutierter Teilnehmerschaft	5. Befragungen mit listen-basierten Stichproben aus Populationen mit einem hohem Abdeckungsgrad
3. Befragungen mit selbstrekrutierten Freiwilligen-Panels	6. Web-Option bei Mixed-Mode-Befragungen
	7. Befragungen mit einem vorrekrutierten Panel von Internetnutzern
	8. Befragungen mit einem vorrekrutierten Panel aus der Gesamtbevölkerung

Wir unterscheiden in erster Linie zwischen nicht-zufallsgesteuerten Befragungen und zufallsgesteuerten Befragungen und identifizieren jeweils eine Reihe von Unterkategorien.[4] Bei nicht-zufallsgesteuerten Befragungen haben Mitglieder der Zielpopulation keine bekannte, von Null verschiedene Auswahlwahrscheinlichkeit. Schlussfolgerungen oder Verallgemeinerungen in Bezug auf die Population, die auf Grundlage von nicht-zufallsgesteuerten Befragungen getroffen werden, basieren daher eher auf Spekulation als auf etablierten statistischen Prinzipien.

1. Nicht-zufallsgesteuerte Ansätze

Typ 1: Web-Befragungen zu Unterhaltungszwecken. Bei dieser Art von Web-Befragungen handelt es sich nicht um Umfragen im wissenschaftlichen Sinne, da sie aber weit verbreitet sind und möglicherweise von einigen Personen mit „echten" Befragungen verwechselt werden, bedürfen sie hier einer kurzen Erwähnung. Viele Websites haben sich darauf spezialisiert, Umfragen durchzuführen und die Antworten der Befragten zu veröffentlichen (z.B. die „Frage des Tages"-Umfragen, die auf vielen Medien-Websites verbreitet sind). Diese Befragungen dienen primär dem Zeitvertreib der Befragten, und es gibt normalerweise keine Kontrolle darüber, wer daran teilnimmt. Der Unterhaltungscharakter solcher Befragungen wird in der Regel ausdrücklich bekannt gegeben und es wird nicht der Anschein erweckt, es handle sich um wissenschaftliche Umfragen mit repräsentativen Stichproben. Das primäre Ziel dieser Websites ist stattdessen, als Forum zum Austausch von Meinungen zu dienen. Sie stellen keine wirkliche „Bedrohung" für legitime (wissenschaftliche) Befragungen dar, solange die erhobenen Daten nicht missbraucht werden.

Typ 2: Befragungen mit uneingeschränkt selbstrekrutierter Teilnehmerschaft. Dieser Ansatz verwendet zur Rekrutierung der Befragungsteilnehmer offene Einladungen auf Portalen, häufig besuchten Websites, oder (in einigen Fällen) speziellen „Befragungs"-Websites. Es handelt sich um die derzeit wahrscheinlich häufigste Form von Web-Befragungen und um eine der potenziell schädlichsten für legitime Befragungsprojekte, weil hier gelegentlich der Anschein wissenschaftlicher Gültigkeit erweckt wird. Normalerweise gibt es bei diesen Umfragen keinerlei Beschränkungen bezüglich der Teilnahme: Jeder kann, und in der Regel beliebig oft, teilnehmen. Selbstrekrutierte Web-Befragungen beruhen auf äußerst selektiven Stichproben. So gibt es Hinweise, dass nur ein sehr geringer Anteil der Personen, die den Befragungsaufruf sehen, tatsächlich zur Teilnahme bewegt werden kann (in der Regel 0,5 Prozent oder weniger, vgl. Tuten, Bosnjak und Bandilla 2000; Alvarez, Sherman und VanBeselaere 2003; Ross et al. 2005). Der Unterschied zwischen den beiden ersten Befragungstypen besteht hauptsächlich darin, dass bei Typ 1 normalerweise kein Anspruch auf Allgemeingültigkeit erhoben wird, bei Typ 2 aber sehr wohl. Ein Beispiel für eine Befragung von Typ 2 liefern Ross et al. (2005) mit einer Studie über das Sexualverhalten in Schweden.

4 Siehe auch die Typologien von Schnell, Hill und Esser (2005: 378) und Schonlau, Fricker und Elliot (2002: Kap. 4).

Um die Auswirkung verschiedener Fehlerquellen auf die Ergebnisse von diesen Surveys zu untersuchen, vergleichen einige Studien selbstrekrutierte Online-Befragte mit den Befragten aus einer Zufallsstichprobe bezüglich ihrer soziodemografischen Zusammensetzung und ihrer Antworten. Die Resultate zeigen, dass sich die selbstrekrutierten Teilnehmer von der Allgemeinbevölkerung unterscheiden. Eine deutsche Studie, beschrieben in Faas (2004) und Schoen (2004), verwendete eine Vergleichstichprobe mit face-to-face-Befragten und einer Teilnahmequote von 68 Prozent. Die selbstrekrutierten Online-Befragten waren gegenüber den Personen in der Vergleichsstichprobe überproportional häufig männlich, gut ausgebildet und auffallend jung. Darüber hinaus konnten Abweichungen auch in der Ausprägung und Stärke der politischen Einstellungen nachgewiesen werden. Beispielsweise äußerten die selbstrekrutierten Online-Befragten andere Wahlabsichten und waren fast fünf Mal so häufig Mitglied einer politischen Partei. Weitere amerikanische Studien berichten ähnliche Abweichungen bei Umfragen zu politischen Themen (vgl. Best et al. 2001; Alvarez et al. 2003). Die Studien liefern jedoch widersprüchliche Ergebnisse dazu, ob der Befragungsmodus die vermuteten Zusammenhänge zwischen den Untersuchungsvariablen beeinflusst.

Zusammenfassend lässt sich sagen, dass sich die Teilnehmer selbstrekrutierter Befragungen auf relevante, jedoch nicht vorhersagbare Weise von der Allgemeinbevölkerung unterscheiden. Die soziodemografischen Eigenschaften solcher Befragten weichen üblicherweise von denjenigen der Allgemeinbevölkerung ab, wobei die Richtung der Unterschiede (z.B. ob Männer oder Frauen überrepräsentiert sind) stark vom Thema der Befragung abzuhängen scheint. Befragte aus selbstrekrutierten Online-Stichproben äußern auch tendenziell ausgeprägtere Meinungen als die Mitglieder der Allgemeinbevölkerung. Manche Fachleute behaupten, man könne dem Einfluss dieser Selektionseffekte auf die Schätzungen von Populationsparametern für die Untersuchungsvariablen entgegenwirken, indem man die Verteilung der demographischen Standardvariablen in der Stichprobe mithilfe von Gewichten den offiziellen Statistiken angleicht. Diese Behauptung basiert auf der Annahme, dass die Übereinstimmung zweier „Stichproben" in Bezug auf eine Reihe von demographischen Charakteristika sicherstellt, dass diese auch in Bezug auf die Untersuchungsvariablen übereinstimmen. Dies ist jedoch in vielen Fällen eine zweifelhafte Annahme (siehe Schnell 1993).

Es gibt viele Befragungen dieses Typs, häufig durchgeführt von Organisationen mit anerkanntem wissenschaftlichen Ruf, die Behauptungen in Bezug auf Gültigkeit und Repräsentativität aufstellen, welche weit über das hinausgehen, was die Daten hergeben. Wir wollen damit nicht behaupten, dass solche Befragungen Zeitverschwendung wären, aber große Vorsicht ist angebracht, wenn man auf Basis so gewonnener Daten Schlussfolgerungen ziehen möchte.

Typ 3: Befragungen mit selbstrekrutierten Freiwilligen-Panels. Bei diesem Ansatz wird ein Panel aus freiwilligen Teilnehmern zusammengestellt, die über Aufrufe auf gut besuchten Websites und Internet-Portalen rekrutiert werden. Grundlegende soziodemographische und Kontaktinformationen werden von den Freiwilligen zum Zeitpunkt der Registrierung hinterlegt, wodurch eine umfangreiche Datenbank potenzieller Teilnehmer für künftige Befragungen entsteht. Die Teilnahme an späteren Befragungen erfolgt ausschließlich auf Einladung und wird über IDs und Passwörter kontrolliert (dies im Gegensatz zu den Befragungstypen 1 und 2, bei denen die Teilnahme nicht eingeschränkt

ist). Wiederholtes Ausfüllen oder die Weitergabe des Fragebogens an andere wird damit verhindert. Web-Befragungen auf Grundlage von Freiwilligen-Panels haben in den letzten Jahren in den Medien und in der Befragungsbranche am meisten Aufmerksamkeit erregt (Göritz 2003; Gräf 2003; Göritz 2004). Befürworter dieser Art von Befragung weisen auf die umfassenden Informationen hin, die über die Panelmitglieder vorhanden sind, sowie auf die Möglichkeit, die Antworten einzelner Personen zwischen verschiedenen Befragungen zu vergleichen und zu validieren. Außerdem müssen, so die Argumentation, weniger Fragen gestellt werden, weil demografische Informationen über die Befragten bereits zur Verfügung stehen.

Die Auswahl der Panelmitglieder zur Teilnahme an den einzelnen Befragungen erfolgt normalerweise per Quoten- oder Zufallsstichprobe. Zwar mag es somit eine gute Kontrolle geben über die Auswahl der Befragten für die einzelnen Stichproben, es mögen viele demographische Merkmale zur Verfügung stehen, an denen sich die Auswahl orientieren kann, und es mag ein großer Pool an potenziellen Teilnehmern zugänglich sein. Dies ändert aber nichts an der grundlegenden Eigenschaft, dass, wie bei den sonstigen nicht-zufallsgesteuerten Ansätzen, das anfängliche Panel aus einer selbstrekrutierten Stichprobe von Freiwilligen besteht. Erwartungsgemäß gibt es bei der Zusammensetzung solcher Panels in Bezug auf soziodemografische Charakteristiken wie Geschlecht, Alter, Ausbildung, Zivilstand usw. erhebliche Abweichungen gegenüber der Allgemeinbevölkerung (Alvarez et al. 2003; Berrens et al. 2003; Hellwig, von Heesen und Bouwmeester 2003; Göritz 2004).

Einiges deutet darauf hin, dass Freiwilligen-Panels das am schnellsten wachsende Segment der Web-Befragungs-Branche sind; Hunderte solcher Panels gibt es bereits in vielen Ländern der Welt. Das bekannteste deutsche Beispiel für diesen Ansatz dürfte „Perspektive Deutschland" sein (www.perspektive-deutschland.de). Die Mitglieder dieses Panel werden einmal pro Jahr über gesellschaftspolitische Themen befragt. Der Website von Perspektive Deutschland zufolge nahmen über 500.000 Personen an der letzten alljährlichen Befragung teil. Das für die USA wohl bekannteste Beispiel ist „Harris Poll Online" (www.harrispollinteractive.com). Auf der Website wird angegeben, dass das Panel mehrere Millionen Mitglieder aus über 200 Ländern umfasst.

Ein wichtiges Merkmal des Ansatzes von Perspektive Deutschland oder Harris Interactive ist die Verwendung der Propensity-Gewichtung oder der Propensity-Score-Anpassung, die Verzerrungen bei der Zusammensetzung von Online-Stichproben ausgleichen soll (Taylor 2000; Schonlau et al. 2003). Die Methode geht auf die Arbeiten von Rosenbaum und Rubin (1983) zurück und wurde ursprünglich entwickelt, um Selektionsverzerrungen bei der statistischen Analyse von Daten aus Beobachtungsstudien auszugleichen (z.B. zum Zusammenhang zwischen Rauchen und Lungenkrebs); vgl. zu der Methode auch Gangl und DiPrete (in diesem Band). Die Methode war jedoch nicht dafür gedacht, in deskriptiven Untersuchungen Verallgemeinerungen auf die Grundgesamtheit zu ermöglichen. Zur Anwendung der Propensity-Gewichtung wird die „Propensity" (Neigung), im Online-Sample zu sein, mithilfe von parallel durchgeführten face-to-face-Interviews (wie bei Perspektive Deutschland) oder Telefonbefragungen (wie bei Harris International) als Funktion von in beiden Erhebungen gemessenen Kovariaten (in der Regel sowohl demografische als auch andere) geschätzt. Die Qualität dieses Ansatzes hängt von der Wahl der Variablen ab, die für die Anpassung

verwendet werden, und von der Güte der Vergleichsmessungen in der Telefon- oder face-to-face-Befragung. Die Details der Variablenwahl werden leider meistens nicht bekannt gegeben, obwohl bei modellbasierten Anpassungsverfahren, so wie sie Perspektive Deutschland oder Harris Interactive verwendet, eine korrekte Spezifikation des Modells (das heißt, die Wahl der angemessenen Variablen für die Anpassung) entscheidend ist. Korrekt angewendete designbasierte Verfahren (Zufallsstichproben) verringern das von solchen Modell-Fehlspezifikationen ausgehende Risiko.[5]

Einer der wenigen veröffentlichten Tests der Güte dieses Gewichtungsverfahrens bei Online-Befragungen findet sich in der Studie von Berrens et al. (2003). Die Autoren verglichen eine Stichprobe aus dem Harris Interactive-Panel (im Folgenden HI-Daten genannt) und die Daten einer gleichzeitig durchgeführten RDD-Befragung mit Populationswerten, die auf dem amerikanischen „Current Population Survey" (CPS) basieren. Die demographischen Eigenschaften der Rohdaten des Panels wichen zwar von den Bevölkerungsdurchschnittswerten in vielen Fällen deutlich ab. Dennoch lagen aber die Propensity- gewichteten HI-Daten bezüglich der demographischen Merkmale dichter an den Populationswerten als die gewichteten RDD-Daten. Die RDD- und die HI-Daten wiesen bei vielen (aber nicht bei allen) politischen Variablen ähnliche Durchschnittswerte auf, die aber zum Teil wiederum vom Bevölkerungsmittel abwichen. Weiterhin fanden die Autoren keine Auswirkungen des Befragungsmodus auf Modelle des Wahlverhaltens und Modelle zur Erklärung von Umwelteinstellungen, sofern demographische Variablen berücksichtigt und die Panel-Daten einer Propensity-Gewichtung unterzogen wurden. Weitere gründliche und unvoreingenommene empirische Evaluierungen dieser Art sind erforderlich, um die Gültigkeit der Propensity-Methode zu belegen.

Im Unterschied zu Perspektive Deutschland oder Harris Interactive scheinen viele Betreiber von Web-Panels keine parallelen Befragungen durchzuführen, um die Ergebnisse aus ihren Online-Umfragen zu evaluieren oder anzupassen. Darüber hinaus variieren auch bei einem Pool von kooperativen Teilnehmern die Rücklaufquoten erheblich: Es wurden Rücklaufquoten von 5 bis 80 Prozent berichtet (Schmidt 2000; Terhanian 2000; Berrens et al. 2003; Göritz 2004).

Wie bereits erwähnt wurde, ist nicht der Umstand Besorgnis erregend, dass ein sehr umfangreiches Panel von Freiwilligen zur systematischen Erfassung von Informationen zu einer Vielzahl von Themen verwendet wird. Problematisch ist, dass die Befürworter behaupten, die Freiwilligen-Panels seien zufallsbasierten Formen der Erfassung von Befragungsdaten ebenbürtig oder gar überlegen (besonders im Vergleich zu RDD-Stichproben bei telefonischen Befragungen). Die Behauptungen gehen so weit, dass die Ergebnisse aus den Web-Panels nicht nur für die Internet-Population repräsentativ seien, sondern vielmehr für die allgemeine Bevölkerung. Ausgegangen wird dabei von der Annahme, dass die Gewichtungsmethoden ihre erhoffte Wirkung zeigen. Wie schon angedeutet, lässt der heutige Stand der Forschung die Frage, ob diese Annahme tatsächlich zutrifft, noch offen.

5 Siehe die Diskussion bei Groves (1989: Kap. 6) zu modell- gegenüber designbasierten Verfahren.

2. Zufallsgesteuerte Ansätze

Im Unterschied zu den zuvor beschriebenen Arten von Web-Befragungen basieren die folgenden Ansätze in irgendeiner Form auf Zufallsstichproben. Dies garantiert zwar noch keine Repräsentativität, da Nonresponse die Zulässigkeit von Schlussfolgerungen aus diesen Befragungen beeinträchtigen kann, aber im Unterschied zu nicht-zufallsgesteuerten Designs erlauben es diese Ansätze aufgrund der Kenntnis der Auswahlgesamtheit sowie der Informationen über das Rekrutierungsverfahren, die Ursachen der Nonresponse zu quantifizieren. Die so gewonnenen Informationen können im Gegensatz zur reinen Poststratifikation zu einer Verbesserung der Ergebnisse designbasierter Anpassungsmethoden beitragen.

Angesichts der Tatsachen, dass der Zugang zum Internet nicht universell ist und keine Auswahlgesamtheit der Internetnutzer existiert, gibt es grundsätzlich zwei Möglichkeiten, zufallsgesteuerte Stichproben von Internetnutzern zu ziehen. Eine erste Möglichkeit besteht darin, die Zielpopulation auf Befragte mit Internetzugang zu begrenzen, wodurch aber auch die Verallgemeinerbarkeit der Ergebnisse eingeschränkt wird. Die zweite Möglichkeit liegt in der Verwendung alternativer Kanäle (z.B. einer RDD-Telefonbefragung), um eine die Population in angemessener Weise widerspiegelnde Stichprobe zu identifizieren. Nachfolgend werden die unterschiedlichen Typen von zufallsgesteuerten Ansätzen genauer besprochen.

Typ 4: Intercept-Befragungen. Die Intercept-basierten Verfahren richten sich an die Besucher einer Website. Ähnlich wie bei Wählerbefragung am Ausgang des Wahllokals („exit polls"), wird beim Intercept-Verfahren ein systematisches Auswahlverfahren verwendet, um jeden n-ten Besucher einer Website zur Teilnahme an der Befragung einzuladen. Zur Vermeidung von Mehrfacheinladungen an dieselbe Person werden zum Beispiel Cookies verwendet. Die Auswahlgesamtheit beschränkt sich per Definition auf die Besucher der Website, was eine vollständige Abdeckung gewährleistet, andererseits sind aber auch keine Rückschlüsse auf eine weiter reichende Grundgesamtheit möglich. Das Verfahren ist nützlich für Befragungen zur Kundenzufriedenheit, Website-Bewertung u.Ä.

Die Teilnahmequoten bei Intercept-Studien sind eher niedrig. Zum Beispiel berichtet Theobald (2003) die folgenden Rücklaufquoten für drei ausgewählte Studien, die in Deutschland von der Rogator AG durchgeführt wurden: 4, 17, bzw. 19 Prozent der Personen, die eine Einladung erhielten, haben die Startseite der Befragung aufgerufen, und von diesen nahmen schließlich 54, 70, bzw. 56 Prozent an der Befragung teil. Die relativ niedrigen Rücklaufquoten bei Intercept-Befragungen deuten auf ein Nonresponse-Problem hin: Es ist nicht auszuschließen, dass die Teilnehmer einer solchen Befragung zum Beispiel andere Ansichten über die zu beurteilende Website haben als die Personen, die sich gegen eine Teilnahme entschieden haben.

Typ 5: Befragungen mit listen-basierten Stichproben aus Populationen mit einem hohem Abdeckungsgrad. Internetzugang bzw. Internetnutzung sind zwar noch lange nicht flächendeckend, es gibt aber durchaus Teilmengen aus der Gesamtbevölkerung, die einen sehr hohen oder gar vollständigen Abdeckungsgrad aufweisen. Auch wenn die mit solchen Gruppen gewonnenen Ergebnisse nur begrenzt verallgemeinerbar sind, so gibt es

doch eine ganze Menge von Teilpopulationen, für die sich solche Befragungen optimal eignen. Mit weiterer Verbreitung des Internets dürften die diesbezüglichen Möglichkeiten außerdem noch zunehmen.

Der Ausgangpunkt für Web-Befragung in Populationen mit hohem Abdeckungsgrad ist eine Auswahlgesamtheit oder Liste von Zielpersonen, die einen Internetzugang haben. Es werden dann zum Beispiel E-Mail-Anfragen an diese Personen verschickt, an einer Befragung teilzunehmen. Der Zugang zur Befragung wird üblicherweise kontrolliert, um zu verhindern, dass der Fragebogen mehrmals ausgefüllt oder an Dritte weitergegeben wird. Solche eingeschränkten Populationen weisen, wie bereits erwähnt, qua definitionem keine oder nur geringe abdeckungsbezogene Probleme auf. Befragungen von Studierenden, Mitarbeitern eines Betriebs oder Mitgliedern von Berufsorganisationen sind gute Beispiele für diese Methode, die in wachsendem Maße angewandt wird.

Zwar spielen abdeckungsbezogene Probleme bei dieser Art von Befragung vielleicht eine geringere Rolle, trotzdem kann die Gültigkeit der Ergebnisse u.U. durch Nonresponse beeinträchtigt werden (auch wenn im Vergleich zu Intercept-basierten Methoden deutlich höhere Teilnahmequoten zu verzeichnen sind). Von hohen Rücklaufquoten (ca. 50 bis 80 Prozent) bei Web-Befragungen unter Studierenden berichten Forsman und Varedian (2002) in Schweden, Heerwegh et al. (2005) in Belgien, McCabe et al. (im Druck) in den USA und Kypri et al. (2004) in Neuseeland. Aber eine Reihe anderer Studentenbefragungen weist auch deutlich niedrigere Rücklaufquoten auf (z.B. Kwak und Radler 2002; Kaplowitz et al. 2004; James, Chen und Sheu 2005). Im deutschen Sprachraum wurden bei einigen Studien mit Stichproben aus nicht-studentischen Populationen mit hoher Abdeckung ebenfalls Rücklaufquoten unter 50 Prozent beobachtet (Batinic 2001; Reips und Franek 2004).

Ergebnisse aus Befragungen von Studierenden deuten darauf hin, dass sich selbst bei Zielpopulationen, die einen hohen Abdeckungsgrad aufweisen und aus jungen und vermutlich technologisch versierten Personen bestehen, die Nonrespondenten von den Teilnehmenden systematisch unterscheiden. So sind bei Online-Befragungen im Vergleich zu parallel durchgeführten postalischen Befragungen Männer häufig übervertreten, genauso wie jüngere Studierende oder Mitarbeitende (z.B. Forsman und Varedian 2002; Kwak und Radler 2002; Reips und Franek 2004). Zusammenfassend lässt sich sagen, dass zwar die Abdeckung bei dieser Art von Befragungen kein wesentliches Problem darstellt und die Population, auf die sich die Schlussfolgerungen beziehen, bekannt ist, das Ausmaß an Nonresponse und die Unterschiede zwischen Nonrespondenten und Befragten möglicherweise aber die Verallgemeinerbarkeit der Resultate selbst auf diese begrenzten Populationen gefährden können.

Typ 6: Mixed-Mode-Befragungen mit Wahl der Teilnahmemethode. Bei diesem Verfahren wird den Befragten das Web als ein Antwortmodus unter anderen angeboten. Das Verfahren wird zum Beispiel bei Befragungen von Betrieben oder sonstigen Organisationen angewandt, bei denen über einen längeren Zeitraum wiederholt mit den Teilnehmern Kontakt aufgenommen wird. So können etwa Firmen in Deutschland und Österreich Informationen für die amtliche Statistik über ihren Warenverkehr innerhalb der EU wahlweise online oder offline abgeben. Es kommt hier vor allem darauf an, den Aufwand und die Kosten für die Befragten zu minimieren. Mixed-Mode-Befragungen

beruhen dabei allerdings auf der Annahme, dass durch die Verwendung verschiedener Befragungsmodi keine substanziellen zusätzlichen Messfehler entstehen.

Es gibt auch Beispiele für die Anwendung von Mixed-Mode-Verfahren bei allgemeinen Bevölkerungsumfragen. Der Ausgangspunkt für solche Studien ist eine Zufallsstichprobe oder sogar ein Zensus der Grundgesamtheit. Der Swiss Census 2000 bot nahezu allen Befragten die Option an, die Frageformulare online auszufüllen. Von dieser Möglichkeit machten aber nur 4 Prozent der Befragten Gebrauch, und das zu einer Zeit, in der fast 26 Prozent der Schweizer Bevölkerung angaben, das Internet regelmäßig zu benutzen (Schweizerisches Bundesamt für Statistik 2005b). Für die Teilnahme per Post trotz eines vorhandenen Internetanschlusses wurden unterschiedliche Gründe angegeben. 63 Prozent meinten, das Ausfüllen per Hand sei einfacher, 12 Prozent erklärten, dass sie über zu wenig Erfahrung mit dem Internet verfügten, 4 Prozent hatten Bedenken wegen des Datenschutzes, 2 Prozent hielten die Verbindungskosten für zu hoch, und 20 Prozent gaben weitere, unterschiedliche Gründe an. Für zusätzliche Beispiele von Mixed-Mode-Befragungen siehe Ballard und Prine (2002) oder Schonlau, Asch und Du (2003). Fricker und Schonlau (2002) geben einen kurzen Überblick über solche Studien.

Der geringe Anteil an Teilnehmern, die in Mixed-Mode-Befragungen von der Web-Option Gebrauch machen, deutet darauf hin, dass Nonresponse bei Web-Befragungen unter Umständen ein größeres Problem darstellen könnte als bei postalischen Umfragen. Zudem dürfte der Mixed-Mode-Ansatz außer für sehr große Strichproben kaum Kosteneinsparungen gegenüber einer rein postalischen Befragung bieten. Zwei parallele Erfassungssysteme sind erforderlich, um den postalischen wie auch den elektronischen Rücklauf zu verarbeiten, und die Fixkosten der elektronischen Variante können, solange der Anteil Online-Teilnehmer weiterhin klein bleibt, nur über verhältnismäßig wenige Fälle verteilt werden. Eine Web-Option scheint auch die Gesamtrücklaufquote nicht zu steigern (Fricker und Schonlau 2002), und es stellt sich die Frage, inwieweit die Messungen über die zwei verschiedenen Medien überhaupt vergleichbar sind. Trotzdem könnten Mixed-Mode-Verfahren bei wiederholten (Panel-)Befragungen und für Befragungen von Unternehmen von zunehmender Bedeutung sein, insbesondere auch, weil sich das Internet im kommerziellen Sektor weiter ausbreitet.

Typ 7: Befragungen mit einem vorrekrutierten Panel von Internetnutzern. Dieses Verfahren ähnelt dem nicht-zufallsgesteuerten Verfahren, bei dem ein Online-Panel erstellt wird (Typ 3). Der Hauptunterschied besteht darin, dass sich bei der nicht-zufallsgesteuerten Methode das Panel aus selbstrekrutierten Freiwilligen zusammensetzt, während hier die anvisierten Panelmitglieder mithilfe eines zufallsgesteuerten Auswahlverfahrens (z.B. einer RDD-Telefonbefragung) bestimmt werden. Dies stellt den Versuch dar, eine Zufallsstichprobe aus Internetnutzern bzw. Personen mit Zugang zum Web zu erstellen. In den Telefoninterviews werden Hintergrundinformationen gesammelt, Personen mit Internetzugang identifiziert und passende Befragte für das Internet-Panel rekrutiert. Nach der Zustimmung zur Aufnahme in das Panel erhalten die Mitglieder per E-Mail eine Aufforderung zur Teilnahme an der Web-Befragung. Die Teilnahme wird mithilfe von Kennzahlen, die in URLs eingebettet werden, persönlichen ID-Nummern und/oder Passwörtern kontrolliert.

So lange die Population, über die eine Aussage getroffen werden soll, den aktuellen Internetnutzern entspricht, ist mangelnde Abdeckung bei diesem Untersuchungstyp kaum von Bedeutung. Nonresponse dürfte auch hier das größere Problem sein, wobei die Nonresponse-Rate allerdings im Gegensatz zur Situation bei den Freiwilligen-Panels messbar ist. Nonresponse kann auf vielen Stufen des Verfahrens auftreten. Wenn eine potentielle Befragungsperson bereits die Teilnahme an der Telefonbefragung verweigert, stehen nur wenige Informationen über ihre Eignung für die Stichprobe, ihre soziodemographischen Merkmale und die Gründe für die Verweigerung zur Verfügung. Weitere Ausfälle können während des Telefoninterviews auftreten, wenn Befragte zum Beispiel (fälschlicherweise) behaupten, sie hätten keinen Internetzugang oder keine gültige E-Mail-Adresse. Schließlich reagiert selbst von den Personen, die über einen Internetzugang verfügen und sich zur Teilnahme an der Web-Befragung bereit erklärten, ein großer Teil nicht auf die Anfragen per E-Mail und fällt somit aus. Allerdings können zumindest für die letzte Gruppe während der Telefonrekrutierung Informationen ermittelt werden, mit deren Hilfe sich Verzerrungen aufgrund von Nonresponse untersuchen lassen. Dabei werden diejenigen, die an der Web-Befragung teilgenommen haben, mit den Nonrespondenten bezüglich demografischer und anderer aus der telefonischen Befragung gewonnener Variablen verglichen. Weiterhin ermöglicht dieser Befragungstyp auch die Messung von Modus-Effekten (z.B. Telefon versus Internet).

Empirische Ergebnisse deuten auf niedrige Gesamtrücklaufquoten bei Befragungen dieser Art hin. Man sieht sich erstens mit dem Problem konfrontiert, dass sich relativ wenige RDD-Befragte mit Internetanschluss für ein Panel rekrutieren lassen. Hellwig et al. (2003) berichten, dass sich nur 4,6 Prozent der Internetnutzer, die per RDD-Telefonbefragung interviewt wurden (2,1 Prozent der gesamten RDD-Stichprobe), zur Teilnahme am Panel bereit erklärten. Bei Weßels und Zimmermann (2001) entsprach der Anteil ungefähr 8 Prozent, bei Sparrow und Curtice (2004) 19 und bei Bandilla, Bosnjak und Altdorfer (2003) 39 Prozent. Die Bereitschaft, Mitglied eines Panels zu werden, bedeutet allerdings noch lange nicht, dass dann auch an den Befragungen teilgenommen wird. Beispielsweise füllten bei Sparrow und Curtice nur 25 Prozent der Panel-Mitglieder tatsächlich den Fragebogen aus, so dass letztlich nur ungefähr 5 Prozent der telefonisch kontaktierten Internetnutzer an der Web-Befragung teilnahmen.

Die Panel-Teilnehmer unterscheiden sich in der Regel bezüglich soziodemografischer Variablen von den Internetnutzern der Zufallsstichproben, aus denen sie rekrutiert werden (z.B. Hellwig et al. 2003; Sparrow und Curtice 2004). Das Panel kann sich aber auch in Bezug auf die Untersuchungsvariablen von den Internetnutzern aus einer Zufallsstichprobe unterscheiden und diese Abweichungen lassen sich auch nicht immer durch eine Gewichtung der Daten beseitigen (Sparrow und Curtice 2004). Des weiteren betonen Sparrow und Curtice (2004: 40), dass „it appears that [those differences] are more the product of the distinctive characteristics of those who are willing to join an internet panel and respond to internet surveys than they are of the social skew of those with access to the internet. This suggests that continued growth in internet access alone will not eliminate the difficulties assessed here ...“

Weiterhin bleiben viele Fragen in Zusammenhang mit Nonresponse im Telefon- und Internet-Modus offen. Aber der Ausgangspunkt des Verfahrens ist zumindest theoretisch eine Zufallsstichprobe aus der gesamten (per Telefon erreichbaren) Internet-Po-

pulation, und wenn man annimmt, dass es keine Verzerrung aufgrund von Nonresponse gibt, erlaubt dies Schlussfolgerungen auf die Population der Internetnutzer.

Typ 8: Befragungen mit einem vorrekrutierten Panel aus der Gesamtbevölkerung. Die letzte Art von Web-Befragungen ist die einzige Methode, bei der *potenziell* eine Zufallsstichprobe aus der Allgemeinbevölkerung ohne Einschränkung auf Personen mit Internetanschluss im Web-Modus befragt werden kann.[6] Gleich wie bei Typ 7 ist der Ausgangspunkt auch hier eine Zufallsstichprobe aus der Zielpopulation, und die Mitglieder der Stichprobe werden anfänglich mit Offline-Methoden (z.B. face-to-face-Befragung oder RDD-Telefonbefragung) zu einer Kooperation zu bewegen versucht. Während Befragungen von Typ 7 jedoch nur Personen mit Internetanschluss berücksichtigen, wird bei diesem Verfahren den Teilnehmern ohne Anschluss die erforderliche technische Ausstattung zur Verfügung gestellt. Es ist somit der einzige Ansatz, bei dem Verallgemeinerungen auf eine über die Gruppe der Internetnutzer hinausgehende Population möglich sind. Wegen den hohen Rekrutierungskosten werden die Stichproben bei diesem Ansatz üblicherweise als Panel konzipiert.

Seinen Ursprung hat der Ansatz in den europäischen Versuchen aus den 1980er Jahren, Befragungen mithilfe von Videotext, Minitel und anderen fernsehbasierten Textinformationssystemen durchzuführen. Das Dutch Telepanel zum Beispiel wurde 1986 ins Leben gerufen und sah die Installation eines einfachen Computers in den Haushalten ausgewählter Befragter vor, wobei die Bedingung gestellt wurde, dass die Haushaltsmitglieder an den regelmäßigen (wöchentlichen) Befragungen teilnehmen würden (Saris 1998). Ein ähnlicher Ansatz wird von A.C. Nielsen, Arbitron und anderen Firmen verfolgt, die in Haushalten Geräte zur Messung des Medienverhaltens installieren.

Die holländische Organisation CentERdata (Universität Tilburg) und das US-Unternehmen Knowledge Networks haben die Vorgehensweise auf Web-Befragungen übertragen. Die Panelmitglieder werden dabei mithilfe von RDD-Telefonbefragungen rekrutiert und dann als Gegenleistung für die Teilnahme am Panel mit einem Web-TV-Gerät und einem kostenlosen Internetzugang ausgerüstet. Dieser Ansatz löst potenziell zwei der größten Probleme von Web-Befragungen: 1. mangelnde Abdeckung und 2. Software- oder Hardware-Kompatibilitätsprobleme. Das erste Problem wird gelöst, indem auch Personen ohne Computer oder Internetanschluss in die Stichprobe aufgenommen werden. Kompatibilitätsprobleme können umgangen werden, wenn alle Panelmitglieder ein identisches Gerät erhalten. Es wird so ermöglicht, die Befragungsinstrumente (einschließlich audiovisuellen Materials) für alle Panelteilnehmer auf konsistente und zuverlässige Weise bereitzustellen.[7]

Nonresponse stellt aber natürlich auch bei diesem Ansatz ein Problem dar, denn Nonresponse kann auf vielen Stufen des Verfahrens auftreten. Einige potenzielle Panel-Teilnehmer fallen aus, weil sie nicht telefonisch kontaktiert werden können oder weil

6 Befragungen des Typs 6 können ebenfalls mit einer Zufallsstichprobe der Allgemeinbevölkerung durchgeführt werden. Allerdings werden nicht alle Teilnehmer im Web-Modus befragt.

7 In der Regel wird jedoch nur den Panelmitgliedern ohne vorhandenen Internetanschluss die Standardausrüstung angeboten, was dazu führt, dass die Kompatibilitätsprobleme eventuell nicht vollständig gelöst werden.

sie die Teilnahme an der Telefonbefragung verweigern. Bei den telefonisch Befragten liegt die Erfolgsrate für die Rekrutierung bei rund 50 Prozent (Berrens et al. 2003; Smith 2003). Gemäß Berrens et al. können bei circa 70 Prozent der Haushalte, die sich zur Teilnahme bereit erklärt haben, die WebTV-Geräte dann tatsächlich installiert werden, und ungefähr 80 Prozent dieser ausgerüsteten Haushalte liefern die erforderlichen Informationen für das Mitgliederprofil. Nur ungefähr 30 (laut Berrens et al.) bis 50 (laut Smith) Prozent der ursprünglich kontaktierten Haushalte werden also schließlich vollwertige Panelmitglieder. Von diesen Panelmitgliedern nehmen 70 bis 80 Prozent an den Befragungen teil. Als Vorteil zu vermerken ist jedoch, dass auf fast jeder Stufe Informationen über die Ausfälle erfasst werden können, was eine detaillierte Untersuchung möglicher Abweichungen aufgrund von Nonresponse und Panel-Auszehrung erlaubt. Im Unterschied zur Methode mit den Freiwilligen-Panels (Typ 3) sind hier die Auswahlwahrscheinlichkeiten auf jeder Stufe bekannt, ebenso die Zielpopulation, was sowohl die Anwendung von Standard-Gewichtungen als auch von Poststratifikation erlaubt (Krotki 2000).

Wie bei den anderen Ansätzen, die auf Panels beruhen, kann die Panel-Konditionierung die Qualität der Ergebnisse beeinträchtigen. Eine von Knowledge Networks vorgeschlagene Strategie, um dem Problem der Panel-Konditionierung zu begegnen, ist die Verwendung von rotierenden Panel-Stichproben. Bei einem solchen Design kann die Stärke des Panel-Konditionierungseffekts, also der Auswirkungen einer wiederholten Teilnahme an den Befragungen, gemessen und – falls erforderlich – mit statistischen Methoden kontrolliert werden (vgl. z.B. Dennis und Li 2003).

Anhaltspunkte zur Qualität des Ansatzes kann auch hier wiederum der Vergleich mit Daten liefern, die mit anderen Methoden gewonnen wurden. So berichten zum Beispiel Berrens et al. (2003), dass die ungewichteten Daten vom Knowledge Networks-Panel (im Folgendem *KN*-Daten genannt) bei vielen demografischen Variablen die Populationswerte besser widerspiegelten als die Daten einer Stichprobe aus dem Harris-Interactive-Panel oder sogar als die Daten einer gleichzeitig durchgeführten RDD-Studie. Obwohl die Werte für einige politische Variablen voneinander abwichen, konnten zwischen den gewichteten RDD- und *KN*-Daten bezüglich der Schätzung von Modellen zum Wahlverhalten und zu Umwelt-Einstellungen keine maßgeblichen Unterschiede festgestellt werden. Weitere Ergebnisse mit Knowledge Networks-Panels präsentieren zum Beispiel Baker et al. (2003), Li et al. (2004) und Smith (2003).

Befragungen, die sich dem Typ 8 zuordnen lassen, haben zwar viele Vorteile, es handelt sich aber wohl auch um die teuersten Arten von Web-Befragungen, weil sowohl für die Rekrutierung als auch für die Pflege des Panels erheblicher Aufwand betrieben werden muss. Ob dieser Zusatzaufwand durch eine bessere Qualität der Ergebnisse gerechtfertigt ist, lässt sich derzeit noch nicht sagen. Trotzdem erscheint der Ansatz viel versprechend, und es ist denkbar, dass er dereinst die hochwertigen zufallsgesteuerten, auf traditionellen Methoden beruhenden Befragungen ablöst.

VI. Ausblick

Das Thema „Online-Befragungen" umfasst eine Vielzahl an Methoden, die unterschiedlichste Zwecke verfolgen und verschiedene Populationen berücksichtigen. Die Methoden reichen von simplen Kurzumfragen zu Unterhaltungszwecken bis zu vollständig zufallsgesteuerten Designs, die Rückschlüsse auf die allgemeine Bevölkerung ermöglichen sollen.

In den fünf Jahren seit Coupers Überblicksartikel (2000) wurde eine Vielzahl neuer Erkenntnisse zur Verbesserung der Qualität von Online-Befragungen gewonnen,[8] und Online-Befragungen haben einen legitimen Platz im Werkzeugkasten der Umfrageforscher eingenommen. Gleichzeitig ist es aber unrealistisch anzunehmen, dass bei der Weiterentwicklung dieser Erhebungsmethoden keine größeren Schwierigkeiten mehr auftreten werden. Forschungsbedarf besteht noch immer, insbesondere um Nonresponsequoten bei Online-Studien zu verringern. Aber auch das Problem unterschiedlicher Internetnutzung bei breit angelegten Bevölkerungsbefragungen und die Möglichkeiten von multimedialen und interaktiven Elementen zur Erhöhung der Qualität der Messungen müssen genauer untersucht werden.

Wir haben nun die Gelegenheit, als Ergänzung zu bestehenden Ansätzen eine leistungsfähige neue Methode zur Datengewinnung zu erforschen und weiter zu entwickeln. Solide Arbeit und ein offener Austausch von Methoden und Ergebnissen sind erforderlich, damit dies in sachkundiger und verantwortungsvoller Weise geschieht. Während sich die spezifischen Methoden der Umfrageforschung auf unvorhersagbare Weise über die Zeit verändern können, bleiben doch die grundlegenden Kriterien, anhand derer die Methoden zu bewerten sind, größtenteils gleich. Die Weiterentwicklung der Methoden der Web-Befragung muss sich an diesen Kriterien orientieren.

Literatur

Alvarez, R. Michael, Robert P. Sherman und *Carla VanBeselaere*, 2003: Subject Acquisition for Web-Based Surveys. Political Analysis 11 (1): 23–43.

Baker, Lawrence, Todd H. Wagner, Sara Singer und *M. Kate Bundorf*, 2003: Use of the Internet and E-Mail for Health Care Information: Results from a National Survey. Journal of the American Medical Association 289 (18): 2400–2406.

Ballard, Chet, und *Rudy Prine*, 2002: Citizen Perceptions of Community Policing: Comparing Internet and Mail Survey Responses. Social Science Computer Review 20 (4): 485–493.

Bandilla, Wolfgang, Michael Bosnjak und *Patrick Altdorfer*, 2003: Survey Administration Effects? A Comparison of Web-Based and Traditional Written Self-Administered Surveys Using the ISSP Environment Module. Social Science Computer Review 21 (2): 235–243.

Batinic, Bernad, 2001: Datenqualität bei internetbasierten Befragungen. S. 116–132 in: *Axel Theobold, Marcus Dreyer* und *Thomas Starsetzski* (Hg.), Online-Marktforschung – Theoretische Grundlagen und praktische Erfahrungen. Wiesbaden: Gabler.

Batinic, Bernad, 2003: Datenqualität bei internetbasierten Befragungen. S. 143–160 in: *Axel Theobald, Marcus Dreyer* und *Thomas Starsetzki* (Hg.), Online-Marktforschung. Wiesbaden: Gabler.

8 Auf viele dieser Arbeiten wird auf der WebSM-Website (www.websm.org) verwiesen.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva und *David L. Weimer*, 2003: The Advent of Internet Surveys for Political Research – A Comparison of Telephone and Internet Samples. Political Analysis 11 (1): 1–22.

Best, Samuel J., Brian Krueger, Clark Hubbard und *Andrew Smith*, 2001: An Assessment of the Generalizability of Internet Surveys. Social Science Computer Review 19 (2): 131–145.

Biemer, Paul P., Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz und *Seymour Sudman* (Hg.), 2004: Measurement Errors in Surveys. New York: Wiley.

Birnholtz, Jeremy P., Daniel B. Horn, Thomas A. Finholt und *Sung Joo Bae*, 2004: The Effects of Cash, Electronic, and Paper Gift Certificates as Respondent Incentives for a Web-Based Survey of Technologically Sophisticated Respondents. Social Science Computer Review 22 (3): 355–362.

Bosnjak, Michael, Tracy L. Tuten und *Wolfgang Bandilla*, 2001: Participation in Web Surveys: A Typology. ZUMA-Nachrichten 48: 7–17.

Bosnjak, Michael, Tracy L. Tuten und *Werner W. Wittman*, 2005: Unit (Non)response in Web-Based Access Panel Surveys – An Extended Planned-Behavior Approach. Pychology & Marketing 22 (6): 489–505.

Carini, Robert M., John C. Hayek, George D. Kuh, John M. Kennedy und *Judith A. Ouimet*, 2003: College Student Responses to Web and Paper Surveys – Does Mode Matter? Research in Higher Education 44 (1): 1–19.

Cobanoglu, Cihan, Bill Warde und *Patrick J. Moreo*, 2001: A Comparison of Mail, Fax and Web-Based Survey Methods. International Journal of Market Research 43 (4): 441–452.

Cook, Colleen, Fred Heath und *Russell L. Thompson*, 2000: A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys. Educational and Psychological Measurement 60 (6): 821–836.

Couper, Mick P., 2000: Web Surveys – A Review of Issues and Approaches. Public Opinion Quarterly 64 (4): 464–494.

Dennis, J. Michael, und *Rick Li*, 2003: Effects of Panel Attrition on Survey Estimates. Vortrag an The Annual Conference of the American Association for Public Opinion Research, Nashville, Tennessee.

Dillman, Don A., 2000: Mail and Internet Surveys – The Tailored Design Method. New York: Wiley.

Dillman, Don A., und *Dennis K. Bowker*, 2001: The Web Questionnaire Challenge to Survey Methodologists. S. 159–178 in: *Ulf-Dietrich Reips* und *Michael Bosnjak* (Hg.), Dimensions of Internet Science. Lengerich: Pabst Science Publishers.

Dillman, Don A., Robert D. Tortora, John Conradt und *Dennis K. Bowker*, 1998: Influence of Plain vs. Fancy Design on Response Rates for Web Surveys. Vortrag an The Joint Statistical Meetings of the American Statistical Association, Dallas, Texas.

Eurostat, 2005: Structural Indicators, Innovation and Research. epp.eurostate.cec.eu.int. Aufgerufen im Juli 2005.

Faas, Thorsten, 2004: Umfragen im Umfeld der Bundestagswahl 2002 – Offline und Online im Vergleich. ZA-Information 52: 120–135.

Flemming, Greg, und *Molly Sonner*, 1999: Can Internet Polling Work? Strategies for Conducting Public Opinion Surveys Online. Vortrag an The 54th Annual Meeting of the American Association for Public Opinion Research, St. Petersburg Beach, Florida.

Forsman, Gösta, und *Maria Varedian*, 2002: Mail and Web Surveys: A Cost and Response Rate Comparison in a Study of Students Housing Conditions. Vortrag an The International Conference on Improving Surveys, Kopenhagen.

Fricker, Ronald D., und *Matthias Schonlau*, 2002: Advantages and Disadvantages of Internet Research Surveys. Field Methods 14 (4): 347–367.

Göritz, Anja, im Druck: Incentives in Web Studies – Methodological Issues and a Review. International Journal of Internet Science.

Göritz, Anja S., 2003: Online-Panels. S. 227–240 in: *Axel Theobald, Marcus Dreyer* und *Thomas Starsetzki* (Hg.), Online-Marktforschung. Wiesbaden: Gabler.

Göritz, Anja S., 2004: The Impact of Material Incentives on Response Quantity, Response Quality, Sample Composition, Survey Outcome, and Cost in Online Access Panels. International Journal of Market Research 46 (3): 327–345.

Göritz, Anja S., 2004: Recruitment for Online Access Panels. International Journal of Market Research 46 (4): 411–425.

Göritz, Anja S., eingereicht: The Long-Term Effect of Material Incentives on Participation in Online Panels.

Gräf, Lorenz, 2003: Internet Access Panels in der Praxis. S. 255–270 in: *Axel Theobald, Marcus Dreyer* und *Thomas Starsetzki* (Hg.), Online-Marktforschung. Wiesbaden: Gabler.

Groves, Robert M., 1989: Survey Errors and Survey Costs. New York: Wiley.

Groves, Robert M., und *Mick P. Couper*, 1998: Nonresponse in Household Interview Surveys. New York: Wiley.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer und *Roger Tourangeau*, 2004: Survey Methodology. New York: Wiley.

Hayslett, Michele M., und *Barbara M. Wildemuth*, 2004: Pixels or Pencils? The Relative Effectivness of Web-Based versus Paper Surveys. Library & Information Science Research 26 (1): 73–93.

Heerwegh, Dirk, Tim Vanhove, Koen Matthijs und *Geert Loosveldt*, 2005: The Effect of Personalization on Response Rates and Data Quality in Web Surveys. International Journal of Social Research Methodology 8 (2): 85–99.

Hellwig, Jörg Otto, Boris von Heesen und *Rene Bouwmeester*, 2003: Rekrutierungsunterschiede bei Online-Panels und ihre Folgen. S. 241–254 in: *Axel Theobald, Marcus Dreyer* und *Thomas Starsetzki* (Hg.), Online-Marktforschung. Wiesbaden: Gabler.

James, Delores C. S., W. William Chen und *Jiunn-Jye Sheu*, 2005: Comparison of Three Tobacco Survey Methods with College Students: A Case Study. The International Electronic Journal of Health Education 8: www.iejhe.org.

Joinson, Adam N., und *Ulf-Dietrich Reips*, im Druck: Personalized Salutation, Power of Sender and Response Rates to Web-Based Surveys. Computers in Human Behavior.

Juran, Juan M. und Frank M. Gryna, 1980: Quality Planning and Analysis. New York: McGraw-Hill.

Kalton, Graham, und *Constance F. Citro*, 1993: Panel Surveys: Adding the Fourth Dimension. Survey Methodology 19 (2): 205–215.

Kaplowitz, Michael D., Timothy D. Hadlock und *Ralph Levine*, 2004: A Comparison of Web and Mail Survey Response Rates. Public Opinion Quarterly 68 (1): 94–101.

Kiernan, Nancy Ellen, Michaela Kiernan, Mary Ann Oyler und *Carolyn Gilles*, 2005: Is a Web Survey as Effective as a Mail Survey? A Field Experiment among Computer Users. American Journal Of Evaluation 26 (2): 245–252.

Krotki, Karol P., 2000: Sampling and Weighting for Web Surveys. Vortrag an The 55th Annual Meeting of the American Association for Public Opinion Research, Portland, Oregon.

Kwak, Nojin, und *Barry Radler*, 2002: A Comparison between Mail and Web Surveys – Response Pattern, Respondent Profile and Data Quality. Journal of Official Statistics 18 (2): 257–273.

Kypri, Kypros, Stephen J. Gallagher und *Martine L. Cashell-Smith*, 2004: An Internet-Based Survey Method for College Student Drinking Research. Drug and Alcohol Dependence 76 (1): 45–53.

Li, Hei, Robert P. Berrens, Alok K. Bohara, Hank Jenkins-Smith, Carol Silva und *David L. Weimer*, 2004: Telephone versus Internet Samples for a National Advisory Referendum: Are the Underlying Stated Preferences the Same? Applied Economic Letters 11: 173–176.

Lozar Manfreda, Katja, Zenel Batagelj und *Vasja Vehovar*, 2002: Design of Web Survey Questionnaires – Three Basic Experiments. Journal of Computer-Mediated Communication 7(3): jcmc.indiana.edu.

Lozar Manfreda, Katja, Michael Bosnjak, Iris Haas und *Vasja Vehovar*, 2005: Web Survey Response Rates Compared to Other Modes – A Meta-Analysis. Vortrag an The 60th Annual Meeting of the American Association for Public Opinion Research, Miami Beach, Florida.

Lukawetz, Gerhard, 2002: Empirically Quantifying Unit-Nonresponse Errors in Online Surveys and Suggestions for Computational Correction Methods. S. 403–415 in: *Bernad Batinic, Ulf-Dietrich Reips* und *Michael Bosnjak* (Hg.), Online Social Sciences. Seattle: Hogrefe & Huber Publishers.

McCabe, Sean Esteban, Mick P. Couper, Scott D. Crawford und *Carol J. Boyd*, im Druck: Comparison of Web and Mail Surveys for Studying Secondary Consequences Associated with Substance Use: Evidence for Minimal Mode Effects. Addictive Behaviors.

National Telecommunications and Information Administration, 2004: A Nation Online – Entering the Broadband Age. www.ntia.doc.gov/reports/anol/NationOnlineBroadband04.pdf. Aufgerufen im Juni 2005.

O'Muircheartaigh, Colm, 1997: Measurement Errors in Surveys – A Historical Perspective. S. 1–25 in: *Lars E. Lyberg, Paul P. Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz* und *Dennis Trewin* (Hg.), Survey Measurement and Process Quality. New York: Wiley.

O'Neil, Kevin M., Steven D. Penrod und *Brian H. Bornstein*, 2003: Web-Based Research – Methodological Variables' Effects on Dropout and Sample Characteristics. Behavior Research Methods Instruments & Computers 35 (2): 217–226.

Reips, Ulf-Dietrich, 2002: Standards for Internet-Based Experimenting. Experimental Psychology 49 (4): 243–256.

Reips, Ulf-Dietrich, und *Lenka Franek*, 2004: Mitarbeiterbefragungen per Internet oder Papier? Der Einfluss von Anonymität, Freiwilligkeit und Alter auf das Antwortverhalten. Wirtschaftspsychologie 1: 67–83.

Rosenbaum, Paul R., und *Donald B. Rubin*, 1983: The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70 (l): 41–55.

Ross, Michael W., Sven-Axel Mansson, Kristian Daneback, Al Cooper und *Ronny Tikkanen*, 2005: Biases in Internet Sexual Health Samples – Comparison of an Internet Sexuality Survey and National Sexual Health Survey in Sweden. Social Science & Medicine 61: 245–252.

Saris, Willem E., 1998: Ten Years of Interviewing without Interviewers: The Telepanel. S. 409–429 in: *Mick P. Couper, Reginald P. Baker, Jelke Bethlehem, Cynthia Z. F. Clark, Jean Martin, William L. Nichols II* und *James O'Reilly* (Hg.), Computer Assisted Survey Information Collection. New York: Wiley.

Schmidt, Julie, 2000: Bill of Rights of the Digital Consumer: The Importance of Protecting the Consumer's Right to Online Privacy. Vortrag an The 55th Annual Conference of the American Association for Public Opinion Research, Portland, Oregon.

Schnell, Rainer, 1993: Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität" und Gewichtungsverfahren. Zeitschrift für Soziologie 22 (1): 16–32.

Schnell, Rainer, Paul B. Hill und *Elke Esser*, 2005: Methoden der empirischen Sozialforschung. München: Oldenbourg.

Schoen, Harald, 2004: Online-Umfragen – schnell, billig, aber auch valide? Ein Vergleich zweier Internetbefragungen mit persönlichen Interviews zur Bundestagswahl 2002. ZA-Information 54: 27–52.

Schonlau, Matthias, Beth J. Asch und *Can Du*, 2003: Web Surveys as Part of a Mixed-Mode Strategy for Populations that Cannot be Contacted by E-Mail. Social Science Computer Review 21 (2): 218–222.

Schonlau, Matthias, Ronald D. Fricker und *Marc N. Elliott*, 2002: Conducting Research Surveys via E-Mail and the Web. Santa Monica: RAND.

Schonlau, Matthias, Kinga Zapert, Lisa Payne Simon et al., 2003: A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey. Social Science Computer Review 22 (1): 128–138.

Schwarz, Stefan, und *Ulf-Dietrich Reips*, 2001: CGI versus JavaScript – A Web Experiment on the Reverse Hindsight Bias. S. 75–90 in: *Ulf-Dietrich Reips* und *Michael Bosnjak* (Hg.), Dimensions of Internet Science. Lengerich: Pabst Science Publishers.

Schweizerisches Bundesamt für Statistik, 2005a: Indikatoren Internetnutzung. www.bfs.admin.ch/bfs/portal/de/index/themen/systemes_d_indicateurs/indi cateurs_de_la/approche_globale.indicator.30106.html. Aufgerufen im Juni 2005.

Schweizerisches Bundesamt für Statistik, 2005b: Nutzung des E-Census. www.bfs.admin.ch/bfs/portal/de/index/themen/systemes_d_indicateurs/indicateurs_de_la/approche_globale.indicator.30302.html. Aufgerufen im Juni 2005.

Sheehan, Kim Bartel, 2001: E-Mail Survey Response Rates: A Review. Journal of Computer-Mediated Communication 6(2): jcmc.indiana.edu.

Smith, Tom W., 2003: An Experimental Comparison of Knowledge Networks and the GSS. International Journal of Public Opinion Research 15 (2): 167–179.

Sparrow, Nick, und *John Curtice*, 2004: Measuring the Attitudes of the General Public via Internet Polls: An Evaluation. International Journal Of Market Research 46 (1): 23–44.

Statistik Austria, 2005: Statistisches Jahrbuch 2005. www.statistik.at/jahrbuch_2005/pdf/k05.pdf. Aufgerufen im Juni 2005.

Statistisches Bundesamt, Deutschland, 2005: Informationstechnologie in Unternehmen und Haushalten 2004. www.destatis.de/download/d/veroe/pb_ikt_04.pdf. Aufgerufen im Juni 2005.

Taylor, Humphrey, 2000: Does Internet Research work? Comparing Online Survey Results with Telephone Surveys. International Journal of Market Research 42 (1): 51–63.

Terhanian, George, 2000: How to Produce Credible, Trustworthy Information Through Internet-Based Survey Research. Vortrag an The 55th Annual Meeting of the American Association for Public Opinion Research, Portland, Oregon.

Theobald, Axel, 2003: Rücklaufquoten bei Online-Befragungen. S. 203–210 in: *Axel Theobald, Marcus Dreyer* und *Thomas Starsetzki* (Hg.), Online-Marktforschung. Wiesbaden: Gabler.

Tourangeau, Roger, Lance J. Rips und *Kenneth A. Rasinski*, 2000: The Psychology of Survey Response. New York: Cambridge University Press.

Tuten, Tracy L., Michael Bosnjak und *Wolfgang Bandilla*, 2000: Banner-Advertised Web Surveys. Marketing Research 11 (4): 17–21.

Tuten, Tracy L., Mirta Galesic und *Michael Bosnjak*, 2004: Effects of Immediate versus Delayed Notification of Prize Draw Results on Response Behavior in Web Surveys: An Experiment. Social Science Computer Review 22 (3): 377–384.

Weßels, Martin, und *Matthias Zimmermann*, 2001: Offline-rekrutierte Access-Panels als „Königsweg" der panelbasierten Online-Forschung? Erkenntnissgewinne durch das „Dialego-Repräsentativpanel". Vortrag an The German Online Research Conference, Göttingen.

Whitcomb, Michael E., und *Stephen R. Porter*, 2004: E-Mail Contacts – A Test of Complex Graphical Designs in Survey Research. Social Science Computer Review 22 (3): 370–376.

Zhang, Yin, 2000: Using the Internet for Survey Research: A Case Study. Journal of the American Society of Information Science 51 (1): 57–68.

IV. Sozialer Kontext und Netzwerke

EGOZENTRIERTE NETZWERKE

Erhebungsverfahren und Datenqualität*

Christof Wolf

Zusammenfassung: Gegenstand dieses Beitrags sind Verfahren zur Erhebung egozentrierter Netzwerke in standardisierten Massenumfragen. Nach einer kurzen Skizzierung der verschiedenen Traditionen sozialwissenschaftlicher Netzwerkforschung folgt eine ausführliche Beschreibung der verschiedenen Methoden, die zur Erhebung egozentrierter Netzwerke entwickelt wurden. Den Schwerpunkt dieser Darstellung bilden Verfahren, die mit Namensgeneratoren arbeiten, also Instrumente, mit denen nach den Namen bestimmter Personen gefragt wird. Im zweiten Teil der Arbeit wird die Validität und Reliabilität egozentrierter Netzwerkdaten diskutiert. Es werden Ergebnisse zur Qualität von Netzwerkgeneratoren, zur Qualität von Proxi-Angaben der Befragten über ihre Netzwerkpersonen sowie zur Qualität von Beziehungsmerkmalen berichtet. Außerdem wird die Auswirkung der Erhebungsmethode und des Interviewerverhaltens auf die Güte egozentrierter Netzwerkdaten untersucht. Der Artikel endet mit Empfehlungen zur Erhebung egozentrierter Netzwerke.

I. Einleitung

Seit Mitte der 1950er Jahre hat die sozialwissenschaftliche Netzwerkforschung stark an Bedeutung gewonnen. Ursache dafür sind in erster Linie gesellschaftliche Veränderungen, die sich in theoretischen, empirischen und methodischen Fortschritten der Forschung niedergeschlagen haben. Zu den wichtigsten theoretischen Entwicklungen zählt die Behauptung und (spätere) empirische Beobachtung eines starken Individualisierungsschubs in den westlichen Gesellschaften (Ester, Halman und Moor 1994), der zu einer Schwächung traditioneller Formen der Vergemeinschaftung führte (Beck 1986: 208ff.). Statt weiterhin „die Familie", „die Klassenzugehörigkeit" oder „die Nachbar-

* Besonderen Dank schulde ich Robert Kecskes. Er hat die hier dargestellten „Kölner" Instrumente mitentwickelt und war stets bereit, über inhaltliche und methodische Fragen zu diskutieren. Für konstruktive Hinweise zur Verbesserung des Manuskripts danke ich Henning Best und den anonymen Gutachtern. Jürgen Friedrichs und Jörg Blasius sei für die Überlassung ihrer Daten gedankt. Für die Unzulänglichkeiten und Fehler in diesem Beitrag trage ich allein die Verantwortung.

schaft“ zu untersuchen, lenkte die gesellschaftliche Individualisierung die Aufmerksamkeit der Forschung immer stärker auf das Netzwerk faktisch bestehender sozialer Beziehungen. Durch die „Befreiung von der Gemeinschaft“ (Wellman 1979) wurde die kontextuelle Zugehörigkeit einer Beziehung zu einer empirisch offenen Frage. Eine weitere Entwicklung in der soziologischen Theorie, die dem Netzwerkkonzept Auftrieb gab und weiterhin gibt, war die Entdeckung sozialer Beziehungen als Ressource, als soziales Kapital (vgl. den Überblick bei Portes 1998). Obwohl sehr verschiedene Definitionen dieses Begriffs existieren, besteht weitgehend Einigkeit darüber, dass soziales Kapital „refers to the importance of resources, which, although possessed by other persons, are available to a given individual through their social relations with these others“ (Flap 2004: 7). Der Stellenwert sozialen Kapitals wird besonders deutlich, wenn man sich vergegenwärtigt, dass es zur Generierung von Humankapital (Coleman 1988), zur Erreichung eines beruflichen Status (Wegener 1989) und zur Erzielung von Einkommen, also von ökonomischem Kapital (Boxman, de Graaf und Flap 1991), verwendet werden kann.

Parallel zu den skizzierten theoretischen Entwicklungen und empirischen Beobachtungen wurde schon bald nach dem Zweiten Weltkrieg an Methoden gearbeitet, um soziale Beziehungen und Netzwerke systematisch empirisch untersuchen zu können. Eine der zentralen Fragen war dabei, wie die Vorzüge der repräsentativen Umfrage und der Wunsch nach Abbildung der Struktur sozialer Beziehungen verknüpft werden können (z.B. Coleman 1958/59). Die heute etablierte Lösung dieses Problems besteht darin, so genannte *egozentrierte Netzwerke* zu erheben.[1] Dabei handelt es sich um die Beziehungen einer Fokusperson – Ego – zu anderen Personen – den Alteri –, wobei die Informationen zu den Kontaktpersonen und zu den Beziehungen zwischen Ego und den Alteri (meist) von Ego stammen. Heute übliche Verfahren zur Erhebung egozentrierter Netzwerke und Angaben zur Qualität der mit diesen Instrumenten erhobenen Daten, sind Gegenstand dieses Beitrags, der folgendermaßen aufgebaut ist:[2]

Im nächsten Abschnitt skizziere ich die Entwicklung entsprechender Erhebungsverfahren und biete einen Überblick über die heute etablierten Formen dieser Messinstrumente. Anschließend wird die Gültigkeit und Zuverlässigkeit egozentrierter Netzwerkdaten aus verschiedenen Perspektiven bzw. auf unterschiedlichen Ebenen untersucht.

1 Dem Konzept des egozentrierten Netzwerks steht das Konzept des Gesamtnetzwerks gegenüber. Darunter wird die Menge der sozialen Beziehungen *aller* Personen (oder anderer Einheiten) einer bestimmten Gruppe verstanden. Beispiele für solche Netzwerke sind die Freundschaftsbeziehungen von Jugendlichen in einem gegebenen Stadtviertel (Kirke 2004) oder die diplomatischen und die Handelsbeziehungen der Mitgliedsländer der UN (Erickson 1975). Die Untersuchung von Gesamtnetzwerken setzt also die Existenz klar definierter, verhältnismäßig kleiner Gruppen voraus. Dies ist bei den eingangs skizzierten Fragestellungen typischerweise nicht der Fall, beziehen diese sich doch auf die Gesellschaft als Ganzes. Wenn beispielsweise die Effekte des gesamtgesellschaftlichen Wandels auf soziale Beziehungen der Bevölkerung oder der Nutzen sozialer Beziehungen untersucht werden soll, bietet es sich an, für einen repräsentativen Querschnitt der Bevölkerung egozentrierte Netzwerke zu erheben.

2 Eine allgemeine Einführung in die Erhebungs- und Analysetechniken von (Gesamt-)Netzwerken bieten Degenne und Forsé (1999 [1994]), Jansen (2003), Schenk (1984), Scott (2000) sowie Wasserman und Faust (1994). Die Geschichte der Netzwerkanalyse kann bei Freeman (2004) nachgelesen werden. Einen Vergleich von egozentrierten und Gesamtnetzwerken liefern Pappi (1987) sowie Gerich und Lehner (2003).

Der Artikel schließt mit Empfehlungen, die die Erhebung möglichst valider und reliabler Daten zu egozentrierten Netzwerken sicherstellen sollen. Auf die Besonderheiten im Zusammenhang mit der Aufbereitung egozentrierter Netzwerkdaten und die Analysemöglichkeiten, die diese Daten bieten, kann hier nicht eingegangen werden (vgl. aber Marsden 1987; Pappi 1987: 22ff.; Wolf 1993).

II. Die Erhebung egozentrierter Netzwerke

Instrumente zur Erhebung egozentrierter Netzwerke bestehen aus drei Teilen: Zunächst bedarf es einer oder mehrerer Fragen nach Personen, mit denen die befragte Person in Beziehung steht; dies sind die *Namensgeneratoren*. Im Anschluss werden Fragen nach Eigenschaften der Netzwerkpersonen und zur Beziehung zwischen Ego und Alter gestellt; diese Fragen werden als *Namensinterpretatoren* bezeichnet. In einem weiteren Teil können schließlich die Beziehungen zwischen den Netzwerkpersonen erhoben werden.

1. Die Identifikation der Netzwerkpersonen

Der erste „moderne" Namensgenerator, der auch als die „Drei-Freunde-Frage" bezeichnet wird, wurde m.W. in einer 1963 von Edward Laumann durchgeführten Studie verwendet.[3] Die Frage lautete damals (1966: 170): „Could you tell me the jobs or occupations of three of your closest friends whom you see most often? Don't tell me their names. They can be relatives of yours." In seiner zweiten größeren Studie greift Laumann auf die Daten der Detroit Area Study von 1965/66 zurück. Hier lautet die entsprechende Frage: „Now would you think of the three men who are your closest friends and whom you see most often. They can be relatives or non-relatives, as you wish" (Laumann 1973: 264). Problematisch an dieser Formulierung ist, dass hier zwei Kriterien vermischt werden, deren Bedeutung zudem in verschiedenen Gruppen von Befragten variiert. Zum einen ist der Begriff „Freund" unpräzise und wird von den Angehörigen verschiedener sozialer Schichten unterschiedlich verstanden (Allan 1977); zum anderen wird der Stimulus Freund mit der Dimension der Häufigkeit des Kontakts vermischt „... whom you see most often". Die Drei-Freunde-Frage weist jedoch noch weitere Unzulänglichkeiten auf. Durch ihre Beschränkung auf maximal drei Personen sind die Angaben zur Netzwerkgröße stark rechtszensiert. Darüber hinaus können auf der Grundlage nur eines Namensgenerators keine Angaben zur Anzahl der verschiedenen Arten von Beziehungen bzw. zum Austausch zwischen Ego und Alter gemacht werden, d.h. die Multiplexität der Beziehung kann nicht bestimmt werden.

3 Ein weiterer, früher Netzwerkgenerator stammt von Wellman (1979: 8f.), der in einer 1968 in East York, Toronto, erhobenen Studie nach „the persons outside your home that you feel closest to" fragte. Eine kurze, persönlich gefärbte Geschichte der Erhebung egozentrierter Netzwerke für Nordamerika und Großbritannien bietet Wellman (1993). In der Sozialforschung der Bundesrepublik Deutschland lassen sich ebenfalls mehrere frühe Ansätze zur Erhebung des „Verkehrskreises" in Umfragen ausmachen (Mayntz 1958; Reigrotzki 1956; Treinen 1965). Eine systematische Aufarbeitung dieser Geschichte steht jedoch noch aus.

Auf Grund dieser Überlegungen entwickelten Lynne McCallister und Claude Fischer Ende der 1970er Jahre ein neues Instrument zur Erhebung egozentrierter Netzwerke, das in der North California Community Study (NCCS) eingesetzt wurde (Fischer 1982).[4] Dieses Instrument hat zwei Besonderheiten: Es besteht aus mehreren Namensgeneratoren und es betrachtet ausschließlich Beziehungen, die auf Interaktionen beruhen, also keine Beziehungen, bei denen es sich ausschließlich um affektive oder normative Beziehungen handelt (McCallister und Fischer 1978: 135f.). Die Interaktionen waren dabei so ausgewählt, dass möglichst viele Netzwerkpersonen aus möglichst unterschiedlichen sozialen Kontexten erfasst werden können. Die ausgewählten Namensgeneratoren lauteten (ebenda: 137):

(1) Who would care for the respondents' homes if they went out of town;
(2) if they work, with whom they talk about work decisions;
(3) who, if anyone, had helped with household tasks in the last three months;
(4) with whom they engaged in social activities (like inviting home for dinner, or going to a movie);
(5) who they talk with about hobbies;
(6) if unmarried, who their fiancé(e) or „best friend" is;
(7) with whom they talk about personal worries;
(8) whose advice they consider in making important decisions;
(9) from whom they would or could they borrow a large sum of money;
(10) enumeration of adult members of the respondents' households.
...
Is there anyone who is important to you who doesn't show up on this list?

Dieses Instrument wurde bei Erhebungen im Raum Mannheim (Pfenning und Pfenning 1987), in Haifa, Israel (Fischer und Shavit 1995), in Jacksonville, Florida, und in Mexico-City (Bernard et al. 1990) eingesetzt. In allen drei Ländern waren die egozentrierten Netzwerke, die mit diesem Instrument erhoben wurden, kleiner als die in der North California Community Study (NCCS) erhobenen (vgl. *Tabelle 1*).

Die Vorteile des von McCallister und Fischer entwickelten Verfahrens liegen auf der Hand: Es werden relativ große Netzwerke erzeugt, die sich beispielsweise nach unterschiedlichen Interaktionsformen oder Beziehungsarten differenzieren lassen. Die Daten ermöglichen zudem, den Grad der Multiplexität sozialer Beziehungen zu untersuchen. Der Nachteil des Verfahrens ist ebenso offensichtlich: Es ist sehr aufwändig und mit hohen Erhebungskosten verbunden – es werden etwa 20 Minuten für diesen Teil der Erhebung veranschlagt (McCallister und Fischer 1978: 136). Daher eignet es sich nur für Studien, deren Schwerpunkt auf der Erhebung sozialer Netzwerke liegt und für die ausreichend Mittel zur Verfügung stehen. Für Untersuchungen, bei denen egozentrierte Netzwerke nicht im Mittelpunkt des Interesses stehen, oder für Mehrthemenumfragen ist das Verfahren ungeeignet. Für diese Fälle hat Ronald Burt (1984) einen Namensgenerator vorgeschlagen, der zuerst im General Social Survey 1985 eingesetzt wurde. Dieser Generator besteht aus einer einzigen Frage, die in der deutschen Fassung

4 Die „Drei-Freunde-Frage" wird jedoch auch weiterhin verwendet; so z.B. in den Erhebungen 1980, 1990 und 2000 des ALLBUS sowie als „Fünf-Freude-Frage" im GSS 1998.

Tabelle 1: Die Namensgeneratoren von McCallister, Fischer und Burt im Vergleich

	McCallister und Fischer					Burt		
	NCCS[a]	ZUMA[b]	JV[c]	MC[c]	Haifa[d]	GSS[e]	ZUMA[b]	WS87[f]
Anzahl Befragte	1.050	239	98	99	262	1.531	240	1.311
∅ Netzwerkgröße	18,5	7,8	21,8	10,1	14,0	2,9	2,6	2,5
Prozent Verwandte	42	39	40	46	44	55	49	60

a North California Community Study (Fischer 1982).
b ZUMA-Methodenexperiment „Egozentrierte Netzwerke“ (Pfenning und Pfenning 1987).
c JV: Jacksonville, Florida, und MC: Mexico-City; Methodenexperiment „Measuring Personal Social Networks“ (Bernard et al. 1990).
d Haifa-Studie (Fischer und Shavit 1995).
e General Social Survey 1985; im GSS konnten bis zu sechs Personen genannt werden; um den Vergleich mit den beiden anderen Studien zu gewährleisten, wurde diese Zahl hier auf maximal fünf festgelegt.
f Wahlstudie 1987 (Pappi und Melbeck 1988: 233).

wie folgt lautet: „Hin und wieder besprechen die meisten Leute wichtige Angelegenheiten mit anderen. Wenn Sie an die letzten sechs Monate zurückdenken: Mit wem haben Sie über Dinge gesprochen, die Ihnen wichtig waren?“ (ZA 1987: 256) Wie Burt anhand der von Fischer erhobenen Daten zeigen konnte, handelt es sich bei diesen Personen um eine Teilgruppe des egozentrierten Netzwerks, die durch relativ nahe Beziehungen (intimacy) gekennzeichnet sind, mit einem im Vergleich zum Fischer-Netz höheren Anteil von Freunden und Verwandten.[5]

Tabelle 1 stellt die Ergebnisse verschiedener Studien gegenüber, in denen entweder der Netzwerkgenerator nach McCallister und Fischer oder der Generator nach Burt verwendet wurde.[6] Erwartungsgemäß sind die mit dem Generator nach McCallister und Fischer erhobenen Netzwerke deutlich größer als die nach der Burt-Methode erfassten.[7] Darüber hinaus zeigt sich für die Netzwerke, die mit dem Namensgenerator nach Burt erhoben wurden, erwartungsgemäß ein größerer Anteil von Verwandten als in den Netzwerken nach Fischer und McCallister, wobei letztere mit rund 40 Prozent einen immer noch sehr hohen Anteil Verwandter aufweisen.

Die bisher vorgestellten Namensgeneratoren zielen auf starke, positiv bewertete Beziehungen. Der Versuch, negativ bewertete Beziehungen zu erheben, findet sich dagegen kaum. Bruce Straits (2000: 127) hat daher untersucht, ob der Burt-Generator mit dem Zusatz „... these may include people that sometimes make you angry or upset“ zu

5 Straits (2000) zeigt, dass eine Variation des Burt Generators, in der nach besonders wichtigen Personen gefragt wird, keine Auswirkung auf die Netzwerkgröße oder die Kontakthäufigkeit hat. Die entsprechende Frage lautete: „Looking back over the last 6 months, who are the people especially significant in your life? By significant, I mean those who are most important.“ (ebenda: 127) Zu einer anderen Einschätzung kommen Bien, Marbach und Neyer (1991: 89) sowie Milardo (1992).

6 Für ähnliche Zusammenstellungen mit teilweise anderen Namensgeneratoren und Studien vgl. Campbell und Lee (1991: Tabelle 2), McCarthy et al. (1997: Tabelle 3) sowie Milardo (1992: Tabelle 1 und 2).

7 Warum die in Deutschland mit dem Generator nach McCallister und Fischer erhobenen Netzwerke so viel kleiner sind als diejenigen der amerikanischen Studie, ist nicht klar. Dass in der ZUMA Variante dieses Generators auf die Nachfrage nach den Haushaltsmitgliedern und dem Lebenspartner verzichtet wird, kann nur einen Teil der Differenz erklären.

anderen Ergebnissen führt. Dies ist anscheinend nicht der Fall. Zwar ergibt sich eine geringfügig höhere Netzwerkgröße – die jedoch nicht signifikant war –, die Zusammensetzung und die Inhalte der Netzwerke (z.B. Gesprächsgegenstände) ändern sich jedoch kaum. Straits (2000: 133) folgert daher: „These results suggest that it is unnecessary to reword core network name generators to encourage ego to report stressful or negative relations." Statt also direkt nach belastenden Beziehungen zu fragen, ist es sinnvoller, im zweiten Teil des Netzwerkinstruments für jede Beziehung zu erheben, wie stark sie Ego belastet. Dies geschah z.B. in der Studie von Christof Wolf (2003), in der zu jeder Netzwerkperson gefragt wurde: „Beziehungen zu anderen Menschen können hilfreich, aber manchmal auch belastend sein. Wie häufig ist Ihre Beziehung zu ... (Name) für Sie belastend?" Wie sich zeigt, geht eine größere Belastung durch das soziale Netzwerk mit einer geringeren psychischen Gesundheit einher.

Mit Fischer und Burt sind zwei allgemeine Strategien zur Generierung von Namen benannt, die je nach Studienziel abgewandelt, verkürzt oder verlängert werden können. So haben z.B. Burt (1997) und Mart van der Poel (1995) in Anlehnung an McCallister und Fischer jeweils ein Netzwerkinstrument mit neun bzw. zehn Generatorfragen entwickelt. Burts Interesse zielt auf die Möglichkeit, Karriereprozesse von Managern innerhalb von Unternehmen untersuchen zu können. Die entsprechenden Namensgeneratoren sind daher spezifisch für diese Zielgruppe formuliert. Van der Poel ist an der Messung sozialer Unterstützung interessiert und entwirft daher in Anlehnung an Fischer einen zehn Fragen umfassenden Namensgenerator, der Beziehungen aus den Bereichen emotionale Unterstützung, instrumentelle Unterstützung und Geselligkeit abbilden soll. Weil er auch Personen erfassen will, mit denen aktuell kein Kontakt besteht, die aber für eine Hilfeleistung zur Verfügung stehen würden, formuliert er die acht Items zu Unterstützungsleistungen als hypothetische Fragen, z.B. „Stellen Sie sich vor, Sie hätten Grippe und müssten einige Tage im Bett verbringen. Wen würden Sie darum bitten, sich um Sie zu kümmern oder Ihre Einkäufe zu erledigen?" (van der Poel 1995: 55; Übersetzung d. Verf.). Für die Erhebung konkreter, aktueller Interaktionen eignet sich dieses Instrument daher nicht.

Eine andere an McCallister und Fischer angelehnte Methode zur Erhebung egozentrierter Netzwerke wurde von Robert Kecskes und Christof Wolf (1996) entwickelt und seither mit leichten Modifikationen in vier weiteren Studien eingesetzt (Friedrichs und Blasius 2000; Friedrichs, Kecskes und Wolf 2002; Kecskes 2000; Wolf 2003; vgl. die Übersicht in Tabelle A1 im Anhang). Ähnlich wie die oben beschriebenen Instrumente von McCallister und Fischer bzw. von van der Poel werden bei unserem Instrument mehrere Namensgeneratoren eingesetzt, die sich jedoch im Gegensatz zu van der Poel auf *konkrete* soziale Interaktionen beziehen. Bei der Auswahl der Interaktionen haben wir uns von Burts (1984) Analysen der NCCS Daten sowie von Überlegungen zur Bedeutung instrumenteller und geselliger Kontakte leiten lassen.[8] Für jeden dieser drei

8 In unserer ersten Studie wollten wir u.a. auch die Reziprozität der Beziehungen messen. Wir haben dies versucht, indem wir jeden Namensgenerator, bei dem dies logisch möglich war, jeweils einmal aus der Sicht Egos und einmal aus der Sicht möglicher Alteri gestellt haben. Wie unsere Analysen und ein Methodenexperiment von Tilburg (1992) zeigen, lässt sich das Ausmaß der Reziprozität sozialer Beziehungen auf diese Weise jedoch nicht messen. Wir haben uns

Bereiche wurden zwei bzw. drei Namensgeneratoren formuliert (Kecskes und Wolf 1996: 40f.):

9. Die meisten Leute besprechen hin und wieder wichtige Angelegenheiten mit anderen Personen. Wenn Sie an die *letzten 14 Tage* zurückdenken:
   a) An wen haben Sie sich gewandt, um Dinge zu besprechen, die *Ihnen* wichtig waren?
   b) Und wer hat sich in den *letzten 14 Tagen* an *Sie* gewandt, um Dinge zu besprechen, die *ihr oder ihm* wichtig waren?
10. Es gibt viele Dinge, die durch die praktische Hilfe anderer erleichtert werden können, wie z.B. Reparaturen im Haushalt oder am PKW, die Erledigung von Besorgungen, die Hilfe im Krankheitsfalle und die Betreuung von Kindern oder alten Menschen.
   a) Wem haben Sie in den *letzten 14 Tagen* in der einen oder anderen Weise praktische Hilfe geleistet?
   b) Wer hat Ihnen in diesem Zeitraum in der einen oder anderen Weise praktische Hilfe geleistet?
11. Kommen wir nun auf die Personen zu sprechen, mit denen Sie Ihre Freizeit verbringen. Wenn Sie wiederum an die *letzten 14 Tage* denken:
   a) Wer hat Sie in dieser Zeit zu Hause besucht?
   b) Und wen haben Sie in den letzten 14 Tagen zu Hause besucht?
   c) Mit wem haben Sie in den letzten 14 Tagen Freizeitaktivitäten außerhalb des Hauses unternommen?

Die Interviewer wurden darauf hingewiesen, bei den Fragen 9 und 10 keine Personen aufzunehmen, denen Ego in seiner Berufsrolle geholfen hat bzw. die Ego in ihrer Berufsrolle geholfen haben, z.B. als Arzt, Priester oder Pfleger. Bei Frage 11c) sollten zudem keine Personen aufgeführt werden, mit denen ausschließlich Aktivitäten im Rahmen von Vereinsveranstaltungen unternommen wurden, wie z.B. die Fußballkameraden. Diese Personen sollten nur dann aufgeführt werden, wenn mit ihnen zudem weitere Freizeitaktivitäten unternommen wurden, wie ein Kinobesuch nach dem Training. Natürlich durften die Netzwerkpersonen auf mehreren Generatoren genannt werden. Wenn man sich z.B. an eine Person gewandt hat, um mit ihr wichtige Dinge zu besprechen, *und* diese Person zu Hause besucht hat, sollte sie bei Frage 9a) *und* bei Frage 11a) genannt und vom Interviewer notiert werden.

Im Gegensatz zu den Namensgeneratoren von McCallister und Fischer bzw. Burt, die einen Zeitraum von drei bzw. sechs Monaten vorgeben, haben wir uns für eine enge Begrenzung auf den Zeitraum der letzten 14 Tage entschieden. Durch die Verwendung eines so überschaubaren Zeitfensters, so unsere Hoffnung, wird die Zuverlässigkeit der Antworten erhöht. Aus diesem Grund wurden auch keine Umschreibungen wie „in letzter Zeit" oder „vor kurzem" verwendet. Personen, die für Ego von zentraler Bedeutung sind, mit denen jedoch in den letzten 14 Tagen kein Kontakt bestand, konnten auf eine letzte Frage genannt werden:[9]

dennoch dafür entschieden, die „doppelten" Namensgeneratoren beizubehalten, weil keiner dieser Fragen vollkommen redundante Informationen liefert.

9 Nach Ansicht eines der Gutachter genügt diese Frage nicht, um die für Ego wichtigen Personen, zu denen jedoch in den letzten 14 Tagen kein Kontakt bestand, zu identifizieren. Die Ergebnisse unserer Studien lassen jedoch vermuten, dass ein erheblicher Teil dieser Personen erfasst wird, da knapp ein Viertel der Netzwerkpersonen auf diese Frage genannt werden.

„Wenn Sie jetzt noch einmal an alle Menschen denken, die Sie mir gerade genannt haben. Fehlen jetzt noch Personen, die Ihnen wichtig sind oder die in Ihrem Leben eine wichtige Rolle spielen?“

Alle bisher genannten Instrumente haben das Ziel, einen mehr oder weniger großen Ausschnitt der sozialen Beziehungen Egos zu erheben, der durch regelmäßige Interaktion gekennzeichnet ist, wobei die Art der Interaktionen durch die Formulierung des Namensgenerators vorgegebenen ist. Diese Form der Erhebung egozentrierter sozialer Netzwerke ist zwar verhältnismäßig aufwändig, aber sehr flexibel. In Abhängigkeit von der Forschungsfrage und bei angemessener Wahl des Namensgenerators bzw. der Namensgeneratoren, können mit dieser Methode unterschiedliche Ausschnitte eines Netzwerks abgebildet werden.

Daneben existiert eine Reihe weiterer Techniken zur Erhebung egozentrierter Netzwerke, die seltener eingesetzt werden. Einer dieser Ansätze wurde von Christopher McCarthy et al. (1997) vorgeschlagen und dient vor allem dazu, valide Punktschätzer für den Umfang, teilweise auch für die Zusammensetzung egozentrierter Gesamtnetzwerke zu erhalten. Bei dieser Methode bekommen die Befragten eine Liste mit 50 „repräsentativ“ ausgewählten Vornamen und sie sollen angeben, ob sie jemand mit diesem Namen „kennen“; dabei kann das Kriterium des Kennens variiert werden – z.B. Ego und Alter *er*kennen sich und können sich mit Namen nennen. Im Gegensatz zu den Generatoren von Fischer oder Burt zielt diese Methode darauf, das gesamte Netzwerk bzw. eine Zufallsauswahl desselben zu erheben (vgl. auch Brewer 1997). Im Vergleich zu den Netzwerkgeneratoren von McCallister und Fischer bzw. Burt werden mit der von McCarthy et al. entwickelten Methode deutlich weniger Verwandte und nahe stehende Personen erhoben, die Kontakthäufigkeit ist geringer und ein größerer Anteil der Netzwerkpersonen lebt außerhalb der Stadt, in der Ego wohnt. All dies sind erwartete Abweichungen zwischen dem Kernnetz auf der einen und dem gesamten Netzwerk auf der anderen Seite. Wie die Autoren selbst schreiben: „it remains the case that many studies of networks require focused, not representative samples of alters. A study of social support among the elderly in a retirement home, for example, might not be based on a representative sample of alters. In contrast, a study of how network alters affect the choice of a physician might usefully be based on a sample of alters from the total network.“ (McCarthy et al. 1997: 321f.) Das Problem dieser Methode liegt in der Auswahl der 50 Vornamen. Ihre Verteilung variiert mit dem Geschlecht, der regionalen Herkunft, der ethnisch-religiösen Zugehörigkeit und dem Alter (Gerhards 2003). Eine für die BRD repräsentative Liste von 50 Namen wird daher für ein bayrisches Dorf, für eine Jugendstudie oder für die Teilgruppe türkischer Befragter nicht geeignet sein. Dennoch ist dies mit Sicherheit ein Ansatz, der in Zukunft weiterverfolgt und auch für den deutschsprachigen Raum angepasst werden sollte.[10]

Während McCarthy et al. (1997: 321) „the total social network of a population“ abbilden wollen, sind Robert Kahn und Toni Antonucci (1980) an den wichtigen Be-

10 Weitere, ähnliche Vorschläge stammen von Freeman und Thompson (1989), Bernard et al. (1989) sowie Killworth et al. (1990), die v.a. dazu dienen, die Größe sozialer Netzwerke zu bestimmen.

zugspersonen und der emotionalen Verbundenheit zu diesen interessiert. Dazu wird den Befragten ein Zielscheibendiagramm mit drei konzentrischen Kreisen vorgegeben. Die sehr eng verbundenen Personen sollen in den ersten Kreis, die eng verbundenen Personen in den zweiten Kreis und weitere wichtige Personen, mit denen jedoch keine enge Verbindung besteht, in den dritten Kreis positioniert werden.[11] Diese meist in der Psychologie eingesetzte Methode wurde in Deutschland z.B. in der Berliner Altersstudie verwendet (Wagner, Schütze und Lang 1996). Zwischen den Netzwerken, die mit der von Kahn und Antonucci vorgeschlagenen Methode erhoben werden, und denen, die mit mehreren, auf Interaktionen basierenden Namensgeneratoren erhoben werden, bestehen sowohl Unterschiede als auch Gemeinsamkeiten: „So werden Verwandte, zu denen eine engere emotionale Beziehung oder eine Hilfebeziehung besteht, durch beide Vorgehensweisen erfaßt. Während ... [durch den Einsatz von Namensgeneratoren] ... Informationen über Familienmitglieder vorliegen, zu denen kein enger Kontakt besteht, werden ... [mit der Methode nach Kahn und Antonucci] auch emotional nahestehende Beziehungen zu entfernteren Verwandten ... und Nicht-Verwandten ... erhoben“ (Wagner, Schütze und Lang 1996: 305). Milardo (1992: 449) behauptet, dass zwischen der Zugehörigkeit zum Netzwerk der wichtigsten Beziehungen – in seiner Terminologie: der signifikant Anderen – und der jeweils aktuellen Häufigkeit von persönlichen Kontakten *kein* Zusammenhang besteht. Diese Aussage ist mit Sicherheit übertrieben. Wie seine eigenen Angaben zeigen, bestehen neben den soeben zitierten Unterschieden auch erhebliche Übereinstimmungen zwischen den Eigenschaften der Netzwerke, die mit den beiden Methoden erhoben werden. Zudem kann angenommen werden, dass die emotionale Nähe zwischen zwei Personen mit der Zeit verblasst, wenn die Beziehung nicht wenigstens hin und wieder durch persönliche Kontakte aktualisiert wird. Daher dürften sich die Unterschiede zwischen den Netzwerken, die nach McCallister und Fischer bzw. nach Kahn und Antonucci erhoben werden, in Grenzen halten.

In allen bisher diskutierten Erhebungsverfahren wird nach den Namen von Personen gefragt bzw. werden solche Namen vorgegeben. Diese Verfahren führen daher zu konkreten Personen, nach deren Eigenschaften und Beziehung gefragt werden kann. Ist man weniger an der Zusammensetzung und Struktur egozentrierter Netzwerke, sondern lediglich an ihrer sozialen Reichweite interessiert, kann der von Nan Lin entwickelte Positionsgenerator verwendet werden (Lin und Dumin 1986; Lin 1999: 476). Bei dieser Methode wird den Befragten eine Liste mit sozial bedeutsamen Positionen vorgelegt – z.B. Berufe, Ämter oder auch Wohnorte – und sie sollen angeben, ob sie Personen kennen, die die entsprechenden Positionen innehaben. Auf der Grundlage dieser Angaben kann die Reichweite eines Netzwerks bzw. der Zugang zu gesellschaftlichen Positionen bestimmt werden. Der Vorteil dieses Verfahrens besteht vor allem in dem vergleichsweise geringen Erhebungsaufwand. Wie Martin van der Gaag, Tom Snijders und Henk Flap (2004: 3f.) zu Recht anmerken, dient der Positionsgenerator in der von Lin vorgeschlagenen Form dazu, denjenigen Ausschnitt des sozialen Kapitals einer Person zu erheben, der vor allem für instrumentelle Ziele, wie dem Finden eines

11 Zu einer verwandten Methode, mit der das Netzwerk signifikant anderer erhoben werden soll, vergleiche Milardo (1992: 448ff.).

neuen Arbeitsplatzes, genutzt werden kann. Soll dagegen ein umfassendes Maß für die Sozialkapitalausstattung einer Person erhoben werden, bietet sich die Verwendung des von Snijders (1999; van der Gaag und Snijders 2004a) vorgeschlagenen Ressourcengenerators an. Dieses Instrument besteht aus den folgenden 35 Items (van der Gaag und Snijders 2004b: 212f.):

Do you know anyone who...

1 can repair a car, bike, etc.
2 owns a car
3 is handy repairing household equipment
4 can speak and write a foreign language
5 can work with a personal computer
6 can play an instrument
7 has knowledge of literature
8 has senior high school (VWO) education
9 has higher vocational (HBO) education
10 reads a professional journal
11 is active in a political party
12 owns shares for at least Dfl.10,000
13 works at the town hall
14 earns more than Dfl.5,000 monthly
15 owns a holiday home abroad
16 is sometimes in the opportunity to hire people
17 knows a lot about governmental regulations
18 has good contacts with a newspaper, radio- or TV station
19 knows about soccer
20 has knowledge about financial matters (taxes, subsidies)
21 can find a holiday job for a family member
22 can give advice concerning a conflict at work
23 can help when moving house (packing, lifting)
24 can help with small jobs around the house (carpenting, painting)
25 can do your shopping when you (and your household members) are ill
26 can give medical advice when you are dissatisfied with your doctor
27 can borrow you a large sum of money (Dfl.10,000)
28 can provide a place to stay for a week if you have to leave your house temporarily
29 can give advice concerning a conflict with family members
30 can discuss what political party you are going to vote for
31 can give advice on matters of law (problems with landlord, boss, municipality)
32 can give a good reference when you are applying for a job
33 can baby-sit for your children
34 can be talked to regarding important matters
35 can be visited socially

Weiterführende Skalenanalysen dieser Items zeigen, dass sie zumindest vier verschiedene, nur schwach miteinander korrelierte Aspekte individuellen sozialen Kapitals abbilden (van der Gaag und Snijders 2004a: 20): Das mit dem Prestige und der Bildung der Netzwerkpersonen assoziierte Kapital, politische und finanzielle Kenntnisse der

Netzwerkpersonen, spezielle Fertigkeiten der Netzwerkpersonen und die soziale Unterstützung durch das Netzwerk. Das Instrument erfasst somit sehr verschiedene Aspekte sozialen Kapitals, ist jedoch in seiner Anwendung sehr aufwändig und damit vergleichbar mit der Verwendung mehrerer Namensgeneratoren. Im Gegensatz zu diesem Verfahren liefert der Ressourcengenerator jedoch keine direkten Informationen über die Größe und Zusammensetzung des Netzwerkes oder über die Beziehungen zwischen den Mitgliedern eines Netzwerks. Daher kommen die Autoren selbst zu der Einschätzung (van der Gaag, Snijders und Flap 2004: 22):

„For the answers to any social capital question, conducting an elaborate Name Generator study with various name interpretation questions (that may include information about any social resource) always remains an option. It can be customised to provide answers to questions about any social capital dimension: alters, relationships, resources, and the availability of resources. Since it is the only method that identifies network members, it is also the only option when specific questions about the influence of network structure on social capital outcomes need to be studied. However, its costs remain considerable."

Abschließend möchte ich noch auf zwei Punkte hinweisen, die unabhängig von der gewählten Methode bei der Erhebung egozentrierter Netzwerke immer bedacht werden sollten. Der erste Punkt betrifft den Personenkreis, der auf die Namensgeneratoren genannt werden darf. Im Zusammenhang mit dem von Kecskes und mir entwickelten Instrument habe ich schon weiter oben darauf hingewiesen, dass wir nur „private" Kontakte zugelassen haben, d.h. weder professionelle Helfer noch Prominente oder Verstorbene, drei Personengruppen, die hin und wieder auf die Nachfrage nach weiteren wichtigen Personen genannt wurden.[12] Die Entscheidung für solche Einschränkungen des Personenkreises müssen von der Forschungsfrage abhängig gemacht werden. Im Zusammenhang anderer Fragestellungen könnten gerade Beziehungen zu professionellen Helfern oder Ratgebern von Interesse sein.

Zusätzliche Kriterien, die die Gruppe der möglichen Alteri weiter einschränken, sind denkbar. Aus den Erfahrungen anderer Netzwerkerhebungen wussten wir beispielsweise, dass ein Teil der Befragten die eigenen Kinder oder andere sehr junge Netzwerkmitglieder nennt. Für diese Netzwerkpersonen lassen sich viele der typischerweise nachfolgenden Fragen, z.B. nach dem Schulabschluss, nicht beantworten. Wir haben daher in unseren Studien immer ein Mindestalter für die Alteri festgelegt. In Abhängigkeit von der Fragestellung sind weitere Einschränkungen, z.B. nach dem Wohnort, der Nationalität, dem Familienstand oder dem Beruf denkbar.

Der zweite Punkt betrifft das Problem des nicht vorhandenen Netzwerks. Personen, die in dem vorgegebenen Zeitraum mit keiner anderen Person Kontakt hatten oder die keine Personen kennen, die ihr wichtig sind, haben ein Netzwerk der Größe null. Um diese Personen von denjenigen unterscheiden zu können, die den Netzwerkteil einer Befragung vollständig verweigern, sollte das Erhebungsinstrument beide Fälle explizit vorsehen. So könnte es neben der Zeile, in der die Namen notiert werden sollen, zwei Kästchen zum Ankreuzen geben, einerseits: „Habe in dieser Zeit mit niemandem ... (z.B. geredet)", andererseits: „keine Angabe". Geschieht dies nicht, kann nicht zwi-

12 Nach unseren Erfahrungen werden manchmal auch Tiere genannt, die wir jedoch ebenfalls nicht erfasst haben.

schen Verweigerern und Personen ohne Netzwerk unterschieden werden (vgl. auch Pappi und Melbeck 1988: 232).

Zusammenfassend kann festgehalten werden, dass die Instrumente von Burt auf der einen und von McCallister und Fischer auf der anderen Seite die Endpunkte eines Kontinuums markieren, auf dem sich interaktionsbasierte Netzwerkgeneratoren, die für allgemeine Bevölkerungsumfragen entwickelt wurden, einordnen lassen. Das von Kecskes und Wolf (1996) entwickelte Verfahren liegt zwischen diesen Endpolen und ist weniger aufwändig als das von McCallister und Fischer vorgeschlagene. Da es aus mehreren Namensgeneratoren besteht und die Anzahl der möglichen Namen nicht eingeschränkt wird, liefert es einen deutlich weniger stark zensierten Schätzer für die Größe der Netzwerke als die Drei-Freunde-Frage. Außerdem erlaubt dieses Verfahren die Analyse von Multiplexität. Damit bietet es sich als echte Alternative zu den Verfahren von Burt oder McCallister und Fischer an.

2. Die Merkmale der Netzwerkpersonen

Um das Netzwerk Egos genauer beschreiben zu können, werden Angaben zu den Netzwerkpersonen benötigt. Soll z.B. etwas über das Ausmaß des sozialen Kapitals ausgesagt werden, müssen die jeweils interessierenden Ressourcen der Alteri erfragt werden – etwa ihre berufliche Stellung, um daraus ihr Berufsprestige abzuleiten (Lin und Dumin 1986). Als Indikatoren der Verfügbarkeit sozialen Kapitals oder sozialer Unterstützungsleistungen werden oft die emotionale Verbundenheit zwischen Ego und Alter, ihre räumliche Nähe und/oder die Häufigkeit ihres Kontakts erhoben. Diese Indikatoren können auch zur Bestimmung der Beziehungsstärke herangezogen werden (Granovetter 1973: 61). Eine weitere, häufig untersuchte Frage bezieht sich auf das Ausmaß der Ähnlichkeit zwischen Ego und Alter, d.h. auf den Grad ihrer Homophilie (Lazarsfeld und Merton 1954; Wolf 1996) bzw. den Grad der differenziellen Assoziation (Reuband 1974). Die Ähnlichkeit von Netzwerkpersonen in Bezug auf sozioökonomische Merkmale wie die berufliche Stellung oder die Bildung wird oft als Gradmesser für die Offenheit von Gesellschaften angesehen (Friedrichs, Kecskes und Wolf 2002: 74ff.). In ähnlicher Weise wird die konfessionelle oder ethnische Homophilie sozialer Beziehungen als Maß für die soziale Bedeutung der Religions- bzw. ethnischen Zugehörigkeit angesehen (Baerveldt et al. 2004; Wolf 1995).

Die Frage, welche Merkmale der Netzwerkpersonen erfasst werden sollen, hängt also von der verfolgten Fragestellung ab und ist unabhängig von der gewählten Methode zur Generierung von Netzwerkpersonen. Typischerweise werden demographische und sozioökonomische Basismerkmale wie das Alter, das Geschlecht, die Bildung und die berufliche Stellung erhoben. Neben solchen personenbezogenen Merkmalen werden meist auch Eigenschaften der Beziehung zwischen Ego und Alter erfragt. So etwa die Kontakthäufigkeit, die Bekanntschaftsdauer, die Art der Beziehung oder die emotionale Nähe. Für diese Fragen zu den Netzwerkpersonen und die Beziehung zwischen Ego und Alter hat sich auch die Bezeichnung Namensinterpretatoren eingebürgert.

Im Zusammenhang mit der Erhebung der Namensinterpretatoren müssen zwei Entscheidungen getroffen werden: Erstens muss entschieden werden, ob die Merkmale

für alle oder nur für einen Teil der genannten Netzwerkpersonen erhoben werden sollen. Bei dem im General Social Survey von 1985 realisierten Verfahren nach Burt konnten die Befragten zwar beliebig viele Personen nennen, mit denen sie wichtige Dinge besprochen hatten, die Namensinterpretatoren wurden jedoch nur für die ersten fünf genannten Personen erhoben. Diese Auswahl führt vermutlich zu einer Selektion eher stärkerer Beziehungen – allerdings muss erwähnt werden, dass überhaupt nur 5,5 Prozent der Befragten im GSS mehr als fünf Netzwerkpersonen nannten.

Die Dringlichkeit, aus der Gruppe der Netzwerkpersonen eine Auswahl zu treffen, steigt mit der Anzahl der Namensgeneratoren. McCallister und Fischer haben zwar einige wenige Merkmale für alle Alteri, die meisten Merkmale jedoch nur für eine Auswahl von höchstens fünf Personen erhoben. Bei dieser Auswahl handelt es sich um diejenigen Alteri, die auf bestimmte, vorher festgelegte Namensgeneratoren jeweils an erster Stelle genannt werden (vgl. Fischer 1982: 332). Bei dem von ZUMA durchgeführten Methodenexperiment zu sozialen Netzwerken (s.u.) wurde als Alternative zu dieser Vorgehensweise die Auswahl der fünf Netzwerkpersonen, für die weitere Informationen gesammelt werden sollten, den Befragten überlassen. Die entsprechende Frage lautete: „Wenn Sie einmal an alle Personen denken, die Sie mir auf die einzelnen Fragen bisher genannt haben: Welche fünf von allen eben genannten Personen sind für Sie persönlich die wichtigsten?“ (Pfenning und Pfenning 1987: 65). Beide Verfahrensweisen führen zu einer selektiven Auswahl der Netzwerkpersonen, bei der enge Beziehungen bevorzugt werden (vgl. Fischer 1982: 145 und Anmerkung 16 auf S. 391). Eine andere Strategie besteht darin, aus der Menge der genannten Netzwerkpersonen eine Zufallsauswahl zu treffen. Der Vorteil einer zufälligen Auswahl der Alteri liegt auf der Hand: Einerseits wird wie bei den anderen Selektionsverfahren auch die Dauer der Datenerhebung verkürzt, andererseits können auf der Basis der ausgewählten Netzwerkpersonen Aussagen über das *gesamte* mit den entsprechenden Namensgeneratoren erhobene Netzwerk gemacht werden. In den von meinen Kollegen und mir durchgeführten Studien wurden zwei verschiedene Zufallsverfahren eingesetzt. Wolf (2003) hat zunächst alle Personen, die auf den verschiedenen Namensgeneratoren genannt wurden, auflisten lassen. Wurden mehr als zehn Netzwerkpersonen genannt, wurden mit Hilfe einer in den Fragebogen eingedruckten, jeweils einzigartigen Zufallszahlentabelle zehn Alteri ausgewählt. Da eine Analyse der Daten zeigte, dass dieses Verfahren nicht von allen Interviewern korrekt verwendet wurde, haben Friedrichs, Kecskes und Wolf (2002) eine einfachere Auswahlmethode verwendet.[13] Wiederum sollen bei mehr als zehn Alteri zehn zufällig ausgewählt werden. Dazu verwendet der Interviewer ein Kartenspiel mit den Nummern 1 bis zur Anzahl der genannten Personen. Die Karten werden gemischt und die befragte Person darf – ohne Zurücklegen – zehn Karten ziehen. Ob dieses Verfahren genauer befolgt und tatsächlich eine Zufallsauswahl aus den genannten Personen gezogen wurde, muss sich noch in weiteren Analysen erweisen. Sicher ist, dass jedes Auswahlverfahren, das dem Interviewer obliegt, eine mögliche Fehlerquelle darstellt. Abhilfe kann hier eine automatisierte Auswahl schaffen, wie sie sich durch den Einsatz eines computergestützten Befragungsverfahrens realisieren lässt.

13 Die Idee zu diesem Verfahren stammt von Andrea Albrecht.

Die zweite Frage, die sich im Zusammenhang mit der Erhebung der Angaben zu den Netzwerkpersonen stellt, ist, ob diese Angaben personen- oder merkmalsbezogen erhoben werden sollen. Bei der personenbezogenen Variante werden zunächst alle Angaben zur ersten Person auf der Liste erfragt, dann zur zweiten, dann zur dritten usw. Bei der merkmalsbezogenen Fassung wird zunächst das erste Merkmal für alle Personen erfasst, dann das zweite Merkmal, dann das dritte Merkmal usw. Im General Social Survey 1985, in der Wahlstudie 1987 sowie in den von meinen Kollegen und mir (Kecskes und Wolf 1996; Friedrichs und Blasius 2000; Friedrichs, Kecskes und Wolf 2002; Wolf 2003) durchgeführten Studien wurde merkmalsbezogen vorgegangen. Im Gegensatz dazu wurden die Netzwerkdaten im ALLBUS 1980, 1990 und 2000 personenbezogen erhoben. Ein neueres Methodenexperiment zeigt jedoch, dass die personenbezogene Vorgehensweise zumindest bei telefonischer Erhebung zuverlässigere Daten liefert (Kogovšek et al. 2002; näheres weiter unten in Abschnitt III.3).

3. Fragen zur Dichte

Ein zentrales Strukturmerkmal egozentrierter Netzwerke ist ihre Dichte. Zur Berechnung dieses Merkmals werden Angaben über die Beziehungen der Alteri untereinander benötigt.[14] Bevor man überlegt, welcher Aspekt der Beziehung zwischen den Alteri in welcher Art erhoben werden soll, sollte man sich bewusst machen, wie viele solcher Beziehungen existieren. Umfasst ein Netzwerk P Netzwerkpersonen, können insgesamt P (P–1) / 2 Beziehungen zwischen diesen Personen bestehen. Bei der in der Studie von Fischer und im GSS vorgenommenen Auswahl von fünf Alteri bedeutet das, dass Fragen zu insgesamt zehn Beziehungen zwischen den Alteri gestellt werden müssen. Bei zehn Alteri hätte dies schon eine Nachfrage zu 45 Beziehungen zur Folge und bei 20 Alteri wären es 190 potenzielle Beziehungen. Hat man sich dazu entschieden, Informationen für alle oder eine große Zahl von Alteri zu erheben, so wird man sich wahrscheinlich bei den Fragen zur Dichte auf eine Auswahl der Alteri beschränken. In den von meinen Kollegen und mir bisher durchgeführten Studien haben wir uns auf fünf Alteri beschränkt. Diese wurden zufällig aus der Menge der genannten Netzwerkpersonen ausgewählt.[15]

---

14 Auf der Basis dieser Angaben können weitere, bisher jedoch kaum verwendete Strukturmerkmale berechnet werden (Marsden 2002). Außerdem lassen sich aus diesen Angaben die Triaden, aus denen die Netzwerke zusammengesetzt sind, bestimmen; letztere werden z.B. von Pappi und Melbeck (1988) analysiert.

15 In einem persönlichen Gespräch hat Wolfgang Sodeur zu Recht darauf hingewiesen, dass es sich bei dieser Vorgehensweise im Zusammenhang mit der Messung der Netzwerkdichte um eine Klumpenauswahl handelt, da die angestrebte Erhebungseinheit die Beziehungen zwischen den Alteri und nicht die Alteri selbst sind. Eine einfache Zufallsstichprobe würde sich nicht auf die Alteri, sondern auf die Beziehungen zwischen diesen beziehen und eine Auswahl aus der Menge der potenziellen Beziehungen treffen. Im Rahmen traditioneller Papierfragebögen ist eine entsprechende Zufallsauswahl jedoch kaum zu realisieren. Würde stattdessen computerunterstützt befragt, wäre es ein Leichtes, eine vorgegebene Anzahl von Beziehungen durch eine geeignete Routine auswählen zu lassen.

Die Fragen zur Beziehung zwischen den Alteri zielen typischerweise darauf ab, ob sich die Alteri kennen oder nicht kennen bzw. auf das Ausmaß der Beziehungsstärke. So fragt Fischer (1982: 145) danach, ob sich die Alteri gut kennen oder nicht. Im GSS 1985 wird die gewünschte Information in einer zweistufigen Abfrage erhoben: Zunächst wird für ein Paar von Alteri gefragt, ob sie sich vollkommen unbekannt sind. Ist dies nicht der Fall, wird weiter gefragt, ob sich die beiden Personen besonders nahe stehen (Davis und Smith 1987: 306). In der dritten Welle der Bundestagsbefragung 1987 wurde eine differenzierte Frage zur Beziehung zwischen den Alteri gestellt: „Stehen diese sich einander sehr nahe (1), nahe (2), einander nicht nahe (3), oder kennen sie sich überhaupt nicht (4)?“ (ZA 1987: Fragebogen zur dritten Welle, S. 12). In den von mir und Kollegen durchgeführten Studien haben wir ähnliche dreistufige Antwortskalen verwendet (Kecskes und Wolf 1996: 196; Friedrichs, Kecskes und Wolf 2002: 240; Wolf 2003: 215). Welchen Einfluss die Verwendung unterschiedlich stark differenzierter Antwortskalen auf die Qualität der Dichtemessung hat, ist bisher nicht untersucht worden.

In der Praxis sind der Analyse struktureller Eigenschaften egozentrierter Netzwerke enge Grenzen gesetzt, da sie typischerweise durch eine verhältnismäßig geringe Netzwerkgröße gekennzeichnet sind. Liegt das Forschungsinteresse primär in der Beschreibung und Analyse der Struktur sozialer Netzwerke eines „Bevölkerungsdurchschnitts“, also keiner abgrenzbaren sozialen Gruppe, deren Gesamtnetz erhoben werden könnte, müssen (wenige) Befragte zu verhältnismäßig vielen Netzwerkpersonen befragt werden. Eine entsprechende Arbeit wurde von McCarthy (2002) vorgelegt, der insgesamt 46 Personen nach 60 Alteri und den 1770 Beziehungen zwischen diesen Alteri gefragt hat.

III. Die Qualität egozentrierter Netzwerkdaten

Noch stärker als bei anderen sozialwissenschaftlichen Daten stellt sich die Frage nach der Validität und Reliabilität in Bezug auf Daten, die mithilfe eines oder mehrerer Namensgeneratoren erhoben werden (vgl. auch Marsden 1990: 450ff.): Wie stabil ist die Menge der erhobenen Netzwerkpersonen? Wie stabil sind die Auskünfte, die Ego über eine mit ihm verbundene Netzwerkperson gibt? Und stimmen die von Ego gemachten Angaben mit den tatsächlichen Eigenschaften Alters überein?

1. Die Qualität der Netzwerkgeneratoren

Eine wichtige deutsche Quelle für die Beurteilung der Qualität von Netzwerkdaten ist das ZUMA-Methodenforschungsprojekt „Egozentrierte Netzwerke in Massenumfragen“ (Hoffmeyer-Zlotnik 1987; Mohler und Pfenning 1987; Schneid 1987). Im Rahmen dieser Studie wurden 720 Personen im Abstand von drei Wochen unter Verwendung verschiedener Namensgeneratoren nach ihren Netzwerkpersonen befragt. Auf der Basis dieser Studie untersucht Astrid Pfenning (1988; Pfenning und Pfenning 1987) die Zuverlässigkeit der Erhebungsinstrumente von McCallister und Fischer sowie von Burt. Für die verschiedenen Netzwerkgeneratoren nach Fischer zeigen sich sehr unterschied-

liche Test-Retest Reliabilitäten (Pfenning 1988: 54). Die höchste Zuverlässigkeit zeigte sich bei den Generatoren „Hobbies besprechen", „persönliche Dinge besprechen" und „gemeinsame Aktivitäten wie z.B. Ausgehen" – auf die zuletzt genannte Frage wurden zudem im Schnitt die meisten Netzwerkpersonen genannt. Die Reliabilität der Netzwerkgröße ist im Instrument von McCallister und Fischer mit einer Test-Retest-Korrelation von 0,77 recht hoch; sie liegt beim Burt-Generator bei lediglich 0,65. Das bessere Abschneiden des umfangreicheren Instruments erklärt Pfenning (1988: 55) so: „wenn eine Person bei einem Item ‚vergessen' wird, wird sie wahrscheinlich bei einem anderen Item doch noch genannt" (Pfenning 1988: 55).

Walter Bien, Jan Marbach und Franz Neyer (1991) berichten von einer Test-Retest-Studie, in der egozentrierte Netzwerke mit 12 Namensgeneratoren erhoben wurden. Im Aggregat zeigte sich in beiden Studien dieselbe Verteilung der Netzwerkgröße. Ebenso ergab sich eine sehr hohe Stabilität der durchschnittlich auf die 12 verschiedenen Namensgeneratoren genannten Personen (vgl. ebenda: 84f.). Auf der Individualebene ergaben sich erwartungsgemäß größere Abweichungen. Nur etwa ein Drittel der Befragten nannte zu beiden Zeitpunkten exakt dieselbe Anzahl von Netzwerkpersonen. Im Durchschnitt wich die Netzwerkgröße zwischen der ersten und der zweiten Erhebung um 1,6 Personen ab (ebenda: 81, Tab. 4; eigene Berechnung). Dennoch wurden drei Viertel aller Netzwerkpersonen zu beiden Erhebungszeitpunkten genannt. Insgesamt kann somit von einer befriedigenden Zuverlässigkeit von Namensgeneratoren ausgegangen werden. Wobei nicht übersehen werden kann, dass es noch zu wenige Studien gibt, um diese Frage abschließend zu beurteilen. Insbesondere sollte genauer untersucht werden, welche Namensgeneratoren besonders gute bzw. schlechte Ergebnisse liefern und in wieweit dies mit der Erhebungsmethode variiert; ein Punkt, auf den ich weiter unten noch einmal genauer eingehe.

2. Die Qualität der Angaben Egos über Alter und die Beziehung

Vorliegende Untersuchungen zeigen, dass Egos Angaben zu den soziodemographischen Merkmalen Alters als gültig anzusehen sind. Laumann (1969: 60) findet in seinen Vergleichen von Egos Angaben über Alter und Alters eigenen Angaben hohe Übereinstimmungen bezüglich des Geburtsjahres ($r \ni 0{,}98$), des sozioökonomischen Status ($r \ni 0{,}89$), der Bildung ($r \ni 0{,}84$) und der Religionszugehörigkeit ($\varphi \ni 0{,}98$). Weniger gut ist die Übereinstimmung der Angaben zur Parteipräferenz Alters, die sich jedoch noch zu 73,4 Prozent miteinander decken (Laumann 1969: 61, Tab. 4; eigene Berechnung). Wird berücksichtigt, dass sowohl Egos als auch Alters Angaben immer Messfehler aufweisen, sind diese Übereinstimmungen als recht hoch einzuschätzen. In einer Replikationsuntersuchung anhand der Altneustadt-Studie finden Franz Pappi und Gunter Wolf (1984) ein geringeres Ausmaß an Übereinstimmung. Zudem weisen sie nach, dass die Angabe zur Parteipräferenz Alters durch Ego systematisch in Richtung auf die eigene Parteipräferenz verzerrt ist.

Eine Untersuchung der Test-Retest-Reliabilität egozentrierter Netzwerkdaten findet sich auf Basis der Daten des ZUMA-Methodenexperiments zu egozentrierten Netzwerken wiederum bei Pfenning (1988). Ihre Ergebnisse verweisen auf eine erstaunliche

Tabelle 2: Reliabilitäten von Eigenangaben der Befragten und Angaben der Befragten über ihre Netzwerkpersonen im Vergleich (Test-Retest-Ergebnisse)

	Ego über Alter (N = 123)		Ego über Ego (N = 86)	
	% übereinstimmende Angaben	Reliabilität[b]	% übereinstimmende Angaben	Reliabilität[b]
		Demographische Angaben		
Geschlecht	94	$\varphi \ni 0{,}88$	95	$\varphi \ni 0{,}90$
Geburtsjahr	68	$r \ni 0{,}96$	96	$r \ni 0{,}99$
Schulabschluss (7)[a]	77	$V \ni 0{,}76$	86	$V \ni 0{,}81$
Familienstand (9)	75	$V \ni 0{,}55$	98	$V \ni 0{,}97$
Haushaltsgröße	71	$r \ni 0{,}72$	96	$r \ni 0{,}97$
Konfession (6)	68	$V \ni 0{,}69$	83	$V \ni 0{,}68$
		Einstellungen		
Religiosität (10)	29	$r \ni 0{,}34$	35	$r \ni 0{,}72$
Parteineigung (2)	84	$\varphi \ni 0{,}67$	74	$\varphi \ni 0{,}46$
Parteiwahl (5)	85	$V \ni 0{,}83$	97	$V \ni 0{,}93$
Autoritarismus				
Disziplin (5)	42	$r = 0{,}55$	43	$r = 0{,}61$
Autorität (5)	35	$r = 0{,}56$	43	$r = 0{,}48$
		Erziehungsziele		
Kritikfähigkeit (7)	41	$r = 0{,}29$	30	$r = 0{,}23$
Selbstvertrauen (7)	34	$r = 0{,}51$	45	$r = 0{,}39$
Fleiß (7)	34	$r = 0{,}57$	39	$r = 0{,}25$
Schulleistungen (7)	34	$r = 0{,}46$	49	$r = 0{,}47$

a In Klammern ist die Anzahl der Kategorien wiedergegeben, mit denen das jeweilige Merkmal erfasst wurde.
b Es wurden folgende Koeffizienten verwendet: Cramer's V (V); Phi (φ); Produkt-Moment-Korrelation (r).

Quelle: Pfenning (1988: 67).

Parallelität der Reliabilität von Egos Angaben über sich und seinen Angaben über die Alteri (s. *Tabelle 2*). „Für fast alle Merkmale gilt: Je höher die Reliabilität der Eigenangaben, um so höher die Reliabilität der Informantenangaben über Netzwerkpersonen" (Pfenning 1988: 72). Liegt ein zuverlässiges Messinstrument für die Erhebung von Egos Merkmalen vor, erzielt dieses Instrument mit einiger Sicherheit auch reliable Informationen über die Netzwerkpersonen.[16]

Bisher wurden lediglich die Zuverlässigkeit und Gültigkeit von Egos Angaben zu den Eigenschaften der Alteri betrachtet. Die Validität relationaler Merkmale wird von Muriel Hammer (1984: 345ff.) untersucht. Sie kann auf Daten zurückgreifen, die einen direkten Vergleich von Egos und Alters Angaben erlauben. Es stellt sich heraus, dass die Übereinstimmung der Angaben bezüglich der Kontakthäufigkeit, der Dauer der Beziehung, des Verwandtschaftsgrades und der Stärke der Beziehung sehr hoch ist,

16 Ein Gutachter hat zu Recht darauf verwiesen, dass der von Pfenning verwendete Koeffizient Cramer's V sich nicht zur Bestimmung der Reliabilität eignet. Da Pfenning keine Test-Retest-Verteilungen der Merkmale präsentiert, können keine anderen Koeffizienten berechnet werden. M.E. lässt die Parallelität der Ergebnisse für Egos Eigenschaften und die Eigenschaften der Alteri jedoch auf eine erhebliche Konsistenz der Angaben schließen.

je nach Merkmal liegt sie zwischen 79 und 100 Prozent. Sie kommt deshalb zu einer positiven Beurteilung der Qualität von Netzwerkdaten: „The high levels of concordance between different respondents about the same relationships suggest that reported relationships do exist, and that they are correctly reported in terms of their durations, their frequencies of contact, and their kinship status" (Hammer 1984: 365).[17] Wie Michael Schenk et al. (1992: 102ff.) zeigen, variiert die Genauigkeit der Proxi-Angaben Egos über die Alteri *nicht* mit der Dauer, der Kontakthäufigkeit, der Intensität oder dem Kontext der Beziehung (ähnlich Pfenning, Pfenning und Mohler 1991: 98). Im Gegensatz dazu zeigen Peter Marsdens (1993) Analysen, dass die Reliabilität von Angaben zur Zusammensetzung egozentrierter Netzwerke für ein „Kernnetzwerk" größer ist als für ein weitergefasstes Netzwerk. Wie sich dieser Widerspruch auflösen lässt, muss hier offen bleiben.

Pfenning untersucht die Reliabilität relationaler Merkmale. Hier zeigt sich eine hohe Zuverlässigkeit der Angaben zur Beziehungsart und zur räumlichen Entfernung zwischen Egos und Alters Wohnort (Pfenning 1988: 78ff.). Dagegen sind die Messungen der Bekanntschaftsdauer, der Kontakthäufigkeit und der Intensität der Beziehung weniger zuverlässig. Nach einer eingehenden semantischen Analyse der Items kommt Pfenning (1988: 82) allerdings zu dem Schluss, dass diese geringe Reliabilität aus der Frageformulierung resultiert, und nicht aus der mangelnden Fähigkeit der Befragten, zuverlässig über relationale Merkmale zu informieren. Zu einem ähnlichen Ergebnis kommen Tina Kogovšek et al. (2002), deren Studie zur Abhängigkeit der Datenqualität von der Erhebungsmethode weiter unten vorgestellt wird.

Zusammenfassend kann festgehalten werden, dass Egos Angaben über soziodemographische und sozioökonomische Merkmale der Alteri als relativ zuverlässig und gültig angesehen werden können. Dasselbe gilt für Merkmale der Beziehungen. In Bezug auf Einstellungen dagegen „soll man sich der großen systematischen Verzerrung bewusst sein und nicht unterstellen, daß man mehr erhoben hat als eben die Perzeption von Aspekten der primären Umwelt" (Pappi und Wolf 1984: 298; vgl. auch die oben in *Tabelle 2* berichteten Ergebnisse zu Reliabilität der Erziehungsziele der Alteri). Ein letzter Punkt bezieht sich auf die Qualität von Merkmalen, die sich auf die Zusammensetzung egozentrierter Netzwerke beziehen – z.B. das Durchschnittsalter oder der Anteil Verwandter. Werden die Merkmale der Alteri zu Merkmalen der Netzwerke aggregiert, steigt die Zuverlässigkeit dieser Netzwerkeigenschaften mit der Anzahl der erhobenen Netzwerkpersonen an (Marsden 1993). Bei Netzwerken von etwa 7 bis 8 Personen erreichen die meisten von Marsden untersuchten Maße der Komposition egozentrierter Netzwerke eine befriedigende Zuverlässigkeit.

17 Im Gegensatz zu Hammer hat eine viel zitierte Studie von Killworth und Bernard (1976) erhebliche Zweifel an der Qualität egozentrierter Netzwerkdaten ausgelöst. Sie vertreten die Auffassung, dass „people simply do not know, with any degree of accuracy, with whom they communicate" (Killworth und Bernard 1976: 283). Diese Schlussfolgerung lässt sich m.E. jedoch aufgrund methodischer Mängel der Studie nicht halten; zudem lag der Fokus der Untersuchung von Killworth und Bernard nicht auf dem „Kernnetzwerk", das wir typischerweise mit den Methoden der egozentrierten Netzwerke erheben (vgl. genauer Wolf 1996: 72f.).

3. Erhebungsmethode und Datenqualität

Die Qualität von Umfragedaten variiert bekanntlich auch mit der Erhebungsmethode (Skjåk und Harkness 2003). Egozentrierte Netzwerkdaten wurden bisher in den allermeisten Fällen in persönlich-mündlichen Interviews erhoben – dies gilt auch für alle oben zitierten Studien zur Reliabilität und Validität von Netzwerkdaten.[18] Die Frage, inwieweit die Qualität egozentrierter Netzwerkdaten mit der Methode der Datengewinnung variiert, ist daher kaum untersucht worden. Seit kurzem liegt jedoch eine Arbeit vor, die dieses Problem ausführlich behandelt (Kogovšek et al. 2002). Die Autoren dieser Studie untersuchen auf der Basis eines Multitrait-Multimethod-Ansatzes die Reliabilität und Validität verschiedener Eigenschaften egozentrierter Netzwerkdaten in Abhängigkeit von der Erhebungsmethode. Die untersuchten Eigenschaften sind die durchschnittliche Häufigkeit des Kontakts, die subjektive Nähe, die Wichtigkeit der Beziehung und die Häufigkeit, mit der Ego sich über Alter ärgert. Diese Merkmale wurden auf drei verschiedene Arten erhoben. Zum einen im Rahmen einer persönlich-mündlichen Befragung, zum anderen mittels telefonischer Befragung. In beiden Fällen handelte es sich um computerunterstützte Interviews. Im Fall der Telefoninterviews wurde noch einmal danach unterschieden, ob die Informationen zu den Alteri (die Namensinterpretatoren) merkmalsbezogen oder personenbezogen erhoben wurde. In der ersten Variante wird die erste Frage zunächst für jede Netzwerkperson gestellt, erst dann wird die zweite Frage gestellt usw. Bei der zweiten Variante werden zunächst alle Fragen zur ersten Netzwerkperson gestellt, dann zur zweiten Netzwerkperson usw.

Wie die in *Tabelle 3* wiedergegebenen Ergebnisse dieses Methodenexperiments zeigen, liefern die Telefoninterviews bessere Validitätswerte als die persönlich-mündlichen Befragungen – und zwar unabhängig davon, ob die Angaben zu den Alteri personen- oder merkmalsbezogen erhoben werden. Als Ursache verweisen Kogovšek et al. (2002: 10) auf die größere Anonymität telefonischer Umfragen. Der Vergleich der beiden telefonischen Befragungen macht deutlich, dass mit der personenbezogenen Abfrage der Namensinterpretatoren eine höhere Gültigkeit erzielt wird als mit dem merkmalsbezogenen Vorgehen.

Zu ähnlichen Ergebnissen kommen die Autoren hinsichtlich der Zuverlässigkeit der Messungen. Auch in Bezug auf dieses Gütekriterium schneidet die telefonische Erhebung der Daten bei einem personenbezogenen Vorgehen am besten ab. Wie ein Vergleich der drei verschiedenen Versuchsbedingungen in *Tabelle 3* zeigt, wird die Reliabilität weniger durch die Methode der Datenerhebung als durch den Modus der Abfrage bestimmt. Bei der personenbezogenen Abfrage der Namensinterpretatoren ist „the salient reference frame ... the current alter, and it is more likely that the respondent would make fewer errors“, wenn dagegen merkmalsbezogen erhoben wird, „it is likely that the current question and also part of the named egocentered network are ‚fighting‘ for the respondent's working memory space, thus increasing the cognitive burden and making errors more likely“ (Kogovšek et al. 2002: 11).

18 Allerdings existieren auch einige Studien, in denen egozentrierte Netzwerke schriftlich (Rippl 1995; Berger und Kriwy 2004) oder telefonisch (Robicheaux 2003) erhoben wurden. Für einen Fragebogen zur schriftlichen Erhebung egozentrierter Netzwerke vergleiche Burt (1998).

Tabelle 3: Reliabilität und Validität verschiedener Netzwerkmerkmale aus einem Multitrait-multimethod-Experiment

Methode	Kontakt[a]	Nähe[b]	Wichtigkeit[c]	Ärger[d]
		Reliabilität		
Face-to-face/personenbezogen	0,96	0,83	0,81	0,79
Telefon/personenbezogen	0,83	0,91	0,92	0,86
Telefon/merkmalsbezogen	0,84	0,83	0,86	0,80
		Validität		
Face-to-face/personenbezogen	0,94	0,86	0,89	0,87
Telefon/personenbezogen	0,96	0,95	0,97	0,95
Telefon/merkmalsbezogen	0,95	0,92	0,94	0,93

a Kontakt: Wie oft haben Sie mit dieser Person Kontakt – einerlei ob persönlich, schriftlich, telefonisch oder über das Internet? Antwort: Jeden Tag (1) bis seltener als einmal im Jahr (6).
b Nähe: Wie nahe fühlen Sie sich dieser Person? Nicht nahe (1) bis sehr nahe (5).
c Wichtigkeit: Wie wichtig ist diese Person in Ihrem Leben? Nicht wichtig (1) bis sehr wichtig (5).
d Ärger: Wie oft ärgern Sie sich über diese Person? Oft (1) bis nie (4).

Quelle: Kogovšek et al. (2002: 17).

Insgesamt zeigt die Studie überzeugend, dass die verhältnismäßig komplizierten Instrumente, die zur Erhebung egozentrierter Netzwerkdaten verwendet werden, auch am Telefon eingesetzt werden können. Zudem zeigt sich, dass diese schnellere und preiswertere Erhebungsmethode mindestens eine ähnlich gute, wahrscheinlich sogar eine bessere Datenqualität erzeugt als das traditionelle persönlich-mündliche Interview.

4. Die Abhängigkeit der Netzwerkgröße vom Interviewerverhalten

Bei der standardisierten Befragung, insbesondere bei der persönlich-mündlichen Befragung, kommt dem Verhalten des Interviewers eine zentrale Bedeutung zu. Während sich Fragereihenfolge und -formulierung sowie Antwortmöglichkeiten durch ein geeignetes Instrument standardisieren lassen, kann das Verhalten des Interviewers oder der Interviewerin nur bedingt kontrolliert werden (Groves 1989: 359; Scheuch 1973: 95). Darüber hinaus variiert die Größe des Interviewereffekts mit den Eigenschaften der verwendeten Items: „If an item is sensitive, non-factual, difficult or open, the question leaves more room for interpretation and is more difficult to answer, meaning that the respondent is more likely to rely on explicit or implicit help from the interviewer. These items therefore produce larger interviewer effects, especially if several of these item properties are present“ (Schnell und Kreuter 2003: 24f.). Im Vergleich zu anderen Fragen handelt es sich bei Namensgeneratoren um offene, relativ schwierige und für eine Reihe von Befragten eventuell auch um heikle Fragen. Namensgeneratoren stellen auch erhöhte Ansprüche an die Interviewer: Die Fragen müssen vorgelesen, genannte Namen notiert und es muss gegebenenfalls nachgefragt werden, ob weitere Personen zu nennen sind. Es ist demnach zu befürchten, dass die Anzahl der erhobenen Namen, d.h. die Netzwerkgröße, vom Verhalten der Interviewer abhängig ist. In der Tat berichtet Marsden (2003) für den General Social Survey von 1998, in dem nach den fünf

besten Freunden gefragt wurde, dass die Anzahl der genannten Personen je nach Interviewer variiert.[19] Im Rahmen einer Mehrebenenanalyse mit den Befragten als unterste und den Interviewern als Aggregatebene berichtet Marsden (2003: 6) eine bivariate Intraklassenkorrelation von 0,15, die sich auch nach Kontrolle mehrerer Befragten- und Interviewermerkmale nicht stark reduziert.

Werden mehrere Namensgeneratoren verwendet, kann mit noch größeren Interviewereffekten gerechnet werden. Tatsächlich berichtet Theo van Tilburg (1998: 315) für ein Instrument mit sieben Namensgeneratoren eine Intraklassenkorrelation von 0,21. Unter Berücksichtigung einer Vielzahl von Eigenschaften der Interviewer und der Interviewsituation konnte dieser Wert auf 0,15 reduziert werden (ebenda: 322). Ein für meine weitere Argumentation wichtiges Ergebnis von Tilburgs Analysen besteht darin, dass Interviewer mit Berufserfahrung aus anderen Projekten deutlich kleinere Netzwerke generierten als weniger erfahrene Interviewer.

Tabelle 4 listet die bivariaten Intraklassenkorrelationen für fünf verschiedene Studien auf, die am Forschungsinstitut für Soziologie der Universität zu Köln von meinen Kollegen und mir durchgeführt wurden (zu den Einzelheiten dieser Studien vgl. *Tabelle A1* im Anhang). Auffällig ist die breite Variation der Koeffizienten. In unserer ersten Studie (RSN: Kecskes und Wolf 1996) lässt sich kaum ein Effekt des Interviewerverhaltens auf die Anzahl der Netzwerkpersonen nachweisen, während die Netzwerkgröße in unserer jüngsten Studie massiv vom Interviewer abhängt (SOWEM: Friedrichs, Kecskes und Wolf 2002). Dabei hatten wir in dieser Studie die Anzahl der Namensgeneratoren im Vergleich zur ersten um zwei reduziert und die Zahl der Netzwerkpersonen, für die Informationen zu erheben waren, auf zehn begrenzt. Diese Maßnahmen waren u.a. ergriffen worden, um den Anreiz für Interviewer, möglichst kleine Netzwerke zu generieren, zu senken. Ein Vergleich zwischen RSN und SUKUG (Wolf 2003) zeigt ebenfalls, dass die Reduktion der Namensgeneratoren und eine Begrenzung der Alterizahl die Interviewereffekte nicht verringert.

Eine alternative Erklärung für die großen Unterschiede des Interviewereffekts auf die Netzwerkgröße kann im Ausmaß der Interviewerkontrolle gesehen werden. Während die komplette Feldsteuerung in unserer ersten Studie von uns selbst geleistet wurde, wurde die Datenerhebung in der SOWEM-Studie einem kommerziellen Institut übertragen. Allerdings waren wir auch in diesem Fall bei der Interviewerschulung anwesend und konnten teilweise noch im Feld korrigierend eingreifen. Trotz dieser ungewöhnlichen Möglichkeit, die Feldsteuerung zu beeinflussen, oblag die Interviewersteuerung und -kontrolle letztlich dem Institut. Im Gegensatz zu einem selbst organisierten und geleiteten Feld entzieht sich somit vieles der direkten Kontrolle. Ein weiterer Unterschied zwischen einer universitär und kommerziell durchgeführten Erhebung betrifft die Auswahl der Interviewer. Es kann mit Sicherheit davon ausgegangen werden, dass die Interviewer für die kommerziellen Institute in der Regel deutlich mehr Erfahrung mit dieser Tätigkeit haben als Interviewer, die an der Universität angeworben werden. Erfahrene Interviewer wissen aber, wie sie Interviews „abkürzen“ können ohne zu fälschen oder in einer anderen Weise ihre vertraglichen Pflichten zu verletzen. So mag

19 Der Namensgenerator im GSS von 1998 lautet: „Many people have some good friends they feel close to. Who are your good friends (other than your spouse)? Just tell me their first names“ (Marsden 2003: 4).

Tabelle 4: Stärke des Interviewereffekts auf die Größe der Netzwerke[a]

	RSN	LBW-D	LBW-T	SUKUG-A	SUKUG-B	SOWEM
∅ Netzwerkgröße	9,3	4,6	1,8	9,1	8,3	7,6
Generatoren Zahl	8	3	3	6	6	6
Zahl der Alteri[b]	alle	alle	alle	10	10	10
Datenerhebung	Uni	Uni	Uni	Uni	kommerz.	kommerz.
Interklassenkorr.	0,05	0,11	0,20	0,08	0,14	0,28

a Die Studiennamen lauten (für nähere Angaben zu den Studien vergleiche Tabelle A1 im Anhang):
RSN: Religion und soziale Netzwerke (Kecskes und Wolf 1996);
LBW: Leben in benachteiligten Wohngebieten, -D: Deutsche, -T: Türken (Friedrichs und Blasius 2000);
SUKUG: Soziale Ungleichheit, Krankheit und Gesundheit, -A: erste Teilstichprobe, -B: zweite Teilstichprobe (Wolf 2003);
SOWEM: Sozialer Wandel einer Mittelstadt (Friedrichs, Kecskes und Wolf 2002).

b In allen Studien wurden alle genannten Alteri aufgeschrieben und aufgelistet. Allerdings wurden nur in den Studien RSN und LBW auch Namensinterpretatoren für alle Netzwerkpersonen gestellt. In den Studien SOWEM und SUKUG war diese Zahl auf 10 beschränkt.

z.B. das Ausmaß des Nachfassens bei einem erfahrenen Interviewer weniger stark ausgeprägt sein, insbesondere in den Fällen, in denen zusätzlich erhobene Informationen weitere Arbeit nach sich zieht, für die der Interviewer nicht bezahlt wird.[20]

Einen unabhängigen Hinweis für den Unterschied zwischen kommerziellen und universitär organisierten Datenerhebungen liefert die Studie SUKUG (Wolf 2003). Die Daten dieser Untersuchung wurden in zwei Wellen erhoben: In der ersten Welle wurde die gesamte Feldsteuerung durch den Autor besorgt; in der zweiten Welle wurde die Datenerhebung an ein kommerzielles Institut abgegeben.[21] Zwar räumte auch dieses Institut einen weitgehenden Zugang zu den Interviewern ein, wiederum oblag die Interviewerkontrolle jedoch letztlich dem Institut. Wie erwartet zeigt sich in der zweiten Erhebungswelle eine größere Variation der Netzwerkgröße zwischen den Interviewern als in der ersten, am Forschungsinstitut für Soziologie erhobenen Welle (vgl. *Tabelle 4*).

Dass die Kontrolle der Interviewer eine zentrale Rolle für die Datenqualität im Zusammenhang egozentrierter Netzwerke spielt, lässt sich auch auf Basis der Ergebnisse der LBW-Studie (Friedrichs und Blasius 2000) vermuten. In dieser Studie wurden deutsche (LBW-D) und türkische (LBW-T) Bewohner benachteiligter Wohngebiete mit einem einheitlichen Erhebungsinstrument, das für die Türken allerdings in türkischer Sprache vorlag, befragt. Nach Rücksprache mit den Projektleitern kann davon ausgegangen werden, dass sich die türkischen Interviewer nicht in derselben Weise kontrollieren ließen wie die deutschen. Ebenso war die Möglichkeit, die Interviewer zu kontrollieren, stark eingeschränkt, weil in den Haushalten der türkischen Befragten oft nicht genug deutsch gesprochen wurde, um zuverlässig über die Anwesenheit eines Interviewers und die Inhalte des Interviews Auskunft zu geben. Dies wäre nur durch den Einsatz einer türkischsprachigen Mitarbeiterin möglich gewesen, die zum damaligen Zeitpunkt nicht verfügbar war.

20 Auf die große Bedeutung des Nachfassens (Probens) im Zusammenhang mit egozentrierten Netzwerkdaten hat zuletzt Brewer (2000) hingewiesen.

21 Es handelt sich nicht um dasselbe Institut, das die Datenerhebung der SOWEM-Studie besorgte.

Zur Sicherung der Datenqualität kommt es – insbesondere im Zusammenhang mit der Erhebung egozentrierter Netzwerkdaten – darauf an, den Spielraum des Interviewers soweit wie möglich einzuschränken. Dies kann einerseits durch *wiederholte* Schulungen geschehen. So haben wir neben der obligatorischen Schulung vor Beginn der Interviews für alle Interviewer eine weitere Schulung nach Abschluss des ersten Interviews durchgeführt. Darüber hinaus wurden alle Interviews laufend kontrolliert und, wenn nötig, sofort nachgeschult.[22] Dieses Verfahren lässt sich jedoch nur durchführen, wenn direkter Zugang zu den Interviewern und jeweils zu den aktuell durchgeführten Interviews besteht. Dies wird meist nur der Fall sein, wenn es sich um eine nichtkommerziell durchgeführte, lokale Studie handelt.

Statt auf die Lernbereitschaft von Interviewern zu vertrauen, und dies scheint vor allem bei erfahrenen Interviewern schwierig zu sein, kann man versuchen, den Verhaltensspielraum der Interviewer durch das Instrument einzuengen. Dies wird wohl am besten durch den Einsatz eines computergestützten Interviews möglich sein, wobei die zentral durchgeführte computergestützte Telefonbefragung durch die Möglichkeit der Supervision noch bessere Ergebnisse liefern sollte (Groves 1989: 552; Schnell, Hill und Esser 1999: 352).[23] Eine computerunterstützte Befragung würde zudem die Zufallsauswahl der Alteri, über die weitere Informationen erhoben werden sollen, sowie die Zufallsauswahl der Beziehungen unter den Alteri, für die Daten gewünscht werden, deutlich erleichtern (vgl. die Diskussion zu diesem Punkt weiter oben). Studien, die egozentrierte Netzwerke computerunterstützt erheben, sind jedoch noch sehr selten (Gerich und Lehner 2003; Podolny und Baron 1997). Noch seltener sind methodische Untersuchungen, die sich mit der Qualität von computerunterstützt erhobenen Netzwerkdaten beschäftigen. Die einzige mir bekannte Untersuchung in diesem Zusammenhang stammt von Steven Corman (1990). Er weist in dieser Studie einen leichten Vorteil für computerunterstützt erhobene Netzwerkangaben im Vergleich zur Papier-und-Bleistift-Methode nach.

5. Das Problem fehlender Werte in Netzwerkdaten

Der letzte Punkt, auf den hier eingegangen werden soll, betrifft das Problem fehlender Werte (item nonresponse). Auch diese können die Qualität von Umfragedaten beeinträchtigen und zwar insbesondere, wenn sie nicht zufällig auftreten. Generell gilt, dass Fragen im Zusammenhang mit der Erhebung sozialer Netzwerke kaum verweigert werden. In den verschiedenen von uns durchgeführten Studien lag der Anteil derjenigen, die die Namensgeneratoren *nicht* beantwortet haben, bei unter einem halben Prozent. Ähnliches gilt für die Namensinterpretatoren, auch hier liegt der Anteil der verweigerten Antworten durchschnittlich bei unter zwei Prozent. Beispielsweise variiert der Anteil von Antworten, die als „keine Angabe“ gekennzeichnet sind, in der SUKUG-Studie

22 Auf die Notwendigkeit einer intensiven Interviewerschulung machen auch Bien, Marbach und Neyer (1991: 86) aufmerksam. Sie berichten aus dem Familiensurvey, dass im Pretest nahezu ein Drittel der Interviews fehlerhaft durchgeführt wurden; in der Hauptuntersuchung waren es trotz intensiver Schulung immer noch 16 Prozent.

23 Auf diese Möglichkeit weist auch Marsden (2003: 13f.) hin.

zwischen 0,2 Prozent für das Geschlecht und 3,1 Prozent für die berufliche Stellung der Netzwerkpersonen. Der Anteil der „weiß nicht"-Antworten variiert in derselben Untersuchung zwischen 0,1 Prozent für Angaben zum Haushaltstyp und zum Wohnort bis zu 7,5 Prozent für den höchsten Schulabschluss der Netzwerkpersonen.[24]

Ersten Analysen nach zu urteilen, ist der Anteil der „k.A."-Antworten einer Netzwerkperson nicht mit der Art oder Qualität der Beziehung zwischen Ego und Alter korreliert. Im Fall der „weiß nicht"-Antworten ist dies anders: Je seltener sich Ego und Alter treffen und je weniger nah sich Ego und Alter stehen, desto häufiger gibt Ego an, keine Angaben über Alter machen zu können. Der Anteil der „weiß nicht"-Antworten jeder Netzwerkperson kann somit als unabhängiges Maß zur Nähe zwischen Ego und Alter herangezogen werden.

Insgesamt ist der Anteil fehlender Werte in Netzwerkdaten sehr gering. Dies spiegelt die große Bereitschaft der Befragten wider, Auskunft über ihr soziales Umfeld zu erteilen. Die negative statistische Beziehung zwischen dem Anteil der „weiß nicht"-Antworten und der Nähe der Beziehung zwischen Ego und Alter weist zudem auf eine gewisse inhaltliche Validität der erhobenen Daten hin.

IV. Schlussfolgerungen

Eine abschließende Bewertung der Qualität egozentrierter Netzwerkdaten ist auf Grund der in den meisten Studien sehr kleinen Fallzahlen und den nicht immer einheitlichen Ergebnissen der verschiedenen Arbeiten nicht leicht. Dennoch dürfte deutlich geworden sein, dass Netzwerkdaten aus Massenumfragen „patterns of social contact and not merely patterns of responding to interview questions" widerspiegeln (Hammer 1984: 368). Auch wer dieser Schlussfolgerung skeptisch gegenübersteht und davon ausgeht, dass diese in Massenumfragen erhobenen Daten allenfalls Egos kognitiv repräsentiertes Netzwerk, nicht aber seine soziale Umgebung widerspiegeln, wird einräumen müssen, dass „die von Ego wahrgenommene Umwelt einen Einfluss auf das Verhalten von Ego ausübt" (Pappi und G. Wolf 1984: 282).

Dass viele empirische Ergebnisse sich unabhängig vom eingesetzten Netzwerkgenerator und der untersuchten Population stabil reproduzieren lassen, spricht ebenfalls für die Qualität dieser Daten. Beispielsweise ist die Netzwerkgröße in allen mir bekannten Studien positiv mit Bildung und anderen Schichtindikatoren korreliert. Dabei steigt vor allem die Größe des Teilnetzwerks der Nicht-Verwandten mit dem sozialen Status an, während die *Anzahl* der Verwandten nicht mit dem sozialen Status variiert. Ähnlich verhält es sich mit den Nachbarn. Ihre *Anzahl* schwankt zwar kaum zwischen niedrigem und hohem Status, aber ihr *Anteil* nimmt mit steigender sozialer Position kontinuierlich ab. Eine Folge dieser Sachverhalte ist, dass die Dichte der Netzwerke umso größer ist, je niedriger die soziale Position einer Person ist. Ein weiterer, oft replizierter Befund der Forschung bezieht sich auf die mit zunehmendem Alter abnehmende Größe der Netzwerke, wobei sich der Umfang der Netzwerke mit zunehmendem Alter immer schneller verringert.

24 Bei Einstellungsfragen oder bei heiklen Fragen wie der Parteipräferenz werden jedoch teilweise deutlich höhere Anteile von „weiß nicht"-Antworten erzielt.

Wie die in diesem Beitrag präsentierten Ergebnisse zu den Determinanten der Qualität egozentrierter Netzwerkdaten zeigen, kann diese durch geeignete Maßnahmen sichergestellt werden. Diese Maßnahmen beziehen sich auf die Erhebungsmethode, damit zusammenhängend auf die Kontrolle der Interviewer, auf den Umfang des Erhebungsinstruments und auf die Art, wie die Daten der Alteri erfasst werden.

In Bezug auf die Erhebungsmethode zeichnet sich ein Qualitätsvorteil für das telefonische Interview ab (Kogovšek et al. 2002). Diese Erhebungsform bietet, wenn sie zentral von einem Telefonstudio aus durchgeführt wird, die besten Möglichkeiten zur Kontrolle der Interviewer (Fuchs 1995). Ein weiterer Vorteil des Telefoninterviews besteht darin, dass sich die Bezahlung der Interviewer auf die geleistete Arbeitszeit bezieht und nicht, wie im Fall der persönlich-mündlichen Befragung, auf die Zahl der durchgeführten Interviews. Die zeitabhängige Bezahlweise minimiert die Anreize dafür, ein Interview möglichst rasch zu beenden, indem der Netzwerkteil „kreativ" verkürzt wird.

Neben dem Qualitätsvorteil für das telefonische Interview zeigte sich ein Vorteil der computergestützten Befragung. Dieser besteht zum einen in der Möglichkeit, das Interviewerverhalten weiter zu standardisieren. Zum anderen kann eine computergestützte Erhebung egozentrierter Netzwerke die Arbeit des Interviewers erheblich erleichtern, vor allem, wenn mehrere Namensgeneratoren eingesetzt werden. So kann die mühsame und fehlerträchtige Zusammenstellung einer Gesamtliste aller genannten Alteri sowie gegebenenfalls die Ziehung einer Stichprobe aus dieser Liste automatisch erledigt werden (Gerich und Lehner 2003: 57).

Die Vorzüge des computergestützten Interviews werden nicht nur im Fall einer telefonischen Befragung, sondern ebenso bei einer persönlich-mündlichen Befragung einen Qualitätsvorteil gegenüber dem traditionellen Papier-und-Bleistift-Interview ergeben. Allerdings zeigen unsere Erfahrungen mit den in Köln durchgeführten Netzwerkstudien, dass eine konsequente Überwachung des Feldes mit sehr zeitnahen Kontrollen und Nachschulungen der Interviewer den Einfluss des Interviewerverhaltens auf die Netzwerkgröße auch im Fall dieser traditionellen Erhebungsform stark reduzieren können.

Weil Netzwerkmerkmale umso zuverlässiger sind, je größer das Netzwerk ist, sollten in Abhängigkeit vom zur Verfügung stehenden Budget möglichst viele Netzwerkpersonen erhoben werden. Mit sieben bis acht Alteri erreichen auch die Angaben zur Beziehungsdauer, zur Treffhäufigkeit sowie weiterer Merkmale eine zufrieden stellende Zuverlässigkeit (Marsden 1993: 412). Da die Netzwerkgröße mit der Zahl der eingesetzten Namensgeneratoren ansteigt, empfiehlt es sich, ein Erhebungsinstrument mit mehreren Namensgeneratoren zu verwenden. Auch unter inhaltlichen Gesichtspunkten dürfte die Verwendung mehrerer dem Einsatz eines einzigen Namensgenerators meist vorzuziehen sein. Allerdings wird der Umfang des Instruments oft auf Grund finanzieller Restriktionen, oder um die Befragungszeit nicht ausufern zu lassen, beschränkt werden müssen.

Unabhängig davon, ob nur ein oder ob mehrere Namensgeneratoren zum Einsatz kommen, muss darauf geachtet werden, dass bei diesen Fragen nachgefasst wird und zwar von allen Interviewern gleichermaßen (Brewer 2000). Darüber hinaus sollten, vor allem wenn nur ein Namensgenerator verwendet wird, alle genannten Netzwerkpersonen erfasst werden. Schließlich sollten die Informationen über die Alteri personenstatt merkmalsbezogen erhoben werden.

Tabelle A1: Übersicht über die am Forschungsinstitut für Soziologie der Universität zu Köln durchgeführten Netzwerkstudien

Studienkürzel	RSN	LBW-D	LBW-T	SUKUG	SOWEM
Studienname	Religion und soziale Netzwerke	Leben in benachteiligten Wohngebieten; Teilstudie: Deutsche	Leben in benachteiligten Wohngebieten; Teilstudie: Türken	Soziale Ungleichheit, Krankheit und Gesundheit	Sozialer Wandel einer Mittelstadt
Grundgesamtheit	Deutsche Einwohner eines Kölner Stadtteils ab 18 Jahre	Deutsche Einwohner in vier armen Vierteln Kölns ab 18 Jahre	Türkische Einwohner in zwei armen Vierteln Kölns ab 18 Jahre	Deutsche Einwohner Kölns im Alter zwischen 30 und 60 Jahren	Die Einwohner Euskirchens ab 18 Jahre
Stichprobe[a]	geschichtete Zufallsstichprobe	einfache Zufallsstichprobe	einfache Zufallsstichprobe	zweistufige Zufallsstichprobe	einfache Zufallsstichprobe
Erhebungsjahr	1993	1996	1996	1999/2000	2001
Ausschöpfung	54,7%	43,7%	76,4%	43,1%	51,1%
Nettostichprobe	692	431	230	695[b]	1204
ZA-Studiennr.	2667	–	–	4055	3636
Veröffentlichung	Kecskes und Wolf (1996)	Friedrichs und Blasius (2000)	Friedrichs und Blasius (2000)	Wolf (2003)	Friedrichs, Kecskes und Wolf (2002)

a Bei allen Studien wurden die Stichproben aus den amtlichen Einwohnerregistern gezogen.
b Diese Studie wurde in zwei Wellen erhoben: $N_1 = 371$, $N_2 = 321$.

Literatur

Allan, Graham, 1977: Class Variation in Friendship Patterns. British Journal of Sociology 28: 389–393.

Baerveldt, Chris, Marijtje A.J. van Duijn, Lotte Vermeij und *Dianne A. van Hemert*, 2004: Ethnic Boundaries and Personal Choice. Assessing the Influence of Individual Inclinations to Choose Intra-ethnic Relationships on Pupils' Networks. Social Networks 26: 55–74.

Beck, Ulrich, 1986: Risikogesellschaft. Auf dem Weg in eine andere Moderne. Frankfurt a.M.: Suhrkamp.

Berger, Roger, und *Peter Kriwy*, 2004: Wer verdient wie viel? Eine Analyse des Verdienstes Münchner Soziologieabsolventen. Sozialwissenschaften und Berufspraxis 27: 133–154.

Bernard, H. Russell, Eugene C. Johnsen, Peter D. Killworth und *Scott Robinson*, 1989: Estimating the Size of an Average Personal Network and of an Event Subpopulation. S. 159–175 in: *Manfred Kochen* (Hg.), The Small World. New Jersey: Ablex.

Bernard, H. Russell, Eugene C. Johnsen, Peter D. Killworth, Christopher McCarthy, Gene A. Shelley und *Scott Robinson*, 1990: Comparing Four Different Methods for Measuring Personal Social Networks. Social Networks 12: 179–215.

Bien, Walter, Jan Marbach und *Franz Neyer*, 1991: Using Egocentered Networks in Survey Research. A Methodological Preview on an Application of Social Network Analysis in the Area of Family Research. Social Networks 13: 75–90.

Boxman, Ed A.W., Paul M. de Graaf und *Hendrik Flap*, 1991: The Impact of Social and Human Capital on the Income Attainment of Dutch Managers. Social Networks 13: 51–73.

Brewer, Devon D., 1997: No Associative Biases in the First Name Cued Recall Procedure for Eliciting Personal Networks. Social Networks 19: 345–353.

Brewer, Devon D., 2000: Forgetting in the Recall-based Elicitation of Personal and Social Networks. Social Networks 22: 29–43.

Burt, Ronald S., 1984: Network Items and the General Social Survey. Social Networks 6: 293–339.

Burt, *Ronald S.*, 1997: A Note on Social Capital and Network Content. Social Networks 19: 355–373.

Burt, Ronald S., 1998: A Social Capital Questionnaire. http://gsbwww.uchicago.edu/fac/ronald.burt/research/QUEST.pdf (04/09/17).

Campbell, Karen E., und *Barrett A. Lee*, 1991: Name Generators in Surveys of Personal Networks. Social Networks 13: 203–221.

Coleman, James S., 1958/59: Relational Analysis: The Study of Social Organizations with Survey Methods. Human Organization 17/4: 28–36.

Coleman, James S., 1988: Social Capital in the Creation of Human Capital. S. S95–S120 in: *Christopher Winship* und *Rosen Sherwin* (Hg.), Organisations and Institutions. Sociological and Economic Approaches to the Analysis of Social Structure (American Journal of Sociology, Vol. 94, Supplement). Chicago: University of Chicago Press.

Corman, Steven R., 1990: Computerized vs. Pencil and Paper Collection of Network Data. Social Networks 12: 375–384.

Davis, James A., und *Tom W. Smith*, 1987: General Social Surveys, 1972–1987. Codebook. Chicago: National Opinion Research Center.

Degenne, Alain, und *Michel Forsé*, 1999 [1994]: Introducing Social Networks. London: Sage.

Erickson, Bonnie H., 1975: International Networks: The Structured Webs of Diplomacy and Trade. Beverly Hills/London: Sage.

Ester, Peter, Loek Halman und *Ruud de Moor* (Hg.), 1994: The Individualizing Society. Value Change in Europe and North America. 2. Auflage. Tilburg: Tilburg University Press.

F*ischer, Claude S.*, 1982: To Dwell Among Friends. Personal Networks in Town and City. Chicago/London: Chicago University Press.

Fischer, Claude S., und *Yossi Shavit*, 1995: National Differences in Network Density: Israel and the United States. Social Networks 17: 129–145.

Flap, Henk, 2004: Creation and Returns of Social Capital: A New Research Program. S. 3–23 in: *Henk Flap* und *Beate Völker* (Hg.), Creation and Returns of Social Capital. A New Research Program. London: Routledge.

Freeman, Linton C., 2004: The Development of Social Network Analysis. A Study in the Sociology of Science. Vancouver, BC: Empirical Press.

Freeman, Linton C., und *Claire R. Thompson*, 1989: Estimating Acquaintanceship Volume. S. 147–175 in: *Manfred Kochen* (Hg.), The Small World. New Jersey: Ablex.

Friedrichs, Jürgen, und *Jörg Blasius*, 2000: Leben in benachteiligten Wohngebieten. Opladen: Leske + Budrich.

Friedrichs, Jürgen, Robert Kecskes und *Christof Wolf*, 2002: Struktur und sozialer Wandel einer Mittelstadt. Euskirchen 1952–2002. Opladen: Leske + Budrich.

Fuchs, Marek, 1995: Die computergestützte telefonsiche Befragung. Antworten auf Probleme der Umfrageforschung? Zeitschrift für Soziologie 24: 284–299.

Gerhards, Jürgen, 2003: Die Moderne und ihre Vornamen. Eine Einladung in die Kultursoziologie. Wiesbaden: Westdeutscher Verlag.

Gerich, Joachim, und *Roland Lehner*, 2003: Egozentrierte Netzwerkerhebung mittels selbstadministrierter Computerinterviews. Österreichische Zeitschrift für Soziologie 28: 46–70.

Granovetter, Mark S., 1973: The Strength of Weak Ties. American Journal of Sociology 78: 1360–1380.

Groves, Robert M., 1989: Survey Errors and Survey Costs. New York: Wiley.

Hammer, Muriel, 1984: Explorations into the Meaning of Social Network Interview Data. Social Networks 6: 341–371.

Hoffmeyer-Zlotnik, Jürgen H.P., 1987: Egozentrierte Netzwerke in Massenumfragen 1: Ein ZUMA-Methodenforschungsprojekt. ZUMA-Nachrichten 20: 37–43.

Jansen, Dorothea, 2003: Einführung in die Netzwerkanalyse. Grundlagen, Methoden, Forschungsbeispiele. 2. Auflage. Opladen: Leske + Budrich.

Kahn, Robert L., und *Toni C. Antonucci*, 1980: Convoys over the Life Course: Attachment, Roles, and Social Support. S. 253–286 in: *Paul B. Baltes* und *Orville G. Brim* (Hg.), Life-span Development and Behavior. Band 3. New York: Academic Press.

Kecskes, Robert, 2000: Abschlussbericht zum Forschungsvorhaben „Lebensbedingungen, Einstellungen und Verhaltensweisen türkischer Jugendlicher". Köln: Forschungsinstitut für Soziologie.

Kecskes, Robert, und *Christof Wolf*, 1996: Konfession, Religion und soziale Netzwerke. Zur Bedeutung christlicher Religiosität in personalen Beziehungen. Opladen: Leske + Budrich.

Killworth, Peter D., und *H. Russell Bernard*, 1976: Informant Accuracy in Social Network Data. Human Organization 35: 269–286.

Killworth, Peter D., *Eugene C. Johnsen, H. Russell Bernard, Gene A. Shelley* und *Christopher McCarthy*, 1990: Estimating the Size of Personal Networks. Social Networks 12: 289–312.

Kirke, Deirdre M., 2004: Chain Reactions in Adolescents' Cigarette, Alcohol and Drug Use: Similarity through Peer Influence or the Patterning of Ties in Peer Networks? Social Networks 26: 3–28.

Kogovšek, Tina, Anuska Ferligoj, Germa Coenders und *Willem E. Saris*, 2002: Estimating the Reliability and Validity of Personal Support Measures: Full Information ML Estimation with Planned Incomplete Data. Social Networks 24: 1–20.

Laumann, Edward O., 1966: Prestige and Association in an Urban Community. An Analysis of an Urban Stratification System. Indianapolis/New York: Bobbs-Merrill.

Laumann, Edward O., 1969: Friends of Urban Men: An Assessment of Accuracy in Reporting Their Socioeconomic Attributes, Mutual Choice, and Attitude Agreement. Sociometry 32: 54–69.

Laumann, Edward O., 1973: Bonds of Pluralism: The Form and Substance of Urban Social Networks. New York: Wiley.

Lazarsfeld, Paul F., und *Robert K. Merton*, 1954: Friendship as Social Process: A Substantive and Methodological Analysis. S. 18–66 in: *Morroe Berger, Theodor Abel* und *Charles H. Page* (Hg.), Freedom and Control in Modern Society. Toronto u.a.: Van Nostrand.

Lin, Nan, 1999: Social Networks and Status Attainment. Annual Review of Sociology 25: 467–487.

Lin, Nan, und *Mary Dumin*, 1986: Access to Occupations through Social Ties. Social Networks 8: 365–385.

Marsden, Peter V., 1987: Core Discussion Networks of Americans. American Sociological Review 52: 122–131.

Marsden, Peter V., 1990: Network Data and Measurement. Annual Review of Sociology 16: 435–463.
Marsden, Peter V., 1993: The Reliability of Network Density and Composition Measures. Social Networks 15: 399–421.
Marsden, Peter V., 2002: Egocentric and Sociocentric Measures of Network Centrality. Social Networks 24: 407–422.
Marsden, Peter V., 2003: Interviewer Effects in Measuring Network Size Using a Single Name Generator. Social Networks 25: 1–16.
Mayntz, Renate, 1958: Soziale Schichtung und sozialer Wandel in einer Industriegemeinde. Stuttgart: Enke.
McCallister, Lynne, und *Claude S. Fischer*, 1978: A Procedure for Surveying Personal Networks. Sociological Methods and Research 7: 131–148.
McCarthy, Christopher, 2002: Structure in Personal Networks. Journal of Social Structure 3 (Issue 1): http://zeeb.library.cmu.edu:7850/JoSS/McCarty/McCarty.htm (04/09/30).
McCarthy, Christopher, H. Russell Bernard, Peter D. Killworth, Gene A. Shelley und *Eugene C. Johnsen*, 1997: Eliciting Representative Samples of Personal Networks. Social Networks 19: 303–323.
Milardo, Robert M., 1992: Comparative Methods for Delineating Social Networks. Journal of Social and Personal Relationships 9: 447–461.
Mohler, Peter Ph., und *Uwe Pfenning*, 1987: Egozentrierte Netzwerke in Massenumfragen 3: Datenorganisation in einer SIR-Datenbank. ZUMA-Nachrichten 20: 51–77.
Pappi, Franz U., 1987: Die Netzwerkanalyse aus soziologischer Perspektive. S. 11–37 in: *Ders.* (Hg.), Methoden der Netzwerkanalyse. München: Oldenbourg.
Pappi, Franz U., und *Christian Melbeck*, 1988: Die sozialen Beziehungen städtischer Bevölkerungen. S. 223–250 in: *Jürgen Friedrichs* (Hg.), Soziologische Stadtforschung. Opladen: Westdeutscher Verlag.
Pappi, Franz U., und *Gunter Wolf*, 1984: Wahrnehmung und Realität sozialer Netzwerke. Zuverlässigkeit und Gültigkeit der Angaben über beste Freunde im Interview. S. 281–300 in: *Heiner Meulemann* und *Karl-Heinz Reuband* (Hg.), Soziale Realität im Interview. Empirische Analysen methodischer Probleme. Frankfurt a.M.: Campus.
Pfenning, Astrid, 1988: Zuverlässigkeit von Netzwerkangaben. Empirische Analysen anhand der ZUMA-Methodenstudie „Egozentrierte Netzwerke in Massenumfragen". Mannheim: Diplomarbeit.
Pfenning, Astrid, und *Uwe Pfenning*, 1987: Egozentrierte Netzwerke: Verschiedene Instrumente – verschiedene Ergebnisse? ZUMA-Nachrichten 21: 64–77.
Pfenning, Astrid, Uwe Pfenning und *Peter Ph. Mohler,* 1991: Zur Reliabilität von egozentrierten Netzwerken in Massenumfragen. ZUMA-Nachrichten 28: 92–108.
Podolny, Joel M., und *J.N. Baron*, 1997: Resources and Relationships: Social Networks and Mobility in the Workplace. American Sociological Review 62: 673–693.
Portes, Alejandro, 1998: Social Capital: Its Origin and Applications in Modern Sociology. Annual Review of Sociology 24: 1–24.
Reigrotzki, Erich, 1956: Soziale Verflechtungen in der Bundesrepublik. Elemente der sozialen Teilnahme in Kirche, Politik, Organisation und Freizeit. Tübingen: Mohr.
Reuband, Karl-Heinz, 1974: Differentielle Assoziation und soziale Schichtung. Hamburg: Diss.
Rippl, Susanne, 1995: Vorurteile und persönliche Beziehungen zwischen Ost- und Westdeutschen. Zeitschrift für Soziologie 24: 273–283.
Robicheaux, Sally, 2003: Perceptions of Social Support within the Context of Religious Homophily: a Social Network Analysis. Louisiana State University and Agricultural and Mechanical College, Department of Sociology: MA Thesis.
Schenk, Michael, 1984: Soziale Netzwerke und Kommunikation. Tübingen: Mohr.
Schenk, Michael, Peter Ph. Mohler, Uwe Pfenning und *Renate Ell*, 1992: Egozentrierte Netzwerke in der Forschungspraxis: Ausschöpfungsquoten und Validität soziodemographischer Variablen. ZUMA-Nachrichten 31: 87–120.
Scheuch, Erwin K., 1973: Das Interview in der Sozialforschung. S. 66–190 in: *René König* (Hg.), Handbuch der empirischen Sozialforschung, Bd. 2. 3. Auflage. Stuttgart: Enke.

Schneid, Michael, 1987: Egozentrierte Netzwerke in Massenumfragen 2: Feldsteuerung mit Computerunterstützung. ZUMA-Nachrichten 20: 44–50.
Schnell, Rainer, und *Frauke Kreuter*, 2003: Separating Interviewer and Sampling-point Effects. UCLA Statistics Electronic Publications Preprint 361.
Schnell, Rainer, Paul B. Hill und *Elke Esser*, 1999: Methoden der empirischen Sozialforschung. 6. Auflage. München: Oldenbourg.
Scott, John, 2000: Social Network Analysis. A Handbook. 2. Auflage. London: Sage.
Skjåk, Knut K., und *Janet Harkness*, 2003: Data Collection Methods. S. 179–193 in: *Janet A. Harkness, Fons J.R. van de Vijver* und *Peter Ph. Mohler* (Hg.), Cross-Cultural Survey Methods. Hoboken, NJ: Wiley.
Snijders, Tom A.B., 1999: Prologue to the Measurement of Social Capital. The Tocqueville Review 20: 27–44.
Straits, Bruce C., 2000: Ego's Important Discussants or Significant People: An Experiment in Varying the Wording of Personal Network Name Generators. Social Networks 22: 123–140.
Tilburg, Theo G. van, 1992: Question Sequence Effects in the Measurement of Reciprocity. Quality and Quantity 26: 395–408.
Tilburg, Theo van, 1998: Interviewer Effects in the Measurement of Personal Network Size. A Non-experimental Study. Sociological Methods and Research 26: 300–328.
Treinen, Heiner, 1965: Symbolische Ortbezogenheit. Eine soziologische Untersuchung zum Heimatproblem. Köln/Opladen: Westdeutscher Verlag.
van Der Gaag, Martin, und *Tom A.B. Snijders*, 2004a: The Resource Generator: Social Capital Quantification with Concrete Items. Amsterdam: Manuskript. http://www.xs4all.nl/~gaag/work/RG_paper.pdf (04/09/30).
van Der Gaag, Martin, und *Tom A.B. Snijders*, 2004b: Proposals for the Measurement of Individual Social Capital. S. 199–218 in: *Henk Flap* und *Beate Völker* (Hg.), Creation and Returns of Social Capital. A New Research Program. London: Routledge.
van Der Gaag, Martin, Tom A.B. Snijders und *Henk D. Flap*, 2004: Position Generator Measures and their Relationship to other Social Capital Measures. Amsterdam: Manuskript. http://www.xs4all.nl/~gaag/work/PG_comparison.pdf (04/09/30)
van der Poel, Mart G.M., 1995: Delineating Personal Support Networks. Social Networks 15: 49–70.
Wagner, Michael, Yvonne Schütze und *Frieder Lang*, 1996: Soziale Beziehungen alter Menschen. S. 301–319 in: *Karl U. Mayer* und *Paul B. Baltes* (Hg.), Die Berliner Altersstudie. Berlin: Akademie Verlag.
Wasserman, Stanley, und *Katherine Faust*, 1994: Social Network Analysis. Methods and Applications. Cambridge: Cambridge University Press.
Wegener, Bernd, 1989: Soziale Beziehungen im Karriereprozeß. Kölner Zeitschrift für Soziologie und Sozialpsychologie 41: 270–297.
Wellman, Barry, 1979: The Community Question: The Intimate Networks of East Yorkers. American Journal of Sociology 84: 1201–1231.
Wellman, Barry, 1993: An Egocentric Network Tale: Comment on Bien et al. (1991). Social Networks 15: 423–436.
Wolf, Christof, 1993: Egozentrierte Netzwerke: Datenorganisation und Datenanalyse. ZA-Information 32: 72–94.
Wolf, Christof, 1995: Religiöse Sozialisation, konfessionelle Milieus und Generation. Zeitschrift für Soziologie 24: 345–357.
Wolf, Christof, 1996: Gleich und gleich gesellt sich. Individuelle und strukturelle Einflüsse auf die Entstehung von Freundschaften. Hamburg: Kovac.
Wolf, Christof, 2003: Soziale Ungleichheit, Gesundheitsverhalten und Gesundheit. Eine soziologische Analyse für Personen im mittleren Erwachsenenalter. Köln: Universität zu Köln, Wirtschafts- und Sozialwissenschaftliche Fakultät (Habilitationsschrift).
ZA – Zentralarchiv für empirische Sozialforschung, 1987: Wahlstudie 1987, Teil 2: Panel-Studie (ZA-Nr. 1537). Maschinenlesbares Codebuch. Köln: ZA.

WAS VERBINDET UNS WANN MIT WEM?

Inhalt und Struktur in der Analyse sozialer Netzwerke*

Frans N. Stokman und Manuela Vieth

Zusammenfassung: Die Analyse sozialer Netzwerke erfreut sich wachsender Beliebtheit. Häufig wird dabei das Konzept des sozialen Kapitals verwendet. Damit wird auf die Möglichkeiten hingewiesen, die soziale Netzwerke Gruppen und Individuen zur Verwirklichung ihrer Ziele bieten. In diesem Artikel wird dargelegt, dass sich die meisten Netzwerkstudien zu sehr auf die Netzwerkstruktur konzentrieren. Diese Studien vernachlässigen, dass ihre Ergebnisse kontextspezifisch sind. Denn die Befunde hängen davon ab, welche Ziele die Akteure gerade anstreben und welche Arten von Beziehungen untersucht werden. Dies führt zu einer Zersplitterung der Netzwerkansätze, ohne dass die Bedingungen angegeben werden, unter denen sie angewandt werden können. In diesem Artikel wird eine Heuristik vorgeschlagen, die es ermöglicht, verschiedene Netzwerkansätze miteinander zu verbinden. Dazu wird der Schwerpunkt der Analyse auf diejenigen Ziele von Akteuren gelegt, die nicht durch eine Person allein verwirklicht werden können und somit zu gemeinschaftlicher Produktion und Nutzung von Gütern führen. Dabei werden ausdrücklich materielle Ziele sowie hedonistische Ziele („sich wohlfühlen") und normative Ziele („sich angemessen verhalten") mit einbezogen. Mit Hilfe dieser Heuristik kann für die Analyse sozialer Netzwerke die *Kombination* von Prozessen und dazugehörigen Beziehungen ausgewählt werden, die in einer gegebenen Situation eine Rolle spielen. Sie führt zu Theorien und Hypothesen darüber, *was* von *welchem Netzwerk* unter *welchen strukturellen Voraussetzungen warum* produziert wird. Die analytische Stärke der Heuristik wird in zwei völlig unterschiedlichen Kontexten vorgeführt, nämlich zum einen bei Netzwerken für Strategieentscheidungen in der Politik und in Organisationen und zum anderen bei Freundschaftsnetzwerken.

Die Analyse sozialer Netzwerke erforscht individuelles Verhalten[1] auf der Mikroebene, Beziehungsmuster (Netzwerkstrukturen) auf der Makroebene und die Wechselwirkungen zwischen beiden. Ziel dieses Artikels ist es, eine Heuristik für die Herausarbeitung

---

* Dieser Artikel basiert auf dem Eröffnungsvortrag von Frans Stokman auf der internationalen Konferenz zu sozialen Netzwerken in Portorož (Slowenien) am 13. Mai 2004. Manuela Vieth hat die deutsche Fassung des Artikels grundlegend neu bearbeitet.

Wir danken Tom Snijders, Stefan Thau, Rafael Wittek, Jacob Dijkstra, Bruce Bueno de Mesquita, Marcel van Assen, David Willer und ganz besonders Siegwart Lindenberg für ihre Anmerkungen zu den Konzeptfassungen dieses Textes. Ihre Kommentare haben beträchtlich dazu beigetragen, die Gedankengänge zu klären. Allgemeine Überlegungen, auf denen die in diesem Artikel dargestellten Ideen basieren, sind in enger Zusammenarbeit von Frans Stokman mit vielen Personen entwickelt worden, von denen insbesondere Jelle van der Knoop, Robert Mokken und Reinier van Oosten erwähnt seien.

1 Wenn von Individuen oder individuellem Verhalten geschrieben wird, sind an vielen Stellen sowohl Personen als auch „korporative Akteure" (Coleman 1990) gemeint, wie etwa Organisationen und organisatorische Einheiten, oder deren Repräsentanten.

der Mikrofundierungen sozialer Netzwerke zu erstellen. Dies soll erstens bei der Entscheidung helfen, welche sozialen Beziehungen untersucht werden, und zweitens bei der Auswahl der bedeutsamen strukturellen Gesichtspunkte sozialer Netzwerke. An einer solchen theoretischen Fundierung fehlt es nur allzu oft. Dies führt dazu, dass die Auswahl der untersuchten Beziehungsarten und die Konzentration auf bestimmte strukturelle Aspekte der untersuchten Netzwerke ad hoc geschehen. Hingegen gehört zur Mikrofundierung erstens eine Analyse der dominanten Ziele der Individuen und korporativen Akteure in unterschiedlichen Zusammenhängen. Zweitens ist sie verbunden mit einer Analyse des instrumentellen Wertes, den die verschiedenen Arten sozialer Beziehungen für die Verwirklichung der betreffenden Ziele besitzen. Dieser instrumentelle Wert bestimmt wesentlich, welche Netzwerkstrukturen und Positionen mit relevanten Möglichkeiten oder Einschränkungen für die Zielerreichung verbunden sind.

I. Soziales Kapital und sozialer Austausch

Soziales Kapital ist das Gefüge von Gelegenheiten, das durch soziale Beziehungen geschaffen wird (Lin 1982; Flap und de Graaf 1986; Coleman 1988; Burt 1992). Es ermöglicht einem Individuum Zugriff auf die Ressourcen anderer, die es für die Verwirklichung seiner Ziele nutzen kann. Die Höhe des sozialen Kapitals hängt von der Menge dieser Ressourcen ab, von ihrem Wert für die Verwirklichung der Ziele des Individuums und von der Bereitschaft anderer, die Ziele zu unterstützen (Flap 1999). Die Idee des sozialen Kapitals, verbunden mit dem Zugriff auf die Ressourcen anderer, ist eng verknüpft mit Austauschtheorien. Wie auch Cook (1995), Molm (1997) und Lawler (2001) seien vier Grundformen des Austauschs mit unterschiedlichen Auswirkungen auf das Netzwerk unterschieden:

1. *verhandlungsbasierter Austausch*, d.h. Tauschhandlungen, die auf ausdrücklichen und verbindlichen Vereinbarungen basieren (z.B. Geschäftsverhandlungen oder Verhandlungen über Gehalt und Aufstiegsbedingungen);
2. *produktiver Austausch*, d.h. Tauschhandlungen, um Ressourcen zur Herstellung von Gemeinschaftsgütern zu kombinieren (z.B. Arbeitsteams, eine Gruppe von Freunden, die eine Feier vorbereiten oder einen Urlaub planen);
3. *reziproker Austausch*, d.h. Tauschhandlungen ohne Absprachen und Verträge zwischen den beteiligten Akteuren (z.B. einem Kollegen[2] bei seiner Arbeit helfen, in der Erwartung, dass er sich später dafür erkenntlich zeigen wird, ohne jedoch sicher zu sein, wann und ob überhaupt eine Gegenleistung erfolgen wird);
4. *generalisierter Austausch*, d.h. Tauschhandlungen, mit denen man einigen Mitgliedern einer Gruppe Ressourcen verschafft, während man darauf hofft, andere Ressourcen von anderen Gruppenmitgliedern zu erhalten (z.B. einem Kollegen oder Freund bei seiner Arbeit helfen in der Erwartung, in anderen Dingen Unterstützung von anderen zu erhalten, ohne jedoch sicher zu sein, wann und ob diese Unterstützung geleistet wird).

2 Überall, wo eine männliche Form für Personen verwendet wird, ist auch die weibliche Form gemeint.

Wenn Individuen Ressourcen unterschiedlich bewerten, können sie ihre eigenen und die Ergebnisse anderer verbessern, indem sie Ressourcen austauschen. Individuen können dann als *interdependent* betrachtet werden. Denn der Wert des Ergebnisses hängt von ihrem eigenen Verhalten und dem Verhalten anderer ab. In Austauschsituationen haben Individuen sowohl *gemeinsame Interessen* als auch *gegensätzliche Interessen*. Die gemeinsamen Interessen ergeben sich aus der Tatsache, dass ein Austausch zu besseren Ergebnissen für die Austauschpartner führt. Hingegen geht die Aufteilung der Gewinne unter den Partnern mit gegensätzlichen Interessen einher. Am deutlichsten ist dies beim *verhandlungsbasierten Austausch*. Die Verhandlungen zwischen den Tauschpartnern führen dazu, dass beide Partner die Bedingungen des Austauschs *gleichzeitig* anerkennen. Damit werden die Vereinbarungen als streng verbindlich vorausgesetzt und die *kooperative Spieltheorie* kann genutzt werden, um die Ergebnisse des Austauschs vorherzusagen. Im Rahmen der kooperativen Spieltheorie untersucht die Netzwerk-Austauschtheorie (*Network Exchange Theory:* NET) vor allem die Einflüsse von Netzwerkstrukturen auf die Wahl zwischen alternativen Tauschhandlungen und auf die Verhandlungsergebnisse, d.h. das ausgehandelte Tauschverhältnis (Willer 1999). Die NET hat theoretisch und experimentell gezeigt, dass insbesondere Individuen, die andere Akteure ausschließen, unverhältnismäßig bessere Ergebnisse erzielen können.

Eine wichtige Erkenntnis aus der NET ist, dass Einflüsse von Netzwerkstrukturen *kontextsabhängig* sind. Somit können sie nicht verallgemeinert werden, ohne den Kontext und die Art der Ressource zu berücksichtigen, über die verhandelt wird. Ein geradezu triviales Beispiel ist verbunden mit der Veränderung der Anzahl Tauschhandlungen, die einem Individuum möglich sind. In NET-Experimenten wird Austausch meist durch die Aufteilung einer Anzahl Punkte dargestellt. Angenommen ein Akteur A kann 24 Punkte zwischen sich und einem von zwei anderen Akteuren (B oder C) aufteilen. Wenn zwei Akteure eine Übereinkunft erreichen, erhalten beide die vereinbarte Anzahl Punkte. Andernfalls verfallen die Punkte. Kann jeder Akteur nur eine einzige Tauschhandlung vornehmen, ist A in der Lage, B und C gegeneinander auszuspielen. Denn A kann von einem seiner Partner einen größeren Anteil verlangen als sein anderer Partner ihm bietet. Die Machtposition von A löst sich jedoch vollständig auf, wenn A zwei Tauschhandlungen vornehmen kann. *Muss* sich A mit beiden Partnern auf einen Anteil einigen, um überhaupt etwas zu erhalten, ist seine Position sogar schwächer als die von B oder C.

Der Kontext spielt auch eine Rolle, wenn wir unsere Aufmerksamkeit, entgegen der NET, nicht auf isolierte Transaktionen beschränken, in denen die Interaktionspartner einander fremd sind. Denn Transaktionen können sowohl in dauerhaftere dyadische Beziehungen als auch in Beziehungen mit Dritten eingebettet sein. Experimente haben gezeigt, dass Individuen geneigt sind, weniger gewinnbringende Tauschhandlungen mit Partnern vorzunehmen, mit denen sie bereits zuvor interagiert haben, als mit neu zugewiesenen, unbekannten Partnern (Kollock 1994). Die Ergebnisse von NET-Studien hängen auch deshalb vom Kontext ab, weil ausschließlich materielle Ressourcen getauscht werden. In der NET sind Individuen zielgerichtet und allein an materiellen Gewinnen interessiert. Ungeachtet der Erkenntnisse durch die NET, ist diese Theorie daher aber nur auf Situationen anwendbar, in denen es um isolierte Transaktionen geht und die individuellen Ziele vollständig auf materiellen Gewinn ausgerichtet sind.

Bei den anderen drei Arten des Austausches sind die Transaktionen in viel höherem Maße miteinander verknüpft, da sie sich im Rahmen dauerhafterer Beziehungen abspielen. Beim *produktiven Austausch* werden Gemeinschaftsgüter hergestellt, die sonst nicht erzeugt werden könnten oder deren Herstellung sonst mit höheren Kosten verbunden wäre. Dauerhafte Interdependenzen ergeben sich dabei aus der Tatsache, dass Individuen ihre Ressourcen zusammenlegen müssen, um diese Güter zu produzieren. Bei einem *reziproken* oder *generalisierten Austausch* werden Angebote gemacht, ohne zu wissen, ob oder wann oder in welchem Maße der andere (oder beim generalisierten Austausch: ein anderer) eine Gegenleistung erbringen wird. In derartigen Situationen ist die Zusammenarbeit der Tauschpartner unsicher und es besteht das Problem des Trittbrettfahrens. Ziel der *nicht-kooperativen Spieltheorie* ist es, die Voraussetzungen anzugeben, unter denen Kooperation zu erwarten ist. Hierbei betonen Netzwerkstudien in der Regel die Informations- und Kontrollmöglichkeiten von Netzwerken, die Kooperation fördern (Raub und Weesie 1990; Flache und Macy 1996; Buskens und Raub 2004). Im Gegensatz zur NET werden dauerhaftere Beziehungen, Vereinbarungen und wechselseitige Abhängigkeiten explizit im Modell erfasst. Außerdem können neben dem materiellen Gewinn auch nicht-materielle Ziele berücksichtigt werden.

II. Netzwerke und Produktion

Managementstudien untersuchen die Voraussetzungen, unter denen Leistungen im Produktionsbereich erbracht werden. Diese Studien zeigen, dass wechselseitige Abhängigkeiten zwischen Individuen hinsichtlich ihrer Aufgaben *(task interdependencies)* und hinsichtlich der Ergebnisse *(outcome interdependencies)* in klare institutionelle Regeln sowie in Gehalts- und Beförderungssysteme eingebettet sein sollten, damit sich *Organizational Citizenship Behavior* (OCB)[3] und besonderes Rollenverhalten (extra-role-behavior) entwickeln kann. In der Literatur wird keine systematische Verbindung zwischen diesen Managementkonzepten und Netzwerken hergestellt. Erkennbar sind jedoch die folgenden drei Verbindungen: Institutionelle Regeln sind verbunden mit Machtnetzwerken; Gehalts- und Beförderungssysteme hängen zusammen mit verhandlungsbasierten Tauschnetzwerken; OCB und besonderes Rollenverhalten gehen einher mit Netzwerken des reziproken oder generalisierten Austauschs.

In der Literatur werden diese drei Netzwerkformen und die damit verbundenen theoretischen und analytischen Perspektiven jedoch weder in einem einheitlichen Rahmen integriert noch auf ihren spezifischen Beitrag für Produktionsbedingungen und Interaktionen hin ausgearbeitet. Die bekannteste Netzwerktheorie, die sich speziell auf die Analyse von Produktionsbedingungen richtet, ist Burts (1992) Maklertheorie *(brokerage theory)*.[4] Makler sind Intermediäre auf Brückenpositionen in Informationsnetz-

3 *Organizational Citizenship Behavior* (Organ 1988) ist die in Managementstudien am häufigsten untersuchte Form kooperativen Verhaltens. Dies sind Verhaltensweisen von Mitarbeitern, die eine allgemeine positive Wirkung auf das Funktionieren der Organisation haben, jedoch nicht vertraglich erzwungen werden können.

4 Siehe Borgatti und Foster (2003) für eine ausführlichere Übersicht zu Netzwerkstudien in der Organisationsforschung.

werken, d.h. an den Verbindungsstellen verschiedener Netzwerksegmente. Auf diese Weise nutzen Makler „strukturelle Löcher“ (Burt 1992), um Informationsvorteile zu erzielen. Im Folgenden wird zunächst dieser Ansatz aufgegriffen. Danach wird ein Integrationsrahmen vorgeschlagen, in dem alle bedeutenden Netzwerkansätze miteinander in Beziehung gesetzt werden können.

Burt (2000) legt umfassendes Beweismaterial für den Wettbewerbsvorteil von Maklerpositionen in Informationsnetzwerken vor. Makler können Informationen von verschiedenen Gruppen auf effiziente Weise kombinieren, um kreative Problemlösungen zu entwickeln. Hingegen kursieren in geschlossenen Netzwerken nur redundante Informationen. Spätere Studien weisen darauf hin, dass der Erfolg eines Teams nicht nur von offenen externen Netzwerken abhängt. Wichtig ist hingegen auch ein dichtes Informationsnetzwerk innerhalb des Teams, das gemeinsame Normen begründet und Koordination erleichtert.

Das Gegenteil einer Maklerposition ist eine Position innerhalb eines geschlossenen Netzwerks. Der einfachste Fall ist eine Triade, in der alle möglichen Beziehungen vorkommen. Simmel (1950) beschäftigt sich eingehend mit der Triade im Vergleich zur Dyade. Das Besondere der Triade wird nur dann deutlich, wenn in der Dreierbeziehung keine vollständige Einheit herrscht. Sowohl gemeinsame als auch gegensätzliche Interessen müssen vorliegen. Nur dann kann die dritte Person von ihren Beziehungen zu den beiden anderen profitieren (*tertius gaudens* – der lachende Dritte). Da jede Beziehung eine Mischung gemeinsamer und gegensätzlicher Interessen ist, spielt die Triade (und auf höherer Ebene das Netzwerksegment) in Bezug auf Normerzeugung und Vertrauen eine wichtige Rolle. Denn mittels Reputation und Kontrolle wird unkooperativem Verhalten Grenzen gesetzt, das in einzelnen dyadischen Beziehungen möglich wäre. Es war ein wichtiger Beitrag Krackhardts (1999), dass er Simmelsche Makler als Sondergruppe erkannte. Diese sind Personen, die an verschiedenen Netzwerksegmenten beteiligt sind. Daher sind sie mit hoher Wahrscheinlichkeit Normkonflikten ausgesetzt.

Die Argumente von Burt und Krackhardt sind hauptsächlich struktureller Art. Podolny und Baron (1997) zeigen hingegen, dass der Wettbewerbsvorteil von Maklern davon abhängt, welches Netzwerk betrachtet wird. So haben Makler Wettbewerbsvorteile in Netzwerken, in denen etwa Aufgabenberatung und Leistungsrückmeldung Inhalt der Beziehungen sind. Eine Maklerposition ist jedoch unvorteilhaft in Netzwerken, die auf Beziehungen zwischen Personen auf der Gefühlsebene beruhen (z.B. bei Büroklatsch oder bei sozialer Unterstützung). Dies bedeutet, dass es nicht nur auf die Struktur ankommt, sondern auch auf den Inhalt von Beziehungen.

Zusammenfassend liegen derzeit sehr deutliche Befunde vor, dass soziale Netzwerke bei gemeinschaftlicher Herstellung von Gütern *(joint production)*[5] sowie bei der Gestaltung von Normen und Vertrauen bedeutsam sind. Die Argumente sind jedoch überwiegend struktureller Art, der Inhalt von Beziehungen spielt eine untergeordnete Rolle. Außerdem verfügen wir über umfassende Forschung auf dem Gebiet sozialer Netzwerke, die sich mit den verschiedenen Typen von Tauschhandlungen beschäftigt. Um die Bedeutung dieser Typen zu verstehen, müssen sie jedoch miteinander in Beziehung gesetzt werden. Dies ist bislang noch nicht geschehen. Dazu muss der Schwerpunkt auf

5 Im Folgenden wird dies „gemeinschaftliche Produktion“ genannt.

der Analyse der Verwirklichung von Zielen liegen, die nicht durch eine Person allein realisiert werden können und somit zu gemeinschaftlicher Produktion und Nutzung von Gütern führen. Diese Ziele sind nicht nur materiell, sondern auch hedonistisch („sich wohlfühlen") und normativ („sich angemessen verhalten") (siehe Lindenberg 2001). Was wir brauchen, ist eine Heuristik, um die *Kombination* von Prozessen und dazugehörigen Beziehungen auszuwählen, die in einer gegebenen Situation eine Rolle spielen. Dazu gehören Überlegungen darüber, *was* von *welchem Netzwerk* unter *welchen strukturellen Voraussetzungen warum* produziert wird. Die im Folgenden dargestellte Heuristik stützt sich auf die Gruppentheorie von Lindenberg (1997).

III. Heuristik zur Ableitung von Netzwerkstrukturen für gemeinschaftliche Produktion

In seinem Überblick über die Literatur zur Gruppentheorie zeigt Lindenberg, dass verschiedene Ansätze drei Typen wechselseitiger Abhängigkeiten hervorheben: funktionale, kognitive und strukturelle Interdependenzen. Funktionale Interdependenzen sind wechselseitige Abhängigkeiten zwischen Individuen, die sich hinsichtlich der Aufgaben und hinsichtlich der Ergebnisse bei gemeinschaftlicher Produktion ergeben. Die kognitiven Interdependenzen bestehen erstens aus den kognitiv vermittelten funktionalen Interdependenzen, d.h. Interdependenzen, die durch die Individuen anerkannt werden. Zum Beispiel kann die Systembetreuung die Probleme mit spezieller Software nicht immer schnell lösen. Nun kann der Bürokollege von speziellen Programmeinstellungen wissen und gerne bereit sein zu helfen. Dies ist dann eine kognitiv vermittelte funktionale Interdependenz, die genutzt werden kann, um die Probleme zu lösen. Zweitens spielen auch externe Zeichen eine Rolle, die zu erkennen geben, was angemessenes Verhalten ist. Dazu gehören zum Beispiel Normen, die die Bedingungen angeben, unter denen vom Bürokollegen Hilfe zur Lösung eines Software-Problems erwartet werden kann. Die strukturellen Interdependenzen betreffen die mehr oder weniger stabilen Beziehungen zwischen Individuen und die wechselseitigen Abhängigkeiten dieser Beziehungen, die sich aus der Struktur des Netzwerks ergeben. Beispielsweise verändert sich, wie bereits erläutert, die Stärke von Machtpositionen in Austauchnetzwerken in Abhängigkeit von den Möglichkeiten, andere Akteure vom Austausch auszuschließen. In den 1950er Jahren wurden in der Feldtheorie von Lewin (1936) (deren bekannteste Vertreter sind Sherif (1966) und Festinger (1954) mit ihrem gruppendynamischen Programm) alle drei Interdependenzen erkannt. Doch Instrumente für eine wirkliche Integration dieser Interdependenzen waren damals nicht vorhanden. Daher ergaben sich rivalisierende Ansätze, die sich jeweils nur auf einen der drei Typen von Interdependenzen konzentrierten. Lindenberg führte den Misserfolg der Feldtheorie auf die Tatsache zurück, dass sie Gruppen wiederholt in Begriffen funktionaler Interdependenzen definierte, während sie sich theoretisch und empirisch hauptsächlich auf die kognitiv vermittelten Folgen funktionaler Interdependenzen konzentrierte (z.B. Gruppenzwang). Dadurch wurde der funktionalen Interdependenz in der Gruppendynamik in Wirklichkeit sehr wenig Aufmerksamkeit zuteil (Lindenberg 1997: 287). Weiterhin war die Integration funktionaler, kognitiver und struktureller Interdependenzen durch die stark dyadische Ausrich-

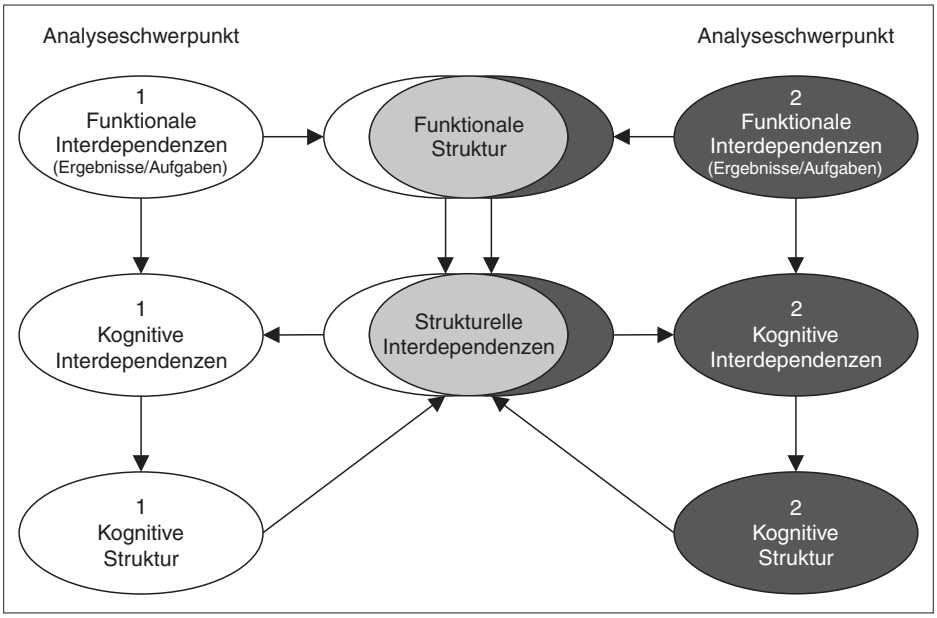
tung der Theorien erschwert, die im Hinblick auf sozialen Austausch entwickelt wurden (Lindenberg 1997: 312).

Mit Lindenberg übereinstimmend sollte gemeinschaftliche Produktion der Ausgangspunkt für die Analyse sozialer Netzwerke sein. Eine solche Perspektive erleichtert die theoretische Integration unterschiedlicher Austauschtheorien. Denn sie spezifiziert den besonderen Beitrag jeder Theorie zu gemeinschaftlicher Produktion. Die hier dargestellte Heuristik legt den Analyseschwerpunkt daher auf die funktionalen Interdependenzen zwischen Individuen, d.h. ihre wechselseitigen Abhängigkeiten in Bezug auf Ergebnisse und Aufgaben. Eine gründliche Analyse dieser Interdependenzen steht an erster Stelle. Diese wechselseitigen Abhängigkeiten zwischen Individuen hinsichtlich ihrer Aufgaben und hinsichtlich der Ergebnisse bestimmen die funktionale Struktur: Mit wem müssen wir kommunizieren, uns koordinieren, Ressourcen austauschen, und in welcher Reihenfolge muss dies geschehen, um das gemeinsame Ziel verwirklichen zu können? Die funktionale Struktur ist eine der beiden wichtigen Determinanten von Beziehungen und den damit verbundenen strukturellen Interdependenzen. Sowohl funktionale als auch strukturelle Interdependenzen erzeugen kognitive Interdependenzen. Die entsprechende kognitive Struktur bildet die zweite wesentliche Determinante struktureller Interdependenzen.

Die Verbindungen, die durch die Analyse sozialer Netzwerke normalerweise untersucht werden, entsprechen in einer produktionsorientierten Organisation in der Regel mehreren, gleichzeitig vorhandenen funktionalen Interdependenzen. In unseren Universitäten kennen wir verschiedene davon in Lehre, Forschung und Verwaltung. Jede dieser funktionalen Interdependenzen lässt unterschiedliche Beziehungen entstehen. Die durch diese Beziehungen verbundenen Personengruppen müssen nicht dieselben sein. Beispielsweise Kollegen, mit denen wir unsere Lehrtätigkeit abstimmen müssen, sind nur zum Teil dieselben Kollegen, mit denen wir an gemeinsamen Forschungsprojekten arbeiten. Unterschiedliche funktionale Interdependenzen führen daher zu Beziehungsstrukturen, die sich nur teilweise überschneiden. Sie können durchaus zu gegensätzlichen kognitiven Strukturen führen. Unterschiedliche Auffassungen können zum Beispiel darüber bestehen, welche der vielen Aufgaben zuerst erledigt werden soll. Dies ist in *Abbildung 1* aus der Perspektive eines Individuums dargestellt. Seine verschiedenen funktionalen Interdependenzen für gemeinschaftliche Produktion bestimmen sowohl seine kognitiven als auch seine strukturellen Interdependenzen. Dabei treten auch Wechselwirkungen auf. In *Abbildung 1* sind einfache Pfeile von den funktionalen Interdependenzen zu den kognitiven und, über die funktionale Struktur, zu den strukturellen Interdependenzen dargestellt. Diese bedeuten nicht, dass bestehende kognitive und strukturelle Interdependenzen nicht genutzt werden können, um damit zusammenhängende oder neue funktionale Interdependenzen aufzubauen. Zusammen mit den bestehenden funktionalen Interdependenzen bestimmen dann auch wiederum die neuen funktionalen Interdependenzen die kognitiven und strukturellen Interdependenzen.

Die Analyse funktionaler Interdependenzen ist nicht auf gewinnorientierte Aktivitäten beschränkt. Wir sind ebenfalls mit anderen funktional interdependent, um Situationen und Aktivitäten zu ermöglichen, in denen wir uns „wohlfühlen“ können (z.B. Ausgehen, Freundschaften, Sport). Und wir haben das Beste erreicht, wenn wir uns bei

Abbildung 1: Heuristik der Interdependenzen in Netzwerken



allen oder manchen Kollegen „wohlfühlen", mit denen wir beruflich zusammenarbeiten.

Es lässt sich also schließen, dass die meisten Beziehungen mit unseren Kollegen multifunktional sind. Diese Beziehungen hängen von funktionalen Interdependenzen ab, die sowohl durch die Ziele der Organisation gesteuert werden als auch durch Interdependenzen außerhalb der Organisation. Solche Beziehungen können zu materiellen sowie hedonistischen Zwecken genutzt werden. Jedes dieser Ziele kann verschiedene kognitive Interdependenzen mit unterschiedlichen und sogar widersprüchlichen Verhaltensnormen erzeugen. Die Analyse sozialer Netzwerke sollte sich darauf konzentrieren, welche Strukturen für welche Ziele wichtig sind und unter welchen Voraussetzungen die verschiedenen funktionalen Interdependenzen miteinander verschmelzen oder einander hindern. *Dies erfordert Untersuchungsdesigns, in denen verschiedene Beziehungsinhalte gleichzeitig analysiert werden, und die es erlauben, die funktionalen und kognitiven Interdependenzen in Beziehungen zu unterscheiden.*

Wann können Probleme bei gemeinschaftlicher Produktion auftreten? Probleme können sich jeweils auf Grund folgender Umstände ergeben:

1. Funktionale Interdependenzen können unbestimmt und unterschiedlich interpretierbar sein. Dann werden Gesten in einer Beziehung unklar („Hilfst du mir oder brüstest du dich mit deinem Wissen?").
2. Es kann sein, dass kognitive und strukturelle Interdependenzen nicht den funktionalen Interdependenzen entsprechen. Funktionale und kognitive Interdependenzen können miteinander in Konflikt stehen oder geraten. (Dieser Punkt wird weiter unten an einem Beispiel erläutert.)

3. Personenpaare können unterschiedliche kognitive Interdependenzen wahrnehmen. Dies ist zum Beispiel der Fall, wenn eine Person jemanden als Freund einstuft, während sie selbst vom anderen als Geschäftspartner betrachtet wird.
4. Möglich ist auch, dass kognitive und strukturelle Interdependenzen nicht miteinander übereinstimmen und daher einander nicht unterstützen.[6]

Der zweite Umstand, der Probleme bei gemeinschaftlicher Produktion erzeugt, lässt sich gut daran illustrieren, was sich im Jahr 1956 an der Ziellinie des Schlittschuhrennens über 200 Kilometer in Friesland ereignet hat (11. „Elf-Städte-Tour" des 20. Jahrhunderts in den Niederlanden). Informelle Teams können zusammenarbeiten, um den besten Läufer in eine Gewinnstellung zu bringen. Dies ist auch der Fall beim Radrennsport. Es ist schwierig, solche Wettkämpfe zu gewinnen, ohne über ein sehr solidarisches und leistungsfähiges Team zu verfügen. Letztendlich wird aber erwartet, dass jeder einzeln um die Siegerposition kämpft. Auf jener „Elf-Städte-Tour" operierten sechs Schlittschuhläufer während des gesamten Wettkampfes als informelles Team. Fünf von ihnen schafften es, die Ziellinie gleichzeitig zu überqueren. Zur Entrüstung des Publikums und der Organisatoren hatten sie beschlossen, die Ziellinie als Team zu überqueren. Die Organisation entschied sich in jenem Jahr, niemandem eine Medaille zu verleihen. Im ganzen 20. Jahrhundert haben nur fünfzehn dieser Wettkämpfe stattgefunden. Alle Gewinner der übrigen vierzehn Wettkämpfe waren für den Rest ihres Lebens Helden. Die Männer jenes elften Schlittschuhrennens erhielten niemals eine Auszeichnung und wurden zu keiner offiziellen Feier eingeladen. Dies ist ein Beispiel dafür, wie starke Solidarität in einer Wettbewerbssituation kontraproduktiv sein kann, Flache und Macy (1996) haben dies auch in ihren Simulationsstudien und Experimenten gezeigt.

Beim Austausch von Informationen können verschiedene funktionale Interdependenzen mit unterschiedlichen optimalen Netzwerkstrukturen einhergehen. Man kann zum Beispiel erwarten, dass sich optimale Netzwerkstrukturen in der Spezifikationsphase von Produkten (z.B. bei Bankprodukten), die eigens auf den Kunden zugeschnitten sind, von denen in der Implementierungsphase grundlegend unterscheiden. Es ist zu vermuten, dass offene Netzwerke in der Spezifikationsphase die individuelle Leistung fördern. Während der Implementierungsphase ist hingegen sorgfältige Koordination und somit ein geschlossenes Netzwerk erforderlich. Erste Hinweise, die diese Überlegungen bestätigen, fanden sich in David Dekkers Netzwerkstudien (Dekker 2001; Dekker et al. im Erscheinen). Er untersuchte Netzwerke von 57 Personen in einer Kontoführungsabteilung und vier Serviceabteilungen im Bankwesen. Wie erwartet, zeigt sich tatsächlich, dass die erfolgreichsten Personen ein relativ offenes Netzwerk bei der Spezifikation mit einem relativ geschlossenen Netzwerk bei der Erbringung der Leistungen kombinieren. Die Personen mit der geringsten Leistung weisen hingegen in beiden Bereichen gegenteilige Netzwerke auf. Entgegen Burts Einschätzung lautet die wesentliche Botschaft also *nicht:* nach außen hin relativ offen und relativ geschlossen innerhalb des eigenen Teams. Hingegen sind innerhalb der eigenen Gruppe offene

6 In der Netzwerkforschung hat sich wiederholt gezeigt, dass kognitive Strukturen wesentlich von den tatsächlichen Interaktionsstrukturen abweichen können (Killworth und Bernard 1976, 1979; Freeman et al. 1987; Kumbasar et al. 1994). Wenig Beachtung fand jedoch die Frage, wie sich dies auf das Verhalten auswirkt.

Netzwerke für die eine Aktivität und geschlossene für eine andere vorteilhaft. Dies hängt von den funktionalen Interdependenzen bei den jeweiligen Aktivitäten ab.

Kognitive Interdependenzen bestehen aus Normen über angemessenes Verhalten. Vertrauen hängt in starkem Maße von früheren Erfahrungen mit angemessenem Verhalten ab. Daher dürften kognitive Interdependenzen am besten anhand von *Vertrauens*netzwerken untersucht werden können. Im Gegensatz zu bestimmten Informationsnetzwerken (z.B. diejenigen in der zuvor erwähnten Spezifikationsphase) erleben Makler in einem Vertrauensnetzwerk Rollenambiguität. Denn sie sind dem Druck verschiedener Gruppen mit unterschiedlichen Normen ausgesetzt. Rollenambiguität kann sich nachteilig auf die Leistungen auswirken, wie die Untersuchung von Dekker gezeigt hat.

Kann man aus dieser Heuristik spezifische Beiträge der verschiedenen Theorien des sozialen Austauschs für gemeinschaftliche Produktion ableiten? Wenn es der Theorie des verhandlungsbasierten Austauschs gelingt, aus dem Bereich der experimentellen Forschung in die wirkliche Welt vorzustoßen, kann sie für die Untersuchung gesetzlicher Arbeitsverträge relevant werden. Ohne Verträge lassen sich langfristige Beiträge zu gemeinschaftlicher Produktion nur schwer durchsetzen. Denn Verträge verknüpfen gemeinschaftliche Produktion mit externen (gesetzlichen) Sanktionssystemen, um Kooperation sicherzustellen. In ähnlicher Weise helfen gesetzliche Regelungen, funktionale Interdependenzen in institutionelle Vereinbarungen einzubetten. Gesetzliche Arbeitsverträge und deren institutionelle Verankerung sind Voraussetzungen dafür, dass genügend Sicherheit darüber besteht, dass nicht-kooperative Individuen effektiv bestraft oder sogar entlassen werden können (siehe Lindenberg 2000). Die Schaffung gemeinsamer Normen und Erwartungen und die damit verbundene Erzeugung von interpersonalem Vertrauen sind wichtige Vorbedingungen dafür, dass sich Individuen an reziprokem oder generalisiertem Austausch beteiligen. Denn dabei riskieren sie, ausgenutzt zu werden, weil sich Gegenleistungen nicht gesetzlich erzwingen lassen. Solche Tauschhandlungen sind jedoch wesentlich für gemeinschaftliche Produktion. Daher sind diese beiden Austauschformen auf direktere Weise mit kognitiven Interdependenzen verknüpft. Vertrauen und wechselseitig anerkannte Normen können sogar zur Folge haben, dass die gesetzlichen Sanktionsmöglichkeiten nie oder selten angewandt werden müssen (siehe Ellickson 1991). Es ist sogar möglich, dass sie untergraben werden, wenn die externen Sanktionssysteme zur Anwendung kommen müssen. Demzufolge können die Beiträge der verschiedenen Formen sozialen Austauschs zu gemeinschaftlicher Produktion nur erklärt werden, indem Interaktionseffekte zwischen diesen Formen untersucht werden. Nur die Kombination von verhandlungsbasiertem und reziprokem Austausch führt zu gemeinschaftlicher Produktion. Untersuchungen dazu fehlen bislang. In ähnlicher Weise müssen soziale Netzwerke sorgfältig mit funktionalen und kognitiven Interdependenzen verknüpft werden. Sonst können wir wichtige Netzwerkdeterminanten für individuelle Beiträge zu gemeinschaftlich hergestellten Gütern verlieren. Trotzdem untersucht kaum eine Netzwerkstudie die Bedingungen, die den Zusammenhang zwischen Netzwerkstruktur und Ergebnissen (z.B. Leistung) beeinflussen. Die meisten Untersuchungen unterstellen, dass die Auswirkung von Netzwerken auf Verhalten nicht vom Kontext abhängt.

IV. Erste Veranschaulichung zur Anwendung der Heuristik: Netzwerke für Strategieentscheidungen in der Politik und in Organisationen (policy networks)

Die Notwendigkeit, zu einem kollektiven Ergebnis zu kommen, kann als ein Sonderfall von funktionaler Interdependenz und gemeinschaftlicher Produktion betrachtet werden. Es gibt ein *gemeinsames Interesse*, nämlich zu einer Entscheidung zu gelangen, die dann für alle verbindlich ist. Ferner bestehen *gegensätzliche Interessen*, da jeder Beteiligte ein Ergebnis wünscht, das seiner eigenen Position möglichst nahe kommt. Demnach kann die beschriebene Heuristik hier angewandt werden. Kollektive Entscheidungsfindung beschränkt sich nicht auf die Politik. Die meisten Prozesse innerhalb von Organisationen können ebenfalls als kollektive Entscheidungsprozesse gesehen werden: Welche Strategie sollte die Organisation verfolgen? Wie sollten verschiedene Produktionsprozesse gestaltet werden? Wie können Prozesse und Gehaltssysteme in einer Weise gestaltet werden, dass sich die richtige Mischung aus Wettbewerb und Solidarität ergibt? Drei Typen von Netzwerken sind relevant, die mit drei grundlegenden Prozessen zusammenhängen, durch welche kollektive Ergebnisse erzeugt werden. Alle drei müssen gleichzeitig untersucht werden. Im Folgenden werden für jeden dieser Prozesse die Voraussetzungen angegeben, unter denen er bei der Entscheidungsfindung dominieren dürfte.

Derzeit werden in der Literatur fünf nebeneinander stehende Ansätze diskutiert, die zur Untersuchung von Strategieentscheidungen in der Politik und in Organisationen angewandt werden können. Vier davon sind Netzwerkansätze, der fünfte ist ein spieltheoretischer Ansatz, der nicht auf Netzwerken basiert. Zu erwähnen sind zunächst die Netzwerkeinflussmodelle. Sie beschreiben den sozialen Einfluss auf individuelle Meinungen und Einstellungen. Diese so genannten Ansteckungsmodelle *(contagion models)* (Friedkin und Johnsen 1990, 1999; Leenders 1995) gehen davon aus, dass Meinungen und Einstellungen von Akteuren in einem sozialen System nur teilweise von individuellen Merkmalen abhängen. Zusätzlich werden sie durch sozialen Einfluss geprägt. Dieser wird mit einem Einflussnetzwerk dargestellt, das den dyadischen Einfluss der Akteure aufeinander widerspiegelt. In technischer Hinsicht werden räumliche Autokorrelationsalgorithmen verwendet, um solche Prozesse zu beschreiben. Die Meinungen und Einstellungen einer Person ergeben sich aus dem gewichteten Mittel der Einstellungen anderer Akteure, die Einfluss auf diese Person ausüben. In der Literatur wird eine Vielzahl an Gewichtungsverfahren vorgeschlagen, aus denen sich unterschiedliche Modelle ergeben (Leenders 2002).

Der zweite Ansatz betrifft Erweiterungen des Austauschmodells von Coleman (1972, 1990) auf Netzwerke. Coleman ging davon aus, dass Akteure an bestimmten Ereignissen Interesse haben, während sie andere Ereignisse kontrollieren. Sie können nun Kontrolle über Ereignisse, an denen ihnen weniger liegt, gegen Kontrolle über Ereignisse austauschen, auf die sie mehr Wert legen. Dadurch können für beide Seiten vorteilhafte Ergebnisse erzielt werden. Das Ausmaß individueller Macht und der Wert von Ereignissen wird nicht ad hoc eingeführt, sondern aus dem Modell abgeleitet. Später wurde Colemans Austauschmodell mit Netzwerken verknüpft, um ungleiche Tauschmöglichkeiten zu berücksichtigen (Marsden und Laumann 1977; Laumann et al. 1987; Braun 1993; König 1997; Pappi und Henning 1998). Diese Modelle wurden angewandt, um Ergebnisse dichotomer Ja/Nein-Entscheidungen vorherzusagen. Daher

konnte Colemans Modell auch auf Ergebnisse von kollektiven Entscheidungsprozessen ausgedehnt werden.

Der dritte und vierte Ansatz beinhalten die Netzwerkaustauschmodelle. Diese wurden bereits in vorangegangenen Abschnitten behandelt. Während Colemans Modell von globalen Gleichgewichten ausgeht, konzentrieren sich die Modelle des verhandlungsbasierten Austauschs der NET darauf, welchen Einfluss Netzwerke auf Austauschraten zwischen den einzelnen Paaren von Akteuren ausüben. Macht entsteht in erster Linie durch die Möglichkeit, andere vom Austausch auszuschließen. Mächtige Akteure können Tauschverhältnisse zum eigenen Vorteil verschieben. Die Modelle des reziproken Austauschs untersuchen unter anderem, wie die Machtsstruktur und die Machtstrategien von Individuen die Entwicklung der Beziehungen zwischen Individuen beeinflussen (Molm 1997).

Der fünfte Ansatz ist von Bueno de Mesquita (1994; Bueno de Mesquita et al. 1985). Er betont, dass Politik im Wesentlichen Konfliktlösung ist, bei der Macht eine größere Rolle spielt als Argumente oder Austausch. Interessenkonflikte zwischen Akteuren drücken sich in unterschiedlichen Positionen hinsichtlich des gewünschten Ergebnisses politischer Entscheidungen aus. In solchen Situationen werden kollektive Ergebnisse durch einen Prozess erzielt, in dem Akteure die Positionen anderer herausfordern *(challenge process)*. In solchen Prozessen spielen Unterschiede in Macht und Interesse *(salience)* eine bedeutende Rolle. Je geringer das eigene Interesse an einem Abstimmungspunkt *(issue)* und je höher das Machtübergewicht der Gegenposition, desto eher wird ein Akteur geneigt sein, seine eigene Position aufzugeben. Dieser Prozess wird durch ein nicht-kooperatives Spiel dargestellt.

Drei der fünf Ansätze bieten nur auf lokaler Ebene Vorhersagen. Der eine umfasst Ansteckungsmodelle, in denen es um individuelle Meinungen geht. Der andere bezieht sich auf die Modelle der NET, die sich auf Tauschverhältnisse in Dyaden konzentrieren. Es ist unklar, unter welchen Bedingungen welcher Ansatz geeignet ist. Für die Autoren selbst ist dies anscheinend kaum von Bedeutung, zumal sie kaum aufeinander Bezug nehmen.

Um die Interdependenzen verstehen zu können, die bei der kollektiven Entscheidungsfindung auftreten, ist eine klare Unterscheidung zwischen Endzielen und instrumentellen Zielen erforderlich. Instrumentelle Ziele können als Mittel zur Realisierung von Endzielen betrachtet werden (z.B. eine effiziente Fischerei-Infrastruktur als Beitrag für den ökonomischen Wohlstand des Landes). Nutzenfunktionen für Endziele sind normalerweise monoton steigend oder fallend. Endziele dominieren in politischen Diskussionen, während die einzelnen Abstimmungspunkte in der Regel instrumentelle Ziele betreffen. Instrumentelle Ziele besitzen typischerweise ein Optimum: zu viel ist ebenso schlecht wie zu wenig. Wo für den Einzelnen das Optimum liegt, ist bedingt durch die Gewichtung der unterschiedlichen Endziele. Diese Gewichtung hängt von den Einschränkungen ab, denen ein Akteur begegnet. Daher ist zu erwarten, dass verschiedene Akteure unterschiedliche Optima anstreben. Nun stehen auch die Endziele bestimmter Akteure stärker auf dem Spiel als die anderer Akteure. Beispielsweise ist der ökonomische Wohlstand von Luxemburg weniger von einer effizienten Fischerei-Infrastruktur abhängig als es etwa in den Niederlanden der Fall ist. Damit unterscheidet sich auch das Interesse der Akteure am Ergebnis einer Abstimmung. Basierend auf die-

sen Überlegungen wird kollektive Entscheidungsfindung über Abstimmungspunkte mit eingipfligen Präferenzfunktionen beschrieben. Problemlösungen können durchaus gleichzeitige Entscheidungen über mehrere Abstimmungspunkte erfordern. Diese verschiedenen Punkte müssen möglichst unabhängige strittige Elemente des Gesamtproblems erfassen. Alle zusammen müssen den gesamten Bereich möglicher Ergebnisse abdecken.

Ein Beispiel ist das Strukturproblem der Fischerei-Industrie, das die Europäische Kommission (KOM) mit dem Verordnungsvorschlag KOM(98) 728 regeln wollte (siehe *Abbildung 2*).[7] Der Vorschlag ist mit mindestens zwei Endzielen verbunden: langfristige Nutzung der Fischbestände und Aufrechterhaltung einer wettbewerbsfähigen europäischen Fischerei-Industrie. Sachverständige stellten im Vorschlag der Europäischen Kommission zwei Abstimmungspunkte fest, über die entschieden werden musste. Der erste betraf das Ausmaß der Zwangsstilllegung bzw. -verschrottung alter Schiffe *(scrap-built penalty)*. Für jedes neue Schiff mit einem Laderaum für 100 Tonnen sollte eine bestimmte Menge alter Schiffe stillgelegt werden. Diese Auflage sollte auch die Nachfrage nach Subventionen für Flottenerneuerungen reduzieren. Die Europäische Kommission schlug in ihrem Verordnungsvorschlag vor, dass jeweils 130 Tonnen an Alttonnage stillgelegt werden sollten. Großbritannien (GB) vertrat die extremste Position: jeweils eine Stilllegung von 150 bis 180 Tonnen. Diese Position ist auf der Skala mit 100 bezeichnet. Das andere Extrem war der damalige Status quo (auf der Skala 0). Dabei wurde eine Stilllegung von 100 Tonnen Alttonnage gefordert, d.h. gleich der Menge neuer Schiffe. Die meisten EU-Mitgliedstaaten (MS) wählten den Status quo als Ausgangsposition. Den befragten Sachverständigen zufolge müsste das von der Kommission bevorzugte Ergebnis (eine Zwangsstilllegung von 130 Tonnen) auf unserer Skala 90 Punkte erhalten, weil diese Alternative der britischen Position viel näher kommt als dem Status quo. Dänemark (DK) und Österreich (AT), erhielten eine Punktzahl zwischen 90 und 100.

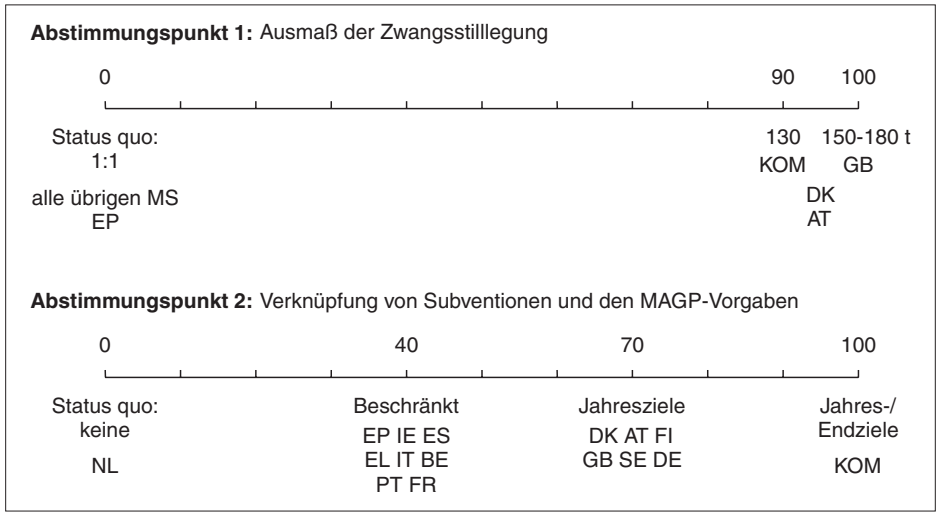
In einem mehrjährigen Ausrichtungsprogramm *(Multi Annual Guidance Program: MAGP)* sind Jahres- und Endziele für die einzelnen Mitgliedstaaten festgeschrieben, die den Umfang der Fischereiflotte beschränken sollen. Der zweite Abstimmungspunkt war nun, in welchem Maße die Mitgliedstaaten ihre MAGP-Vorgaben erreichen müssen, um Subventionen zu erhalten. Die Europäische Kommission vertrat den Standpunkt, dass Subventionen für die Flottenerneuerung von der strikten Befolgung aller MAGP-Ziele abhängen sollten (100 auf der Skala). Die Niederlande (NL) hatten die größten Schwierigkeiten, die MAGP-Vorgaben zu erfüllen. Dies veranlasste sie dazu, die extremste Gegenposition einzunehmen. Sie hätten am liebsten, dass Subventionen über-

7 Von einer internationalen Projektgruppe wurde eine größere Studie über Entscheidungsfindung in der Europäischen Union (EU) durchgeführt. Durch Interviews mit Experten und der Analyse von Dokumenten wurden 162 Abstimmungspunkte in 66 Vorschlägen der Europäischen Kommission im Zeitraum 1996–2000 festgestellt. Diese Abstimmungspunkte wurden so spezifiziert, dass auf einer Skala von 0 bis 100 alle Positionen, das tatsächliche Ergebnis und nach Möglichkeit auch der Status quo dargestellt werden konnten. Das Interesse der Akteure an den Abstimmungspunkten wurde auf einer Skala von 0 Prozent (kein Interesse) bis 100 Prozent (größtes Interesse) angegeben (für Einzelheiten siehe Stokman und Thomson 2004; Thomson et al. 2005). Das Beispiel wird auch in van Assen et al. (2003) angeführt.

haupt nicht mit der Einhaltung der MAGP-Vorgaben verknüpft würden (damals der Status quo).

Die Entscheidung über diese beiden Abstimmungspunkte fiel letztlich so aus, dass die 1:1-Stilllegung (Skalenpunkt 0) beibehalten wurde, aber Subventionen mit den jährlichen Zielvorgaben verknüpft wurden (Skalenpunkt 70). Die Niederlande und Großbritannien enthielten sich bei der Endabstimmung ihrer Stimmen, während alle anderen Akteure dafür stimmten.

Abbildung 2: Zwei Abstimmungspunkte in KOM(98) 728 zum Fischerei-Strukturproblem



Beim Abstimmungspunkt 1 nimmt NL keine Position ein.

Abkürzungen: Europäisches Parlament (EP), Europäische Komission (KOM), Mitgliedstaaten (MS); Österreich (AT), Belgien (BE), Deutschland (DE), Dänemark (DK), Griechenland (EL), Spanien (ES), Finnland (FI), Frankreich (FR), Großbritannien (GB), Irland (IE), Italien (IT), Niederlande (NL), Portugal (PT), Schweden (SE).

Jeder Akteur versucht unterschiedlich intensiv die eigene Position durchzusetzen und nutzt dazu verschiedene Möglichkeiten. Es kann jedoch nur ein einziges Ergebnis erzielt werden. Dadurch entsteht die Dynamik im Entscheidungsprozess. In einer komplexen Situation mit vielen Beteiligten werden sich die Akteure darum bemühen, möglichst große Koalitionen zu bilden. Die Position einer solchen Koalition muss dabei entweder der eigenen Ausgangspositionen entsprechen oder dieser möglichst nahe kommen. Auf diese Weise hoffen die Akteure, die Positionen derjenigen zu beeinflussen, die letztlich die Entscheidung treffen. Denn damit erzielen sie ein kollektives Ergebnis, das ihre Interessen so weit wie möglich widerspiegelt. Die Dynamik der Entscheidungsfindung basiert daher vor allem auf Prozessen, durch die Akteure dazu veranlasst werden, ihren Standpunkt zu ändern. Dies kann freiwillig geschehen oder durch Zwang. Drei grundlegende Prozesse können solche Verschiebungen von Positionen bewirken: *Überzeugen (persuasion), Stimmentausch* (*logrolling*) und *Erzwingung (enforcement)*. Jeder dieser Prozesse ist mit spezifischen Interdependenzen verbunden. *Tabelle 1*

Tabelle 1: Grundlegende Prozesse, dominante Netzwerke, Ansätze und Bedingungen, unter denen Prozesse dominieren

Grundlegende Prozesse	***Dominante Netzwerke***	***Derzeitige, nicht integrierte Ansätze***	***Integrierter Ansatz***	***Bedingungen, unter denen ein Prozess dominiert***
Überzeugen	Informations-netzwerke	(1) Ansteckungs-modelle (2) Netzwerke des reziproken Austauschs	kooperative Nash-Verhand-lungs-Lösung für alle relevanten Akteure	(a) keine Über-einkunft *(reversal point)* sehr unattraktiv (b) Gesamtkoalition möglich, aber Teilkoalitionen schwierig zu bilden (c) risiko-aversive Akteure
Stimmen-tausch	Netzwerke des verhandlungs-basierten Austauschs	(3) Colemans Austausch-modell (4) Netzwerk-Austausch-theorie (NET)	Stimmen-tauschmodell (kooperative Lösungen für einige Akteure mit positiven und/oder negativen Externalitäten für andere)	gegensätzliche Positionen und komplementäre Interessen
Erzwingung	hierarchische Netzwerke/ Machtnetzwerke	(5) nicht-kooperative Modelle	(nicht-kooperatives) Herausforderungsmodell	gegensätzliche Positionen, aber nicht-komplementäre Interessen

gibt eine Übersicht über diese drei Prozesse. Dargestellt werden die dazugehörigen Typen von Netzwerken, die jeweiligen theoretischen Ansätze, der hier vorgeschlagene integrierte Ansatz und die Voraussetzungen, unter denen diese Prozesse bei kollektiver Entscheidungsfindung vorherrschen.[8]

Durch *Überzeugen* versuchen die Akteure, sowohl die Ausgangspositionen anderer als auch deren Interesse an der jeweiligen Ausgangsposition zu verändern (Stokman et al. 2000). Die Voraussetzungen, unter denen durch überzeugende Informationen eine Entscheidung erzielt werden kann, sind in *Tabelle 1* dargestellt. Erstens müssen gemeinsame Interessen vorherrschen. Dann ist eine Entscheidung erwünscht, die durch alle beteiligten Akteure getragen wird. Wenn eine solche Entscheidung nicht erreicht werden kann *(reversal point)*, sind das Ergebnis und alle damit verbundenen Konsequenzen sehr unattraktiv. Es kann wesentlich schlechter sein als der Ausgangspunkt (Status quo). Zweitens muss die Bildung einer großen Koalition aller Akteure möglich

8 In seiner soziologischen Kritik an ökonomischen Modellen der Politik leitet Udehn (1996) aus der Literatur dieselben drei grundlegenden Prozesse ab.

sein. Drittens muss die Bildung kleinerer Koalitionen schwierig sein. Dies ist typischerweise der Fall, wenn die Koalitionen von Abstimmungspunkt zu Abstimmungspunkt anders liegen und sich keine Koalition bilden lässt, die über längere Zeit stabil bleibt.[9]

Akteure müssen darauf vertrauen können, dass die erhaltenen Informationen aufrichtig und nicht strategisch manipuliert sind. Eigene, persönliche Vorteile zu verfolgen ist erlaubt, solange dies anderen nicht schadet und solange die persönlichen Vorteile und die gemeinsamen Interessen miteinander vereinbar sind. Solche Bedingungen lassen sich nur schaffen, wenn die Akteure in ein dichtes Vertrauensnetzwerk eingebettet sind oder bei Abweichungen von den gemeinsamen Interessen mit schweren Strafen rechnen müssen. Innerhalb eines solchen Kontextes können sich Akteure darauf verlassen, Gegenleistungen zu erhalten. Sie machen anderen Akteuren Zugeständnisse, die großes Interesse an den aktuellen Abstimmungspunkten haben. Wenn sie nun später an einem anderen Abstimmungspunkt selbst großes Interesse haben, können sie eine entsprechende Kompensation von anderen erwarten. Reziproke und generalisierte Tauschhandlungen sind daher ein wesentlicher Bestandteil der Entscheidungsfindung durch Überzeugen (und nicht, wie die Bezeichnung vielleicht vermuten ließe, beim Stimmentausch, der auf verhandlungsbasierten Austausch basiert).

Stimmentausch und Erzwingung betreffen in der Regel nicht die Ausgangspositionen und die Interessen. Stimmentausch kann als ein Prozess verhandlungsbasierten Austauschs betrachtet werden. Das Ergebnis ist, dass Akteure eine andere Position bei einem Abstimmungspunkt unterstützen, der für sie von relativ geringem Interesse ist. Im Gegenzug erhalten sie dafür Unterstützung von einem anderen Akteur bei einem Abstimmungspunkt, der ihnen wichtiger ist. In ähnlichem Sinne können Akteure angesichts eines Übergewichts an Macht gezwungen sein, eine andere Position einzunehmen. Diese beiden Prozesse sind vor allem dann zu erwarten, wenn sich die Ausgangspositionen grundsätzlich voneinander unterscheiden. Dies ist der Fall, wenn Akteure die Endziele sehr unterschiedlich gewichten. In solchen Situationen helfen keine Argumente, um die Ausgangspositionen einander anzunähern. Koalitionen können dann nur durch Prozesse gebildet werden, die die Endpositionen oder das Stimmverhalten der Akteure betreffen.

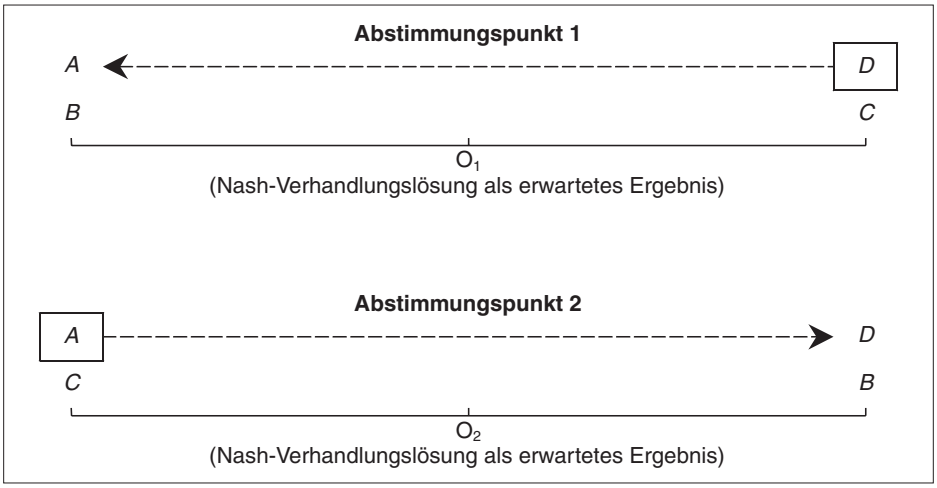
Beim Überzeugen sind Informations- und Vertrauensnetzwerke vorherrschend. Hingegen bestimmen Netzwerke des verhandlungsbasierten Austauschs die Möglichkeiten beim *Stimmentausch*. Während die NET statische Netzwerke untersucht, werden Austauschnetzwerke im integrierten Ansatz aus der Verteilung von Positionen und Interessen der Beteiligten an den Abstimmungspunkten abgeleitet (siehe Abbildung 3).[10]

Mit Positionsverschiebungen sind Gewinne und Verluste verbunden. Diese hängen von der Veränderung der Ergebnisse der beiden Abstimmungspunkte ab, die aus sol-

9 Achen (2006) zeigte, dass der Durchschnitt der politischen Positionen, jeweils gewichtet mit der Macht und dem Interesse eines Akteurs, unter diesen Umständen eine Näherung ersten Grades für die kooperative Nash-Verhandlungslösung ist. Dabei ging er von quadratischen Verlustfunktionen bezüglich der Abstimmungspunkte aus (d.h. von risikoscheuen Akteuren).

10 Colemans Austauschmodell weist als Modell kollektiver Entscheidungsfindung einige Nachteile auf (siehe Stokman und van Oosten 1994). Der wichtigste besteht darin, dass das Modell von Coleman auf dem Austausch von Kontrolle basiert. Dabei wird angenommen, dass der Nutzen von Kontrolle monoton steigend ist. Dies ist jedoch eher eine Eigenschaft von Endzielen als von instrumentellen Zielen.

Abbildung 3: Auswirkungen eines Stimmentauschs zwischen zwei Akteuren *A* und *D*



Die Pfeile bedeuten, dass Akteur *D* seine Position beim ersten Abstimmungspunkt in *A*'s Richtung verschiebt, während sich Akteur *A* beim zweiten Abstimmungspunkt Akteur *D* annähert.

chen Verschiebungen resultieren.[11] Zwei Akteure aus Gruppen mit entgegengesetzten Positionen können vom Stimmentausch profitieren, wenn das relative Interesse der beiden Gruppen an den Abstimmungspunkten unterschiedlich ist (siehe *Tabelle 1*) (Stokman und van Oosten 1994). Ein Stimmentausch ist dann für beide vorteilhaft, hat jedoch auch Nebenwirkungen oder externe Effekte auf den Nutzen anderer (siehe *Abbildung 3*). Angenommen Akteur *A* hat im Vergleich zu Akteur *D* ein größeres Interesse am ersten Abstimmungspunkt als am zweiten. Abstimmungspunkt 1 ist dann *A*'s Nachfragepunkt und *D*'s Angebotspunkt. Die beiden Akteure *A* und *D* können nun ihre Positionen tauschen. Dazu verschiebt *D* seine Position beim ersten Abstimmungspunkt in *A*'s Richtung, während sich *A* beim zweiten Punkt *D* annähert. Dadurch bewegen sie sich weg von Akteur *C* und unterstützen *B*'s Position. In diesem Beispiel ist *C* doppelt gestraft und *B* doppelt belohnt, obwohl beide nicht am Austausch mitgewirkt haben. Nun können *A* und *D* auch Gruppen von Akteuren mit gleicher Position repräsentieren. Positive und negative Externalitäten können sich dann auch innerhalb dieser Gruppen durch einen solchen Austausch ergeben. Für positive Externalitäten müssen die Interessen an den Abstimmungspunkten innerhalb jeder Gruppe relativ homogen sein. Andernfalls entstehen negative Externalitäten. Im Extremfall möchte ein Mitglied der Gruppe *A* beispielsweise den ersten Abstimmungspunkt als Angebots-

11 Durch das Ergebnis eines Abstimmungspunktes kann ein Akteur einen Nutzenverlust erleiden. Es wird angenommen, dass sein Verlust auf dem Kontinuum eines Abstimmungspunktes dem Abstand zwischen seiner politischen Position und dem eingetretenen Ergebnis entspricht. Diese Differenz wird dann mit dem Interesse des Akteurs am Abstimmungspunkt gewichtet. Zusätzlich geht das Modell davon aus, dass die anfänglichen Ergebniserwartungen der Akteure dem Mittelwert der Ausgangspositionen der *n* Akteure entsprechen. Dieser wird mit dem Produkt ihrer jeweiligen Möglichkeiten und Interessen gewichtet (siehe Fußnote 9).

punkt nutzen, während ein anderes Mitglied der Gruppe *A* darin einen Nachfragepunkt sieht.

Aus der Kombination der beiden Abstimmungspunkte im Verordnungsvorschlag KOM(98) 728 bezüglich der Fischereiverordnung ergeben sich vier Gruppen (siehe *Abbildung 4*). Die Mitgliedstaaten, die Europäische Kommission und das Europäische Parlament können entsprechend der Abstimmung in diese Gruppen eingeteilt werden. In *Abbildung 4* ist die Gruppe *C* leer. Damit sind Tauschhandlungen für die Gruppe *B* wegen Mangel an Tauschpartnern ausgeschlossen. Gruppe *A* besteht aus 8 Akteuren, Gruppe *D* aus 4. Daraus ergibt sich ein Austauschnetzwerk, in dem 32 Tauschhandlungen möglich sind. Eine davon muss jedoch ausgeschlossen werden, weil die relativen Interessen an den Abstimmungspunkten für das Europäische Parlament und Großbritannien dieselben sind.[12] Verglichen mit den Mitgliedern der Gruppe *A* besitzen alle Mitglieder der Gruppe *D* ein relativ größeres Interesse an der MAGP-Frage. Daher ist der erste Abstimmungspunkt (Zwangsstilllegungen) der Angebotspunkt von Gruppe *D* und der Nachfragepunkt von Gruppe *A*. Aus allen Tauschhandlungen ergeben sich Verschiebungen in dieselbe Richtung, nämlich hin zur Ausgangsposition der Gruppe *B*.

Tabelle 2: Stimmentausch mit ausschließlich positiven Externalitäten in KOM(98) 728

		Abstimmungspunkt 2: **Verknüpfung von Subventionen und MAGP-Vorgaben**	
		Keine oder beschränkte Verknüpfung	*Verknüpfung mit Jahres- und/oder Endzielen*
Abstimmungspunkt 1: → **Ausmaß der Zwangsstilllegung**	*1:1*	*A:* EP; IE ES EL IT BE PT FR →	*B:* DE SE FI
	130 Tonnen oder mehr	*C:* (leer)	*D:* KOM; DK AT GB ↑

Die Pfeile bedeuten, dass die Akteure der Gruppe *D* ihre Positionen beim ersten Abstimmungspunkt in Richtung der *A*-Akteure verschieben, während sich die Akteure der Gruppe *A* beim zweiten Abstimmungspunkt den *D*-Akteuren annähern. Beide Gruppen nähern sich dann auf beiden Abstimmungspunkten der Position der Akteure der Gruppe *B* an.

NL ist nicht angegeben, weil sie beim Abstimmungspunkt 1 keine Position einnehmen.

Abkürzungen: Europäisches Parlament (EP), Europäische Komission (KOM), Mitgliedstaaten (MS); Österreich (AT), Belgien (BE), Deutschland (DE), Dänemark (DK), Griechenland (EL), Spanien (ES), Finnland (FI), Frankreich (FR), Großbritannien (GB), Irland (IE), Italien (IT), Niederlande (NL), Portugal (PT), Schweden (SE).

12 Gruppe *D* kann Mitglieder der Gruppe *A* ausschließen. Daher sagt die NET große Macht für Gruppe *D* vorher. Allerdings trifft dies nicht zu. Denn *alle* Akteure sind von Verschiebungen der Ergebnisse durch bilaterale Tauschhandlungen betroffen.

Daher stimmt das Endergebnis mit den Ausgangspositionen der Mitglieder der Gruppe *B* überein.

Nun kann spezifiziert werden, unter welchen Bedingungen Stimmentausch auf Grundlage bilateraler Tauschhandlungen mit kognitiven Interdependenzen vereinbar ist, die auf Überzeugen und Konsensbildung basieren:

1. Eine der vier Gruppen sollte leer sein. Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass Gruppe *C* leer ist (wie in *Abbildung 3* und *Tabelle 2*).
2. Die relativen Interessen der beiden tauschfähigen Gruppen erlauben eine Verschiebung in Richtung der nicht-leeren und nicht-tauschfähigen Gruppe (z.B. Austausch zwischen *A* und *D* mit Verschiebung in Richtung *B*).
3. Die relativen Interessen der beiden tauschfähigen Gruppen sind so beschaffen, dass keine negativen Externalitäten innerhalb der eigenen Gruppe auftreten.

Für den dritten Punkt müssen folgende Voraussetzungen gegeben sein. Angenommen und ohne Beschränkung der Allgemeinheit, dass ein Akteur der Gruppe *A* im Vergleich zu allen übrigen Akteuren der Gruppen *A* und *D* das größte relative Interesse am ersten Abstimmungspunkt besitzt. Es kommt nicht zu negativen Externalitäten, wenn das Tauschverhältnis zwei Bedingungen erfüllt. Erstens muss es niedriger sein als das relative Interesse desjenigen Akteurs der Gruppe *A*, der für diesen Abstimmungspunkt das geringste relative Interesse hat. Zweitens muss es aber auch höher sein, als das relative Interesse für denjenigen Akteur aus Gruppe *D*, der für diesen Abstimmungspunkt das größte relative Interesse hat. Negative Externalitäten innerhalb der eigenen Gruppe sind demnach unvermeidlich, wenn sich Interessen der Akteure verschiedener Gruppen überschneiden.

Jeder bilaterale Stimmentausch, für den mindestens eine der drei Bedingungen nicht erfüllt ist, ist mit negativen Externalitäten verbunden. Negativen Externalitäten verhindern die Konsensbildung. Eine Lösung ist dann, dass Akteure, die durch negative Externalitäten benachteiligt sind, mit Hilfe dritter Abstimmungspunkte entschädigt werden.

Die Verteilung der Positionen und Interessen, wie sie sich hinsichtlich der Fischereiverordnung zeigt, ist eher eine Ausnahme unter den 66 untersuchten Anträgen der Europäischen Kommission mit strittigen Abstimmungspunkten. Insgesamt sind die negativen Externalitäten der bilateralen Tauschhandlungen, die vom Modell vorhergesagt wurden, etwa doppelt so hoch wie die positiven Externalitäten. Dies zeigt, dass bilaterale Tauschhandlungen bei diesen Anträgen tendenziell *nicht* zur Konsensbildung beitragen. Überzeugen und Stimmentausch stehen hier offensichtlich im Konflikt miteinander (Arregui et al. 2006).

Erzwingung, nicht Überzeugen, kennzeichnet den Machtprozess bei kollektiver Entscheidungsfindung (siehe *Tabelle 1*). Die Beteiligten zeigen, dass sie genügend Macht besitzen, um eine Entscheidung durchzusetzen oder Alternativen zu verhindern. Damit versuchen sie, hinter ihrer eigenen Position für das gewünschte Ergebnis eine möglichst große Koalition aufzubauen. Lösungen für wesentliche Probleme werden nicht mit Argumenten gesucht. Vielmehr versuchen die Beteiligten zu zeigen, dass sie über genügend Unterstützung verfügen, um die Entscheidung mit Hilfe formeller Verfahren oder informeller Machtmittel zu erzwingen. Die Einwilligung in Lösungen spielt eine größere Rolle als ihre Billigung. Wenn Strategieentscheidungen in Organisationen überwie-

gend durch Erzwingung erzielt werden, spielt Hierarchie auch in der Vorbereitungsphase zur Entscheidungsfindung eine größere Rolle als Argumente. In einer solchen Situation werden die Organisationsziele weniger als gemeinsame Ziele betrachtet, sondern vielmehr als von der Organisationsspitze gewünscht und vorgegeben. Diese Aspekte führen zu kognitiven Interdependenzen, in denen persönliche Beziehungen vorrangig durch die hierarchische Position der Individuen bestimmt werden. Mit anderen Worten, Machtnetzwerke sind für die Ergebnisse kollektiver Entscheidungsprozesse ausschlaggebend. Die Ergebnisse von Entscheidungsprozessen können als Resultat eines nicht-kooperativen Spiels betrachtet werden, in dem keine verbindlichen Vereinbarungen möglich sind (Bueno de Mesquita et al. 1985; Bueno de Mesquita 1994).

Bilateraler verhandlungsbasierter Austausch kann auf Konsens beruhen oder Konsensbildung sogar verstärken. Voraussetzung ist, dass durch den Austausch umfangreiche positive und keine negativen Externalitäten entstehen. Überzeugen auf der Grundlage gemeinsamer Interessen erfordert auch klare institutionelle Regeln und davon abgeleitete klare Verantwortlichkeiten. Formelle institutionelle Regeln verknüpfen gemeinschaftliche Produktion mit externen (gesetzlichen) Sanktionierungssystemen, um Kooperation durchzusetzen. Dies gewährleistet, dass genügend Sicherheit darüber besteht, dass nicht-kooperative Akteure effektiv bestraft werden können. Informelle soziale Normen können besagen, dass Entscheidungen auf einem Konsens beruhen sollten. Wenn kognitive Interdependenzen mit solchen Normen verknüpft sind, bieten formelle Regeln (z.B. gesetzliche Verträge) den Akteuren Absicherungen für den Fall, dass schwerwiegende Probleme bei der Entscheidungsfindung auftreten oder andere Akteure von den informellen Regeln abweichen. Doch je öfter man auf formelle Absicherungen zurückgreifen muss, desto stärker wird die Norm der Konsensbildung unter Druck geraten. Gewöhnlich beendet es eine Beziehung, wenn vertragliche Vereinbarungen vor Gericht geltend gemacht werden müssen. Ebenso dürfte eine häufige Anwendung formeller Regeln die Auffassung untergraben, dass die gemeinsamen Interessen wesentlich genug sind, um Konsens zu erreichen. Bereits das bloße Vorhandensein der Regeln sollte Übereinstimmung und Einhaltung von Absprachen herbeiführen.

In zukünftiger Forschung sollten die Übergänge von einem dominierenden Prozess zum anderen bei der Entscheidungsfindung untersucht und modelliert werden. Dabei kann sowohl Lindenbergs Theorie über den Wechsel kognitiver Frames (*frame switch*) (Lindenberg und Frey 1993; Lindenberg 1998, 2000) als auch Essers (1997, 2000) Modell über Veränderungen der Definition von Situationen hilfreich sein.

Welcher dieser drei Netzwerktypen (Informations-, Austausch- oder Machtnetzwerk) im Kontext der Europäischen Union die wichtigste Rolle spielt, wurde anhand der Vorhersagegenauigkeit dreier Modelle beurteilt. Die Genauigkeit eines Modells ergibt sich aus dem Abstand zwischen den vom Modell vorhergesagten Ergebnissen und den tatsächlichen Ergebnissen auf den Skalen der jeweiligen Abstimmungspunkte (Stokman und Thomson 2004). Die kooperativen Modelle, die die Positionen aller Entscheidungsträger in der EU berücksichtigen, führen zu den genauesten Vorhersagen. In der EU ist *„Einstimmigkeit, wo immer möglich"* eine sehr wichtige Norm, auch wenn eine qualifizierte Mehrheit gesetzlich erlaubt ist (siehe auch Mattila und Lane 2001). In die Entscheidungsergebnisse der EU fließen meist die wesentlichen Interessen von allen Beteiligten mit ein. Außerdem vermeiden die Akteure, die wesentlichen Interessen

anderer zu schädigen. Dies bedeutet, dass Informationsnetzwerke, in denen Prozesse des Überzeugens vorherrschen, im europäischen Kontext überwiegen. Wie bereits geschrieben, sind Netzwerke des verhandlungsbasierten Austauschs in der Europäischen Union für die Konsensbildung häufig nicht förderlich. Dies liegt am großen Ausmaß negativer Externalitäten, die mit bilateralem Austausch verbunden sind. Austauschnetzwerke spielen im europäischen Kontext keine dominante Rolle, da Konsensbildung eine vorherrschende Norm ist. Dies zeigt sich daran, dass die Vorhersagen mit Austauschmodellen schlechter sind als die mit dem Modell des Überzeugens. Dasselbe gilt für Machtnetzwerke. Denn nicht-kooperative Verhandlungsmodelle schneiden sogar noch schlechter ab.[13] Daher lässt sich schließlich festhalten, dass auch im europäischen Kontext nicht die formellen Regeln das Verhalten bestimmen. Sie setzten vielmehr nur die Grenzen fest, innerhalb derer die Akteure handeln. Der Leser sollte sich darüber im Klaren sein, dass diese Schlüsse über die Entscheidungsfindung in der Europäischen Union nur gezogen werden können, indem eine vergleichende Analyse der drei Prozesse und der dazugehörigen Netzwerke durchgeführt wird.[14]

V. Zweite Veranschaulichung zur Anwendung der Heuristik: Struktur und Entstehung von Freundschaftsnetzwerken

Trotz aller Vieldeutigkeit des Wortes „Freund" besteht Übereinstimmung in der Vorstellung, dass ein Freund Unterstützung bietet, vertrauenswürdig ist, Rücksicht nimmt, Achtung und wirkliches Interesse zeigt, in Gesprächen aufgeschlossen ist und sich als Kamerad erweist. Somit sind Freundschaften mit hedonistischen Zielen verbunden. In der Literatur wird Anziehungskraft als wichtigste treibende Kraft in Freundschaften beschrieben. Dabei werden verschiedene Arten von Anziehungskräften erwähnt. Die vorherrschende Grundlage ist Ähnlichkeit. Individuen bevorzugen Freunde, die ihre Einstellungen, Werte und Überzeugungen teilen (u.a. Urberg et al. 1998; Duck 1991; Hallinan und Teixeira 1987; Brehm 1985; Werner und Parmelee 1979). Andere Anziehungskräfte können jedoch Freundschaften zwischen Individuen fördern, die einander nicht ähnlich sind. Davison and Jones (1976) berichten zum Beispiel von einer allgemein zu beobachtetenden Anziehungskraft, die von Individuen mit höherem Status ausgeht.

Freundschaft kann nun auch aus Sicht der Heuristik betrachtet werden. Dann stellen sich folgende zwei Fragen:

---

13 Dies ist auch bei den Verfahrensmodellen *(procedural models)* der Fall. Diese gehen einfach davon aus, dass die Ergebnisse von Entscheidungen das Resultat der Wechselwirkung zwischen institutionellen Regeln und Präferenzen der Akteure sind. Sie lassen jedoch die informellen Verhandlungsprozesse und die damit verbundenen Netzwerke außer Acht.

14 Die Gültigkeit dieses Ansatzes hat sich nicht nur in der wissenschaftlichen Forschung bewiesen, sondern findet auch in kommerziellen Projekten Anwendung. Dieser Ansatz ist ein erfolgreiches Instrument, das ein strategisches Eingreifen in Prozesse der Entscheidungsfindung ermöglicht. Dadurch werden Entscheidungen erzielt, die der Position des Klienten nahe kommen und doch genügend von andern Beteiligten unterstützt werden, um zu erfolgreichen Ergebnissen zu führen. Siehe www.decide.nl für nähere Einzelheiten.

1. Welche anderen funktionalen Interdependenzen können Freundschaften fördern?
2. Welche funktionalen Interdependenzen werden durch Freundschaften erzeugt?

Im Hinblick auf die erste Frage muss berücksichtigt werden, dass sich Freundschaft nur zwischen Individuen entwickeln kann, die eine Gelegenheit hatten, miteinander zu interagieren (Newcomb 1956; van Duijn et al. 2003). Bevor man Freundschaft schließen kann, muss man sich begegnet sein. Das Begegnen ist im hohen Maße von den Gelegenheitsstrukturen abhängig, die durch andere funktionale Interdependenzen und räumliche Nähe erzeugt werden. Freundschaften entstehen daher fast immer in Situationen, in denen Individuen verschiedene Aktivitäten gemeinsam verrichten. Dies verstärkt die Multifunktionalität der Beziehungen. Eine interessante Frage ist dann, durch welche anderen funktionalen Interdependenzen die Entwicklung von Freundschaften gefördert wird. In der Sprache sozialer Netzwerke ausgedrückt, geht es darum, unter welchen Voraussetzungen es wahrscheinlich ist, dass informelle Verbindungen zwischen Personen entstehen.

Im Hinblick auf die zweite Frage sollte untersucht werden, ob wir es nur mit einer einzigen Art von Beziehung zu tun haben oder mit mehreren. Es gibt empirische Hinweise darauf, dass in Bekanntschaften andere funktionale Interdependenzen erzeugt werden als in den wenigen starken Freundschaften. In Bekanntschaften bezeichnen sich die Personen durchaus als „Freunde". Wirkliche Freundschaftsbeziehungen müssen sich jedoch erst daraus entwickeln. Zunächst seien die Bekanntschaften betrachtet. Diese hängen mit drei funktionalen Interdependenzen zusammen: mit dem Bedürfnis nach Verhaltensbestätigung, nach Status und nach Anregungen. Im Folgenden wird ein allgemeines Modell vorgeschlagen, in dem diese drei Interdependenzen integriert sind.

Freunde zu haben, ist ein allgemeines Bedürfnis. Dennoch unterscheidet sich die Anzahl Freunde, die sich ein Individuum wünscht. Dies ist unter anderem bedingt durch Persönlichkeitsunterschiede und zeitliche Einschränkungen. Zudem unterscheiden sich Individuen in zweierlei Hinsicht darin, welche Eigenschaften sie bei ihren Freunden bevorzugen. Erstens kann eine *Merkmalsdimension*, die einer Person wichtig ist, für jemand anderes bedeutungslos sein. Beispielsweise ist manchen wichtig, dass ihre Freunde einer bestimmten Altersgruppe angehören oder bestimmte politische Ansichten vertreten. Hingegen ist anderen das Alter oder die politische Orientierung ihrer Freunde gleichgültig. Die Wichtigkeit einer Merkmalsdimension für eine Person wird als Interesse *(salience)* an dieser Merkmalsdimension bezeichnet. Zweitens unterscheiden sich Individuen darin, welche *Merkmale* sie bei ihren Freunden bevorzugen. Manche Personen bevorzugen Freunde im gleichen Alter. Andere suchen Freunde, die älter bzw. jünger sind als sie selbst. Ein solches bevorzugtes Merkmal bei Freunden ist der „Idealfreund-Wert" eines Individuums. Wie dieses Beispiel zeigt, kann dieser dem eigenen Wert des Individuums auf der Merkmalsdimension entsprechen, muss es jedoch nicht notwendig.

Diesen Überlegungen zufolge kann der Grundprozess der Freundschaftsbildung mit mehreren eindimensionalen Abstimmungspunkten modelliert werden. Die Individuen weisen dabei eingipflige Präferenzen für ihren Idealfreund-Wert bei der jeweiligen Merkmalsdimension auf. Je größer dabei der Abstand zwischen dem tatsächlichen Wert eines (potentiellen) Freundes und dem Idealfreund-Wert des Individuums, desto weniger attraktiv ist es, eine Freundschaft mit der anderen Person aufzubauen oder zu er-

halten. Ein solcher Abstand wiegt umso schwerer, je größer das Interesse eines Individuums an einer Merkmalsdimension ist. Dies entspricht dem Vorschlag von Jones (1982). Er hat die Stärke einer Ähnlichkeitsbeziehung als Umkehrfunktion des Abstandes zwischen dem impliziten Idealpunkt eines Individuums und dem Wert des anderen bei der jeweiligen Merkmalsdimension modelliert.

Es mag überraschen, dass Freundschaft somit auf eine ähnliche Weise untersucht werden kann wie Strategieentscheidungen in der Politik und in Organisationen (siehe auch Stokman 2004). Die Grundprozesse und die damit verbundenen Netzwerke sind jedoch grundlegend verschieden. Denn anders als bei Strategieentscheidungen in der Politik und in Organisationen ist gemeinschaftliche Produktion bei Freundschaft kein kollektives und für alle verbindliches Ergebnis. Gegenseitige Freundschaftsbeziehungen als Ergebnis gemeinschaftlicher Produktion sind für die Beteiligten nicht verbindlich. Ein Individuum sucht Freunde mit Eigenschaften, die den Idealfreund-Werten auf den Merkmalsdimensionen genügend nahe kommen, an denen das Individuum interessiert ist. Dies ermöglicht dem Individuum, (Selbst-)Bestätigung seines Verhaltens zu finden und zu erhalten. Von besonderem Interesse sind in diesem Zusammenhang daher drei Fragen: Welche Dimensionen sind wichtig? Welche Arten von Idealfreund-Werten sind verbreitet? Welchen Einfluss haben Dimensionen und Idealfreund-Werte auf die Struktur und die Entwicklung von Freundschaftsnetzwerken? Aufbauend auf früheren Forschungsergebnissen lassen sich drei Grundlagen für Idealfreund-Werte unterscheiden. Diese hängen mit den folgenden drei grundlegenden Anziehungskräften zusammen (Stokman und Zeggelink 1996):

– *Ähnlichkeit (similarity)* ergibt Idealfreund-Werte, die den eigenen Werten eines Individuums bei relevanten Dimensionen nahe kommen. Damit verbunden ist der Wunsch, von Freunden Verhaltensbestätigung zu erhalten.
– *Ehrgeiz (aspiration)* führt zu Idealfreund-Werten, die den Statusmerkmalen der Referenzgruppe entsprechen, zu denen wir gehören möchten (Festinger 1954).
– *Komplementarität (complementarity)* hat Idealfreund-Werte zufolge, die deutlich von den eigenen Merkmalen des Individuums abweichen. Damit verbunden ist der Wunsch, Anregung von Freunden zu erhalten.

Komplementäre Freundschaftsbeziehungen haben anscheinend nur dann Bestand, wenn reziproke Statusunterschiede vorhanden sind. Dabei gleicht der höhere Status eines Individuums bei einer Dimension den höheren Status aus, den der Freund bei einer anderen Dimension aufweist (Tesser 1988, 2000). Freundschaftsnetzwerke, die auf Verhaltensbestätigung beruhen, dürften eine hohe Transitivität besitzen. Denn Verhaltensbestätigung wird vor allem innerhalb von Gruppen erzeugt. Die anderen Anziehungskräfte dürften verstärkt Netzwerke erzeugen, die sich überlagern *(layered networks)*.

Es gibt Hinweise darauf, dass sich *starke Freundschaften* von den oben beschriebenen Bekanntschaften grundsätzlich unterscheiden, da sie nicht transitiv sind (van de Bunt 1999), obwohl sie zu den so genannten „starken Beziehungen" gehören (Granovetter 1973). Starke Freundschaftsbeziehungen sind dyadische Beziehungen, die mit ausgeprägter Multifunktionalität und der gemeinschaftlichen Produktion von Erinnerungen und von Identität verbunden sind (Lindenberg 2001). Es ist wahrscheinlich,

dass sie sich aus Bekanntschaften heraus entwickeln, aber sie besitzen besondere funktionale Interdependenz.

Im Allgemeinen sind insbesondere spieltheoretische Modelle bei der Analyse von *Netzwerken für Strategieentscheidungen in der Politik und in Organisationen* und Prozesssimulationsmodelle hilfreich. Neue statistische Modelle leisten einen hervorragenden Beitrag, um die Entwicklung und die Struktur von Freundschaftsnetzwerken zu analysieren. Zwei beliebige Akteure A und B können dabei als Punkte in einem Eigenschaftsraum dargestellt werden. Die meisten neueren statistischen Modelle basieren auf der Annahme, dass die Beziehung zwischen den Akteuren durch vier Zustände beschrieben werden kann: reziproke Wahl, einseitige Wahl von *A* durch *B*, einseitige Wahl von *B* durch *A* und gar keine Wahl. Diese vier Zustände können als Realisierung eines stochastischen Prozesses betrachtet werden. Das so genannte p_1-Modell schätzt für jeden Punkt einen Senderparameter (Aktivität) und einen Empfängerparameter (Popularität) sowie für das gesamte Netzwerk einen Dichte- und einen Reziprozitätsparameter (Holland und Leinhardt 1981; Wasserman und Weaver 1985). Eine Art logistische Regression ermöglicht, die Varianz der Parameter zwischen Individuen und Teilgruppen zu erklären (das p_2-Modell, van Duijn et al. 2004). Die so genannten p*-Modelle beziehen auch die Schätzung von Parametern höheren Ranges mit ein, wie etwa triadische Abhängigkeiten (z.B. Transitivität, siehe Wasserman und Pattison 1996). Dem liegt der Gedanke zu Grunde, dass Transitivität von Freundschaften erstens durch Gelegenheitsstrukturen entsteht. Zweitens kommt der Wunsch nach einem Freundeskreis hinzu, statt einer Reihe untereinander nicht verbundener Freunde.

Holland und Leinhardt (1977) führten Markov-Prozesse als allgemeinen Rahmen für stochastische Modelle der *Netzwerkentwicklung* ein. Im Wesentlichen drücken Markov-Modelle aus, wie die Struktur eines sozialen Netzwerks über die Zeit betrachtet von einem Zustand in einen anderen übergeht. Die Analyseeinheit ist in der Regel die Dyade mit ihren vier möglichen Zuständen (siehe oben zu p_1-Modellen). Die Parameter, die den Prozess lenken, betreffen die Wahrscheinlichkeit eines Übergangs von einem dieser vier Zustände in einen anderen. Die ursprünglichen Markov-Modelle unterstellen für den gesamten Prozess zeithomogene Parameter sowie Homogenität der Population (Wasserman 1980). Allgemeinere Abhängigkeiten zwischen Dyaden können mit einem Vorschlag von Snijders (1996, 2001, 2005) berücksichtigt werden. Er verknüpft Theorie und statistische Prüfung in viel stärkerem Maße, indem er Markov-Modelle mit Zufallsnutzenmodellen *(random utility models)* verbindet. Netzwerkmerkmale oder Wechselwirkungen zwischen diesen und individuellen Merkmalen erzeugen dabei einen bestimmten Nutzen. Von solchen stochastischen Nutzenmodellen kann abgeleitet werden, welche Merkmale und Wechselwirkungen einen hohen Nutzen erzeugen und somit für die Veränderungen eines Netzwerks maßgeblich sein dürften. Für jede dieser Nutzenkomponenten werden Parameter geschätzt, die ihren jeweiligen Einfluss anzeigen. Van de Bunt (1999; van de Bunt et al. 1999) und van Duijn et al. (2003) wenden diesen Ansatz in Studien über Freundschaftsnetzwerke innerhalb von Organisationen und unter Soziologie-Studierenden im ersten Semester an. Van Duijn et al. zeigen, dass Ähnlichkeit bei zwei beobachtbaren Dimensionen (Geschlecht und Studiengang) für die Gesamtstruktur des Freundschaftsnetzwerks der Erstsemester von

Anfang an in hohem Maße entscheidend ist. Außerdem zeigen sie, dass insbesondere die Gelegenheiten in einem Netzwerk die weitere Entwicklung von Freundschaften innerhalb der Grenzen dieser zwei beobachtbaren Dimensionen erklären.[15]

VI. Schlussfolgerungen

Im Laufe der letzten Jahrzehnte hat die Analyse sozialer Netzwerke zu zahlreichen neuen Erkenntnissen geführt. Sie hat völlig neue Forschungslinien eröffnet, die für unsere Einsichten in soziale Phänomene von grundlegender Bedeutung sind. Die Entwicklungen haben dazu beigetragen, dass die Beziehungen zwischen Netzwerken und Verhalten nun empirisch untersucht werden können. Dadurch hat sich der Begriff des sozialen Netzwerks von einer bloßen Metapher zu einem Konzept mit klarem empirischen Bezug entwickelt. Es sind gerade diese Entwicklungen, die es Forschern ermöglichen, komplexere Beziehungen zwischen Strukturen sowie zwischen Strukturen und Verhalten zu untersuchen. Gleichzeitig haben diese Entwicklungen jedoch viele, die soziale Netzwerke erforschen, dazu verführt, ihre Aufmerksamkeit auf die Struktur zu konzentrieren und die Wechselwirkung zwischen Struktur und Inhalt von Beziehungen zu vernachlässigen. Dies ist der Grund, warum in diesem Artikel den grundsätzlichen Fragen so viel Aufmerksamkeit gewidmet wurde. Die heutige Methodologie der Netzwerkanalyse ist in der Lage, komplexe Zusammenhänge zu behandeln. Gerade deshalb müssen wir mehr Aufmerksamkeit auf die grundlegenden Fragen richten: Auf welche Beziehungen kommt es in welchen Kontexten an? Welche strukturellen Merkmale schaffen in welchen Kontexten welche Gelegenheiten? Solche Fragen können nur durch eine gründliche Analyse der Mikrofundierungen sozialer Netzwerke beantwortet werden. Natürlich gibt es bereits viele Studien, die Mikrofundierungen ernst nehmen. Es sollte jedoch deutlich geworden sein, dass noch immer viele Erkenntnisse gewonnen werden können, wenn die Erforschung der Mikrofundierungen wesentlich und in verschiedene Richtungen erweitert wird:

– Andere Ziele außer materielle Gewinne sollten mit einbezogen werden (insbesondere „richtiges Verhalten“ und „Sich-Wohlfühlen“).
– Die Beziehungen, die in den unterschiedlichen Kontexten zu betrachten sind, sollten auf der Grundlage der wichtigsten Prozesse explizit abgeleitet werden.
– Studien sollten sich auf die Interaktionen zwischen unterschiedlichen Arten von Beziehungen und Netzwerkstrukturen konzentrieren.
– Das Gleichgewicht zwischen Kooperation und Wettbewerb sollte berücksichtigt werden (auf der Ebene der Beziehung selbst, zwischen Beziehungen sowie zwischen Gruppen oder Netzwerken).

Hauptziel dieses Artikels war es, Forschern eine Heuristik für eine erweiterte Analyse zu verschaffen. Diese Heuristik legt den Schwerpunkt auf die Analyse der Verwirklichung von Zielen, die nicht durch eine Person allein realisiert werden können: die gemeinschaftliche Produktion und Nutzung von Gütern. Außerdem basiert diese Heuris-

15 Die in diesem Abschnitt vorgestellten statistischen Modelle sind in dem statistischen Softwarepaket StOCNET erhältlich (Huisman und van Duijn 2003, 2005; Boer et al. 2003).

tik darauf, dass die meisten Beziehungen multifunktional sind und der gleichzeitigen Erfüllung unterschiedlicher Ziele dienen. Die Ausgangspunkte finden sich in Lindenbergs Gruppentheorie (Lindenberg 1997). Zum einen legt Lindenberg den Analyseschwerpunkt auf funktionale Interdependenzen, d.h. auf wechselseitige Abhängigkeiten zwischen Individuen, die sich hinsichtlich ihrer Aufgaben und hinsichtlich der Ergebnisse bei gemeinschaftlicher Produktion ergeben. Zum anderen stellt er systematisch die Wechselwirkungen zwischen funktionalen, kognitiven und strukturellen Interdependenzen dar. Drei Beispiele haben gezeigt, wie die Heuristik angewandt werden kann, und dass sie zu grundlegenden neuen Erkenntnissen führt.

Literatur

Achen, Christopher H., 2006: Institutional Realism and Bargaining Models. Kap. 4 in: *Robert Thomson, Frans N. Stokman, Christopher Achen* und *Thomas König* (Hg.), The European Union Decides. Cambridge: Cambridge University Press (im Erscheinen).

Arregui, Javier, Frans N. Stokman und *Robert Thomson*, 2006: Exchange and Compromise in the European Union. Kap. 5 in: *Robert Thomson, Frans N. Stokman, Chris Achen* und *Thomas König* (Hg.), The European Union Decides. Cambridge: Cambridge University Press (im Erscheinen).

Boer, Peter, Mark Huisman, Tom A. B. Snijders und *Evelien P.H. Zeggelink,* 2003: StOCNET: An Open Software System for the Advanced Statistical Analysis of Social Networks. Version 1.4. Groningen: ProGAMMA/ICS.

Borgatti, Stephen P., und *Pacey C. Foster,* 2003: The Network Paradigm in Organizational Research: A Review and Typology. Journal of Management 29: 991–1013.

Braun, Norman, 1993: Socially Embedded Exchange. Frankfurt a.M.: Peter Lang.

Brehm, Sharon S., 1985: Intimate Relationships. New York: McGraw-Hill.

Bueno de Mesquita, Bruce, 1994: Political Forecasting: An Expected Utility Method. S. 71–104 in: *Bruce Bueno de Mesquita* und *Frans N. Stokman* (Hg.), European Community Decision Making. Models, Comparisons, and Applications. Yale University Press.

Bueno de Mesquita, Bruce, David Newman und *Rabushka Alvin,* 1985: Forecasting Political Events. The Future of Hong Kong. New Haven: Yale University Press.

Burt, Ronald S., 1992: Structural Holes: The Social Structure of Competition. Cambridge: Harvard University Press.

Burt, Ronald S., 2000: The Network Structure of Social Capital. Research in Organizational Behaviour 22: 345–423.

Buskens, Vincent, und *Werner Raub,* 2004: Soziale Mechanismen rationalen Vertrauens. Eine theoretische Skizze und Resultate aus empirischen Studien. S. 183–216 in: *Andreas Diekmann* und *Thomas Voss* (Hg.), Rational Choice Theorie in den Sozialwissenschaften. Anwendungen und Probleme. Festschrift für Rolf Ziegler, München: Oldenbourg.

Coleman, James S., 1972: Systems of Social Exchange. Journal of Mathematical Sociology 2: 145–163.

Coleman, James S., 1988: Social Capital in the Creation of Human Capital. American Journal of Sociology 94: 95–120.

Coleman, James S., 1990: Foundations of Social Theory. Cambridge: The Belknap Press.

Cook, Karin S., 1995: Social Relationships and Group Processes. S. 203–235 in: *Karin S. Cook, Gary A. Fine* und *James S. House* (Hg.), Sociological Perspectives on Social Psychology. Boston, MA: Allyn and Bacon.

Davison, Mark L., und *Lawrence E. Jones,* 1976: A Similarity-Attraction Model for Predicting Sociometric Choice from Perceived Group Structure. Journal of Personality and Social Psychology 33: 601–612.

Dekker, David J., 2001: Network Perspectives on Tasks in Account Management. Rotterdam: Tinbergen Institute Research Series Nr. 259.

Dekker, David, Frans N. Stokman und *Philip H. Franses*, im Erscheinen: Effectiveness of Brokering within Management Organizations.

Duck, Steve W., 1991: Friends, for Life: The Psychology of Personal Relationships. New York: Harvester.

Ellickson, Robert C., 1991: Order without Law. Cambridge, MA: Harvard University Press.

Esser, Hartmut, 1997: Die „Definition" der Situation und die Rationalität des Handelns. S. 69–90 in: *Tamás Melegny, Heinz-Jürgen Niederzu, Max Preglau, Franz Traxler* und *Bettina Scheikal* (Hg.), Soziologie im Konzert der Wissenschaften. Opladen: Westdeutscher Verlag.

Esser, Hartmut, 2000: Normen als Frames: Das Problem der „Unbedingtheit" des normativen Handelns. S. 137–155 in: *Regina Metze, Kurt Mühler* und *Karl-Dieter Opp* (Hg.), Normen und Institutionen: Entstehung und Wirkungen. Leipzig: Leipziger Universitätsverlag.

Festinger, Leon, 1954: A Theory of Social Comparison Processes. Human Relations 7: 117–140.

Flache, Andreas, und *Michael W. Macy,* 1996: The Weakness of Strong Ties: Collective Action Failure in Highly Cohesive Groups. Journal of Mathematical Sociology 21: 3–28.

Flap, Henk D., und *Nan Dirk de Graaf,* 1986: Social Capital and Attained Occupational Status. The Netherlands Journal of Sociology 22: 145–161.

Flap, Henk D., 1999: Creation and Returns of Social Capital. A New Research Program. La Revue Tocqueville 20: 1–22.

Freeman, Linton C., A. Kimball Romney und *Sue C. Freeman,* 1987: Cognitive Structure and Informant Accuracy. American Anthropologist 89: 310–325.

Friedkin, Noah E., und *Eugene C. Johnsen,* 1990: Social Influence and Opinions. Journal of Mathematical Sociology 15: 193–205.

Friedkin, Noah E., und *Eugene C. Johnsen,* 1999: Social Influence Networks and Opinion Change. Advances in Group Processes 16: 1–29.

Granovetter, Mark S., 1973: The Strength of Weak Ties. American Journal of Sociology 78: 1360–1380.

Hallinan, Maureen T, und *Ruy A. Teixeira,* 1987: Opportunities and Constraints: Black-White Differences in the Formation of Interracial Friendships. Child Development 58: 1358–1371.

Holland, Paul W., und *Samuel Leinhardt,* 1977: A Dynamic Model for Social Networks. Journal of Mathematical Sociology 5: 5–20.

Holland, Paul W., und *Samuel Leinhardt,* 1981: An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association 77: 33–50.

Huisman, Mark, und *Marijtje A.J. van Duijn,* 2003: StOCNET: Software for the Statistical Analysis of Social Networks. Connections 25 (1): 7–26.

Huisman, Mark, und *Marijtje A.J. van Duijn,* 2005: Software for Social Network Analysis. S. 270–316 in: *Peter Carrington, John Scott* und *Stanley Wasserman* (Hg.), Models and Methods in Social Network Analysis. New York: Cambridge University Press.

Jones, Lawrence E., 1982: Construal of Social Environments: Multidimensional Models of Interpersonal Perception and Attraction. S. 61–85 in: *Nancy Hirschberg* und *Lloyd G. Humphreys* (Hg.), Multivariate Applications in the Social Sciences. Hillsdale, NJ: Erlbaum.

Killworth, Peter D., und *H. Russell Bernard,* 1976: Informant Accuracy in Social Network Data. Human Organization 35: 269–286.

Killworth, Peter D., und *H. Russell Bernard,* 1979: Informant Accuracy in Social Network Data III: A Comparison of Triadic Structure in Behavioral and Cognitive Data. Social Networks 2: 19–46.

Kollock, Peter, 1994: The Emergence of Exchange Structures: An Experimental Study of Uncertainty, Commitment, and Trust. American Journal of Sociology 100: 313–45.

König, Thomas, 1997: Europa auf dem Weg zum Mehrheitssystem. Gründe und Konsequenzen nationaler und parlamentarischer Integration. Opladen: Westdeutscher Verlag.

Krackhardt, David, 1999: The Ties that Torture: Simmelian Tie Analysis in Organizations. Research in the Sociology of Organizations 16: 183–210.

Kumbasar, Ece A., Romney A. Kimball und *William H. Batchelder,* 1994: Systematic Biases in Social Perception. American Journal of Sociology 100: 477–505.

Laumann, Edward O., David Knoke und *Yong-Hak Kim,* 1987: Event Outcomes. S. 343–373 in: *Edward O. Laumann* und *David Knoke* (Hg.), The Organizational State. Social Choice in National Policy Domains. Madison: The University of Wisconsin Press.

Lawler, Edward J., 2001: An Affect Theory of Social Exchange. American Journal of Sociology 107: 321–352.

Leenders, Roger T. A. J., 1995: Structure and Influence. Statistical Models for the Dynamics of Actor Attributes. Network Structure and Their Interdependence. Amsterdam: Thela Thesis Publishers.

Leenders, Roger T. A. J., 2002: Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix. Social Networks 24: 21–48.

Lewin, Kurt, 1936: Principles of Topological Psychology. New York: McGraw Hill.

Lin, Nan, 1982: Social Resources and Instrumental Action. S. 131–145 in: *Peter V. Marsden* und *Nan Lin* (Hg.), Social Structure and Network Analysis. London: Sage.

Lindenberg, Siegwart, 1997: Grounding Groups in Theory. Advances in Group Processes 14: 281–331.

Lindenberg, Siegwart, 1998: Solidarity: Its Microfoundations and Macro-Dependence. A Framing Approach. S. 61–112 in: *Patrick Doreian* und *Thomas J. Fararo* (Hg.), The Problem of Solidarity: Theories and Models. Amsterdam: Gordon and Breach.

Lindenberg, Siegwart, 2000: It Takes Both Trust and Lack of Mistrust: The Working of Solidarity and Relational Signaling in Contractual Relationships. Journal of Management and Governance 4: 11–33.

Lindenberg, Siegwart, 2001: Social Rationality versus Rational Egoism. S. 635–668 in: *Jonathan H. Turner* (Hg.), Handbook of Social Theory. New York: Kluwer Academic/Plenum.

Lindenberg, Siegwart, und *Bruno S. Frey,* 1993: Alternatives, Frames, and Relative Prices. A Broader View of Rational Choice. Acta Sociologica 36: 191–205.

Marsden, Peter V., und *Edward O. Laumann,* 1977: Collective Action in a Community Elite: Exchange, Influence, Resources and Issue Resolution. S. 199–250 in: *Roland J. Liebert* und *Allen W. Imershein* (Hg.), Power, Paradigms, and Community Research. London: Sage.

Mattila, Mikko, und *Jan-Erik Lane*, 2001: Why Unanimity in the Council? A Roll Call Analysis of Council Voting. European Union Politics 2 (1): 31–52.

Molm, Linda D., 1997: Coercive Power in Social Exchange. Cambridge: Cambridge University Press.

Newcomb, Theodore M., 1956: The Prediction of Interpersonal Attraction. The American Psychologist 11: 575–581.

Organ, Dennis W., 1988: Organizational Citizenship Behavior: The Good Soldier Syndrome. Lexington, MA: Lexington Books.

Pappi, Franz U., und *Christian H.C.A. Henning,* 1998: Policy Networks: More than a Metaphor? Journal of Theoretical Politics 10: 553–575.

Podolny, Joel M., und *James N. Baron,* 1997: Resources and Relationships: Social Networks and Mobility in the Workplace. American Sociological Review 62: 673–693.

Raub, Werner, und *Jeroen Weesie,* 1990: Reputation and Efficiency in Social Interactions: An Example of Network Effects. American Journal of Sociology 96: 626–654.

Sherif, Muzafer, 1966 [1953]: Groups in Harmony and Tension: An Integration of Studies on Intergroup Relations. New York: Organon Books.

Simmel, Georg, 1950 [1923]: The Sociology of Georg Simmel. New York: Free Press.

Snijders, Tom A. B., 1996: Stochastic Actor-Oriented Models for Network Change. Journal of Mathematical Sociology 21: 149–172.

Snijders, Tom A. B., 2001: The Statistical Evaluation of Social Network Dynamics. S. 361–395 in: *Michael E. Sobel* und *Mark P. Becker* (Hg.), Sociological Methodology-2001. Boston and London: Basil Blackwell.

Snijders, Tom A. B., 2005: Models for Longitudinal Network Data. Kap. 11 in: *Peter Carrington, John Scott* und *Stanley Wasserman* (Hg.), Models and Methods in Social Network Analysis. New York: Cambridge University Press.

Stokman, Frans N., 2004: Frame Dependent Modeling of Influence Processes. S. 113–127 in: *Andreas Diekmann* und *Thomas Voss* (Hg.), Rational-Choice-Theorie in den Sozialwissenschaften. Anwendungen und Probleme. Festschrift für Rolf Ziegler, München: Oldenbourg.

Stokman, Frans N., und *Robert Thomson,* 2004: Winners and Losers in the European Union. European Union Politics 5: 5–23.

Stokman, Frans N., und *Reinier van Oosten,* 1994: The Exchange of Voting Postions: An Object-Oriented Model of Policy Networks. S. 105–127 in: *Bruce Bueno de Mesquita* und *Frans N. Stokman* (Hg.), European Community Decision Making. Models, Comparisons, and Applications. New Haven, CT: Yale University Press.

Stokman, Frans N., und *Evelien P. H. Zeggelink,* 1996: ‚Self-Organizing' Friendship Networks. S. 385–418 in: *Wim B. G. Liebrand* und *David M. Messick* (Hg.), Frontiers in Social Dilemmas Research. Berlin: Springer.

Stokman, Frans N., Marcel A. L. M. van Assen, Jelle van der Knoop und *Reinier C. H. van Oosten,* 2000: Strategic Decision Making. Advances in Group Processes 17: 131–153.

Tesser, Abraham, 1988: Toward a Self-Evaluation Maintenance Model of Social Behavior. S. 181–227 in: *Leonard Berkowitz* (Hg.), Advances in Experimental Social Psychology, Bd. 21. New York: Academic Press.

Tesser, Abraham, 2000: On the Confluence of Self-Esteem Maintenance Mechanisms. Personality and Social Psychology Review 4: 290–299.

Thomson, Robert, Frans N. Stokman, Christopher Achen und *Thomas König* (Hg.), 2006: The European Union Decides. Cambridge: Cambridge University Press (im Erscheinen).

Udehn, Lars, 1996: The Limits of Public Choice: A Sociological Critique of the Economic Theory of Politics. London: Routledge.

Urberg, Kathryn A., Serdar M. Degirmencioglu und *Jerry M. Tolson*, 1998: Adolescent Friendship Selection and Termination: The Role of Similarity. Journal of Social and Personal Relationships 15: 703–710.

van Assen, Marcel A. L. M., Frans N. Stokman und *Reinier van Oosten,* 2003: Conflict Measures in Cooperative Exchange Models of Collective Decision Making. Rationality and Society 15: 85–112.

van de Bunt, Gerhard G., 1999: Friends by Choice. An Actor-Oriented Statistical Network Model for Friendship Networks through Time. Amsterdam: Thela Thesis.

van de Bunt, Gerhard G., Marijtje A. J. van Duijn und *Tom A. B. Snijders,* 1999: Friendship Networks through Time: An Actor-Oriented Dynamic Statistical Network Model. Computational and Mathematical Organization Theory 5: 167–192.

van Duijn, Marijtje A. J., Tom A. B. Snijders und *Bonne H. Zijlstra,* 2004: P_2: A Random Effects Model with Covariates for Directed Graphs. Statistica Neerlandica 58: 234–254.

van Duijn, Marijtje A. J., Mark Huisman, Frans N. Stokman et al., 2003: Evolution of Sociology Freshmen into a Friendship Network. Journal of Mathematical Sociology 27: 153–191.

Wasserman, Stanley S., 1980: Analyzing Social Networks as Stochastic Processes. Journal of the American Statistical Association 75: 280–294.

Wasserman, Stanley, und *Philippa Pattison,* 1996: Logit Models and Logistic Regressions for Social Networks. I. An Introduction to Markov Graphs and p*. Psychometrika 61: 401–425.

Wasserman, Stanley, und *Sheila O'Leary Weaver,* 1985: Statistical Analysis of Binary Relational Data: Parameter Estimation. Journal of Mathematical Psychology 29 (4): 406–427.

Werner, Carol M., und *Pat Parmelee,* 1979: Similarity of Activity Preferences among Friends: Those Who Play Together Stay Together. Social Psychology Quarterly 42: 62–66.

Willer, David, 1999: Network Exchange Theory. Westport: Praeger.

SOZIALER KONTEXT IN DER MEHREBENENANALYSE

Uwe Engel und Julia Simonson

Zusammenfassung: Der Beitrag behandelt Fragestellungen und Grundlagen der Kontextanalyse und zeigt, wie das Instrumentarium der statistischen Mehrebenenanalyse für Kontextanalysen eingesetzt werden kann. Zunächst werden Ziele und Einsatzgebiete der Mehrebenenanalyse beschrieben, es folgt die Darstellung einiger Grundlagen der hierarchischen linearen Regression, darauf aufbauend wird die strukturelle Analyse und Identifizierung von Kompositionseffekten behandelt. Modelle für Querschnitts- und Paneldaten werden vorgestellt. Abschließend wird ein Ansatz für die Analyse von kategorialen abhängigen Variablen dargestellt. Zur Illustration der Modelle werden Statusinkonsistenz- und relationale Inkongruenzeffekte auf verschiedene Indikatoren sozialer Integration mit Daten aus vier Wellen des Sozioökonomischen Panels aus den Jahren 1998 bis 2001 geschätzt.

I. Einführung

1. Soziale Kontexte

„Die Kontextanalyse ist jene Variante der empirischen Sozialforschung, in der *systematisch* auch die soziale und personale Umgebung in die Variablen aufgenommen wird, mit denen ein bestimmtes Verhalten erklärt werden soll" (Esser 1999: 428). Mit ihr verbindet sich die Vorstellung, dass das Handeln von Personen nicht nur von individuellen Neigungen, Zielsetzungen und psychologischen Dispositionen bestimmt wird, sondern dass die soziale Umgebung davon unabhängige Wirkungen erzeugen kann. Solche „strukturellen Effekte" auf die Orientierungen und Handlungen von Akteuren können sich so auswirken, dass sich Personen trotz gleicher *individueller* Dispositionen unterschiedlich verhalten, wenn sie unterschiedlichen sozialen Umgebungen ausgesetzt sind.[1]

Eine zentrale Aufgabe der Kontextanalyse ist entsprechend darin zu sehen, die individuellen Effekte von denen der sozialen Umgebung zu trennen. Dies kann – wie in der klassischen Analyse von Blau (1960) – bedeuten, die Wirkung individueller Wertorientierungen von der des vorherrschenden Werteklimas in der Gruppe zu trennen, um im nächsten Schritt diese Einflussgrößen in ihrem statistischen Zusammenhang zu analysieren. Soziale Umgebungen können sich aber nicht nur über die in ihnen vor-

1 Vgl. Esser (1993: 423–425). Der Begriff des strukturellen Effektes stammt von Peter Blau (1960). Dieser wurde früh von Davis et al. (1961) aufgegriffen, die in ihrem stärker statistisch und weniger soziologisch ausgerichteten Beitrag vorschlagen, ihn durch den Begriff des „Kompositionseffektes" zu ersetzen. Siehe zu diesem Begriff unten die Ausführungen in Abschnitt III.4.

herrschenden *Werte* auf die Personen auswirken, die in ihnen leben, sondern auch über ihre immanente *Sozialstruktur*. Sich inmitten eines sozialen Umfeldes zu befinden, bedeutet nicht nur, dadurch in einen Wertekontext eingebunden zu sein. Es bedeutet auch, sich in einem Strukturkontext zu befinden, der den Akteuren Handlungschancen eröffnet, ihnen diese aber auch einschränken kann. Ein Grundgedanke der Kontextanalyse besagt, dass sich diese Struktureigenschaften der sozialen Umgebung aus den Häufigkeitsverteilungen der Eigenschaften *der Personen* ableiten, die in ihr leben (und aus Sicht einer jeden dieser Personen Teil ihrer sozialen Umgebung sind). Eine Region kann nicht wohlhabend sein, wenn in ihr nur arme Leute wohnen. Die Arbeitslosenrate einer Region kann nicht hoch sein, wenn es in ihr kaum Arbeitslose gibt. Das Einkommensgefälle einer Region kann nicht ausgeprägt sein, wenn alle Haushalte in ihr in etwa über dasselbe Äquivalenzeinkommen verfügen etc. Und dennoch können solche Umgebungsmerkmale eigenständige Effekte entfalten. So lässt es die Menschen ggf. nicht unberührt, in einer Region mit hoher Arbeitslosigkeit zu leben, auch wenn sie selbst nicht arbeitslos sind. Es mag auch zu Wechselwirkungen kommen: Armut inmitten von Überfluss mag für die Menschen eine andere Bedeutung besitzen als Armut inmitten von Armut oder Wohlstand inmitten von Wohlstand. Weitere Beispiele ließen sich anfügen. Entscheidend ist nur der Gedanke, dass sich über die Verteilung von Individualmerkmalen *Gruppen- bzw. Bevölkerungszusammensetzungen* ergeben, die als Aggregatmerkmale auf die Personen zurückwirken, die Teil des jeweiligen Kontextes sind. Eine Aufgabe der Kontextanalyse besteht entsprechend darin, solche *Kompositionseffekte* zu identifizieren.

Geht man der Frage nach handlungsrelevanten sozialen Umgebungen nach, so ist an die *Nahumwelten* zu denken, in denen sich die Menschen bewegen: der gemeinsame Haushalt, die Familie, die Schulklasse, der Freundes- oder Kollegenkreis etc. Diese konstituieren unmittelbare Handlungskontexte. Ebenso bedeutsam sind im Prinzip auch *mittelbare* Handlungskontexte: Nachbarschaften, Gemeinden, Orte und Städte, Regionen etc. Ein sozialer Kontext wäre entsprechend über das Kriterium des unmittelbar oder mittelbar umgebenden Raumes definiert, weniger darüber, dass zwischen den Personen eines Kontextes auch soziale Beziehungen bestehen. Mit Kontextanalysen kann sich geradezu die Frage verbinden, wie Strukturkontexte die Gelegenheiten zum sozialen Kontakt etwa *über Bevölkerungskategorien hinweg* beeinflussen und die Wahrscheinlichkeit sozialer Beziehungen zur Funktion der sozialen Zusammensetzung des Kontextes machen. Die Menge der Personen, die einen sozialen Kontext konstituiert, wird daher im Allgemeinen auch eher als ein soziales „Aggregat" denn als „Gruppe" aufzufassen sein, denn soziale Kontexte konstituieren sich nicht ausschließlich über soziale Beziehungen. Insofern unterscheiden sich Zugang und typische Fragestellung einer Kontextanalyse von denen einer Netzwerkanalyse.

2. Individuelle und kollektive Eigenschaften

Eine Möglichkeit, konzeptuell zwischen den bei der Berücksichtigung sozialer Kontexte auftauchenden unterschiedlichen Analyseebenen zu differenzieren, besteht in der Anwendung der von Lazarsfeld und Menzel (1969) entwickelten Typologie, in der zwi-

schen den Eigenschaften von Kollektiven bzw. Aggregaten und den Eigenschaften der Mitglieder dieser Aggregate differenziert wird. Während einige Merkmale direkt auf der Analyseebene, auf die sie sich beziehen, gemessen werden können, lassen sich andere Merkmale auf einer höheren oder niedrigeren Ebene messen und dann durch Aggregation oder Disaggregation auf eine andere Ebene beziehen. Ein Vorteil der in *Tabelle 1* im Anschluss an Galtung (1970: 40) zusammengefassten Lazarsfeld-Menzel-Typologie besteht darin, dass Variablen und deren Auswirkungen bestimmten Ebenen zugeordnet werden können.

Tabelle 1: Variablentypologie nach Lazarsfeld und Menzel

	Absolute (genuine) Eigenschaften	**Eigenschaften, basierend auf ...**		
		Verteilung	**Struktur**	**Inklusion**
Ebene n (Mitglieder eines Aggregats)	*absolut*	*komparativ*	*relational*	*kontextuell*
Ebene n +1 (Aggregat)	*global*	*analytisch*	*strukturell*	

Quelle: Galtung (1970: 40), leicht veränderte Darstellung.

Um individuelle Einheiten, die Mitglieder eines Aggregats sind, zu beschreiben, können *„absolute", „komparative", „relationale"* und *„kontextuelle"* Eigenschaften herangezogen werden. Absolute Eigenschaften können dabei ohne Verwendung von Informationen über Charakteristika des Aggregats oder die Beziehungen der Mitglieder eines Aggregats untereinander ermittelt werden. Sie beziehen sich jeweils nur auf die Ebene, für die sie definiert wurden und sind damit genuine Merkmale der Mitglieder eines Aggregats. Betrachtet man etwa Länder als Aggregate und die Bewohner dieser Länder als Mitglieder der Aggregate, so wäre z.B. die individuelle Parteipräferenz der Bewohner, bezogen auf die Mitgliederebene, ein absolutes Merkmal. Komparative Eigenschaften eines Mitglieds bezeichnen dagegen Merkmale, die durch Vergleich mit den Eigenschaften der anderen Mitglieder des Aggregats zustande kommen. Die individuelle Stellung in einer Rangreihe wäre ein solches Merkmal: Die Messung erfolgt zwar jeweils für ein Aggregatmitglied, wird jedoch zu den Messungen der anderen Mitglieder in Bezug gesetzt. Relationale Charakteristika von Mitgliedern beziehen sich auf die zwischen den Mitgliedern eines Aggregats bestehenden Relationen, beispielsweise auf Kommunikationsnetzwerke. Kontextuelle Merkmale hingegen sind mit der jeweils übergeordneten Einheit verknüpft. Alle Mitglieder eines Aggregats erhalten denselben Wert in einer Variablen, der sie als Mitglied eben dieses Aggregats kennzeichnet. Bei Haushalten als Aggregaten und Personen als Mitgliedern dieser Aggregate könnte jede dieser Personen z.B. durch die Ausprägung des Inkongruenzgrades des Haushaltes beschrieben werden. Personen, die innerhalb einer bestimmten Region leben, könnten durch Ausprägungen dieser Region gekennzeichnet werden, etwa durch das regionale Wohlstandsniveau oder das Wohlstandsgefälle.

Den absoluten, komparativen, relationalen und kontextuellen Eigenschaften von Aggregatmitgliedern stehen auf der Ebene der Aggregate „*analytische*", „*strukturelle*" und „*globale*" Eigenschaften gegenüber. Analytische Eigenschaften basieren auf Infor-

mationen über die Mitglieder eines Aggregates, spiegeln also beispielsweise die Verteilung eines bestimmten Merkmals innerhalb eines Aggregats wider. Eine analytische Eigenschaft eines Haushaltes wäre etwa der aus individuellen Statusinkonsistenzwerten abgeleitete Inkongruenzgrad des Haushaltes. Als analytisches Merkmal einer Schulklasse wäre anzusehen, wenn zu ihrer Charakterisierung aus Individualmerkmalen z.B. der Criss-Cross-Grad[2] in zwei konfliktrelevanten Statusdimensionen berechnet wird (Engel 1988). Ein im Sinne der Typologie analytisches Merkmal wäre z.B. auch dadurch gegeben, dass das Wohlstandsgefälle einer Region über den Gini-Index aus individuellen Einkommenswerten von in der Region lebenden Personen berechnet wird. Strukturelle Eigenschaften eines Aggregats basieren dagegen auf Informationen über die Beziehungen zwischen den Mitgliedern eines Aggregats, könnten also z.B. die aggregatspezifische Dichte privater Netzwerke wiedergeben. Globale Eigenschaften von Aggregaten schließlich beruhen auf Informationen, die sich *nicht* auf die Eigenschaften der Mitglieder eines Aggregats beziehen, sondern genuine Aggregateigenschaften sind. So ist beispielsweise die Staatsform eines Landes ein globales Aggregatmerkmal, das nicht auf aggregierten Informationen einer tiefer liegenden Ebene beruht. Ebenso wären Infrastrukturmerkmale als globale Merkmale der betrachteten Aggregateinheiten anzusehen. Es ist nicht zuletzt diese Möglichkeit, solche aus ganz anderen Quellen kommenden Daten in die Auswertung von Umfrage- und Paneldaten einbeziehen und damit für die empirische Sozialforschung nutzbar machen zu können, die das Instrumentarium der Mehrebenenanalyse für die empirische Sozialforschung interessant macht.

3. Statistische Mehrebenenanalyse

Der Begriff „Mehrebenenanalyse“ steht sowohl für eine soziologische Methodologie, nach der es möglich ist, in soziologischen Analysen auf sinnvolle Weise unterschiedliche Analyseebenen zu unterscheiden (so etwa im Zusammenhang mit der vorgestellten Lazarsfeld-Menzel-Typologie oder im Zusammenhang mit Erklärungen in den Sozialwissenschaften, vgl. Coleman 1990), als auch für *statistische* Methoden zur Auswertung hierarchisch strukturierter Datensätze in der empirischen Sozialforschung. Solche statistischen Mehrebenenanalysen ermöglichen es, sowohl die von Eigenschaften auf der Ebene der Aggregate als auch die von individuellen Eigenschaften ausgehenden Effekte in statistisch adäquater Weise zu schätzen. Darüber hinaus können auch Effekte, die durch die Interaktion zwischen gruppen- und individuenbezogenen Variablen zustande kommen, also so genannte „Cross-level“-Wechselwirkungen, berücksichtigt werden (vgl. z.B. Hox und Kreft 1994).

2 Das von Galtung (1966) zur Charakterisierung von Aggregaten entwickelte Criss-Cross-Konzept gibt Aufschluss über die Konfliktbindungskapazität der Sozialstruktur von sozialen Systemen. Ein hoher Criss-Cross-Grad indiziert ein hohes Konfliktbindungspotenzial, ein geringer Grad das Gegenteil. Das Criss-Cross-Potenzial eines Aggregats ist dabei eine Funktion der relativen Häufigkeit, mit der Statusinkonsistenz im betreffenden sozialen System auftritt. D.h. der Criss-Cross-Grad hängt von der Verteilung der Mitglieder eines Aggregats auf inkonsistenten Positionen ab, da angenommen wird, dass den Trägern solch inkonsistenter Statusverknüpfungen eine Brückenfunktion zwischen ansonsten disparaten Konfliktgruppen zukommen kann (Engel 1998: 53ff.).

Eine besondere Bedeutung kommt dem statistischen Mehrebenenansatz unter anderem deshalb zu, weil er bei der Vermeidung ökologischer Fehlschlüsse im Sinne Robinsons (1950) hilft. Ein ökologischer Fehlschluss liegt dann vor, wenn auf der Basis von Aggregatdaten von einer Kollektivhypothese fälschlicherweise auf eine Individualhypothese geschlossen wird. So muss beispielsweise ein auf der Aggregatebene von Landkreisen gefundener Zusammenhang zwischen der Höhe des Arbeiteranteils und dem Stimmenanteil für rechtsextreme Parteien nicht unbedingt implizieren, dass es tatsächlich die Arbeiter sind, die zur Wahl rechtsextremer Parteien neigen. Korrelationen auf der Makroebene müssen also nicht zwangsläufig mit denen auf der Mikroebene identisch sein.[3] Umgekehrt kann die Verallgemeinerung von Korrelationen auf der Mikroebene in Richtung höherer Ebenen zu einem individualistischen oder atomistischen Fehlschluss führen. So muss beispielsweise ein auf individueller Ebene festgestellter Zusammenhang zwischen der Zugehörigkeit zur Arbeiterklasse und der Neigung zur Wahl sozialdemokratischer Parteien nicht implizieren, dass die Regierungen in Ländern mit hohem Arbeiteranteil eher von sozialdemokratischen Parteien gebildet werden. Zwar bestehen in gewissen Grenzen durchaus Möglichkeiten, im Sinne einer „Cross-Level"-Inferenz (Achen und Shively 1995) auf der Basis von Aggregatdaten Schlussfolgerungen über die Individualebene zu treffen, jedoch kann sich die Nützlichkeit solcher Verfahren wie z.B. der Duncan-Davis-Methode im konkreten Anwendungsfall dann doch als recht begrenzt erweisen.[4]

Eine frühe mehrebenenanalytische Untersuchung stellt der „Coleman-Report" über die Gleichheit der Bildungschancen in den USA dar (Coleman et al. 1966). Bei der Untersuchung von schulischen Leistungen und deren Ursachen wurden auch schulbezogene Einflüsse beachtet. Schulen können hinsichtlich der Infrastruktur unterschiedlich gut ausgestattet sein und dadurch bessere oder schlechtere Lernbedingungen schaffen; sie können sich auch in der Qualität des Lehrpersonals unterscheiden und dadurch unterschiedliche Lernbedingungen offerieren. In der Studie wurde darüber hinaus noch ein dritter Typus schulbezogener Einflussfaktoren berücksichtigt, und zwar Charakteristika der Schülerschaft. Diese entstehen über Prozesse der sozialen Rekrutierung in die Schulen bzw. Schulklassen und die dadurch geschaffenen Häufigkeitsverteilungen relevanter Bildungsfaktoren. D.h. gefragt wurde, welche Lernumwelten und -bedingungen daraus resultieren, dass die Schüler ihre jeweiligen Ressourcen in die Schulen und Schulklassen tragen und dadurch als kollektive Merkmale der jeweiligen Schülerschaften zu verstehende Ressourcenverteilungen hervorbringen. Im Ergebnis war der durch diese Charakteristika erklärbare Varianzanteil recht hoch, wohingegen die Erklärungsbeiträge der anderen Schul- und Lehrercharakteristika eher gering waren (Coleman et al. 1966: 304).

Auch heute noch stellt die Bildungssoziologie ein bedeutsames Anwendungsfeld der Mehrebenenanalyse dar. So wurden erst kürzlich im Rahmen der viel diskutierten Schülerleistungsstudie PISA 2000 zumindest einige wenige Mehrebenenanalysen durchgeführt (so z.B. von Baumert und Schümer 2001).

---

3 Weitere Fälle solcher „naiver Disaggregation" mit der Folge des ökologischen Fehlschlusses zeigt beispielsweise Falter (1991) für die historische Wahlforschung der Weimarer Republik auf.

4 Zur „Method of bounds" vgl. Duncan und Davis (1953).

Mehrebenenanalytische Verfahren werden jedoch auch in anderen Feldern der empirischen Sozialforschung eingesetzt. Eine in gängigen sozialwissenschaftlichen Datenbanken durchgeführte Literaturrecherche[5] verdeutlicht, dass ein Hauptanwendungsgebiet der Mehrebenenanalyse in der Epidemiologie und der Gesundheitswissenschaft liegt. 27 Prozent der in den untersuchten Datenbanken aufgeführten Publikationen stammten aus diesem Bereich. Aber auch in der Arbeitsmarktsoziologie (12 Prozent), der Jugend- und Bildungssoziologie (16 Prozent), der Soziologie abweichenden Verhaltens und sozialer Probleme (13 Prozent) sowie der Familiensoziologie (13 Prozent) kommen statistische Verfahren der Mehrebenenanalyse häufiger zum Einsatz. Deutlich seltener konnten dagegen Publikationen mit mehrebenenanalytischen Anwendungen aus den Bereichen der Sozialstrukturanalyse und der politischen Soziologie (jeweils 5 Prozent), der Sozialpsychologie und der Migrationssoziologie (jeweils 2 Prozent) identifiziert werden.

Erkennbar ist, dass Mehrebenenanalysen längst nicht mehr nur in speziellen Feldern, bei denen von einem strukturtheoretischen Hintergrund ausgegangen werden kann, eingesetzt werden. Auch jenseits von strukturtheoretischen Ansätzen werden die Bedeutung sozialer Kontexte und die Relevanz der Mehrebenenanalyse durchaus erkannt.[6]

4. Voraussetzungen

Mit Hilfe der Mehrebenenanalyse lassen sich hierarchisch strukturierte Daten untersuchen. Hierarchische Strukturen entstehen, wenn berücksichtigt wird, dass der Objektbereich sozialwissenschaftlicher Studien nicht nur aus atomisierten Individuen besteht, sondern dass die Menschen zu Gruppenbildungen der verschiedensten Art neigen und die Sozialwissenschaften insofern auch als Mitglieder dieser so konstituierten sozialen Kontexte interessieren.

Statistische Mehrebenenanalysen werden möglich, wenn zwischen diesen unter- und übergeordneten Analyseeinheiten ein Verhältnis der Inklusion besteht, wie es aus mehrstufigen Auswahlverfahren bekannt ist. Eine Menge von Personen bildet zum Bei-

5 Als Suchwörter wurden – in differierenden Schreibweisen – verwendet: Mehrebenenanalyse, Multi Level Analysis, Multi Level Model(s), Mehrebenenmodell(e). Gewertet wurden Einträge in den Datenbanken wiso-net Sowi (BLISS, IHSLIT, SOLIS, DZI SoLit, WAO), Sociological Abstracts, PSYNDEXplus und BELIT ab 1985 bis Ende 2003, bei denen aus dem in der Datenbank ausgewiesenen Abstract klar auf die Anwendung der Mehrebenenanalyse geschlossen werden konnte. Beiträge, die zwar einen theoretischen Mehrebenenbezug aufwiesen, in denen aber keine mehrebenenanalytischen Verfahren im statistischen Sinne zum Einsatz kamen, wurden ebenso wie rein methodologische Beiträge zur Mehrebenenanalyse ohne Anwendungsbezug nicht berücksichtigt.

6 Einen wesentlichen Anteil an der Einsatzfähigkeit dieses Instrumentes hat sicherlich auch die Entwicklung und Weiterentwicklung geeigneter Auswertungssoftware, die inzwischen die Schätzung auch komplexer Mehrebenenmodelle auf relativ anwenderfreundliche Weise ermöglicht. Anzuführen sind in erster Linie die Programmpakete „MLwiN“, derzeit in der Version 2.0, und „HLM“, derzeit in der Version 5.05. LISREL 8.54 enthält ein Modul zur Schätzung von Mehrebenenmodellen (mit bis zu 3 Ebenen). Innerhalb des Strukturgleichungsansatzes bietet insbesondere auch Mplus, zurzeit in der Version 3.01, gute Modellierungsoptionen.

spiel eine Nachbarschaft. Eine Menge solcher Nachbarschaften einen Ortsteil, eine Menge solcher Ortsteile einen Ort, eine Menge von Orten eine Region, eine Menge solcher Regionen ein Land usw. Das Prinzip ist immer dasselbe: Eine Menge von Einheiten der Ebene *n* konstituiert eine Analyseeinheit auf der Ebene *n*+1 und eine Menge solcher Einheiten ihrerseits eine übergeordnete Einheit auf der Ebene *n*+2 etc. Wie die Aggregation der Daten vorgenommen wird, ob Personen also als Mitglieder von Organisationen oder als Mitglieder von Nachbarschaften, Schulklassen, Haushalten, Regionen oder anderen Kontexten studiert werden, ist eine Frage des Erkenntnisinteresses und des theoretischen Hintergrundes der Analyse. Damit verknüpft aber ist unweigerlich auch die Frage nach der Wahl des *adäquaten Aggregierungsgrades*. „Adäquat" kann bedeuten, dass dadurch die Stufen einer zugrunde liegenden *mehrstufigen Auswahl* bei der Stichprobenziehung widergespiegelt werden, es kann aber ebenso bedeuten, dadurch eine *im Sinne einer substanzwissenschaftlichen Theorie* relevante Kontexteinheit abzubilden, um deren Einfluss auf Einstellungen und Verhalten der Menschen abzuschätzen, die in diesem Kontext leben. Gerade für Kontextanalysen drängt sich daher die Frage auf, ob es die in einem Datensatz verfügbaren Informationen eigentlich erlauben, Kontexte so abzubilden, dass von ihnen auch erwartet werden kann, dass sie sich auf Einstellungen und Verhalten der in ihnen lebenden Menschen auswirken. Dies wird mit Blick auf individuelle Eigenschaften insbesondere dann nicht zu erwarten sein, wenn die abbildbaren Kontexte *zu hoch aggregiert* sind: Wenn aus theoretischen Gründen die „Nachbarschaft" oder die „Gemeinde" als adäquate Kontexteinheit anzusehen ist, jedoch nur sehr viel höher aggregierte „Regionen" als übergeordnete Analyseeinheiten gebildet werden können, dann ist es keinesfalls überraschend, wenn diese Regionen bei der Varianzzerlegung kaum Varianzanteile binden, die im weiteren Verlauf der Analyse durch Charakteristika *dieser Regionen* erklärbar wären. Warum sollten, nur um ein illustratives Beispiel anzuführen, die 97 Regionen, die für das Gebiet der Bundesrepublik mit den Daten des Sozioökonomischen Panels gebildet werden können, im Allgemeinen nennenswerte Varianzanteile eines Individualmerkmals binden? Umgekehrt kann der Fall eintreten, dass die in einem Datensatz verfügbaren oder bildbaren Aggregateinheiten *nicht hoch genug* aggregiert sind, um sinnvolle Kontextanalysen durchzuführen. Dies betrifft beispielsweise Haushaltsstudien, in denen für Kontextanalysen zwar der Haushalt verfügbar ist, nicht aber ggf. benötigte, höher aggregierte *Meso*-Ebenen. Schließlich ist zu beachten, dass gerade Kontextanalysen im Allgemeinen nur in großen Samples durchgeführt werden können. An die Stelle eigener Primärforschung tritt schon aus Kostengründen entsprechend verstärkt die Sekundäranalyse vorhandener Datensätze. Das aber hat zur Folge, dass die *Nützlichkeit der Kontextanalyse als Instrument empirischer Sozialforschung* mit der Verfügbarkeit geeigneter Datensätze sowie der Verfügbarkeit von Informationen steht und fällt, die es erlauben, innerhalb dieser Datensätze adäquate Analyseebenen bilden zu können.

II. Statistische Mehrebenenanalyse

1. Hierarchische lineare Regression

Mehrebenenanalysen können unter Rückgriff auf unterschiedliche statistische Ansätze durchgeführt werden. Am prominentesten dürfte der Ansatz der hierarchischen linearen Regression sein. Deren Grundgedanke lässt sich ausgehend von der linearen Einfachregression,

$$y_i = b_0 + b_1 x_{1i} + e_i \qquad i = 1,...,n \qquad (1)$$

bei der b_0 und b_1 als jeweils für alle Personen des Samples gleich angenommen werden, aufzeigen. Eine 2-Ebenen-Variante dieses Modells würde genau an dieser Annahme ansetzen und berücksichtigen, dass beide Koeffizienten im Prinzip variable Größen darstellen können. Denkbar ist zum Beispiel, dass ihre Ausprägung von Kontext zu Kontext *j* variiert. Um diese Möglichkeit zum Ausdruck zu bringen,[7] ersetzen wir b_0 durch b_{0j} und b_1 durch b_{1j} und schreiben für (1)

$$y_{ij} = b_{0j} + b_{1j} x_{1ij} + e_{ij} \qquad i = 1,...n_j;\ j = 1,...J \qquad (2)$$

Sofern die Variabilität von b_{0j} und b_{1j} nun allein Ausdruck *zufälliger* Unterschiede von Kontext zu Kontext ist, so kann diese Annahme für die Konstante und den Steigungskoeffizienten über die beiden Ebene-2-Gleichungen

$$b_{0j} = b_0 + u_{0j} \qquad j = 1,...J \qquad (3)$$

$$b_{1j} = b_1 + u_{1j} \qquad j = 1,...J \qquad (4)$$

spezifiziert werden. Durch Einsetzen von (3) und (4) in (2) resultiert nach Auflösen der Klammern und Umstellung der Terme mit

$$y_{ij} = b_0 + b_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij} \qquad (5)$$

eine Schätzgleichung, die sich von einer gewöhnlichen Regressionsgleichung darin unterscheidet, dass sie nicht nur *e* als Zufalls- bzw. Residualterm für die Individualebene umfasst, sondern darüber hinaus die u's als Zufallseffekte für die Aggregat- bzw. Kontextebene. Gleichung (5) beschreibt y_{ij} entsprechend als Summe eines aus fixen Effekten, den b's, bestehenden Teils und eines aus zufälligen Effekten, den u's und e, bestehenden Teils. b_1 ist dabei wie ein üblicher unstandardisierter Regressionskoeffizient im Sinne eines erwarteten Unterschieds bzw. einer erwarteten Veränderung interpretierbar.

Prinzipiell könnten Kontexteffekte natürlich auch in eine Ein-Ebenen-OLS-Regression einbezogen werden, allerdings führt dies z. B. das Problem mit sich, dass keine angemessenen Signifikanztests mehr durchgeführt werden können, da Standardfehler nicht korrekt geschätzt werden. Darüber hinaus wird die relativ weitreichende und

7 Im Allgemeinen indiziert das doppelte Subskript *ij* Variabilität über Ebene-1-Einheiten innerhalb einer Ebene-2-Einheit (z.B. über die *i* Personen eines Kontextes *j*). Das einfache Subskript *j* indiziert Variabilität über Ebene-2-Einheiten (Kontexte), wobei alle Ebene-1-Einheiten innerhalb *j* den gleichen Wert haben. Ohne *i*- oder *j*-Subskript handelt es sich um eine Konstante für alle Ebene-1- und Ebene-2-Einheiten.

nicht ohne Prüfung zu übernehmende Annahme getroffen, dass sowohl Effekte als auch Fehlerterme über alle Aggregateinheiten hinweg gleich sind.

2. Annahmen

e_{ij} wird als Zufallsvariable aufgefasst, deren mittlerer Wert gleich Null ist. Zugleich wird üblicherweise angenommen, dass e_{ij} normalverteilt ist mit über die Kontexte konstanter bzw. homogener Varianz, und dass die Ebene-1-Residuen unabhängig (unkorreliert) sind, zusammenfassend:[8]

$$e_{ij} \sim N\left(0, \sigma_e^2\right) \qquad \rho\left(e_{ij}, e_{ij-1}\right) = 0 \tag{6}$$

Darüber hinaus wird angenommen, dass die Ebene-1-Residuen e_{ij} nicht mit den Ebene-2-Residuen u_{0j} bzw. u_{1j} korreliert sind. Auch diese *u*'s werden als normalverteilte Zufallsvariablen aufgefasst, jeweils mit Erwartungswert gleich Null, und Varianz/Kovarianz

$$\text{var}\left(u_{0j}\right) = \sigma_{u0}^2 \qquad \text{var}\left(u_{1j}\right) = \sigma_{u1}^2 \qquad \text{cov}\left(u_{0j}, u_{1j}\right) = \sigma_{u01} \tag{7}$$

Mit den beiden σ^2 werden die Varianzen der mit b_0 (Konstante) bzw. mit b_1 (Steigung) assoziierten Zufallseffekte bezeichnet, mit σ die Kovarianz zwischen Konstante und Steigung.

3. Varianten

Gegenüber einem aus den Gleichungen (2), (3) und (4) gebildeten Modell sind sowohl einfachere als auch komplexere Modellvarianten denkbar. Spielraum in der Modellbildung entsteht zunächst durch die Möglichkeit, *unterschiedliche Annahmen über die Zufallseffekte* zu treffen, die mit b_{0j} und b_{1j} assoziiert sind. So würde die Annahme, dass u_{0j} *ungleich* Null ist, im Vergleich der Kontexte zu einem Bild paralleler Regressionsgeraden führen, wenn dem Modell zugleich die Restriktion auferlegt wird, dass u_{1j}, also der mit dem Steigungskoeffizienten assoziierte Zufallseffekt u_{1j}, *gleich* Null ist. Wird hingegen angenommen, dass auch u_{0j} gleich Null ist, würden alle parallelen Linien auf einer einzigen Regressionslinie aufeinander fallen. Das wäre der in der linearen Einfachregression unterstellte Standardfall (Gleichung 1). Werden im Zuge der Modellbildung aber *sowohl* u_{0j} *als auch* u_{1j} als ungleich Null angenommen, so entsteht im Vergleich der Kontexte die Möglichkeit, dass Konstante und Steigung kovariieren und somit eine Ansammlung *nicht*-paralleler Regressionsgeraden entsteht.

Neben solchen Modellierungsoptionen entsteht Spielraum dadurch, dass die Gleichungen selbst verkürzt oder erweitert werden können. Eine Verkürzung führt uns zu einem so genannten „leeren Modell" ohne erklärende Variable, eine Erweiterung zu

8 Vgl. z.B. Bryk und Raudenbush (1992: 12, 29f., 67f.); siehe auch Goldstein (1995: Kap. 2).

Modellen, in denen zusätzliche erklärende Individual- und/oder Aggregatmerkmale in die Gleichung aufgenommen werden.

4. Varianzzerlegung

Das „leere Modell“[9] entsteht aus (2), (3) und (4) durch Herausnahme der einzigen erklärenden Variable x_1 aus dem Modell. Dies führt zu den Gleichungen (8) und (9) und infolge dessen zur Schätzgleichung (10):

$$y_{ij} = b_{0j} + e_{ij} \qquad i = 1,...n_j;\ j = 1,...J \tag{8}$$

$$b_{0j} = b_0 + u_{0j} \qquad j = 1,...J \tag{9}$$

$$y_{ij} = b_0 + u_{0j} + e_{ij} \tag{10}$$

Das leere Modell schätzt qua b_0 den mittleren y_{ij}-Wert sowie die Varianzkomponenten der Zufallseffekte u_{0j} und e_{ij}. Es zerlegt die y_{ij}-Varianz in eine Komponente innerhalb der Kontexte sowie eine Komponente zwischen den Kontexten und reflektiert damit die Varianzanteile auf den Ebenen 1 und 2. Eine solche Varianzzerlegung ist nützlich, da auf ihrer Grundlage die Intraklassenkorrelation geschätzt werden kann. Außerdem lassen sich anhand des leeren Modells die auf den einzelnen Ebenen jeweils prinzipiell erklärbaren Varianzanteile ermitteln.

Das leere Modell ist zudem hilfreich bei der Bestimmung von Anteilen erklärter Varianz,[10] indem es der *PRE*-Logik der proportionalen Fehlerreduktion folgend als Bezugsmodell M_0 dienen kann.

5. Erklärte Varianz

Im Modell der proportionalen Reduktion des Vorhersagefehlers werden die Residualkomponenten zweier Modelle M_1 und M_0 über

$$\frac{\sigma^2_{(M0)} - \sigma^2_{(M1)}}{\sigma^2_{(M0)}} \tag{11}$$

zueinander in Beziehung gesetzt. Dabei unterscheiden sich die beiden Modelle in der Menge der erklärenden Variablen: M_1 enthält eine Teilmenge bestehend aus einer oder mehreren erklärenden Variablen, die in M_0 fehlen, um genau deren Erklärungsbeitrag über den PRE-Vergleich festzustellen. M_0 ist dabei häufig das leere Modell. Zu beachten ist nun, dass die Frage erklärter Varianz in Mehrebenenmodellen für jede der Ebenen gestellt werden kann. Folgen wir dafür dem von Bryk und Raudenbush (1992) be-

9 Den Begriff des „empty model“ übernehmen wir von Snijders und Bosker (1999). Das Modell wird auch als Random ANOVA (Varianzanalyse mit Zufallseffekten), „intercept-only model“ oder „variance decomposition model“ bezeichnet.

10 Vgl. hierzu Bryk und Raudenbush (1992: Kap. 4), Hox (1995: Kap. 2), Engel (1998: 3.1), Snijders und Bosker (1999: Kap. 7).

schriebenen Ansatz, so wäre (11) separat auf die Ebenen anwendbar und würde dann z.B. eine der folgenden Formen annehmen:

$$\frac{\sigma^2_{e(M0)} - \sigma^2_{e(M1)}}{\sigma^2_{e(M0)}} \qquad \frac{\sigma^2_{u0(M0)} - \sigma^2_{u0(M1)}}{\sigma^2_{u0(M0)}} \qquad \text{(12a) (12b)}$$

Dieser Ansatz liefert jedoch bestenfalls Approximationen (Hox 1995). Einen anderen Ansatz schlagen Snijders und Bosker (1999: 102–105) für Modelle vor, in denen für die Regressionskonstante, nicht aber für Steigungskoeffizienten Zufallseffekte angenommen werden. Die erklärte Varianz *auf Ebene 1* ergibt sich danach durch Anwendung von (11) auf die Schätzung der jeweiligen Gesamt- bzw. Gesamtresidualvarianz

$$\sigma^2_e + \sigma^2_{u0} \qquad \text{(13a)}$$

während die erklärte Varianz *auf Ebene 2* durch Anwendung von (11) auf

$$\frac{\sigma^2_e}{n} + \sigma^2_{u0} \qquad \text{(13b)}$$

ermittelt werden kann. *n* steht dabei für einen repräsentativen Wert für die Gruppen- bzw. Kontextgröße. Variiert diese Zahl n_j der Level-1-Einheiten pro Gruppe bzw. Kontext *j*, dann empfehlen die Autoren die Berechnung des harmonischen Mittels

$$\frac{N}{\left\{\sum_j \left(\frac{1}{n_j}\right)\right\}} \qquad \text{(N Anzahl der Kontexte [Level-2-Einheiten]).} \qquad \text{(14)}$$

Während die erklärten Varianzanteile damit leicht berechnet werden können, ist die Schätzung in Modellen, die zusätzlich auch Zufallseffekte für die Steigungskoeffizienten beinhalten, komplizierter. Die berechneten Werte erklärter Varianz unterscheiden sich Snijders und Bosker (1999: 105) zufolge jedoch kaum von denen, die sich bei Ignorierung der Zufallskomponenten der Steigungen ergeben. Daher schlagen die beiden Autoren zur Bestimmung der erklärten Varianzen von Modellen *mit* Zufallseffekten für die Steigungskoeffizienten vor, diese Modelle parallel auch als Modelle *ohne* zufällig variierende Steigungskoeffizienten zu schätzen und die erklärten Varianzen zu übertragen.

Alternativ ist sowohl für einfache als auch komplexere Modelle die Berechnung erklärter Varianzen auf der Grundlage der –2*Log-Likelihood oder Devianz eines Modells möglich. Langer (2004: 120) schlägt hierfür die Berechnung des Maximum Likelihood-R^2 nach Maddala (1986) über

$$R^2 = 1 - \exp\left(\frac{-2 * \log(lh)_{M0} - \left(-2 * \log(lh)_{M1}\right)}{n_{ij}}\right) \qquad \text{(15)}$$

(n_{ij} Stichprobenumfang)

vor.

III. Strukturelle Analyse

1. Aggregatmerkmale

Mehrebenenanalysen ermöglichen es, zusammen mit dem Einfluss von Individualmerkmalen auch denjenigen globaler, struktureller und analytischer Aggregatmerkmale abzuschätzen, indem kontextuelle Merkmale in die Analyse einbezogen werden. Ein „kontextuelles" Merkmal im Sinne der Lazarsfeld-Menzel-Typologie charakterisiert eine Person durch eine Eigenschaft des Kontextes, in dem sie lebt. Als Mitglieder eines solchen Kontextes aufgefasst, erhalten somit alle Personen, die Teil dieses Kontextes sind, in der zu analysierenden Datenmatrix denselben Wert der Kontexteigenschaft zugewiesen.

Um solche Aggregatmerkmale in eine Kontextanalyse einzubeziehen, sind die Ebene-2-Gleichungen (3) und (4) um Terme für entsprechende systematische Effekte zu erweitern. Denkbar wäre zum Beispiel, dass die Gleichung für b_{0j} um G und die Gleichung für b_{1j} um L erklärende Aggregatcharakteristika x_j erweitert wird:

$$b_{0j} = b_0 + \sum_{g=1}^{G} b_g x_{gj} + u_{0j} \qquad j = 1,...J \tag{16}$$

$$b_{1j} = b_1 + \sum_{l=1}^{L} b_l x_{lj} + u_{1j} \qquad j = 1,...J \tag{17}$$

Dabei kann es sich je nach Aggregateinheit und Untersuchungsfragestellung um unterschiedlichste Aggregatmerkmale handeln. Beispiele wären das aus den Individualdaten ermittelte mittlere Einkommensniveau, um damit das Wohlstandsniveau einer Region zu charakterisieren, oder die am Gini-Koeffizienten gemessene Einkommensungleichheit, um damit das in einer Region gegebene Wohlstandsgefälle abzubilden. In bildungs- bzw. schulsoziologischen Studien kann es beispielsweise darum gehen, neben individuellen Determinanten des Schulerfolgs auch Kontexteigenschaften von Schulklassen und/oder Schulen einzubeziehen, beispielsweise den mittleren sozioökonomischen Status oder den Grad ethnischer Heterogenität. In Studien, welche die Gemeinde als Meso-Ebene beinhalten, könnte Galtungs (1966) Ungleichheitstheorie folgend der Criss-Cross-Grad in konfliktrelevanten Dimensionen der kommunalen Sozialstruktur ein relevantes Aggregatmerkmal darstellen. Neben solchen analytischen Aggregatmerkmalen im Sinne der Lazarsfeld-Menzel-Typologie ließen sich aber auch globale Merkmale einbeziehen, zum Beispiel Infrastruktureigenschaften von Regionen.

2. Sozialstrukturanalyse

Speziell für die Sozialstrukturanalyse können sich die Erkenntnismöglichkeiten, die die Einbeziehung von Eigenschaften sozialer Aggregate in Individual-Level-Analysen bieten, als nützlich erweisen. Dies zumindest dann, wenn man sich Blaus (1981: 9) Position zueigen macht, derzufolge ein gemeinsamer Nenner in ansonsten unterschiedlichen Sozialstrukturkonzeptionen in der *Annahme emergenter Eigenschaften* gesehen werden kann. Denn danach würde sich der Begriff der Sozialstruktur gerade auf solche Ei-

genschaften eines Aggregats beziehen, die emergent[11] sind und folglich nicht die separaten Elemente charakterisieren können, aus denen das Aggregat gebildet wird. Dies sind unter anderem Eigenschaften, die die *Zusammensetzung* eines Aggregats reflektieren. Um diese zu berücksichtigen, müssten Studien, die Einflüssen der Sozialstruktur nachgehen wollen, die gängige Praxis erweitern, den Einfluss der Sozialstruktur im Wesentlichen auf die Frage *positionaler* Zugehörigkeiten zu beschränken. Darüber hinaus wäre vielmehr zu beachten, welche Binnenverteilungen solcher positionalen Zugehörigkeiten bestehen, da sich genau darüber Struktureigenschaften dieses Aggregats konstituieren, und zwar als spezielle analytische Eigenschaften im Sinne der Lazarsfeld-Menzel-Typologie. Im einfachsten Fall ginge es dann nicht nur darum zu berücksichtigen, welchen Einkommens- oder Bildungsstatus, welche berufliche Stellung oder welche ethnische Zugehörigkeit etc. jemand aufweist, sondern auch darum, wie sich die Mitglieder eines sozialen Aggregates über diese Dimensionen verteilen, welche Gruppen- bzw. Bevölkerungszusammensetzungen dadurch erzeugt werden und über welche Aggregatkonzepte diese beschrieben und gemessen werden können.[12] Während die Ausprägungen einfacher analytischer Aggregatkonzepte über Mittelwerte und durchschnittliche Streuungswerte in Individualvariablen berechnet werden können, beruhen komplexere Aggregatkonzepte auf der Verarbeitung von Zusammenhangsinformationen zwischen zwei und mehr Variablen einer tiefer liegenden Analyseebene. So wird z.B. die oben angesprochene, von Galtung als „Criss-Cross" bezeichnete Konfliktbindungskapazität der Sozialstruktur über eine Formel berechnet, in die für jede Aggregateinheit (z.B. Gemeinde) die gemeinsame Häufigkeitsverteilung von zwei oder mehr Individualvariablen eingeht (z.B. ökonomischer und sozialer Status).

In ähnlicher Weise stellt die in Abschnitt III.4 beschriebene „relationale Statusinkongruenz" ein Aggregatmerkmal dar, dessen Ausprägungen nicht aus den univariaten Verteilungen eines einzelnen Individualmerkmals hervorgehen, sondern jeweils aus dem statistischen Zusammenhang zwischen zwei Individualmerkmalen.

3. Kompositionseffekte

Bei einer Mehrebenenanalyse ist zwischen dem *Haupteffekt*, den ein Aggregatmerkmal auf y_{ij} ausübt, und der *Wechselwirkung* zu unterscheiden, die in Bezug auf y_{ij} zwischen diesem Aggregatmerkmal und einem Individualmerkmal entsteht.

Um zu verdeutlichen, wie beide Arten von Effekten bei der Modellspezifikation berücksichtigt werden können, greifen wir auf die Ebene-1-Gleichung (2) zurück, die wir hier unverändert als

11 Blau (1981: 13–15) unterscheidet vier grundlegende Arten von emergenten Eigenschaften, und zwar: die über die Zahl der individuellen Mitglieder oder die Zahl von Subaggregateinheiten messbare *Größe* des Aggregats, *soziometrische* Eigenschaften der Beziehungen in kleinen Gruppen und makrosoziologischen Netzwerken, die in Unterschieden (Ungleichheit und Heterogenität) zwischen seinen Elementen reflektierte *Zusammensetzung* des Aggregats, sowie Struktureigenschaften höherer Ordnung.

12 Dies ist ein Kerngedanke der „strukturellen Analyse". Siehe Galtung (1966), Blau und Schwartz (1984), Blau (1994); vgl. auch Müller (1993: Kap II, 4), Wuggenig (1990), Engel (1998: Kap. 2).

$$y_{ij} = b_{0j} + b_{1j}x_{1ij} + e_{ij} \qquad i = 1,...n_j;\ j = 1,...J \qquad (18)$$

wiederholen. Erweitern wir von den beiden zugehörigen Ebene-2-Gleichungen (3) und (4) Gleichung (4) um einen Term zur Schätzung des Effektes eines Aggregatmerkmals x_{2j}, während (3) unverändert bleibt, so resultieren als Ebene-2-Gleichungen

$$b_{0j} = b_0 + u_{0j} \qquad j = 1,...J \qquad (19)$$

$$b_{1j} = b_1 + b_2x_{2j} + u_{1j} \qquad j = 1,...J \qquad (20)$$

die in (18) eingesetzt zu einer Gleichung führen,

$$y_{ij} = b_0 + b_1x_{1ij} + b_2x_{2j}x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij} \qquad (21)$$

die über b_2 eine etwaige Cross-Level-Wechselwirkung zwischen Aggregatmerkmal x_{2j} und Individualmerkmal x_{1ij} zu schätzen erlaubt. Dies bedeutet, dass unter einer Cross-Level-Wechselwirkung der Einfluss verstanden werden kann, den ein Aggregatmerkmal *auf den Effekt* ausübt, den ein Individualmerkmal auf y_{ij} besitzt. Werden die Ebene-2-Gleichungen hingegen so respezifiziert, dass der Effekt des Aggregatmerkmals in die Gleichung für die Konstante b_{0j} aufgenommen wird,

$$b_{0j} = b_0 + b_2x_{2j} + u_{0j} \qquad j = 1,...J \qquad (22)$$

$$b_{1j} = b_1 + u_{1j} \qquad j = 1,...J \qquad (23)$$

so resultiert durch Einsetzen dieser Gleichungen (22) und (23) in (18) mit (24)

$$y_{ij} = b_0 + b_1x_{1ij} + b_2x_{2j} + u_{0j} + u_{1j}x_{1ij} + e_{ij} \qquad (24)$$

eine Gleichung, die nun den *Haupteffekt* des Aggregatmerkmals enthält. Dabei wird deutlich, dass es sich um keinen Effekt handelt, der einen anderen Effekt moderiert, sondern um einen Effekt, der sich auf die Höhe des mittleren kontextspezifischen y_{ij}-Niveaus auswirkt. Schließlich erhalten wir das umfassendste Modell, wenn *beide* Ebene-2-Gleichungen um einen Term für x_{2j} ergänzt werden.

$$b_{0j} = b_0 + b_{20}x_{2j} + u_{0j} \qquad (25)$$

$$b_{1j} = b_1 + b_{21}x_{2j} + u_{1j} \qquad (26)$$

$$y_{ij} = b_0 + b_1x_{1ij} + b_{20}x_{2j} + b_{21}x_{2j}x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij} \qquad (27)$$

Davon, dass der Haupteffekt eines Aggregatmerkmals oder eine Wechselwirkung zwischen diesem Aggregatmerkmal und einem Individualmerkmal Varianten von *Kompositionseffekten* darstellen, ist die Rede, wenn das Aggregatmerkmal als *analytisches* Merkmal im Sinne der Lazarsfeld-Menzel-Typologie aufgefasst werden kann und folglich eine Entsprechung auf der Individualebene aufweist.[13] Dann nämlich kann nicht nur nach dem eigenständigen Effekt eines Individualmerkmals gefragt werden, sondern zugleich nach den Effekten, die sich daraus ergeben, wie häufig oder selten *genau dieses Merkmal* innerhalb der Aggregateinheiten vorkommt, welche Gruppen- bzw. Aggregatzusammensetzungen daraus resultieren und wie sich diese Kompositionen auswirken.

13 In diesem Sinne z.B. Davis et al. (1961) oder Boyd und Iversen (1979: 56f.), die darin das Paradigma der Kontextanalyse sehen.

Beispielsweise kann die Wirkung, die der Verlust des Arbeitsplatzes auf die Lebenszufriedenheit entfaltet, davon abhängen, ob jemand in einer Region mit geringer oder hoher Arbeitslosigkeit lebt (Wechselwirkung). Unabhängig davon, ob jemand individuell arbeitslos ist oder nicht, mag sich die durchschnittliche Zufriedenheit in Regionen mit hoher oder weniger hoher Arbeitslosigkeit voneinander unterscheiden (Haupteffekt). Ob bzw. wie sehr ein Leben in Armut bestimmte Formen der Kriminalität beispielsweise im Jugendalter begünstigt, mag davon abhängen, ob Armut ein eher seltenes oder häufiges Phänomen in den Nachbarschaften und Wohnquartieren der in Armut lebenden Personen ist, so dass es weniger Armut per se (Individualeffekt), als vielmehr Armut *inmitten von Überfluss* und der damit verbundene Kontrast der Lebenschancen wäre, von dem ein Effekt zu erwarten ist (Wechselwirkung). Dabei kann das analytische Aggregatmerkmal im Sinne unserer Ausführungen in Abschnitt III 2. auch komplexerer Natur sein, d.h. nicht aus den univariaten Verteilungen eines einzelnen Individualmerkmals hervorgehen, sondern aus der Kombination mehrerer Individualmerkmale.

4. Anwendungsbeispiel: Statusinkonsistenz und relationale Statusinkongruenz

Zur Illustration möchten wir ein Problem aus der Sozialstrukturanalyse aufgreifen. Es betrifft die Frage nach möglichen Individualisierungseffekten, wenn Individualisierung als Strukturphänomen begriffen wird. Dann stellen Individualisierungs*effekte* Einflüsse dar, die in einem multidimensionalen Ungleichheitsgefüge dem jeweiligen strukturellen Differenzierungsgrad zuzurechnen sind.[14]

Bemühungen, Effekte struktureller Differenzierung empirisch zu identifizieren, sind nicht neu. Sie können vielmehr schon auf eine längere, wenngleich nicht mehrebenenanalytisch angelegte Forschungstradition zurückblicken, und zwar auf die Statusinkonsistenzforschung. „Statusinkonsistenz" ist ein Konzept zur Beschreibung des „Statussets" individueller Personen. Es unterstellt, dass dieses Statusset aus mehr als einem relevanten Faktor gebildet wird. Meist werden dies die Elementarfaktoren sein, die auch in die Standarddemographie eingehen bzw. in der Sozialstrukturanalyse eine wichtige Rolle spielen (Bildung, berufliche Stellung, ggf. Berufsprestige, Erwerbsstatus, Einkommen); aber je nach untersuchter Teilpopulation oder Fragestellung können durchaus auch andere Statusfaktoren relevant sein (vgl. z.B. Engel 1988).

Mit dem Statusinkonsistenzkonzept verbindet sich nun der Kerngedanke, die einzelnen Werte einer Person in diesen Elementarfaktoren sozialer Ungleichheit nicht zu einem Indexwert zu verbinden, der dann beispielsweise die Schichtzugehörigkeit oder Klassenzugehörigkeit oder ein vergleichbares Konstrukt widerspiegeln soll. Vielmehr bleiben die Einzeleinstufungen getrennt, um der Frage nachgehen zu können, inwie-

14 Vorliegende Analysen entstanden im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projektes „Integrationsleistungen der Sozialstruktur. Statusinkonsistenz und strukturelle Bindungen im Wandel westlicher Gesellschaften". Im Zentrum des Projektes steht das Ziel herauszuarbeiten, in welcher Weise sich im Zuge sozialen Wandels die Integrationsleistungen der Sozialstruktur der Bundesrepublik Deutschland und anderer westlicher Gesellschaften verändert haben und verändern. Wir bedanken uns bei der Deutschen Forschungsgemeinschaft für die finanzielle Förderung des Projekts.

weit diese in ihrem Binnenverhältnis stimmig sind. Sind sie stimmig, wird gesagt, dass dies einen Zustand der Statuskonsistenz indiziert, sind sie es nicht, wird von Status-*in*konsistenz gesprochen. Wann aber stehen die Einstufungen in einem stimmigen Binnenverhältnis zueinander? Der Theorie zufolge ist dies eine Frage der Konsistenznorm. Dies könnte z.B. die Gleichheit sein. Statuskonsistenz läge danach vor, wenn die Einstufungen in z.B. Bildung und beruflicher Stellung gleich sind, ungleiche Platzierungen würden Statusinkonsistenz anzeigen. Abgesehen von Messproblemen kann gegen diese Gleichheitsdefinition von Statusinkonsistenz auch ein Argument ins Felde geführt werden, das alternativ zur *Erwartungsdefinition von Statusinkonsistenz* führt: In einer Gesellschaft, in der niemand die Gleichheit als „normalen" Zustand erwartet, werden Abweichungen von diesem Zustand keine besonderen Effekte erzeugen, da die Abweichungen gar nicht als Anomalien wahrgenommen werden. Eher schon ist mit besonderen Effekten zu rechnen, wenn eine Statuskonfiguration von dem Zustand abweicht, der in einer Gesellschaft zu einer gegebenen Zeit erwartet werden kann.

Um diese Erwartung empirisch bestimmen zu können, stützen wir uns in unseren Analysen von Daten der 1998 bis 2001er Wellen des Sozioökonomischen Panels auf eine Hilfshypothese, die besagt, dass in der Bevölkerung erwartet wird, was *typischerweise* auftritt. Dies könnte durchaus ein Zustand sein, der am Gleichheitskriterium gemessen als „inkonsistent" einzustufen wäre. Zudem stützen wir uns auf eine Hilfshypothese, die besagt, dass sich das, was als normal oder typischerweise erwartbar angesehen wird, weniger an unbedingten als vielmehr an *bedingten* Erwartungen als impliziter Konsistenznorm orientiert. Dabei vermuten wir, dass als übergreifende Bedingungen insbesondere Investitionsdimensionen – wie der Bildungsstatus einer Person – relevant sind, und dass die Bewertung der zwischen Investitions- und Belohnungsstatus auftretenden Relationen bezugsgruppenspezifisch erfolgt: Als normal würde entsprechend angesehen werden, was für die eigene komparative oder normative Bezugsgruppe normal ist.

In leistungsorientierten Gesellschaften müssen insbesondere Investitionen in die eigene Bildung als zentrale Größe angesehen werden, um darüber vermittelt in berufliche Positionen zu gelangen und die mit diesen Positionen assoziierten Gratifikationen zu erwerben. Operational über die Anzahl der Schul- bzw. Bildungsjahre definiert, geht der Bildungsstatus als Investitionsmerkmal ein, um die bildungsbedingte Gratifikationserwartung über eine lineare oder logistische Regression zu bestimmen. Um dem oben skizzierten Bezugsgruppenargument Rechnung zu tragen, werden die Erwartungswerte über konditionale Haupteffekte der Bildungsvariable, also getrennt für einzelne Bevölkerungskategorien, geschätzt. Insgesamt wurden 12 Gruppen gebildet, und zwar durch Kombination von Nationalität (deutsch: j/n), Erwerbstatus (erwerbstätig; Rentner/in; andere Nichterwerbstätige) und Geschlechtszugehörigkeit (m/w), um den Bildungseffekt spezifisch für die jeweiligen Gruppen zu schätzen.

Das hier verwendete Statusinkonsistenz-Maß setzt die Anzahl der Bildungsjahre mit den monatlichen Einkünften in Beziehung, um Bildungs-Einkommens-Statusinkonsistenz operational über den Grad der Abweichung von der bildungsbedingten Einkommenserwartung zu bestimmen. Dazu wurde ein Index der Gesamteinkünfte gebildet und in einer Regression konditional für die einzelnen Gruppen auf die Anzahl der Bildungsjahre zurückgeführt. Das Maß bestimmt Statusinkonsistenz dabei spezifisch für

jede Gruppe über die bei der Regression anfallenden Residuen, also über die Differenz zwischen beobachteten und geschätzten Einkommenswerten, wobei jedes Residuum durch die Standardabweichung der Verteilung der Residuen der jeweiligen Gruppe dividiert wird. Der Index nimmt positive Werte an, wenn die Gesamteinkünfte über der bildungsbedingten Einkommenserwartung liegen, und negative Werte, wenn die Gesamteinkünfte unter dieser Erwartung liegen.

„Statusinkonsistenz" ist eine Eigenschaft von Statussets individueller Personen. Nun bilden solche individuellen Statussetträger im Kontext unserer Analysen Gruppen in Gestalt von Haushalten. Für Haushalte mit zwei und mehr Personen besteht folglich die Möglichkeit, die zwischen individuellen Statussets bestehenden Relationen einer Analyse zu unterziehen. Dazu werden – orientiert an der klassischen Idee der „Statusinkongruenz" – Paare von Haushaltsmitgliedern daraufhin betrachtet, ob sie in gleicher Weise statuskonsistent oder -inkonsistent sind, so dass sich die Dyade selbst in einem Zustand der Kongruenz oder schwächer oder stärker ausgeprägter Inkongruenz befindet. Die Relationalmaße setzen entsprechend die Idee um, die zwischen Paaren von Statussetträgern bestehenden Relationen über die Distanzen zwischen den individuellen Statusinkonsistenzwerten zu bestimmen. Dabei wird der Inkongruenzbestimmung die absolut gesetzte Differenz als Maß zugrunde gelegt, da es im Grunde arbiträr ist, wer innerhalb eines Haushaltes als 1. Person und wer als 2. Person in einen betreffenden paarweisen Vergleich eingebunden wird.

Um Haushalte in diesem Sinne nach ihrem Inkongruenzgrad zu klassifizieren, sind zwei variable Größen zu beachten: die Haushaltsgröße und die Zahl der innerhalb eines jeden Haushalts klassifizierbaren Personen. Beide Größen wären nur identisch, wenn jedes Haushaltsmitglied im Sinne des jeweiligen Statusinkonsistenzkonzeptes klassifizierbar wäre und bei keiner der klassifizierbaren Person fehlende Werte auftreten. Die Relationalmaße folgen der Konstruktionslogik, jeweils alle $k(k-1)/2$ Distanzen zu berechnen, die zwischen k inhaltlich klassifizierbaren Haushaltsmitgliedern bestimmt werden können, um schließlich den jeweiligen Inkongruenzgrad auf dieser Grundlage über die innerhalb eines Haushaltes bestehende *durchschnittliche Distanz* zu ermitteln.

IV. Modelle für Querschnitts- und Paneldaten

1. Modelle für Querschnittsdaten

Die Erweiterung von Mehrebenenmodellen um weitere Analyseebenen ist im Prinzip unkompliziert möglich, solange die Ebenen über das beschriebene Inklusionsprinzip miteinander verbunden bleiben. So kann ein 2-Ebenen-Modell mit Personen auf der Ebene 1 und Haushalten auf der Ebene 2 beispielsweise um eine Ebene 3 erweitert werden, die Regionen repräsentiert. Die hierarchische Datenstruktur „Personen i innerhalb von Haushalten j" wäre dazu lediglich auf die Struktur „Personen i innerhalb von Haushalten j innerhalb von Regionen k" zu erweitern. Eine solche Struktur liegt beispielsweise der Analyse mit Daten der 2001er Welle des *SOEP* zugrunde, die der oben in Abschnitt III 4. aufgeworfenen Fragestellung folgt und über deren Ergebnisse in *Ta-*

belle 2 berichtet wird. Diesen liegen Modelle mit Ebene-1-, Ebene-2- und Ebene-3-Gleichungen folgenden Typs zugrunde

$$y_{ijk} = b_{0jk} + b_1 x_{1ijk} + b_2 x_{2ijk} + e_{ijk} \tag{28}$$

$$b_{0jk} = b_{0k} + b_3 x_{3jk} + b_4 x_{4jk} + u_{0jk} \tag{29}$$

$$b_{0k} = b_0 + u_{0k} \tag{30}$$

Werden (30) in (29) und (29) in (28) eingesetzt, resultiert mit (31)

$$y_{ijk} = b_0 + b_1 x_{1ijk} + b_2 x_{2ijk} + b_3 x_{3jk} + b_4 x_{4jk} + u_{0k} + u_{0jk} + e_{ijk} \tag{31}$$

eine Gleichung mit Zufallseffekten für die Konstante b_0 auf den Ebenen 1, 2 und 3, jedoch ohne Zufallseffekte für die b's der einbezogenen erklärenden Variablen. Dabei handelt es sich auf der Ebene 1 der Person um ein Maß für Statusinkonsistenz (x_1) und einen Bildungsindex (x_2). Auf der Ebene 2 (Haushalt) wurde ein Maß für relationale Statusinkongruenz (x_4) einbezogen sowie ein begleitendes Maß (x_3) zur statistischen Kontrolle, das indiziert, ob es sich um einen Single-Haushalt handelt, in dem keine relationale Statusinkongruenz auftreten kann, oder ob aufgrund fehlender Werte in den individuellen Statusinkonsistenzwerten keine relationale Statusinkongruenz berechnet werden konnte.

Als Explananda wurde eine Auswahl einschlägiger, im *SOEP* verfügbarer Indikatoren sozialer Integration betrachtet, um die Auswirkungen von Statusinkonsistenz und -inkongruenz auf die Bereiche Ökonomie, Kultur und Politik sowie Mobilität zu untersuchen. Pro Bereich wurden jeweils zwei Indikatoren einbezogen: der Grad der Zufriedenheit mit der eigenen sozioökonomischen Lage (Indikatoren: Zufriedenheit mit Haushaltseinkommen und Lebensstandard), der Grad der Partizipation am kulturellen und politischen Geschehen (Indikatoren: Besuch kultureller Veranstaltungen und politisches Interesse) sowie für den Bereich der Mobilität die subjektive Erwartung einer beruflichen Veränderung innerhalb der nächsten 2 Jahre (Indikatoren: Aufnahme einer bezahlten Arbeit bzw. von sich aus Suche einer neuen Stelle (Jobsuche) und sich durch Lehrgänge oder Kurse weiterzuqualifizieren (Weiterbildung)).

Gemäß der Statusinkonsistenztheorie (vgl. z.B. Galtung 1966) wird davon ausgegangen, dass Statusinkonsistenz ein Ungleichgewichtszustand ist, den Akteure nach Möglichkeit zu vermeiden suchen und der aufgrund von Deprivationserfahrungen verschiedene psychosoziale Folgen mit sich bringen kann. Konkret wird für das vorliegende Modell davon ausgegangen, dass Statusinkonsistenz im Sinne eines unter der Bildungserwartung liegenden Einkommens negativ auf die Zufriedenheit mit dem Haushaltseinkommen und dem Lebensstandard wirkt, Statusinkonsistenz, bei der das tatsächliche Einkommen das aufgrund der Bildung erwartete Einkommen übersteigt, die Zufriedenheit mit Einkommen und Lebensstandard dagegen steigert.

Der Theorie zufolge (Geschwender 1967) würden statusinkonsistente Individuen zunächst versuchen, ihre Lage durch Mobilität zu ändern. Zu erwarten sind daher positive Statusinkonsistenzeffekte auf die beiden betrachteten Mobilitätsindikatoren, sofern das tatsächliche Einkommen unterhalb der bildungsspezifischen Erwartung liegt.

Bei Mobilitätsbarrieren wird der Theorie zufolge angenommen, dass inkonsistente Personen entweder versuchen, durch Anschluss an soziale bzw. politische Bewegungen

Tabelle 2: Bildungsstatus und Einkünfte (EK): Effekte von Statusinkonsistenz und -inkongruenz*

	Zufriedenheit mit ...				Besuchsfrequenz kultureller Veranstaltungen		Politisches Interesse		Subjektive Wahrscheinlichkeit von ...			
	... Haushalts-einkommen		... Lebens-standard						... Jobsuche		... Weiterbildung	
	b	s.e	b	s.e	b	s.e	b	s.e	b	s.e	b	s.e
Konstante	5,343	0,094	6,371	0,077	0,740	0,027	1,123	0,031	17,760	1,586	–9,683	1,615
Individualebene: Statusinkonsistenz (SI)												
EK: real über (+) / unter (–) bildungsbedingter Erwartung (x_1)	0,254	0,014	0,120	0,012	0,029	0,005	0,074	0,006	–5,772	0,315	0,016	0,315
Haushaltsebene: Statusinkongruenz												
(Single-HH oder kein Vgl. möglich [1]) vs. Kongruenz [0] (x_3)	–0,183	0,049	–0,083	0,040	0,075	0,015	0,053	0,017	5,822	0,844	2,556	0,860
Inkongruenzgrad (x_4)	0,202	0,029	0,122	0,023	0,059	0,009	0,022	0,009	1,363	0,445	0,944	0,458
Kontrolle des Haupteffektes												
Anzahl der Bildungsjahre (x_2)	0,096	0,006	0,067	0,005	0,086	0,002	0,094	0,002	0,404	0,125	3,717	0,126
Erklärte Varianz	R^2		R^2		R^2		R^2		R^2		R^2	
Gesamtmodell	6,3		3,4		14,2		10,4		3,1		8,1	
ohne Haupteffekt (nur SI-Maße)	4,6		2,0		1,6		1,4		2,9		0,5	
ohne Haupteffekt und HH-Maße (nur SI auf Individualebene)	3,2		1,3		0,4		0,8		2,5		0,0	
	Varianzzerlegung („empty model") (in Prozent)											
Level 3 : Region	4,5		4,9		2,2		1,7		1,5		2,0	
Level 2 : Haushalt	57,9		54,2		55,5		30,3		9,9		21,6	
Level 1 : Person	37,6		40,9		42,3		68,0		88,6		76,4	
Fallzahl n	17.154		17.154		17.154		17.154		12.288		12.288	

* Bezug: Gleichung (31); s.e: Standardfehler.

oder Gruppen Veränderungen herbeizuführen, von denen sie eine Verbesserung der gesellschaftlichen oder eigenen Lage erwarten, oder dass sie sich immer mehr aus dem sozialen und kulturellen Leben zurückziehen. Für das vorliegende Modell hieße dies, dass ein über der Bildungserwartung liegendes Einkommen zu einem Rückzug aus dem kulturellen Bereich (im Sinne einer abnehmenden Besuchsfrequenz kultureller Veranstaltungen) und eventuell auch zu einer Abnahme des politischen Interesses führen sollte.

Die in *Tabelle 2* ausgewiesenen Ergebnisse zeigen, dass leichte Statusinkonsistenz- und Inkongruenzeffekte unter den Bedingungen des Modells auch bei Kontrolle des zugehörigen Ungleichheitseffektes nachgewiesen werden können. Die Erklärungsleistung des Modells ist jedoch gering. Dies trifft in verstärktem Maße zu, wenn der Haupteffekt der Bildung statistisch nicht kontrolliert wird. Der relativ größte Anteil erklärter Varianz – unter Kontrolle des Haupteffektes – lässt sich mit einem Wert von 4,6 Prozent für die Zufriedenheit mit dem Haushaltseinkommen feststellen. Diese kann also durch die einbezogenen Statusinkonsistenz- und -kongruenzmaße vergleichsweise gut erklärt werden.

Festzustellen ist ein negativer Effekt individueller Statusinkonsistenz auf die subjektive Wahrscheinlichkeit, innerhalb der nächsten zwei Jahre einen neuen Job zu suchen. Das heißt: Liegt das Einkommen über der bildungsbedingten Erwartung, sinkt die Wahrscheinlichkeit der Jobsuche, liegt das Einkommen dagegen unter der bildungsbedingten Erwartung, steigt die Wahrscheinlichkeit der Jobsuche, was auch der theoretischen Annahme entspricht.

Bei den übrigen Indikatoren ist die Lage wie folgt: Liegt das Einkommen über der bildungsbedingten Einkommenserwartung, steigt die erwartete Zufriedenheit mit dem Haushaltseinkommen und dem Lebensstandard, und auch die Besuchsfrequenz kultureller Veranstaltungen und das politische Interesse werden positiv beeinflusst. Die eingangs formulierten Vermutungen zur Wirkung von Statusinkonsistenz können damit zunächst bestätigt werden.

Relevante Varianzanteile der abhängigen Variablen sind nicht nur auf der Individualebene, sondern auch auf der Haushaltsebene zu finden. Hier können zum Teil über 50 Prozent der Varianz erklärt werden. Die Vernachlässigung der Haushaltsebene hätte somit zur Folge gehabt, die Effekte struktureller Differenzierung auf dieser Ebene zu übersehen und damit den haushaltsbezogenen Struktureinfluss zu unterschätzen. Nur marginale Varianzanteile entfallen dagegen auf die Ebene der 97 Regionen, was – wie oben beschrieben – auf den zu hohen Aggregierungsgrad der im SOEP verfügbaren Regionalvariable zurückzuführen sein dürfte.

2. Modelle für Paneldaten

Die Erweiterung von Modellen, die zwei Analyseebenen unterscheiden, auf Modelle mit drei und mehr Ebenen kann nicht nur dazu genutzt werden, dadurch *mehr als eine* Kontextebene in der Analyse zu berücksichtigen. Vielmehr kann dadurch auch die Zeitdimension einbezogen werden.

Dies ermöglicht ein Ansatz, mit dem für Daten, in denen in interessierenden Variablen wiederholte Messungen vorliegen, Wachstumskurven (growth curves) modelliert

werden. In den Sozialwissenschaften fallen solche Daten insbesondere in *Panel-Studien* an, also in Studien, in denen ein Sample von Haushalten und/oder Personen in zumeist regelmäßigen Zeitabständen wiederholt befragt wird.

Dabei liegt dem hier vorzustellenden Modellierungsansatz eine rearrangierte Datenstruktur zugrunde. Diese wird für das vorliegende Beispiel in *Tabelle 3* ausgewiesen[15] und besteht im Wesentlichen darin, dass die üblicherweise in einer Datenmatrix *nebeneinander* stehenden Werte einer Person in den t_1-, t_2- und t_3-Variablen nun *untereinander* stehen. Eine Datenzeile repräsentiert dann einen Zeitpunkt „innerhalb" einer Person, nicht mehr die Person selbst. Entsprechend gibt „ID1" den jeweiligen Messzeitpunkt an. Die Identifikationsnummer der Personen ist in „ID2" erfasst. Diese stellt eine laufende Nummer innerhalb der Kontexte dar. Die Kontext-ID selbst liefert „ID3". Sie indiziert im vorliegenden Beispiel, dass die ersten 3 Personen Mitglieder des Kontextes mit der ID3=1 und die folgenden 2 Personen Mitglieder des Kontextes mit der ID3=2 sind. Erst solche Kontext-IDs ermöglichen es, Mehrebenenanalysen durchzuführen, da über sie die Elemente der Ebene n mit den umgebenden Einheiten der Ebene n+1 verknüpft werden.

Wir haben damit eine hierarchische Datenstruktur „Messzeitpunkt *i* innerhalb Person *j* innerhalb Kontext *k"* erzeugt, die nun einer Wachstumskurvenanalyse unterzogen werden kann.[16] Zeitkonstante Merkmale wie x_0 in vorliegendem Beispiel werden dabei wie in der rechten Spalte von *Tabelle 3* angedeutet über die Zeitpunkte *innerhalb der jeweiligen Person* „ausgedehnt", d.h. entsprechend häufig repliziert.

Im Ansatz zur Modellierung von Wachstumskurven repräsentiert Ebene 1 üblicherweise die Zeitdimension und enthält entsprechend alle *zeitveränderlichen* Merkmale einer Person. Ebene 2 verkörpert die Person, deren Merkmale als *zeitkonstante* Größen in die Analyse eingehen. Die dritte Ebene, *k,* repräsentiert den Kontext mit seinen ebenfalls als *zeitkonstante* Größen in die Analyse eingehenden Aggregatmerkmalen, während *zeitveränderliche* Aggregatmerkmale zusammen mit den zeitveränderlichen Individualmerkmalen als Kovariate auf der Ebene 1 berücksichtigt werden können.

So könnte es sich in *Tabelle 3* bei y_1 um die Zufriedenheit mit dem Haushaltseinkommen handeln, bei x_1 um ein zeitveränderliches Individualmerkmal und bei x_2 um ein zeitveränderliches Aggregatmerkmal. Neben diesen Größen könnte zudem x_0 als ein für den Beobachtungszeitraum zeitkonstantes Individualmerkmal in die Analyse eingehen.

Ein entsprechendes 3-Ebenen-Beispiel, das wieder der in III 4. aufgeworfenen Frage nach möglichen Einflüssen struktureller Differenzierung folgt, haben wir mit Daten der vier Wellen von 1998 bis 2001 des *SOEP* berechnet. Dabei geht es wieder um den Einfluss von individueller Bildungs-Einkommens-Statusinkonsistenz (x_1) und relationa-

15 Beispielsweise bietet die für Mehrebenenanalysen entwickelte MLwiN-Software die Möglichkeit, den erforderlichen Umbau der Datenmatrix unkompliziert über eine entsprechende Routine zu bewerkstelligen.

16 Der Ansatz erlaubt flexible Modellierungen (vgl. z.B. Maas und Snijders 2003) und über Polynome auch unkompliziert die Schätzung von nichtlinearen Effekten der Zeit. Er ist zudem von Vorteil, da er das Problem fehlender Werte dadurch leicht entschärft, dass pro Person nicht dieselbe Zahl wiederholter Messungen vorliegen muss (vgl. Engel 1998: Kap. 4.1 und 6). Einer Analyse von Marsh und Hau (2002) zufolge erweist er sich aber als anfällig gegenüber Artefakten bei der Veränderungsmessung.

Tabelle 3: Für die Modellierung von Wachstumskurven rearrangierte Datenstruktur

Lfd.	ID3	ID2	ID1 (t)	y_1	x_1	x_2	x_0
1	1	1	1	y_{11}	x_{11}	x_{21}	x_0
2	1	1	2	y_{12}	x_{12}	x_{22}	x_0
3	1	1	3	y_{13}	x_{13}	x_{23}	x_0
4	1	2	1	y_{11}	x_{11}	x_{21}	x_0
5	1	2	2	y_{12}	x_{12}	x_{22}	x_0
6	1	2	3	y_{13}	x_{13}	x_{23}	x_0
7	1	3	1	..	..	..	..
8	1	3	2	..	..	..	..
9	1	3	3	..	..	..	..
10	2	1	1	..	..	..	..
11	2	1	2	..	..	..	..
12	2	1	3	..	..	..	..
13	2	2	1	..	..	..	..
14	2	2	2	..	..	..	..
15	2	2	3	..	..	..	..

Lfd.: Laufende Nummer, ID3: Kontext, ID2: Person, ID1 (t): Messzeitpunkt.

ler Statusinkongruenz (x_4). Einbezogen ist auch hier wieder ein Bildungsindex (x_2). Wie in der Querschnittsanalyse des vorangehenden Abschnitts wird außerdem auch hier wieder zur statistischen Kontrolle ein begleitendes Maß (x_3) einbezogen, das indiziert, ob es sich um einen Single-Haushalt handelt, in dem keine relationale Statusinkongruenz auftreten kann, oder ob diese aufgrund fehlender Werte in den individuellen Statusinkonsistenzwerten nicht berechnet werden konnte. Schließlich wurde der für diesen Modelltyp erforderliche Zeitindex (t) aufgenommen. Die zu erklärende Größe *y* ist die Zufriedenheit mit dem Haushaltseinkommen.

Tabelle 4 informiert in ihrer rechten Hälfte über das Ergebnis der mit dem Programm MLwiN 2.0 durchgeführten Parameterschätzungen und liefert zum Vergleich in ihrer linken Hälfte das korrespondierende Modell mit vier unterschiedenen Analyseebenen. Die zugehörigen Ebene-1-, Ebene-2- und Ebene-3-Gleichungen

$$y_{ijk} = b_{0jk} + b_t t_{ijk} + b_1 x_{1\,ijk} + b_3 x_{3\,ik} + b_4 x_{4\,ik} + e_{ijk} \tag{32}$$

$$b_{0jk} = b_{0\,k} + b_2 x_{2jk} + u_{0jk} \tag{33}$$

$$b_{0\,k} = b_0 + u_{0\,k} \tag{34}$$

beschreiben dabei wieder ein Modell, das Zufallseffekte und folglich entsprechende Varianzkomponenten für die Konstante, nicht aber für die Steigungskoeffizienten vorsieht. Werden (34) in (33) und (33) in (32) eingesetzt, resultiert mit (35)

$$y_{ijk} = b_0 + b_t t_{ijk} + b_1 x_{1\,ijk} + b_2 x_{2jk} + b_3 x_{3\,ik} + b_4 x_{4\,ik} + u_{0\,k} + u_{0jk} + e_{ijk} \tag{35}$$

die dem in *Tabelle 4* ausgewiesenen Befund zugrunde liegende Schätzgleichung für die Modellvariante, die drei Analyseebenen unterscheidet.

Über die Subskribierung wird dabei verdeutlicht, dass die Aggregatmerkmale x_3 und x_4 über die Zeit *i* und die Kontexte *k* variieren, nicht aber über die Personen *j* innerhalb eines Kontextes. Als „kontextuelle Eigenschaft" im Sinne der Lazarsfeld-Menzel-Typologie aufgefasst, würden alle Personen eines Kontextes denselben Wert der Kontexteigenschaft zugewiesen bekommen.

Tabelle 4: Lineares Wachstumskurvenmodell mit Kovariaten auf Individual- und Haushaltsebene*

	Zufriedenheit mit dem Haushaltseinkommen			
	4-Ebenen-Modell		3-Ebenen-Modell	
	b	s.e	b	s.e
Konstante	5,638	0,089	5,667	0,075
Verstrichene Zeit seit 1. Messung (t)	0,107	0,006	0,108	0,006
Individualebene: Statusinkonsistenz (SI)				
EK: real über (+) / unter (–) bildungsbedingter Erwartung (x_1)	0,046	0,010	0,047	0,010
Haushaltsebene: Statusinkongruenz				
(Single-HH oder kein Vgl. möglich [1]) vs. Kongruenz [0] (x_3)	–0,330	0,039	–0,302	0,040
Inkongruenzgrad (x_4)	0,087	0,014	0,102	0,014
Kontrolle des Haupteffektes				
Anzahl der Bildungsjahre (x_2)	0,044	0,006	0,037	0,006
	Varianzzerlegung („empty model") (in Prozent)			
Level 4 Region [n_4 = 97]:	4,5			
Level 3 Haushalt [n_3 = 6.582]:	49,0		53,6	
Level 2 Person [n_2 = 11.273]:	5,7		5,6	
Level 1 Zeit [n_1 = 44.062]:	40,9		40,8	

* Bezug: Gleichung (35), s.e: Standardfehler.

In der Tabellenreihe „Konstante" liefert b_0 (=5,667) den geschätzten *y*-Wert für den Fall, dass *t* und die *x*-Variablen gleich Null sind. Um dem Schätzwert die Bedeutung einer *anfänglichen Schätzung* zu geben, dessen erwartete Veränderung pro Zeiteinheit von b_t abgelesen werden kann, startet der Zeitindex *t* mit Null für t_1 (t=0, 1, 2, 3). Da bei der Interpretation allerdings auch die Kovariate mit ihren Ausprägungen zu beachten sind, liefert b_0 diesen anfänglichen y-Wert genau genommen für statuskonsistente Personen (x_1=0) aus kongruenten Haushalten (x_3=0, x_4=0) mit Null Jahren Schulbildung (x_2=0). Während Statuskonsistenz und -kongruenz empirisch auftritt, trifft dies auf die Null Jahre Schulbildung nicht zu, so dass b_0 = 5,667 um 7 × 0,037 = 0,259 auf 5,926 zu erhöhen wäre, um den anfänglichen Schätzwert in der Gruppe mit der Mindestbildung von 7 Schuljahren zu bekommen. Eine Alternative wäre gewesen, die Bildungsvariable zentriert in das Modell einzubeziehen, da die „Null" damit die Bedeutung durchschnittlicher Bildung erhalten hätte.

Für die Zeit selbst ist mit b_1 (=0,108) ein geringfügig positiver Effekt festzustellen, die durchschnittliche erwartete Zufriedenheit mit dem Haushaltseinkommen steigt also im Zeitverlauf leicht an. Die Effekte der einbezogenen Statusinkonsistenzindikatoren weisen in dieselbe Richtung wie in der oben präsentierten Querschnittsanalyse: Liegt das Einkommen über der bildungsspezifischen Erwartung, steigt die vorhergesagte Zufriedenheit mit dem Haushaltseinkommen.

Den Einfluss von zeitveränderlichen Kovariaten auf Individual- und Aggregatebene wie hier spezifiziert in die Analyse einzubeziehen, erscheint sinnvoll, wenn „gleichlaufende Variation von vermuteter Ursache und Wirkung" als Kriterium herangezogen

werden soll, um auf den Grad zu schließen, in dem eine Beziehung als Kausalbeziehung angesehen werden kann.

V. Kategoriale abhängige Variable

Vieles von dem, was für die bislang vorgestellten Modelle beschrieben wurde, lässt sich im Prinzip auch auf Modelle mit kategorialen abhängigen Variablen beziehen. Zu beachten ist allerdings, dass das Modell für solche *y*-Variable geeignet sein muss. Wir beschränken uns hier auf den Fall einer *dichotomen*, 0/1-kodierten Variable y_{ijk} für das Beispiel eines mehrebenenanalytischen logistischen 3-Ebenen-Modells, das ansonsten wie oben in Gleichung (31) spezifiziert ist. Den Ausgangspunkt bildet dann eine Gleichung

$$y_{ijk} = p_{ijk} + e_{ijk} \tag{36}$$

die den y-Wert von Person *i* aus Haushalt *j* aus Region *k* als Funktion einer Wahrscheinlichkeit p_{ijk}, in *y* den Wert eins zu haben, und einer Level-1-Residualkomponente e_{ijk} auffasst, wobei sich mit y_{ijk} bzw. e_{ijk} die Annahmen verbinden, dass diese binomialverteilt sind

$$y_{ijk} = Bin\left(p_{ijk}, n_{ijk} \right) \tag{37}$$

mit Varianz gleich

$$\text{var}\left(y_{ijk} \middle| p_{ijk} \right) = \frac{p_{ijk}\left(1 - p_{ijk}\right)}{n_{ijk}} \tag{38}$$

Dabei stellt n_{ijk} den Nenner für den Anteils- bzw. Wahrscheinlichkeitswert dar. Dieser nimmt für den Fall, dass individuelle, d.h. nicht bereits aggregierte bzw. tabellierte Daten modelliert werden, den Wert eins an.

Für den Anteilswert selbst wird (parallel zu Gleichung 31) mit

$$\ln\left(\frac{p_{ijk}}{1 - p_{ijk}} \right) = b_0 + b_1 x_{1ijk} + b_2 x_{2ijk} + b_3 x_{3jk} + b_4 x_{4jk} + u_{0jk} + u_{0k} \tag{39}$$

eine Gleichung formuliert, die neben den b's für die erklärenden *x* Variablen Zufallseffekte für die Konstante auf den Ebenen 2 und 3 enthält und durch Anwendung der Exponentialfunktion leicht in eine Gleichung überführt werden kann, in der das Verhältnis in den Komplementärwahrscheinlichkeiten selbst (und nicht dessen Logarithmus) als Funktion der Einflussfaktoren e^{b1}, e^{b2}, ... ausgedrückt wird.

Snijders und Bosker (1999: 225ff.) schlagen für mehrebenenanalytische logistische Regressionen ein Verfahren zur Bestimmung der erklärten Varianz vor. Angewandt auf vorliegendes 3-Ebenen-Beispiel würde dies dazu führen, den Anteil erklärter Varianz über

$$R^2_{dicho} = \frac{\sigma^2_F}{\sigma^2_F + \left[\sigma^2_{u0jk} + \sigma^2_{u0k}\right] + \sigma^2_e} \tag{40}$$

zu berechnen. Dabei steht die Varianzkomponente mit Subskript *F* für die Varianz des unter Rückgriff auf die b's berechneten linearen Prädiktors und die übrigen Varianzkomponenten für die Residualvarianzen auf den drei Ebenen. Für Level 1 ist diese Varianz auf einen festen Wert fixiert, da sie eine Funktion des vorhergesagten Wertes für p ist (Snijders und Bosker (1999: 225ff., vgl. auch Hox 1995: 78).[17]

VI. Schluss

Der Beitrag hat versucht, mit Fragestellungen und Grundlagen der Kontextanalyse vertraut zu machen, und aufzuzeigen, wie das Instrumentarium der statistischen Mehrebenenanalyse für Kontextanalysen eingesetzt werden kann. Einem einführenden Teil folgte dazu in Teil II die Darstellung einiger elementarer Grundlagen der hierarchischen linearen Regression, bevor sich Teil III darauf aufbauend der strukturellen Analyse und Identifizierung von Kompositionseffekten zuwandte. Teil IV erweiterte die Modelle auf drei Analyseebenen, um aufzuzeigen, wie diese Erweiterung für die Analyse von Querschnittsdaten und Paneldaten genutzt werden kann. Abschließend wandte sich Teil V einer Möglichkeit der Modellierung kategorialer abhängiger Variablen zu.

Auch wenn damit die Modellierungsoptionen alles andere als erschöpfend beschrieben wurden, konnte vielleicht ansatzweise verdeutlicht werden, *dass* Kontext- und Mehrebenenanalysen vielfältige Optionen bieten.[18] Je komplexer die Modellbildung wird, desto mehr Spielraum entsteht allerdings für Fehlspezifikationen der Modelle.

---

17 Die Frage des adäquaten Schätzverfahrens stellt sich für Mehrebenenmodelle für binäre y-Variablen in besonderer Weise. Guo und Zhao (2000) gehen auf die Problematik verzerrter Schätzungen ein und vergleichen verschiedene Quasi-Likelihood-Verfahren miteinander (MQL-1, MQL-2, PQL-1, PQL-2), die das Programm MLn bietet. Sie finden nur minimale Unterschiede zwischen PQL-1 und PQL-2 und werten ihren Befund dahingehend, dass er konsistent mit Goldstein und Rasbashs (1996) Beobachtung sei, dass das PQL Modell erster Ordnung erwartungsgemäß gut funktioniere, wenn die Varianzen in einem Mehrebenenmodell für binäre Daten ungefähr 0,5 nicht übersteigen. Guo und Zhao (2000) beziehen in ihren Vergleich auch das Pseudo-Likelihood-Schätzverfahren ein, welches das *SAS* Macro GLIM-MIX bietet, und stellen fest, dass dessen Schätzwerte annähernd mit denen von *MLn*s PQL-1 übereinstimmen, ohne allerdings mit diesen identisch zu sein. Die Frage adäquater Schätzer wird in Mehrebenenmodellen jedoch nicht nur im Falle diskreter Response-Variablen gestellt. MLwiN bietet z.B. die Möglichkeit, Modelle auf der Basis des RIGLS/IGLS (Restricted) Iterative Generalized Least Squares-Kriteriums zu schätzen, ebenso sind MCMC, Quasilikelihood und Bootstrap-Schätzungen möglich (siehe dazu Goldstein 1995: 21–24; zum IGLS Verfahren, vgl. auch Blien 2001: 159–163). Blien und Wiedenbeck (2002) diskutieren eine Verallgemeinerung des üblichen IGLS-Schätzverfahrens nach Goldstein, und zwar den CIGLS „Conditioned Iterative Generalized Least Squares Estimator" nach Rice, Jones und Goldstein. Der interessierte Leser sei an diese Quellen verwiesen.

18 Als eine weitere, hier nicht diskutierte Option wäre die Überführung eines Wachstumskurvenmodells in ein latentes Wachstumsmodell zu nennen, welche die Schätzung eines Modells mit multiplen Indikatoren ermöglicht.

Die verfügbaren Programmpakete bieten entsprechend Diagnosemöglichkeiten an, um etwaige Verletzungen von Anwendungsvoraussetzungen erkennen zu können. Auch werden die einsetzbaren Schätzverfahren selbst diskutiert. Da diese Punkte in vorliegendem Beitrag nicht oder nur am Rande behandelt werden konnten, bleibt nur, interessierte Leser an die jeweilige Spezialliteratur zu verweisen.

Gleiches gilt, wenn die Methode der statistischen Mehrebenenanalyse für andere Zwecke als die Durchführung einer Kontextanalyse genutzt wird. Dabei denken wir insbesondere an die Mehrebenenanalyse als Instrument zur Durchführung von Metaanalysen, zur Schätzung von Standardfehlern in mehrstufigen Samples und als Missing-Data-Technik.

Literatur

Achen, Christopher H., und *W. Phillips Shively*, 1995: Cross-Level Inference. Chicago: University of Chicago Press.

Baumert, Jürgen, und *Gundel Schümer*, 2001: Schulformen als selektionsbedingte Lernmilieus. S. 454–467 in: *Deutsches PISA-Konsortium* (Hg.), PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Blau, Peter M., 1960: Structural Effects. American Sociological Review 25: 178–193.

Blau, Peter M., 1981: Diverse Views of Social Structure and Their Common Denominator. S. 1–23 in *Peter M. Blau* und *Robert K. Merton* (Hg.), Continuities in Structural Inquiry. London/Beverly Hills: Sage.

Blau, Peter M., 1994: Structural Contexts of Opportunities. Chicago: University of Chicago Press.

Blau, Peter M., und *Joseph E. Schwartz*, 1984: Crosscutting Social Circles. Orlando: Academic Press.

Blien, Uwe, 2001: Arbeitslosigkeit und Entlohnung auf regionalen Arbeitsmärkten. Heidelberg: Physica.

Blien, Uwe, und *Michael Wiedenbeck,* 2002: Mehrebenenanalyse. In: *Gerhard Kleinhenz* (Hg.), IAB Kompendium Arbeitsmarkt- und Berufsforschung. Beiträge zur Arbeitsmarkt- und Berufsforschung 250. Nürnberg.

Boyd, Lawrence H., und *Gudmund R. Iversen*, 1979: Contextual Analysis: Concepts and Statistical Techniques. Belmont: Wadsworth Publ. Co.

Bryk, Anthony S., und *Stephen W. Raudenbush*, 1992: Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park: Sage.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld und *Robert L. York*, 1966: Equality of Educational Opportunity. Washington D.C.: U.S. Government Printing Office.

Coleman, James S., 1990: Foundations of Social Theory. Cambridge: Belknap Press of Havard University Press.

Davis, James A., Joe L. Spaeth und *Carolyn Huson,* 1961: A Technique for Analyzing the Effects of Group Composition. American Sociological Review 21: 215–225.

Duncan, Otis Dudley, und *Beverly Davis*, 1953: An Alternative to Ecological Correlation. American Sociological Review 18: 665–666.

Engel, Uwe, 1988: Status Inconsistency and Criss-Cross in an Adolescent Society. International Sociology 3: 283–300.

Engel, Uwe, 1998: Einführung in die Mehrebenenanalyse. Grundlagen, Auswertungsverfahren und praktische Beispiele. Opladen/Wiesbaden: Westdeutscher Verlag.

Esser, Hartmut, 1993: Soziologie. Allgemeine Grundlagen. Frankfurt a.M./New York: Campus.

Esser, Hartmut, 1999: Soziologie. Spezielle Grundlagen. Band 1: Situationslogik und Handeln. Frankfurt a.M./New York: Campus.

Falter, Jürgen W., 1991: Hitlers Wähler. München: Beck.

Galtung, Johan, 1966: Rank and Social Integration: A Multidimensional Approach. S. 145–198 in: *Joseph Berger, Morris Zelditch* und *Bo Anderson* (Hg.), Sociological Theories in Progress, Band 1. Boston: Houghton Mifflin Company.

Galtung, Johan, 1970 (1967): Theory and Methods of Social Research. London: George Allen and Unwin Ltd.

Geschwender, James A., 1967: Continuities in Theories of Status Consistency and Cognitive Dissonance. Social Forces 46: 160–171.

Goldstein, Harvey, 1995: Multilevel Statistical Models. London: Edward Arnold.

Goldstein, Harvey, und *Jon Rasbash*, 1996: Improved Approximations for Multilevel Models with Binary Responses. Journal of the Royal Statistical Society, Series A, 159: 505–513.

Guo, Guang, und *Hongxin Zhao*, 2000: Multilevel Modeling for Binary Data. Annual Review of Sociology 26: 441–462.

Hox, Joop J., 1995: Applied Multilevel Analysis. Amsterdam: TT-Publikaties.

Hox, Joop J., 2000: Multilevel Analyses of Grouped and Longitudinal Data. S. 15–32 in: *T. D. Little, K. U. Schnabel, J. Baumert* (Hg.), Modeling Longitudinal and Multilevel Data. Mahwah, N. J.: Erlbaum.

Hox, Joop J., und *Ita G. Kreft*, 1994: Multilevel Analysis Methods. Sociological Methods & Research 22: 283–299.

Langer, Wolfgang, 2004: Mehrebenenanalyse. Eine Einführung für Forschung und Praxis. Wiesbaden: VS Verlag.

Lazarsfeld, Paul F., und *Herbert Menzel*, 1969 (1961): On the Relation Between Individual and Collective Properties. S. 449–516 in: *Amitai Etzioni* (Hg.), A Sociological Reader on Complex Organisations. 2. Auflage. New York.

Maas, Cora J.M., und *Tom A.B. Snijders*, 2003: The Multilevel Approach to Repeated Measures for Complete and Incomplete Data. Quality and Quantity 37: 71–89.

Maddala, G. S., 1986: Limited-dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

Marsh, Herbert W., und *Kit-Tai Hau*, 2002: Multilevel Modeling of Longitudinal Growth and Change: Substantive Effects or Regression Toward the Mean Artifacts? Multivariate Behavioral Research 37: 245–282.

Müller, Hans-Peter, 1993: Sozialstruktur und Lebensstile. Frankfurt a.M.: Suhrkamp.

Rasbash, Jon, William Browne, Harvey Goldstein, Min Yang, Michael Healy, Geoff Woodhouse, David Draper, Ian Longford und *Toby Lewis*, 2000: A User's Guide to MLwiN. Multilevel Models Project. Institute of Education. University of London.

Raudenbush, Stephen, Anthony Bryk, Yuk Fai Cheong und *Richard Congdon*, 2001: HLM 5. Lincolnwood: Scientific Software International.

Robinson, W.S., 1950: Ecological Correlations and Behavior of Individuals. American Sociological Review 15: 351–357.

Snijders, Tom A. B., und *Roel J. Bosker*, 1999: Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling. London: Sage.

Wuggenig, Ulf, 1990: Eine strukturelle Version der Theorie der Statusinkonsistenz. S. 37–69 in: *Karl-Dieter Opp* und *Reinhard Wippler* (Hg.), Empirischer Theorienvergleich. Opladen: Westdeutscher Verlag.

V. Zeitbezogene Daten

METHODEN ZUR ANALYSE VON SEQUENZDATEN*

Josef Brüderl und Stefani Scherer

Zusammenfassung: Dieser Artikel gibt einen Überblick über Methoden zur Analyse von Sequenzdaten. Er behandelt die Darstellung von Lebensverläufen als Sequenzen, Optimal-Matching-Analyse und soziologische Anwendungen der Sequenzdatenanalyse. Zwei exemplarische Anwendungen zum Übergang von der Schule in den Beruf und zur Pluralisierung partnerschaftlicher Lebensformen geben einen Einblick in das Analysepotenzial dieser Verfahren. Zum Schluss geht der Beitrag auf die Diskussion um das Optimal-Matching Verfahren ein.

I. Einleitung

Eine Sequenz ist eine geordnete Liste von Elementen. Die klassische Sequenz ist die DNS, in der in einer Folge von nur vier Basen (A, C, G und T) die Erbinformation aller Lebewesen festgehalten ist. Spätestens seit das Humangenomprojekt in den 1990er Jahren mit der Entzifferung der genauen Sequenz der drei Milliarden Basen der menschlichen DNS begann, ist die DNS der Inbegriff einer Sequenz (vgl. den Titel des Buches von Kevin Davies 2001: „Die Sequenz"). Aber auch Texte und Computerprogramme sind Sequenzen, denn hier sind Buchstaben entlang der Textzeile angeordnet. Ein wesentliches Problem – unabhängig von der spezifischen Fragestellung – ist der Vergleich von Sequenzen, um deren Ähnlichkeit zu bestimmen. Molekularbiologen, Linguisten und Informatiker entwickeln seit Jahrzehnten Verfahren, die den Vergleich von Sequenzen ermöglichen (klassisch: Sankoff und Kruskal 1983). Linguisten setzen sie zur Spracherkennung und Informatiker zur Mustererkennung ein. Dem Biologen ermöglichen sie es, die Ähnlichkeit der DNS verschiedener Lebewesen zu bestimmen und darauf aufbauend Stammbäume zu erstellen. Ohne diese Verfahren wäre das Humangenomprojekt nicht möglich gewesen.

Seit etwa 15 Jahren werden einige dieser Verfahren – insbesondere „Optimal-Matching" (OM) – auch in der soziologischen Forschung eingesetzt. Andrew Abbott hat den Transfer von OM in die Soziologie entscheidend angeregt. Er wies darauf hin, dass Lebensverläufe als Sequenzen konzipiert und mit OM analysiert werden können.

* Für hilfreiche Kommentare danken wir Andreas Diekmann, Irena Kogan und einem anonymen Gutachter.

OM erlaubt die Bestimmung der Ähnlichkeit von Lebensverläufen und darauf aufbauend die Typisierung derselben. Den „Durchbruch" brachte eine interessante Anwendung auf die Lebensverläufe von deutschen Musikern des 18. Jahrhunderts (Abbott und Hrycak 1990).

Neben OM gibt es noch einfache deskriptive Verfahren, mit denen man Lebensverläufe beschreiben kann. Optimal-Matching und die deskriptiven Verfahren fassen wir hier unter dem Begriff „Sequenzdatenanalyse" zusammen. Inzwischen gibt es bereits einige Übersichtsartikel zu OM (z.B. Abbott und Tsay 2000; Billari 2001; ein kurzes Lehrbuch ist Taris 2000). Auch einige deutschsprachige Einführungen existieren (Erzberger und Prein 1997; Aisenbrey 2000). Der vorliegende Artikel gibt eine Einführung in die Sequenzdatenanalyse und betont dabei insbesondere anwendungsbezogene Aspekte.

II. Lebensverläufe als Sequenzen

Ein Lebensverlauf lässt sich als Abfolge von typischen Lebensphasen darstellen. Aus analytischen Gründen unterscheidet man meist nach Lebensbereich: Bildungsverlauf, Erwerbsverlauf, Partnerschaftsverlauf usw. Einen Wechsel der Lebensform bezeichnet man als „Ereignis" oder „Übergang". Das Ereignis „Heirat" etwa bewirkt den Übergang von der Lebensform „Ledig" in die Lebensform „Verheiratet". Lebensverläufe werden inzwischen überwiegend mit den Verfahren der Ereignisdatenanalyse untersucht. Dabei sucht man nach den bedingenden Faktoren der einzelnen Ereignisse des Lebensverlaufs. Dies ist eine klassisch regressionsanalytische Herangehensweise. Allerdings im Vergleich mit der klassischen Regression mit dem Unterschied, dass man die zeitliche Ordnung der Ereignisse kennt. (Kausal-)Schlüsse sind deshalb mit höherer Sicherheit möglich.

Jedoch „zerlegt" die Ereignisdatenanalyse den Lebensverlauf in einzelne Ereignisse. Damit verliert sie den Lebensverlauf als Ganzes aus dem Auge. „Holistische" Fragestellungen an Lebensverläufe kann sie nicht beantworten. Solche Fragestellungen sind in der Soziologie jedoch häufig anzutreffen: Welche Lebensverlaufstypen gibt es? Wie homogen sind die Lebensverläufe einer Gruppe? Wie wirken sich der historische und soziale Kontext auf Lebensverläufe als Ganzes aus? Auf diese Fragen kann die Sequenzdatenanalyse eine Antwort liefern, weil sie die Lebensverläufe insgesamt in den Blick nimmt.

Der erste Schritt einer Sequenzdatenanalyse ist die Darstellung der Lebensverläufe als Sequenzen. Dazu müssen die Zeitachse und die möglichen Lebensformen (technisch: der Zustandsraum) definiert werden. Die unterschiedenen Lebensformen bilden dabei die Elemente der Sequenz, deren Ordnung durch die Zeitachse (z.B. das Lebensalter) definiert ist. Je mehr Zustände unterschieden werden, umso komplexer werden die Sequenzen. Bei der Definition des (diskreten) *Zustandsraumes* sollte man daher nur so differenziert vorgehen wie nötig, denn je mehr Lebensformen man unterscheidet, desto schwieriger wird die Interpretation. Leitlinie sollte sein, nur die Lebensformen zu unterscheiden, die auch von den betroffenen Individuen als unterschiedlich wahrgenommen werden. Für die Definition der *Zeitachse* muss zunächst ein Beobachtungsfenster, d.h. der Beobachtungsbeginn und das Ende festgelegt werden. Aus Gründen

der Vergleichbarkeit empfiehlt es sich, das Beobachtungsfenster zu standardisieren, wodurch die Sequenzen auf dieselbe Länge gebracht werden.[1] Die Zeitachse der Sequenz kann dabei unterschiedlich definiert werden: etwa über das Alter der Befragten (wenn man sie z.B. ab ihrem 18. Lebensjahr beobachtet) oder über das Eintreten eines bestimmten Ereignisses (wie z.B. dem Eintritt in den Arbeitsmarkt). Die Zeitachse wird in diskrete Zeiteinheiten unterteilt. Ausschlaggebend ist hierfür die Genauigkeit, mit der die Ereignisse gemessen wurden. Hat man nur Jahresangaben erhoben, so wird man ein Jahr als Zeiteinheit festlegen; hat man monatsgenau gemessen, so wird man in Monatsschritten arbeiten. Dabei ist zu beachten, dass mit kleineren Zeiteinheiten die Länge der Sequenzen zunimmt und die erforderliche Rechenzeit mit dem Quadrat der Länge steigt.

Nach diesen Definitionsschritten kann man aus den üblicherweise biographisch abgelegten Lebensverlaufsdaten (1. Job, 2. Job usw.) die Sequenzen bilden. In den festgelegten Zeitschritten wird für jeden Zeitpunkt der entsprechende Zustand, in dem sich ein Individuum befindet, festgehalten.[2] Untersucht man beispielsweise die Erwerbsverläufe über die ersten fünf Jahre im Arbeitsmarkt in Monatsschritten, so resultieren Sequenzen der Länge 60. Eine weitere Entscheidung ist zu treffen, wenn sich Individuen gleichzeitig in mehreren Zuständen befinden können. Dann muss man entweder festlegen, welcher Zustand dominiert (etwa Vollzeiterwerbstätigkeit über eine geringfügige Beschäftigung), oder man erweitert den Zustandsraum um Doppelzustände.[3]

III. Sequenzdatenanalyse

Sind die Lebensverläufe in Sequenzform gebracht, so hat man vielfältige Möglichkeiten, die Lebensverläufe unter verschiedenen Aspekten zu beschreiben (Beispiele findet man bei Brückner und Rohwer 1996; Rohwer und Trappe 1997; Brüderl und Klein 2003). Grundsätzlich sind zwei Herangehensweisen möglich.

1. Man kann die Sequenzen aggregieren und die Zustandsverteilung über die Zeit bestimmen. Damit erhält man eine Makro-Beschreibung der Lebensverläufe, sozusa-

1 Die Verfahren der Sequenzdatenanalyse sind auch bei ungleich langen Sequenzen anwendbar, allerdings beeinflusst die Sequenzlänge das Ergebnis (Sequenzen unterschiedlicher Länge sind sich ceteris paribus unähnlicher). Ungleich lange Lebensverläufe resultieren aus (Rechts-)Zensierung, d.h. unterschiedlich langer Beobachtungsdauer. Die Probleme mit Zensierung vermeidet man, wenn man alle Sequenzen auf gleiche Länge bringt, d.h. künstlich zensiert.

2 Man beachte, dass durch dieses Vorgehen die Dauer der Zustände in den Sequenzen abgebildet wird. Man könnte die Dauer ignorieren und nur die Abfolge der Zustände festhalten. Dies wurde früher häufig so gemacht, um die Analyse einfach zu halten (z.B. Rindfuss et al. 1987; Berger et al. 1993). Mit diesem Vorgehen verliert man aber viel Information, und es ist angesichts der Verfügbarkeit von OM auch nicht mehr nötig.

3 Bei fünf Zuständen und einer Länge von 60 gibt es 5^{60} verschiedene Sequenzen. Reale Lebensverläufe schöpfen diese Vielfalt zwar bei weitem nicht aus, weil Menschen ihre Lebensform nicht jeden Monat wechseln. Dennoch werden insbesondere die deskriptiven Auswertungen schnell unübersichtlich, wenn man viele Zustände zulässt. Praktikabel scheinen Zustandsräume bis etwa 15. OM arbeitet aber auch mit umfangreichen Zustandsräumen problemlos. Abbott und Hrycak (1990) etwa unterscheiden in ihrer Anwendung 35 Zustände. Die Folge ist, dass ihre Karrieretypologie 20 Typen umfasst – eine nicht leicht zu interpretierende Fülle.

gen die „Sozialstruktur“ der Lebensverläufe. Macht man dies für verschiedene Teilgruppen (z.B. Männer-Frauen, Kohorten oder auch verschiedene Länder), so kann man die aggregierten Zustandsverteilungen vergleichen (z.B. mit dem Dissimilaritätsindex).

2. Man kann für die individuellen Sequenzen Kennzahlen berechnen und dann die Verteilung dieser Kennzahlen betrachten. Beispiele für solche Kennzahlen sind die mittlere verbrachte Zeit in den Zuständen und die Zahl der Ereignisse. Auch hier kann man die Verteilungen der Kennzahlen für verschiedene Teilgruppen vergleichen.

Komplizierter wird es, wenn man die Lebensverläufe typisieren will. Das Standardverfahren zur Typenbildung ist die Clusteranalyse, die wiederum auf einer Distanzmatrix der Untersuchungseinheiten aufbaut. Eine Distanzmatrix gibt für jedes Paar der Untersuchungseinheiten (in unserem Fall der Lebensverläufe) deren „Unähnlichkeit“ an. Distanzmaße haben üblicherweise ein Minimum von Null (das Paar ist identisch), sind ratioskaliert und symmetrisch. Man könnte nun daran denken, ein in der clusteranalytischen Literatur verbreitetes Distanzmaß für nominale Daten zu verwenden (z.B. Martens 1991; oder man metrisiert die Zustände und verwendet dann die euklidische Distanz, s. Buchmann und Sacchi 1995). Im Prinzip vergleicht man damit die Sequenzen elementweise. Die Distanz ist dann die Zahl der unterschiedlich besetzten Positionen. Dieses Vorgehen ist bei Sequenzen aber nicht sinnvoll, wie folgendes Beispiel zeigt:

	Vergleich 1	Vergleich 2	Vergleich 3
Sequenz 1	AAABBB	AAABBB	GBCDEF
Sequenz 2	CDEFGH	BBBAAA	BCDEFA
Naive Distanz	6	6	6
Levenshtein-Distanz (Subst=1, Indel=0,5)	6	3	1

Man vergleicht jeweils Sequenzen der Länge 6. Es sind 8 Zustände möglich. Bei allen drei Vergleichen unterscheiden sich alle Elemente. Deshalb ist die „naive“ Distanz bei allen drei Vergleichen gleich 6. Die naive Distanzberechnung entspricht einem Substitutionsalgorithmus: Die Distanz ist die Zahl der Substitutionen, die man braucht, um Sequenz 1 in Sequenz 2 überzuführen (oder umgekehrt wegen der Symmetrie).[4] Beim ersten Vergleich ersetzt man das erste A durch ein C, das zweite A durch ein D usw. Es sind jeweils 6 Substitutionen nötig.

Dies erscheint aber nicht sinnvoll, denn die „Unähnlichkeit“ der Sequenzen unterscheidet sich. Während das Paar in Vergleich 1 komplett unterschiedlich ist, sind die Sequenzen der anderen beiden Vergleiche mehr oder weniger „verschoben“. Im Vergleich 3 etwa ist die Subsequenz BCDEF in beiden Sequenzen zu finden, nur um eine Stelle verschoben. Wären dies Erwerbsverläufe, so würden beiden Personen die Berufskarriere BCDEF durchlaufen. Person 1 hätte davor vielleicht noch gejobbt (Zustand G). Es wäre offensichtlich nicht sinnvoll, diese beiden Erwerbsverläufe als maximal unterschiedlich zu klassifizieren.

4 Das ist der grundsätzliche Ansatz aller Distanzmaße bei Sequenzen: Wie viele Operationen sind nötig, um eine Sequenz in die andere zu überführen (um ein so genanntes „Alignment“ zu erreichen)?

Dies lässt sich beheben, wenn man zwei weitere Operationen zulässt: Einfügen und Löschen („insertion“ und „deletion“ oder kurz: Indel). Es sind dann nur zwei Operationen nötig, um im Vergleich 3 ein Alignment zu erreichen: Lösche bei Sequenz 1 das G und füge hinten ein A an. Eine Substitution kann man durch zwei Indel-Operationen ersetzen (füge den neuen Zustand ein und lösche den alten). Deshalb wird man Indel-Operationen nur mit dem Gewicht 0,5 versehen. Die Summe der so gewichteten Operationen, die mindestens für ein Alignment nötig sind, wird nach ihrem Erfinder Levenshtein-Distanz genannt. Bei Vergleich 1 bleibt es bei der maximalen Distanz: Es sind 6 Substitutionen nötig (oder 6mal Einfügen und 6mal Löschen). Bei Vergleich 2 löscht man vorne dreimal A und fügt es hinten dreimal ein. Die Levenshtein-Distanz ist 3. Bei Vergleich 3 sind zwei Indel-Operationen erforderlich, die Distanz ist also 1. Diese Art der Distanzberechnung erscheint wesentlich sinnvoller.

Bei kurzen Sequenzen (wie in unserem Beispiel) ist es recht einfach, die minimale Zahl der nötigen Operationen zu bestimmen. Bei längeren Sequenzen wird das aber sehr schwierig. Ein Algorithmus ist nötig. OM besteht im Kern aus einem Algorithmus, der auch bei langen Sequenzen die Bestimmung der Levenshtein-Distanz leistet. Eine ausführliche Darstellung dieses Minimierungs-Algorithmus findet man bei Abbott und Hrycak (1990).

Dieses Grundprinzip kann man flexibilisieren, indem man die Gewichte der Operationen variiert. Zunächst ist es üblich, eine Substitution mit dem Gewicht 2 und eine Indel-Operation mit dem Gewicht 1 zu versehen. Dies ändert nichts, sondern verdoppelt nur die Metrik. Weiterhin kann man die Substitutionsgewichte unterschiedlich wählen, je nach den „Kosten“ der Ersetzung. Ersetzt man einen Zustand durch einen anderen sehr ähnlichen, so kann man die Substitutionskosten niedriger als 2 ansetzen. Dadurch wird die Distanz bei Sequenzen, die aus unterschiedlichen, aber sehr ähnlichen Zuständen bestehen, kleiner. Will man von dieser Flexibilität Gebrauch machen, muss man eine (symmetrische) Substitutionskostenmatrix festlegen. Das Problem ist allerdings, dass man eine Theorie braucht, die sagt, welche Zustände sich in welchem Ausmaß ähnlich sind. Die hat man meist aber nicht. Eine „clevere“ Lösung bietet TDA („Transition Data Analysis“, ein weit verbreitetes Programm zur Ereignis- und Sequenzdatenanalyse) an, das die Substitutionskostenmatrix automatisch aus den Daten errechnet. TDA berechnet die in den Daten beobachtete Übergangswahrscheinlichkeit zwischen den beiden Zuständen und zieht sie von dem Substitutionsgewicht 2 ab. Ein mustergültiges Beispiel dafür, wie man anhand theoretischer Überlegungen und der Übergangswahrscheinlichkeiten zu einer Substitutionskostenmatrix gelangt, findet man bei Stovel und Bolan (2004).[5]

5 Einige weitere Details für Experten: Indel-Kosten = 1 und Substitutionskosten = 2 ist die Voreinstellung in TDA. Substitutionskosten ≥ 2 haben zur Folge, dass keine Substitution mehr durchgeführt wird, weil eine Substitution billiger durch Einfügen und Löschen erreichbar ist. In diesem Fall ist die Distanz zweier Sequenzen gegeben durch: (Länge der Sequenz – Länge der längsten gemeinsamen Subsequenz) · 2. Will man dies vermeiden, so sollte man (zumindest einige) Substitutionskosten < 2 ansetzen. Andererseits sollte man nicht dem Beispiel von Abbott und Hrycak (1990) folgen und die Indel-Kosten höher als die Substitutionskosten ansetzen. Dadurch kann man leicht die Vorteile der Zulassung der Indel-Operationen zunichte machen (s. Abbott und Tsay 2000). Bei Indel-Kosten von 1 sollten sich also die Substitutionskosten zwischen 1 und 2 bewegen.

Mittels OM berechnet man für jedes Sequenzpaar die Levenshtein-Distanz (wegen Symmetrie n · (n – 1) / 2 Distanzen). Die so gewonnene Distanzmatrix kann man dann als Input für Klassifizierungsverfahren – insbesondere Clusteranalyse – nutzen. Praktische Erfahrung hat gezeigt, dass das Ward-Verfahren die besten Ergebnisse erbringt. Das Ward-Verfahren erzeugt aus einer Levenshtein-Distanzmatrix homogene Cluster, während andere Verfahren eher ungleich große Cluster produzieren (zu den verschiedenen Verfahren der Clusteranalyse s. Bacher 1996). Ein Problem der Clusteranalyse ist die Bestimmung der „optimalen" Anzahl der Cluster, weil es hierfür kein eindeutiges Kriterium gibt. Einen Anhaltspunkt liefert das Elbow-Kriterium.[6] Aber die endgültige Entscheidung wird man inhaltlich treffen müssen: Ziel sind in sich möglichst homogene Cluster, die sich möglichst deutlich unterscheiden. Die „Interpretierbarkeit" der resultierenden Typologie der Lebensverläufe ist deshalb in den meisten Anwendungen das Hauptkriterium zur Festlegung der Clusterzahl.

Dies ist das typische zweistufige Vorgehen der meisten Anwendungen von OM. Zuerst berechnet man die Matrix der Levenshtein-Distanzen, dann klassifiziert man die Lebensverläufe mittels Clusteranalyse. Die zweite Stufe – also die Clusteranalyse – ist allerdings nicht für alle Fragestellungen nötig. Manchmal gilt das Interesse den Distanzen an sich. Eine Fragestellung, die nur Stufe eins erfordert, ist der Vergleich von Lebensverläufen mit einem vorgegebenen Referenzlebensverlauf. Man kann etwa jeden Lebensverlauf eines Datensatzes mit einer kulturell definierten „Normalbiographie" vergleichen. Die Distanz zur Normalbiographie kann man dann zur abhängigen Variable einer Regression machen. Damit kann man der Frage nachgehen, welche Gruppen besonders stark von der Normalbiographie abweichen. Man beachte, dass bei dieser Fragestellung der Rechenaufwand deutlich geringer ist, da man für jede Sequenz nur eine Distanz berechnet (insgesamt also nur n Distanzen).

Einen Überblick über die Möglichkeiten der Sequenzdatenanalyse gibt die folgende Darstellung:

(1) Aggregation der Sequenzen
- Aggregierte Zustandsverteilung
  Beispiele: Verteilung der Erwerbszustände 1, 2, ... 60 Monate nach Verlassen der Schule
  Verteilung der Familienform im Alter von 15, ... 35 Jahren

(2) Kennzahlen für die individuellen Sequenzen
- Verweildauer in einzelnen Zuständen
  Beispiel: kumulierte Arbeitslosigkeitsdauer (in Monaten)
- Anzahl der Zustände und der Zustandswechsel (Ereignisse)
  Beispiel: Anzahl beruflicher Wechsel

Weitere Flexibilisierung ist möglich (s. Sankoff und Kruskal 1983): Man könnte die Indel-Kosten mit dem eingefügten bzw. gelöschten Zustand variieren lassen, weiterhin könnte man die Kosten von den in einer Sequenz vorangegangenen Zuständen abhängig machen usw. Diese Komplikationen sind allerdings in TDA nicht implementiert.

6 Man erstellt einen Plot, in dem man die Distanz der fusionierenden Cluster für die letzten Fusionsschritte abträgt. Dort, wo die resultierende Kurve einen „Knick" (Ellbogen) zeigt, ist die „optimale" Clusterzahl.

(3) Levenshtein-Distanz (OM)
- Vergleich zu einer Referenzsequenz
  Beispiel: Abweichung von durchgängiger Vollzeiterwerbstätigkeit
- Paarweiser Vergleich → Distanzmatrix
  (als Ausgangsbasis für Klassifizierungsverfahren)
  Beispiel: Bildung von Lebensverlaufstypen

IV. Soziologische Anwendungen der Sequenzdatenanalyse

In diesem Abschnitt werden wir anhand einer knappen Übersicht einen Einblick in die Fragestellungen und das Analysepotenzial der Sequenzdatenanalyse geben. Anschließend werden unsere Überlegungen durch zwei exemplarische Anwendungen konkretisiert (zu weiteren Anwendungen siehe Abbott und Tsay 2000; Aisenbrey 2000). Im Anhang werden einige Hinweise auf die programmtechnische Umsetzung einer Sequenzdatenanalyse gegeben.

1. Fragestellungen und Analysepotenzial der Sequenzdatenanalyse

Die meisten bisherigen soziologischen Anwendungen setzen OM zur Typisierung von Erwerbsverläufen ein. Es seien nur zwei Beispiele angeführt: Die klassische Anwendung (Abbott und Hrycak 1990) identifizierte anhand der Berufsverläufe deutscher Musiker des 18. Jahrhunderts 20 typische Karriereverläufe. Halpin und Chan (1998) transformieren das Berufsschema in ein Klassenschema und können so typische Muster von „Klassenkarrieren" identifizieren. Anwendungen auf andere Arten von Lebensverläufen sind selten. Mowitz-Lambert (2001) arbeitet die Muster des Übergangs von einer beruflichen Ausbildung in die Erwerbstätigkeit heraus. Eine Anwendung auf Partnerschaftsverläufe wird unten vorgestellt. Typische Muster räumlicher Mobilität untersuchen Stovel und Bolan (2004).

Hat man mittels OM und Clusteranalyse Lebensverlaufstypen identifiziert, kann man weitere Analysen anschließen: Die Lebensverlaufstypen können sowohl als abhängige wie auch als unabhängige Variable verwendet werden. Beispielsweise ist es möglich, ein multinomiales logistisches Regressionsmodell auf die Lebensverlaufstypen zu berechnen. Damit kann der Frage nach den Bedingungen nachgegangen werden, unter denen bestimmte Lebensverlaufstypen auftreten (z.B. in Abhängigkeit vom Geschlecht, der Geburtskohorte, der Bildung usw.). Umgekehrt kann man auch fragen, welche Auswirkungen bestimmte Lebensverläufe auf zukünftiges Geschehen haben (z.B. auf den weiteren Lebensverlauf, die Gesundheit usw.).

Dieses Analysepotenzial wurde aber bisher nur in wenigen Studien genutzt. Einige Ausnahmen: Han und Moen (2001) untersuchen mittels log-linearer Modelle, wie der Berufsverlauf mit der Erwerbsgeschichte des (Ehe-)Partners zusammenhängt. Kogan (2003) geht mit einer multinomialen logistischen Regression der Frage nach, ob sich die Berufsverläufe der Immigranten von denen der Deutschen unterscheiden. Windzio und Grotheer (2002) untersuchen, wie sich der bisherige Berufsverlauf auf die Bereit-

schaft zur Teilnahme an einer Panelbefragung auswirkt. In allen drei Studien werden mittels OM und Clusteranalyse gewonnene Berufsverlaufstypen verwendet.

Schließlich sei ergänzend erwähnt, dass Sequenzdatenanalyse natürlich nicht nur im Kontext der Lebensverlaufsforschung einsetzbar ist. Sie kann auf alle Daten angewendet werden, die die Abfolge von Ereignissen festhalten. Insbesondere Handlungsmuster sind einschlägig. Es existieren Anwendungen, die mittels OM den typischen Ablauf von Riten, Tänzen und politischen Prozessen untersuchen. Eine besonders ausgefallene Anwendung untersucht historische Muster des Lynchens im Süden der USA (Stovel 2001). Auch die Abfolge rhetorischer Elemente in soziologischen Artikeln war bereits Gegenstand einer Sequenzdatenanalyse (Abbott und Barman 1997). Eine weitere interessante Anwendung ist die Klassifikation von Tagesabläufen anhand von Zeitbudgetdaten (Hertkorn et al. 2001).

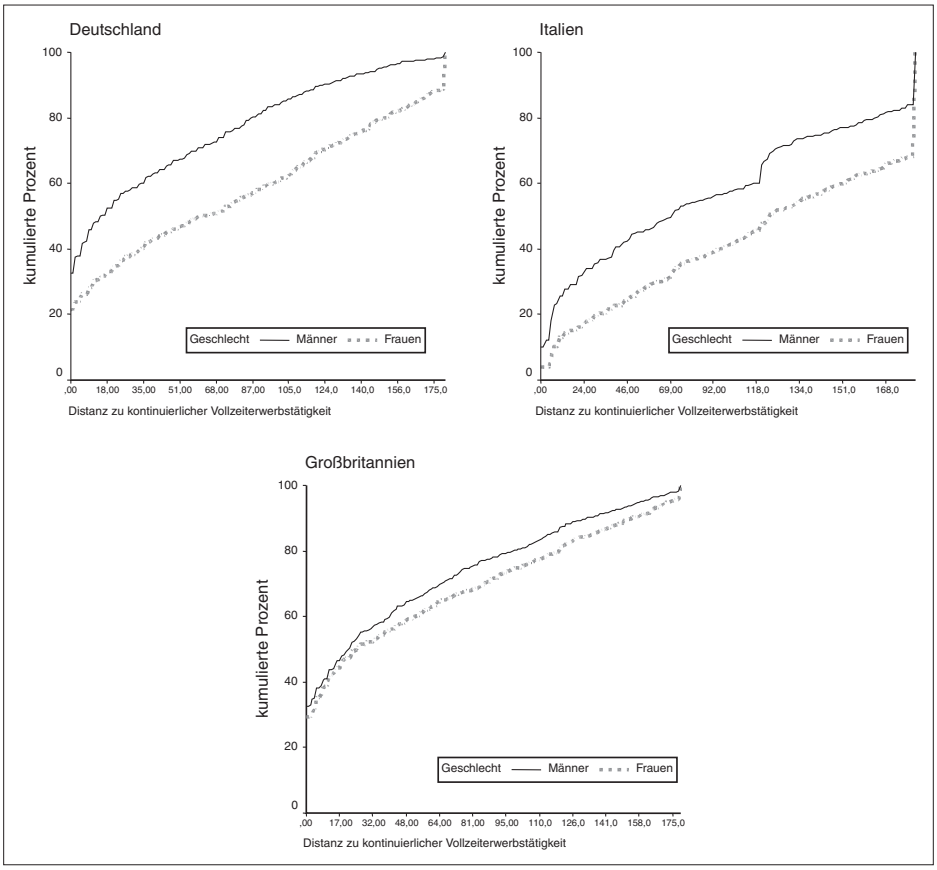
2. Anwendung I: Der Übergang von der Schule in den Beruf

Der Übergang von der Schule in den Beruf ist eine entscheidende Statuspassage. Infolge von Pfadabhängigkeiten kommt diesem Übergang für den weiteren Lebensverlauf eine große Bedeutung zu. In einer internationalen Perspektive variiert die Bedeutung des Erwerbseintritts mit der Rigidität des Arbeitsmarktes, denn auf sehr unflexiblen Arbeitsmärkten ist so etwas wie „Job-Shopping" nicht möglich, man ist durch den ersten Job bereits (lebenslang) festgelegt. Insofern sollten die Schulabgänger sorgfältig und lange suchen. Italien ist so ein Fall. Das andere Extrem ist Großbritannien, wo auf Grund hoher Arbeitsmarktflexibilität lange Jobsuche nicht lohnt. Die Schulabgänger sollten sofort in das Erwerbsleben einsteigen, denn sie können ja später den Job wechseln. Deutschland sollte irgendwo zwischen diesen beiden Extremen liegen.

Dieser Frage geht Scherer (2001, 2004) unter Verwendung der Daten des Sozio-oekonomischen Panels (SOEP, N=1.997, nur Westdeutschland), der Indagine Longitudinale sulle Famiglie Italiane (ILFI, N=1.326) und des British Household Panel Survey (BHPS, N=1.674) nach. Für alle Schulabgänger wird für die ersten 60 Monate eine Erwerbssequenz konstruiert (Zustände: Vollzeit erwerbstätig, Teilzeit erwerbstätig, selbstständig, in Bildung, arbeitslos, nicht im Arbeitsmarkt). Nach dem oben Gesagten sollte man für Italiener lange Phasen der Arbeitslosigkeit (wegen Jobsuche) erwarten, während die Briten schnell erwerbstätig sein sollten. Die Deutschen sollten in der Mitte liegen. Dies kann man überprüfen, indem man jede Sequenz mit der Referenzsequenz „kontinuierliche Vollzeiterwerbstätigkeit" vergleicht und die Levenshtein-Distanz errechnet. In Italien sollten die Lebensverläufe am stärksten von der Referenzsequenz abweichen.

Dass dem tatsächlich so ist, erkennt man in *Abbildung 1*. Dargestellt sind für die drei Länder die kumulierten Häufigkeitsverteilungen der Distanzen getrennt nach Geschlecht. Der Wertebereich der Levenshtein-Distanz ist in dieser Anwendung 0 bis 180. Eine kontinuierlich vollzeiterwerbstätige Person unterscheidet sich nicht von der Referenzsequenz und hat insofern eine Distanz von 0. Dagegen erreicht eine Person, die 60 Monate nicht im Arbeitsmarkt war, die maximal mögliche Distanz von 180, da die Substitutionskosten „Inaktivität-Erwerbstätigkeit" auf 3 festgelegt wurden (60 × 3).

Abbildung 1: Kumulierte Verteilungen der Levenshtein-Distanzen zur kontinuierlichen Vollzeiterwerbstätigkeit

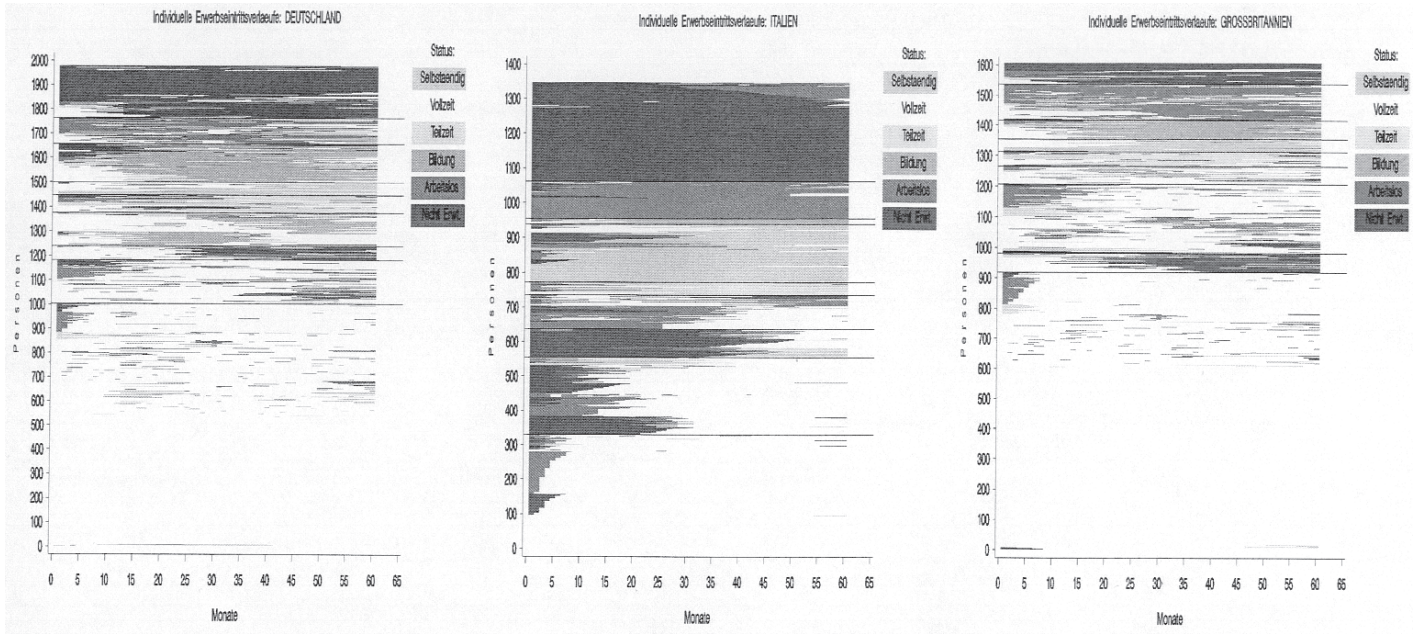


Quelle: SOEP, ILFI, BHPS; eigene Berechnungen.

Vergleicht man zuerst die Verteilungen der Männer, so fällt auf, dass sich Großbritannien und Deutschland sehr ähnlich sind. In beiden Ländern sind etwa 30 Prozent der Männer kontinuierlich vollzeiterwerbstätig (Distanz 0). Praktisch niemand erreicht die maximale Distanz. Ganz anders in Italien: Hier sind nur 10 Prozent der Männer durchgehend und fast 20 Prozent nie erwerbstätig. Bezüglich der weiblichen Berufseintrittsverläufe zeigt sich, dass sie in allen drei Ländern größere Distanzen zur Referenzsequenz haben (die kumulierte Verteilung liegt niedriger). In Großbritannien allerdings sind die Unterschiede zu den Männern eher gering. In einem nächsten Schritt können die Distanzen zur Referenzsequenz „kontinuierliche Vollzeiterwerbstätigkeit" als abhängige Variable in einer Regressionsgleichung berücksichtigt werden (s. Scherer 2004: 104).

Als nächstes sollen typische Eintrittsmuster bestimmt werden. Dazu werden für jedes Land mittels OM die paarweisen Distanzen berechnet und anschließend mit dem Ward-Verfahren einer Clusteranalyse unterzogen. In jedem Land erwies sich eine Lö-

Abbildung 2: Individuelle Erwerbseintrittsverläufe nach Clusterzugehörigkeit



Quelle: SOEP, ILFI, BHPS; eigene Berechnungen. Eine Farbversion ist zu finden unter http://www.sowi.uni-mannheim.de/lehrstuehle/lessm/papers.htm.

sung mit 9 Clustern als sinnvoll. Ein entscheidender Punkt ist die informative Beschreibung der Cluster. Scherer (2001; s.a. Kogan 2003) hat hierfür ein einfaches graphisches Instrument vorgeschlagen. Jede Sequenz wird als horizontale Linie gezeichnet, wobei die Zustände farblich unterschieden werden. Die Linien werden nach der Clusterzugehörigkeit sortiert und übereinander gestapelt (Sequenz-Indexplot). Das Ergebnis ist in *Abbildung 2* festgehalten. Die Cluster sind durch die horizontalen schwarzen Linien getrennt. In einem solchen Sequenz-Indexplot ist eine Menge an Informationen enthalten, allerdings durch die vorangegangene OM-Analyse in einer bereits geordneten Form, so dass man die Information leicht verarbeiten kann. Nur einige Ablesebeispiele: Alle drei Länder haben ein Vollzeiterwerbstätigkeits-Cluster, das aber in Italien deutlich schwächer besetzt ist. Dafür sticht Italien mit drei anderen Clustern hervor: Selbständige, Langzeitarbeitslose und Nichterwerbstätige. In Deutschland ist besonders auffällig, dass es viele Lebensverläufe gibt, die eine Rückkehr in das Bildungssystem aufweisen. Damit wird die implizite Annahme eines Großteils der soziologischen Mobilitätsforschung, dass das Bildungsniveau mit Verlassen des Schulsystems festgelegt ist, problematisiert.

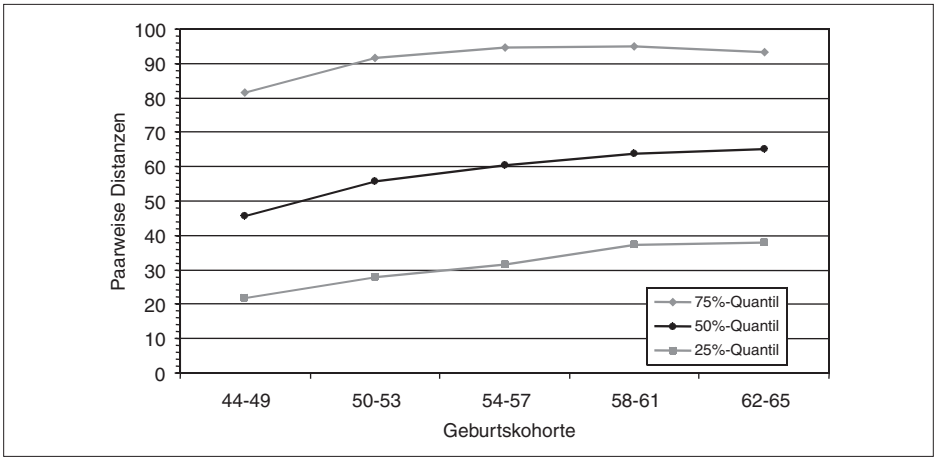
3. Anwendung II: Pluralisierung partnerschaftlicher Lebensformen?

In einer weiteren Anwendung beschäftigen wir uns mit der Frage, ob es in den letzten Jahrzehnten zu einer Pluralisierung partnerschaftlicher Lebensformen gekommen ist (Brüderl und Klein 2003; Brüderl 2004). Pluralisierung wird hier verstanden als zunehmende Heterogenität (qualitative Varianz) von Lebensverläufen. Diese klassisch soziologische Frage lässt sich mit herkömmlichen Varianzmaßen nicht beantworten.

Die Analyse basiert auf den Partnerschaftsbiographien des Familiensurveys 2000. Folgende partnerschaftlichen Lebensformen können mit diesen Daten unterschieden werden: ledig (partnerlos vor einer ersten Ehe), nichteheliche Lebensgemeinschaft (vor einer ersten Ehe; NEL), verheiratet (erste Ehe), getrennt (partnerlos nach einer Ehe, die durch Trennung/Scheidung beendet wurde), verwitwet (partnerlos nach einer Ehe, die durch Tod des Ehepartners beendet wurde), nacheheliche NEL, wiederverheiratet. Aus den Partnerschaftsbiographien aller Westdeutschen wurde jeweils eine Partnerschaftssequenz in der Altersspanne von 14 bis 35 Jahren erzeugt (N=2.988). Die Zeiteinheit ist 3 Monate, so dass die Sequenzen die Länge 85 haben (14,00; 14,25; 14,50; ...; 35,00). Die Indel-Kosten werden auf 1, die Substitutionskosten auf 2 minus Übergangswahrscheinlichkeit festgelegt. Damit hat die Levenshtein-Distanz einen Wertebereich von 0 bis 170 ($85 \cdot 2$).

Wie kann man messen, ob es zu einer Zunahme der Vielfalt der Lebensverläufe gekommen ist? Das ist gar nicht so einfach, denn es ist ziemlich unklar, was mit „zunehmender Vielfalt der Lebensverläufe“ genau gemeint ist. Das Problem ist, dass für Lebensverläufe kein einfaches Varianzmaß definiert ist. Einen ersten möglichen Ansatzpunkt sehen wir darin, die kohorteninterne Heterogenität der Partnerschaftsverläufe zu berechnen. Dazu vergleichen wir alle Partnerschaftssequenzen einer Geburtskohorte und berechnen für jedes Paar die Levenshtein-Distanz. Das Mittel der paarweisen Distanzen einer Kohorte ist dann eine Kennzahl der Heterogenität. Pluralisierung würde

Abbildung 3: Quartile der paarweisen Levenshtein-Distanzen innerhalb der fünf Geburtskohorten



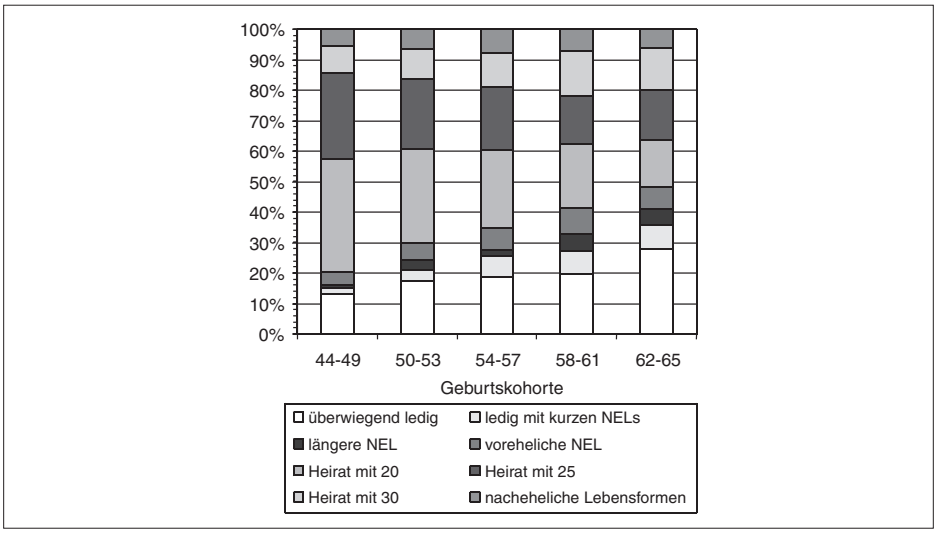
Quelle: Familiensurvey 2000; eigene Berechnungen.

sich dann in einem Anstieg dieser mittleren Distanz über die Kohorten manifestieren. In *Abbildung 3* sind die Quartile der paarweisen Distanzen für fünf Geburtskohorten abgetragen. Man erkennt, dass alle Quartile monoton ansteigen (Ausnahme: das 75%-Quantil der jüngsten Kohorte). Der Median steigt von 46 auf 65. Dies ist ein Beleg dafür, dass die Heterogenität der Partnerschaftssequenzen zugenommen hat.

Ein zweiter Ansatzpunkt ist die Bestimmung typischer Lebensverläufe. Im Anschluss wird die Varianz der Verteilung der Lebensverlaufstypen über die Kohorten verglichen. Basierend auf den Levenshtein-Distanzen wurde eine Clusteranalyse (Ward-Verfahren) gerechnet. Es erwies sich als sinnvoll (statistisch und inhaltlich), acht verschiedene Lebensverlaufstypen zu unterscheiden: Der erste Typ umfasst Personen, die bis 35 (überwiegend) ledig blieben (also auch keine NEL eingingen). Der zweite und dritte Typ sind Personen, die eine NEL eingingen, aber keine Ehe. Beim zweiten Typ sind dies kürzere Phasen von nichtehelichen Lebensgemeinschaften. Personen des dritten Typs dagegen weisen (meist) eine längere NEL-Phase auf. Als nächstes werden vier Ehetypen unterschieden. Typ vier hat vor der Eheschließung eine längere NEL-Phase. Die Typen fünf, sechs und sieben unterscheiden sich hauptsächlich durch den Zeitpunkt der Eheschließung: Heirat mit etwa 20, 25 bzw. 30. Schließlich gibt es noch einen achten Lebensverlaufstypus, bei dem eine Person bis 35 längere Phasen in nachehelichen Lebensformen verbringt.

In *Abbildung 4* sind die so bestimmten Verteilungen der Lebensverlaufstypen für die fünf Geburtskohorten abgetragen. Man erkennt deutliche Veränderungen in der Verteilung. Drastisch ist der Rückgang der drei „reinen" Ehetypen (5, 6, 7) von 75 Prozent in der ältesten auf 45 Prozent in der jüngsten Kohorte. Zählt man den Typ „voreheliche NEL" – der zugenommen hat – hinzu, so fallen die Ehetypen immer noch von 79 auf 52 Prozent. Zugenommen haben dagegen die beiden NEL-Typen: von 3 auf 13 Prozent. Dieser Anstieg reicht allerdings nicht aus, um den Rückgang der Ehetypen zu kompensieren. Folglich stieg der Typus „ledig" von 13 auf 28 Prozent.

Abbildung 4: Verteilung der acht Lebensverlaufstypen



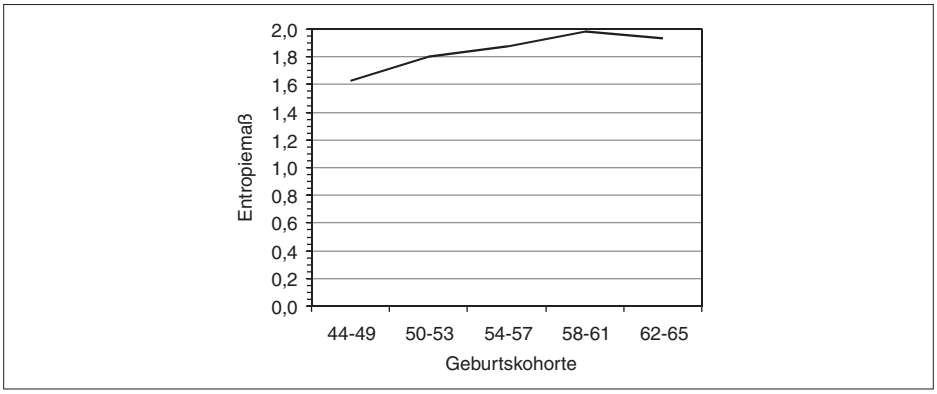
Quelle: Familiensurvey 2000; eigene Berechnungen.

Ist das aber nun Pluralisierung? Optisch wirken die Verteilungen „bunter" (insbesondere, wenn man die *Abbildung 4* farbig darstellt).[7] Aber ob die Heterogenität zugenommen hat, kann nur durch eine geeignete statistische Maßzahl festgestellt werden. Die „qualitative Varianz" dieser Verteilungen messen wir mit dem so genannten Entropiemaß. In *Abbildung 5* sind die Werte abgetragen. Die Varianz der Verteilungen steigt von 1,63 in der ältesten auf 1,98 in der zweitjüngsten Kohorte. Das Entropiemaß kann bei acht Typen maximal den Wert 2,08 erreichen (wenn jeder Typ gleich häufig vorkommt). Standardisiert man auf diesen Maximalwert, so beobachten wir einen Anstieg der Varianz von 0,78 auf 0,95. Der hohe Ausgangswert zeigt, dass es bereits in der ältesten Kohorte eine hohe Vielfalt der Lebensverlaufstypen gab. Es ist also eine Fehlwahrnehmung, wenn man älteren Kohorten eine völlige Homogenität unterstellt. Aber dennoch kam es zu einem weiteren Anstieg der Heterogenität – zu einer Pluralisierung. In der jüngsten Kohorte beobachten wir allerdings einen Rückgang der Entropie, sie fällt auf 1,93. Dieses Ergebnis mag unerwartet sein. Aber es ist erklärbar. Wie man in *Abbildung 4* erkennt, wird in der jüngsten Kohorte der Typus „ledig" dominant. Deshalb sinkt die Varianz, der Pluralisierungstrend kehrt sich um.[8]

7 Auf den folgenden Internet-Seiten der Zeitschrift findet man die farbigen Darstellungen der Abbildungen: <http://www.uni-koeln.de/kzfss/materialien/KS-57-4-Bruederl-scherer-abb.pdf>.

8 Ob dies eine Trendumkehr ist, bleibt allerdings noch unklar. Insbesondere die inferenzstatistische Absicherung kann nicht gegeben werden, weil die Clusteranalyse keine Inferenzstatistik kennt.

Abbildung 5: Varianz der Verteilungen der Lebensverlaufstypen



Quelle: Familiensurvey 2000; eigene Berechnungen.

V. Die Diskussion um Optimal-Matching

Die inzwischen vorliegenden vielfältigen Anwendungen und nicht zuletzt unsere oben angeführten Analysen haben das Potenzial der Sequenzdatenanalyse eindrücklich demonstriert. Dennoch bestehen bezüglich der Sequenzdatenanalyse eine Reihe von Kritikpunkten. Sie reichen von generellen Zweifeln an der Fruchtbarkeit der Sequenzdatenanalyse für die Soziologie bis zu umsetzungstechnischen Problemen.

Wu (2000) etwa ist vom Erkenntnisgewinn bisheriger Anwendungen nicht beeindruckt. Abbott gibt diesem pragmatischen Argument teilweise Recht, wenn er davon spricht, dass bisher noch eine „blockbuster application" fehlt (Abbott 2000: 75). Wir teilen diese Einschätzung nicht, denn die oben vorgestellten Anwendungen erweitern unser Wissen über Lebensverläufe deutlich.

Wus (2000) substanzieller Hauptkritikpunkt ist, dass OM zwar für die DNS-Analyse taugen mag, nicht aber für die Analyse von Lebensverläufen. Einfügen, Löschen und Substitution beschrieben in der Molekularbiologie mehr oder weniger reale Prozesse (Mutationen), hätten aber in Lebensverläufen keine realen Entsprechungen. Darüber hinaus ignoriere OM die Zeit – das zentrale Konzept der Lebensverlaufsanalyse –, weil ja DNS-Sequenzen keine zeitliche Ordnung haben. Das zugrunde liegende Argument ist also, OM bilde keine realen Lebensverlaufsprozesse ab. Dieser Argumentation kann entgegengehalten werden, dass kaum ein statistisches Verfahren reale Prozesse abbildet (Abbott 2000), es auch gar nicht muss. Die (Sozial-)Wissenschaft ist an Modellen der Wirklichkeit interessiert, die uns ihr Verständnis ermöglichen, nicht aber an ihrer exakten Abbildung. Statistische Verfahren dienen dabei einem Zweck (z.B. bedingte Verteilungen zu vergleichen, Sequenzen zu typisieren), und es ist die Sinnhaftigkeit und der Zielerreichungsgrad dieses Zwecks, nach dem sie beurteilt werden. Kaum jemand wird bestreiten, dass die Typisierung von Lebensverläufen sinnvoll ist. Und die vorliegenden Anwendungen zeigen, dass OM dieses Ziel erreicht.

Ein eher anwendungsbezogener Kritikpunkt bezieht sich auf die zu treffenden Annahmen bezüglich der Indel- und Substitutionskosten. Nun sehen die Kritiker zwar,

dass jede statistische Analyse auf Annahmen beruht, aber bei OM habe der Anwender keine klare Anweisung, wie er vorzugehen habe. Das ist aber ein Missverständnis: Die TDA-Voreinstellung beispielsweise stellt eine klare Anweisung dar. OM funktioniert mit dieser Voreinstellung hervorragend. Nur wenn der Anwender seine Analyse „verfeinern" will, muss er Abweichungen von den Standardannahmen vornehmen. Dies ist analog etwa zur Regressionsanalyse: Die meisten Anwender geben sich mit den Standardannahmen zufrieden, selbst wenn für ihre Analysen komplexere Annahmen sinnvoller wären. Darüber hinaus zeigen die bisherigen praktischen Erfahrungen mit OM und Simulationen (Abbott und Hrycak 1990: 164), dass die Ergebnisse von OM ziemlich robust gegenüber massiven Veränderungen der Indel- und Substitutionskosten sind. Dies gilt auch für Veränderungen des Zustandsraums. Wenn man den Zustandsraum beispielsweise verfeinert und den Subzuständen geringe Substitutionskosten zuweist, wird mittels OM die Ähnlichkeit der Sequenzen erkennbar. Ein wichtiger Punkt für die Replizierbarkeit von Optimal-Matching-Analysen ist allerdings die Dokumentation der Kostenmatrix.

Ein weiterer Kritikpunkt bezieht sich auf den zweiten Schritt einer typischen Sequenzdatenanalyse: die Clusteranalyse. Eine Clusteranalyse liefert selbst dann eine Klassifikation, wenn in den Daten keine sinnvolle Struktur besteht. Weiterhin gibt es kein eindeutiges Kriterium für die Auswahl der Clusteranzahl, d.h. welche Cluster-Lösung die beste ist. Und: verschiedene Clusteranalyseverfahren produzieren meist auch verschiedene Lösungen. Es ist deshalb umso wichtiger, die Entscheidungen im Rahmen der Clusteranalyse auch inhaltlich bzw. theoretisch zu untermauern. Der ausführlichste Teil der Ergebnispräsentation sollte deswegen der Cluster-Validierung gewidmet sein. Das Vertrauen in eine Clusterlösung steigt, wenn man homogene Cluster gefunden hat, die sich deutlich von den anderen Clustern unterscheiden. Das beste Instrument hierfür ist ein Sequenz-Indexplot (s. *Abbildung 2*). Ergänzend kann man für jeden Cluster die aggregierten Zustandsverteilungen und die mittlere Zeit in den Zuständen ausweisen. Damit hat man recht anschauliche Instrumente zur Validierung der vorgeschlagenen Cluster-Lösung zur Hand.

Zu guter Letzt gibt es eine Diskussion um den „Zielerreichungsgrad": Ist OM wirklich das optimale Verfahren für die Bestimmung der Ähnlichkeit von Lebensverläufen? Zwei alternative Verfahren wurden bisher vorgeschlagen. Dijkstra und Taris (1995) schlugen einen Algorithmus (SEQUENCE) vor, bei dem erst nicht gemeinsame Elemente und Wiederholungen gelöscht werden. Dann werden die so gekürzten Sequenzen durch Vertauschungen („transpositions") ineinander überführt. Die Distanz ergibt sich aus der Zahl der Löschungen und Vertauschungen. Dieser Algorithmus ist heftig umstritten (Abbott 1995; Driel und Oosterveld 2001; Elzinga 2003). Vielversprechender erscheint ein von Elzinga (2003) auf axiomatischer Basis entwickelter Algorithmus. Die Ähnlichkeit zweier Sequenzen bestimmt sich hier aus der Häufigkeit, mit der geordnete k-Tupel der Zustände in beiden Sequenzen auftreten.[9] Allerdings ist

9 Ein 1-Tupel ist einfach ein Zustand. Für k=1 wird also die Zahl der übereinstimmenden Zustände gezählt. Ein 2-Tupel ist die Anordnung von zwei Zuständen, z.B. AB. 2-Tupel gibt es m · m (m = Zahl der Zustände). Für k=2 zählt man, wie viele 2-Tupel in beiden Sequenzen auftreten (egal wie viele Elemente zwischen A und B liegen!). Und so weiter, bis k die Länge der kürzeren Sequenz erreicht. Der Elzinga-Algorithmus berücksichtigt somit insbesondere die Ordnung der Zustände (A vor B), aber nicht, in welchem zeitlichen Abstand sie auftreten.

der Elzinga-Algorithmus sehr rechenintensiv und für lange Sequenzen ungeeignet. Deshalb erscheint er insbesondere dann sinnvoll, wenn man sich nur für die Ordnung – nicht die zeitliche Dauer – der Zustände interessiert.

Neben den Kritikern gibt es natürlich auch die Verfechter dieser Methode. Andrew Abbott etwa sieht in der Sequenzdatenanalyse sogar einen Paradigmenwechsel verkörpert: von den Ursachen zu den Ereignissen („from causes to events", Abbott 1992). Damit meint er, dass sich die soziologische Forschung von der „Variablensoziologie" weg bewegt, hin zu einer Sichtweise, die Ereignisse und Handlungen in ihrer zeitlichen Einbettung betrachtet. Der Lebensverlauf wird nicht als eine Verkettung von Ursachen und Wirkungen konzipiert, sondern als eine Sequenz von Zuständen (und damit Ereignissen). Wie wir oben gesehen haben, entspricht dies der Herangehensweise der Sequenzdatenanalyse. Aber im Gegensatz zu Abbott werden die meisten Soziologen nicht bereit sein, die Frage nach den Ursachen aufzugeben. Hier werden auch die Grenzen der Sequenzdatenanalyse deutlich: Sie kann uns zwar verdeutlichen, wie Lebensverläufe typischerweise gestaltet sind, aber warum sie diese Form annehmen, darüber kann man mit ihr keine Aussage machen. Dazu muss man die einzelnen Ereignisse des Lebensverlaufs und die dahinter stehenden Entscheidungen analysieren (typischerweise mit Ereignisdatenanalyse). Nun wäre es aber übertrieben zu erwarten, eine Methode müsse alles leisten können. Wie wir gezeigt haben, können mit dem Instrumentarium der Sequenzdatenanalyse klassische soziologische Fragestellungen über die Vielfalt der Lebensverläufe, ihre Bedingtheit durch historische und soziale Kontexte und ihre Folgen nun weitaus besser beantwortet werden. Der Sequenzdatenanalyse gebührt deshalb ein Platz im „Instrumentenkoffer" der modernen Lebensverlaufsforschung.

Anhang
Programmtechnische Umsetzung einer Sequenzdatenanalyse

Der aufwendigste Teil ist die Konstruktion der Sequenzen aus den Biographiedaten. Im Prinzip programmiert man eine Schleife über das Beobachtungsfenster, wobei die Schrittweite durch die gewählte Zeiteinheit festgelegt ist. In dieser Schleife wird für jedes Individuum zu jedem Zeitpunkt der Zustand festgestellt und in jeweils einer Variablen abgelegt. Sind die Sequenzen 60 Elemente lang, so muss man 60 neue Variablen bilden. Für zeitpunktbezogene deskriptive Auswertungen muss man den resultierenden Datensatz „umdrehen" (Variablen in die Zeilen, Personen in die Spalten). Für die Datenaufbereitung und die deskriptiven Auswertungen kann man jedes konventionelle Statistikpaket verwenden (TDA bietet allerdings auch hier einige hilfreiche Prozeduren an).

Optimal-Matching dagegen ist in keinem konventionellen Statistikpaket implementiert. Andrew Abbotts Programm OPTIMIZE kann nur wenige hundert Sequenzen verarbeiten. Diese Begrenzung führt dazu, dass OM in vielen Anwendungen nur mit einem Subsample gerechnet wird (z.B. noch Stovel und Bolan 2004). Für Sozialwissenschaftler bietet sich deshalb die Verwendung von TDA an (Rohwer und Pötter 1999: Kap. 3.4, 6.6, 6.7). TDA ist kostenlos unter http://www.stat.ruhr-uni-bochum.de/tda.html erhältlich. TDA ist sehr effizient programmiert und kann selbst große Mengen von Sequenzen in relativ kurzer Zeit verarbeiten. In der Pluralisierungsanwendung waren 2.988 Sequenzen der Länge 85 zu vergleichen, also 4.462.578 Levenshtein-Distanzen zu berechnen. Eine gewaltige Rechenaufgabe, für die man vor 15 Jahren noch einen Supercomputer benötigt hätte. TDA schafft dies in 14 Minuten (Pentium IV, 2,8 GHz)! Die resultierende Distanzmatrix (in Listenform) ist dann der Input für einen Clusteralgorithmus. Konventionelle Statistik-

pakete haben mit einer solch großen Distanzmatrix eher Probleme. Nicht so TDA: Die Clusteranalyse mit Ward dauert für das Anwendungsbeispiel 10 Minuten. Allerdings ist eine Clusteranalyse in TDA gewöhnungsbedürftig (Rohwer und Pötter 1999: Kap. 3.6, 7.5).

Im Folgenden sind die TDA-Kommandos für das Pluralisierungsbeispiel aufgeführt:

```
nvar (dfile=sequenz.dat, noc=2988, ID=c1, Y{0,84}=c2);
seqdef = Y0,,Y84;
seqm (icost = 1, scost = 2, v = ID) = dist.dat.
```

Mit nvar liest man den Datensatz ein, der die Sequenzen enthält. Der Datensatz heißt sequenz.dat und hat 2.988 Fälle (noc). In der ersten Spalte (c1) steht die Personennummer (ID), in der 2. bis 86. Spalte folgen die Elemente der Sequenz (Y0 bis Y84). Mit seqdef definiert man die Sequenz. Mit seqm berechnet man die Levenshtein-Distanzen paarweise. Die Indel-Kosten (icost) sind 1, die Substitutionskosten werden von TDA anhand der Übergangswahrscheinlichkeiten datenbasiert berechnet (scost=2). Mit der Option v werden mit jeder Distanz die beiden Personennummern abgespeichert und zwar im Datensatz dist.dat.

Zur Erzeugung der Sequenz-Indexplots (*Abbildung 2*) werden den wie oben beschrieben konstruierten Sequenzen die Information der Clusterzugehörigkeit sowie weitere Kennzahlen (TDA bietet mit dem seqgc Kommando eine Reihe interessanter Möglichkeiten) zugespielt. Auch das ist mit jedem Standardstatistikpaket möglich. Für eine übersichtliche Darstellung sind die Sequenzen nach sinnvollen Kriterien zu sortieren. Im dargestellten Beispiel wurde nach Clusterzugehörigkeit, Verweildauer in verschiedenen Statuszuständen und der Anzahl an Ereignissen sortiert, und zwar in dieser Reihenfolge. Der eigentliche Plot wird in SAS mit dem Kommando proc gplot erstellt. Ein Beispielprogramm findet man unter http://www.sowi.uni-mannheim.de/lehrstuehle/lessm/papers.htm.

Literatur

Abbott, Andrew, 1992: From Causes to Events: Notes on Narrative Positivism. Sociological Methods & Research 20: 428–455.

Abbott, Andrew, 1995: A Comment on „Measuring the Agreement between Sequences". Sociological Methods & Research 24: 232–243.

Abbott, Andrew, 2000: Reply to Levine and Wu. Sociological Methods & Research 29: 65–76.

Abbott, Andrew, und *Emily Barman*, 1997: Sequence Comparison via Alignment and Gibbs Sampling: A Formal Analysis of the Emergence of the Modern Sociological Article. Sociological Methodology 27: 47–87.

Abbott, Andrew, und *Alexandra Hrycak,* 1990: Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers: American Journal of Sociology 96: 144–185.

Abbott, Andrew, und *Angela Tsay*, 2000: Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. Sociological Methods & Research 29: 3–33.

Aisenbrey, Silke, 2000: Optimal Matching Analyse: Anwendungen in den Sozialwissenschaften. Opladen: Leske + Budrich.

Bacher, Johann, 1996: Clusteranalyse: Anwendungsorientierte Einführung. München: Oldenbourg.

Berger, Peter, Peter Steinmüller und *Peter Sopp*, 1993: Differentiation of Life-Courses? European Sociological Review 9: 43–61.

Billari, Francesco, 2001: Sequence Analysis in Demographic Research. Canadian Studies in Population 28: 439–458.

Brückner, Hannah, und *Götz Rohwer,* 1996: Geschlechtsspezifische Unterschiede in den Erwerbsverläufen der Geburtskohorte 1960. München und Berlin: unveröffentlichtes Manuskript (http://www.stat.ruhr-uni-bochum.de/papers.html).

Brüderl, Josef, 2004: Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland und Europa. Aus Politik und Zeitgeschichte B 19: 3–10.

Brüderl, Josef, und *Thomas Klein*, 2003: Die Pluralisierung partnerschaftlicher Lebensformen in Westdeutschland 1960–2000. S. 189–217 in: *Walter Bien* und *Jan Marbach* (Hg.), Partnerschaft und Familiengründung. Opladen: Leske + Budrich.

Buchmann, Marlies, und *Stefan Sacchi*, 1995: Mehrdimensionale Klassifikation beruflicher Verlaufsdaten. Kölner Zeitschrift für Soziologie und Sozialpsychologie 47: 413–442.

Davies, Kevin, 2001: Die Sequenz: Der Wettlauf um das menschliche Genom. München: Hanser.

Dijkstra, Wil, und *Toon Taris*, 1995: Measuring the Agreement between Sequences. Sociological Methods & Research 18: 37–68.

Driel, Kees van, und *Paul Oosterveld*, 2001: Nonoptimal Alignment. Sociological Methods & Research 29: 524–531.

Elzinga, Cees, 2003: Sequence Similarity: A Nonaligning Technique. Sociological Methods & Research 32: 3–29.

Erzberger, Christian, und *Gerald Prein*, 1997: Optimal-Matching-Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe. ZUMA-Nachrichten 21 (Heft 40): 52–80.

Halpin, Brendan, und *Tak Wing Chan*, 1998: Class Careers as Sequences. European Sociological Review 14: 111–130.

Han, Shin-Kap, und *Phyllis Moen*, 2001: Coupled Careers: Pathways through Work and Marriage in the United States. S. 201–231 in: *Hans-Peter Blossfeld* und *Sonja Drobnic* (Hg.), Careers of Couples in Contemporary Societies. Oxford: University Press.

Hertkorn, Georg, *Claudia Hertfelder* und *Peter Wagner*, 2001: Klassifikation von Zeitverwendungstagebüchern. S. 78–90 in: *Manfred Ehling* und *Joachim Merz* (Hg.), Zeitbudget in Deutschland. Wiesbaden: Statistisches Bundesamt.

Kogan, Irena, 2003: A Study of Employment Careers of Immigrants in Germany. Mannheimer Zentrum für Europäische Sozialforschung: Arbeitspapier Nr. 66.

Martens, Bernd, 1991: Explorative Analyse zeitlicher Verläufe. Berlin: Duncker & Humblot.

Mowitz-Lambert, Joachim, 2001: Übergangsmuster in der Statuspassage von beruflicher Ausbildung in die Erwerbstätigkeit. S. 199–217 in: *Reinhold Sackmann* und *Matthias Wingens* (Hg.), Strukturen des Lebenslaufs: Übergang – Sequenz – Verlauf. Weinheim: Juventa.

Rindfuss, Ronald, Gray Swicegood und *Rachel Rosenfeld*, 1987: Disorder in the Life Course: How Common and Does it Matter? American Sociological Review 52: 785–801.

Rohwer, Götz, und *Ulrich Pötter*, 1999: TDA User's Manual. Bochum: unveröffentlichtes Manuskript (http://www.stat.ruhr-uni-bochum.de/tman.html).

Rohwer, Götz, und *Heike Trappe,* 1997: Describing Life Courses. An Illustration Based on NLSY Data. Berlin: POLIS Project Paper (http://www.stat.ruhr-uni-bochum.de/papers.html).

Sankoff, David, und *Joseph Kruskal* (Hg.), 1983: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading: Addison-Wesley.

Scherer, Stefani, 2001: Early Career Patterns: A Comparison between Great Britain and West Germany. European Sociological Review 17: 119–144.

Scherer, Stefani, 2004: Erwerbseintritt und Berufsverlauf: Westdeutschland, Italien und Großbritannien im Vergleich. Frankfurt a.M.: Peter Lang.

Stovel, Katherine, 2001: Local Sequential Patterns: The Structure of Lynching in the Deep South, 1882–1930. Social Forces 79: 843–880.

Stovel, Katherine, und *Marc Bolan*, 2004: Residential Trajectories: Using Optimal Alignment to Reveal the Structure of Residential Mobility. Sociological Methods & Research 32: 559–598.

Taris, Toon, 2000: A Primer in Longitudinal Data Analysis. London: Sage.

Windzio, Michael, und *Michael Grotheer*, 2002: Bleiben die Erfolgreichen übrig? Zeitschrift für Soziologie 31: 514–528.

Wu, Lawrence, 2000: Some Comments on „Sequence Analysis and Optimal Matching Methods in Sociology". Sociological Methods & Research 29: 41–64.

VERWEILDAUERN UND ÜBERGANGSRATEN BEI MEHREREN FOLGEZUSTÄNDEN

Götz Rohwer

Zusammenfassung: Der Beitrag beschäftigt sich mit der Frage, wie bei Episoden, die durch Ereignisse unterschiedlicher Art („konkurrierende Risiken") enden können, Verweildauern ermittelt und dargestellt werden können. Es wird gezeigt, dass die Verwendung des Kaplan-Meier-Verfahrens bei Episoden dieser Art zu Pseudo-Survivorfunktionen führt, die keine sinnvoll interpretierbaren Darstellungen von Verweildauern liefern. Als Alternative werden Anteilsfunktionen vorgeschlagen, die ebenfalls mit teilweise zensierten Daten berechnet werden können. Zur Illustration dient ein Simulationsmodell für die Dauer von Ehen, die durch eine Scheidung oder den Tod eines Ehepartners enden können.

I. Einleitung

In vielen Bereichen der Sozialforschung interessiert man sich für Zeitdauern, während der sich Menschen oder andere Untersuchungseinheiten in bestimmten Zuständen aufhalten, z.B. für Ausbildungsdauern oder Dauern der Arbeitslosigkeit oder Ehedauern. Oft interessiert auch nicht nur die Dauer in einem Ausgangszustand, sondern die Zeitdauer bis zum Übergang in einen bestimmten Folgezustand, z.B. Arbeitslosigkeitsdauern bis zum Erreichen einer neuen Beschäftigung. Dann muss offenbar berücksichtigt werden, dass es mehrere mögliche Folgezustände gibt, in die gewechselt werden kann.

In der Literatur wird oft von „konkurrierenden Risiken" gesprochen. Zur Darstellung geeignete statistische Methoden wurden zuerst in der Demographie entwickelt: Sterbetafeln, bei denen zwei oder mehr unterschiedliche Arten von Todesfällen unterschieden werden. In der englischsprachigen Literatur wird von „multiple-decrement (life) tables" gesprochen (man vgl. etwa Namboodiri und Suchindran 1987: 93ff.; Smith 1992: 139ff.); als kurze deutsche Bezeichnung hat Ulrich Mueller (1993: 205) vorgeschlagen, von „Multi-Exit-Tafeln" zu sprechen. Solche Tafeln können natürlich nicht nur zur Untersuchung von Lebensdauern verwendet werden, sondern lassen sich immer anwenden, wenn Episoden durch unterschiedliche Arten von Ereignissen beendet werden können.

Der Grundgedanke der meisten statistischen Methoden für konkurrierende Risiken besteht darin, von zustandsspezifischen Übergangsraten auszugehen, die sich jeweils gesondert auf den Übergang in einen bestimmten Folgezustand beziehen. Hier setzen auch Regressionsmodelle an, die diese zustandsspezifischen Raten als Funktionen von Kovariablen konzeptualisieren. Darauf soll in diesem Beitrag jedoch nicht näher eingegangen werden. Vielmehr konzentrieren wir uns auf die Frage, wie bei Episoden mit mehreren möglichen Folgezuständen von Verweildauern gesprochen werden kann. In

vielen praktischen Anwendungen werden hierfür Pseudo-Survivorfunktionen verwendet, bei deren Berechnung davon ausgegangen wird, dass alle Episoden, die nicht in dem jeweils interessierenden Folgezustand enden, „zensiert" sind. Es wird also angenommen, dass diese Episoden gleichwohl noch in dem interessierenden Folgezustand enden könnten. Diese Annahme ist aber offenbar problematisch, insbesondere in denjenigen Fällen, in denen man weiß oder wissen könnte, dass eine Episode bereits durch einen anderen Folgezustand tatsächlich beendet worden ist. Wenn man z.B. schon weiß, dass eine Ehe durch den Tod eines Ehepartners beendet wurde, wäre es wenig plausibel anzunehmen, dass sie später noch durch eine Scheidung enden könnte.

In den folgenden Ausführungen diskutieren wir dieses Problem. Dabei geht es nicht nur um Begriffsbildungen, sondern wir versuchen auch, eine Vorstellung der Fehler und ihrer Größenordnungen zu gewinnen, die durch das übliche Verfahren bei der Berechnung von Pseudo-Survivorfunktionen entstehen können. Um zu zeigen, dass beliebig große Fehler entstehen können, verwenden wir zunächst ein fiktives Beispiel. Dann untersuchen wir das Problem anhand von Ehen, die durch Scheidungen oder durch den Tod eines Ehepartners enden können.

Als Alternative zur Berechnung von Pseudo-Survivorfunktionen schlagen wir Anteilsfunktionen vor, die sich auf die Anteile an einer Ausgangsgesamtheit beziehen, die in die unterschiedlichen Folgezustände wechseln. Es wird gezeigt, dass auch solche Anteilsfunktionen mit teilweise zensierten Daten geschätzt werden können.

II. Der begriffliche Rahmen

Wir gehen von einer Prozesszeitachse aus, die wir uns als eine Folge von Zeitstellen $t = 0,1,2,3,...$ vorstellen. Zeitstellen können z.B. Tage, Wochen, Monate oder Jahre sein. Die Zeitachse soll dazu dienen, Verweildauern zu erfassen, während der sich Menschen oder andere Untersuchungseinheiten in einem bestimmten Zustand befinden. Im Folgenden sprechen wir von Personen, nehmen also an, dass wir uns auf eine Gesamtheit von Personen beziehen können, die zum Beginn der Prozesszeit in einen gleichen Ausgangszustand geraten sind, z.B. durch ihre Geburt oder ihre Heirat oder dadurch, dass sie arbeitslos geworden sind. Diese Gesamtheit repräsentieren wir durch eine Menge $\Omega := \{\omega_1,...,\omega_n\}$. Jedes Element ω_i repräsentiert eine Person aus der Gesamtheit Ω. Insgesamt gibt es n Personen.

Um die Verweildauern im Ausgangszustand zu erfassen, wird eine statistische Variable $T : \Omega \rightarrow \{0,1,2,3,...\}$ definiert, deren Wertebereich durch die Prozesszeitachse gegeben ist. Wenn ω eine Person aus Ω ist, soll also $T(\omega)$ die Zeitstelle angeben, in der diese Person den Ausgangszustand verlässt. So kann man auch sagen, dass $T(\omega)$ die Anzahl der vollendeten Zeitstellen angibt, während der sich ω im Ausgangszustand aufgehalten hat.

Weiter sei n_t die Anzahl der Personen, die sich zum Beginn der Zeitstelle t noch im Ausgangszustand befinden, so dass also $n_0 = n$ ist. Somit erhält man die Survivorfunktion, die die Verteilung der Verweildauervariablen T darstellt, durch $G(t) := n_t/n$. (Eine alternative Möglichkeit besteht darin, die Survivorfunktion durch $1 - F$ zu definieren, wobei F die Verteilungsfunktion ist. Die oben angegebene Definition vereinfacht je-

doch bei einer diskreten Zeitachse die Notationen.) Definiert man außerdem durch $w_t := n_t - n_{t+1}$ die Anzahl der Personen, die während der Zeitstelle t den Ausgangszustand verlassen, kann man die Rate für das Verlassen des Ausgangszustands in der Zeitstelle t durch $r(t) := w_t/n_t$ definieren. Offenbar kann man Raten aus der Survivorfunktion berechnen:

$$r(t) = \frac{w_t}{n_t} = \frac{p(t)}{G(t)} = 1 - \frac{G(t+1)}{G(t)}$$

wobei $p(t) := w_t/n$ die relative Häufigkeit für einen Wechsel in der Zeitstelle t ist. Umgekehrt kann man aber auch die Survivorfunktion aus den Raten berechnen, denn es gilt der Zusammenhang

$$G(t) = \prod_{k=0}^{t-1} (1 - r(k))$$

Allerdings benötigt man hierfür die gesamte Ratenfunktion r, die jeder Zeitstelle t die entsprechende Rate $r(t)$ zuordnet. (Eine ausführliche Darlegung dieser formalen Zusammenhänge findet man bei Rohwer und Pötter 2001, Kap.12)

III. Mehrere Folgezustände

Solange man sich nur für die Aufenthaltsdauer im Ausgangszustand interessiert, sind die Begriffsbildungen unproblematisch. In den meisten Anwendungen richtet sich die Aufmerksamkeit jedoch auf bestimmte Folgezustände, z.B. interessiert man sich für Ehedauern bis zu einer Scheidung oder für Arbeitslosigkeitsdauern bis zur Aufnahme einer neuen Beschäftigung. Dann muss berücksichtigt werden, dass es stets mehrere mögliche Folgezustände gibt. Insbesondere muss auch berücksichtigt werden, dass die Verweildauer im Ausgangszustand durch den Tod einer Person beendet werden kann.

Wir betrachten im Weiteren zwei Folgezustände, die durch a und b indiziert werden. Es sei also $w_t = w_t^a + w_t^b$, wobei w_t^a und w_t^b die Anzahlen der Personen angeben, die in der Zeitstelle t in den Folgezustand a bzw. b wechseln. Somit kann man die Häufigkeiten $p^a(t) := w_t^a/n$ und $p^b(t) := w_t^b/n$ definieren, außerdem die zustandsspezifischen Raten

$$r^a(t) := \frac{w_t^a}{n_t} = \frac{p^a(t)}{G(t)} \text{ und } r^b(t) := \frac{w_t^b}{n_t} = \frac{p^b(t)}{G(t)}$$

Offenbar gilt $r(t) = r^a(t) + r^b(t)$, wobei $r(t)$ wie im vorangegangenen Abschnitt die Rate für das Verlassen des Ausgangszustands ist.

Survivorfunktionen. Wie kann man aber jetzt von Verweildauern sprechen? Man könnte versuchen, eine Verweildauerverteilung durch eine Survivorfunktion

$$\tilde{G}^a(t) := \prod_{k=0}^{t-1}(1-r^a(k))$$

zu definieren. Diese Funktion liefert aber nicht die Verteilung der Verweildauern bis zum Erreichen des Folgezustands *a*. Denn ist man an diesen Verweildauern interessiert, kann man sich offenbar nur auf Personen beziehen, die diesen Folgezustand auch tatsächlich erreichen. Man muss also zunächst die Gesamtheit Ω in zwei Teilgesamtheiten zerlegen, wobei sich die Unterscheidung an dem Folgezustand orientiert, der schließlich erreicht wird. Beziehen wir uns auf die Teilgesamtheit Ω^a, die aus denjenigen Personen besteht, die schließlich den Folgezustand *a* erreichen, ist die Anzahl ihrer Mitglieder

$$n^a := \sum_{t=0}^{\infty} w_t^a .$$

Bezugnehmend auf diese Gesamtheit kann man nun erneut eine Verweildauervariable $T^a : \Omega^a \rightarrow \{0,1,2,3...\}$ definieren, die die Verweildauer bis zum Erreichen des Folgezustands *a* erfasst. Ihr entsprechen die relativen Häufigkeiten w_t^a / n^a und die zustandsspezifische Survivorfunktion

$$G^a(t) := \sum_{k=1}^{\infty} \frac{w_k^a}{n^a} = 1 - \sum_{k=0}^{t-1} \frac{w_k^a}{n^a} = \prod_{k=o}^{t-1} \left(1 - \frac{w_k^a \Big/ n^a}{G^a(k)} \right)$$

die sich ersichtlich von der Survivorfunktion $\tilde{G}^a$ unterscheidet. Wenn man G^a berechnen möchte, genügt es auch nicht, nur die Ratenfunktion r^a zu kennen; man benötigt außerdem die Ratenfunktion r^b oder die allgemeine Ratenfunktion $r = r^a + r^b$ oder, damit äquivalent, die Survivorfunktion *G*.

Bei dieser Formulierung ist wichtig, dass man jeweils die gesamten Funktionen bis zum Ende der Prozesszeit kennen muss, oder anders formuliert: Zur Berechnung der Survivorfunktion G^a muss man den Anteil der Personen kennen, die schließlich in den Folgezustand *a* wechseln. Da dieser Anteil in den meisten praktischen Anwendungen – insbesondere dann, wenn Retrospektivdaten verwendet werden – nicht bekannt ist, wird stattdessen fast immer die Survivorfunktion $\tilde{G}^a$ berechnet. Wie wir im nächsten Abschnitt anhand eines Beispiels zeigen werden, bezieht sich diese Funktion jedoch nicht auf sinnvoll interpretierbare Verweildauern, $\tilde{G}^a$ wird deshalb oft eine Pseudo-Survivorfunktion genannt.

Anteilsfunktionen. Wenn man sich für Übergänge in einen Folgezustand *a* interessiert und nicht bekannt ist, wie viele Personen schließlich in diesen Folgezustand wechseln werden, kann man die Survivorfunktion G^a nicht berechnen. Man kann aber sinnvoll interpretierbare Anteilsfunktionen

$$H^a(t) := \frac{1}{n}\sum_{k=0}^{t-1} w_k^a = \sum_{k=0}^{t-1} p^a(k) = \sum_{k=0}^{t-1} r^a(k)G(k)$$

berechnen. $H^a(t)$ gibt den Anteil der Personen an der Ausgangsgesamtheit an, die bis zur Zeitstelle t in den Folgezustand a gewechselt haben. Bezieht man sich z.B. auf Ehen, die durch eine Scheidung (a) oder durch den Tod eines Ehepartners (b) enden können, würde $H^a(t)$ den Anteil der bis zur Ehedauer t geschiedenen Ehen angeben.

In der Praxis wird stattdessen oft die Pseudo-Survivorfunktion $\tilde{G}^a$ berechnet und so interpretiert, *als ob* es sich um eine Anteilsfunktion handeln würde. Diese Interpretation ist aber bei mehreren Folgezuständen nicht korrekt, da $1 - \tilde{G}^a$ keineswegs mit H^a identisch ist (es sei angemerkt, dass eine solche Interpretation auch in Teilen der Lehrbuchliteratur nahe gelegt wird, man vgl. z.B. Blossfeld, Hamerle und Mayer 1986: 135f., und Blossfeld und Rohwer 2002: 84); man sieht das, wenn man $H^a(t)$ analog zur Definition von $\tilde{G}^a$ in der Form

$$H^a(t) = 1 - \prod_{k=0}^{t-1}\left(1 - \frac{w_k^a}{\overline{n}_k^a}\right)$$

schreibt, wobei $\overline{n}_k^a$ die Anzahl der Personen ist, die bis zur Zeitstelle k *nicht* in den Folgezustand a gewechselt haben. Bei der Berechnung der Pseudo-Survivorfunktion werden anstelle der Quotienten $\frac{w_k^a}{\overline{n}_k^a}$ die Raten $\frac{w_k^a}{n_k}$ verwendet. Im Nenner steht dann nicht die Anzahl der Personen, die bisher nicht in den Folgezustand a gewechselt haben, sondern die Anzahl der Personen, die sich noch im Ausgangszustand befinden. Da $n_k \leq \overline{n}_k^a$ ist, findet man

$$H^a(t) \leq 1 - \tilde{G}^a(t)$$

Wenn also die Pseudo-Survivorfunktion $\tilde{G}^a$ so interpretiert wird, als ob es sich um eine Anteilsfunktion handelt, kann es – abhängig von den Anteilen der Personen, die in den Folgezustand b wechseln – zu beliebig großen Überschätzungen der Wechsel in den Folgezustand a kommen.

Es ist auch bemerkenswert, dass Anteilsfunktionen im Unterschied zu Pseudo-Survivorfunktionen additiv sind: $H(t) = H^a(t) + H^b(t)$ liefert den Anteil der Personen, die bis zur Zeitstelle t in irgendeinen Folgezustand gewechselt haben. Schließlich sei auch noch angemerkt, dass es einen einfachen Zusammenhang zwischen der Anteilsfunktion H^a und der zustandsspezifischen Survivorfunktion G^a gibt:

$$G^a(t) = 1 - \sum_{k=0}^{t-1}\frac{w_k^a}{n^a} = 1 - \sum_{k=0}^{t-1}\frac{w_k^a}{n}\frac{n}{n^a} = 1 - \frac{n}{n^a}H^a(t)$$

Würde man den Anteil der Personen, die schließlich in den Folgezustand a wechseln, kennen, könnte man G^a aus H^a berechnen.

IV. Eine numerische Illustration

Zur Illustration der Begriffsbildungen und Verdeutlichung des Problems betrachten wir ein Zahlenbeispiel mit fiktiven Daten. Wir gehen von einer Gesamtheit von 2500 arbeitslos gewordenen Personen aus, und es gibt zwei mögliche Folgezustände: (a) Übergang in eine neue Beschäftigung und (b) Verlassen des Arbeitsmarktes (einschließlich Sterbefälle). *Tabelle 1* zeigt die fiktiven Daten. Die Zeitstellen der Zeitachse $t = 0,...,19$ kann man sich als Quartale vorstellen. w_t^a und w_t^b geben die Anzahl der Personen an, die in der Zeitstelle t in den Zustand a bzw. b wechseln.

Tabelle 1: Fiktive Daten für ein Beispiel mit zwei Folgezuständen a und b

t	w_t^a	w_t^b	t	w_t^a	w_t^b	t	w_t^a	w_t^b	t	w_t^a	w_t^b
0	100	50	5	80	60	10	60	70	15	40	80
1	100	50	6	80	60	11	60	70	16	20	80
2	100	50	7	80	60	12	40	80	17	20	90
3	100	50	8	60	70	13	40	80	18	20	90
4	80	60	9	60	70	14	40	80	19	20	

Da in diesem Beispiel vollständige Daten gegeben sind, kann man alle Begriffsbildungen nachvollziehen:

a) Man findet, dass von den 2500 Personen 1200 (48 Prozent) eine neue Beschäftigung annehmen und 1300 (52 Prozent) aus dem Arbeitsmarkt ausscheiden.
b) Man kann die Survivorfunktionen G^a und G^b für die beiden Folgezustände berechnen (*Abbildung 1*).
c) Man kann die durchschnittlichen Verweildauern berechnen. Die durchschnittliche Verweildauer im Ausgangszustand beträgt 8.5, bis zu einer neuen Beschäftigung 6.8 und bis zum Ausscheiden aus dem Arbeitsmarkt 10.0 Quartale. (Natürlich könnte man berücksichtigen, dass unsere Prozesszeitachse bei 0 beginnt und jeweils ein halbes Quartal hinzurechnen.)

Jetzt berechnen wir eine Survivorfunktion für den Übergang in eine Beschäftigung, wobei wir diejenigen Episoden, die nicht mit einer neuen Beschäftigung enden, als rechts zensiert betrachten. Dafür kann das Kaplan-Meier-Verfahren verwendet werden. *Tabelle 2* zeigt die Rechenschritte. Als Ergebnis erhält man die Pseudo-Survivorfunktion $\tilde{G}^a$.

Aber welche Bedeutung kann dieser Survivorfunktion gegeben werden? Sie liefert jedenfalls keine sinnvollen Aussagen über die Dauer der Arbeitslosigkeit, und zwar weder über die Dauer im Ausgangszustand noch über die Dauer bis zum Erreichen einer Beschäftigung. So liegt zum Beispiel der Median von $\tilde{G}^a$ bei 13 bis 14 Quartalen und der Mittelwert beträgt 11.9 Quartale. Beide Werte sind deutlich größer als die durchschnittliche Dauer bis zum Erreichen einer neuen Beschäftigung (6.8 Quartale) und sogar größer als die durchschnittliche Verweildauer im Ausgangszustand (8.5 Quartale).

Die Funktion $\tilde{G}^a$ liefert aber auch keine Information über die Anteile der Personen, die noch nicht in eine Beschäftigung gewechselt haben. Wie im vorangegangenen Abschnitt ausgeführt wurde, erhält man diese Anteile durch die Funktion $H^a \neq 1 - \tilde{G}^a$.

Abbildung 1: Survivorfunktionen G^a und G^b für die durch *Tabelle 1* gegebenen Daten

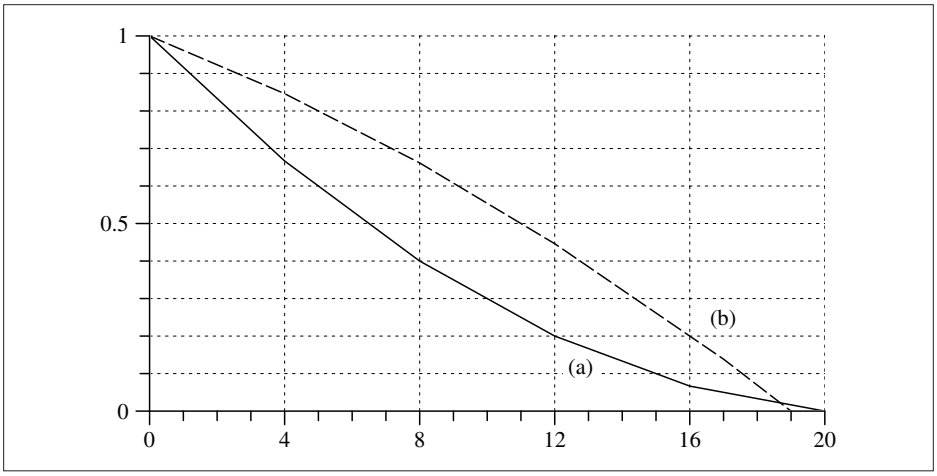


Abbildung 2: Darstellung der Funktionen H^a (durchgezogene Linie) und $1 - \tilde{G}^a$ (gestrichelte Linie) für die durch *Tabelle 1* gegebenen Daten

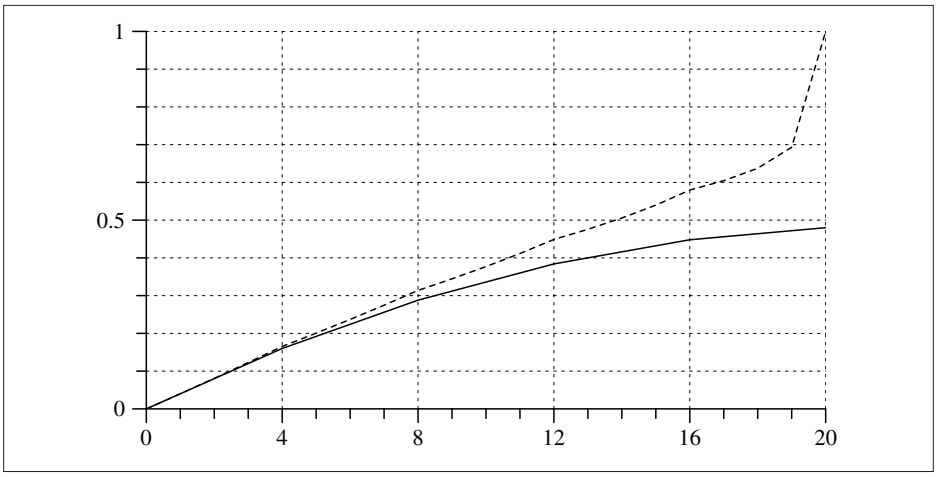


Abbildung 2 illustriert die Differenz für unser Beispiel. Es ist deutlich sichtbar, dass die Pseudo-Survivorfunktion keine adäquaten Aussagen über die Anteile derjenigen Personen macht, die in eine Beschäftigung wechseln.

Tabelle 2: Berechnung der Survivorfunktion $\tilde{G}^a$ mit dem Kaplan-Meier-Verfahren aus den Daten in *Tabelle 1*

t	n_t	w_t^a	w_t^b	$r^a(t)$	$\tilde{G}^a(t)$
0	2500	100	50	0.0400	1.0000
1	2350	100	50	0.0426	0.9600
2	2200	100	50	0.0455	0.9191
3	2050	100	50	0.0488	0.8774
4	1900	80	60	0.0421	0.8346
5	1760	80	60	0.0455	0.7994
6	1620	80	60	0.0494	0.7631
7	1480	80	60	0.0541	0.7254
8	1340	60	70	0.0448	0.6862
9	1210	60	70	0.0496	0.6555
10	1080	60	70	0.0556	0.6230
11	950	60	70	0.0632	0.5884
12	820	40	80	0.0488	0.5512
13	700	40	80	0.0571	0.5243
14	580	40	80	0.0690	0.4944
15	460	40	80	0.0870	0.4603
16	340	20	80	0.0588	0.4202
17	240	20	90	0.0833	0.3955
18	130	20	90	0.1538	0.3626
19	20	20	0	1.0000	0.3068
20	0	0	0		0.0000

V. Verwendung zensierter Daten

Manchmal wird die Verwendung einer Pseudo-Survivorfunktion mit dem Argument begründet, dass bei ihrer Berechnung zensierte Beobachtungen berücksichtigt werden können (z.B. Diekmann und Klein 1991). Dies ist zwar richtig, das eigentliche Problem besteht jedoch darin, dass zugleich unterstellt wird, dass alle zensierten Episoden schließlich in dem speziellen Folgezustand enden werden, auf den sich die Berechnung der Pseudo-Survivorfunktion bezieht. Wenn z.B. eine Pseudo-Survivorfunktion für Ehedauern bis zu einer Scheidung berechnet wird, wird zugleich angenommen, dass alle zensierten Ehe-Episoden schließlich in einer Scheidung enden werden. Eine solche Annahme ist aber sicherlich fragwürdig.

Andererseits ist es auch nicht richtig, dass bei der Berechnung von Anteilsfunktionen zensierte Beobachtungen grundsätzlich nicht verwendet werden können. Es gilt nämlich folgender Zusammenhang, der bereits in Abschnitt III erwähnt wurde:

$$H^a(t) = \sum_{k=0}^{t-1} p^a(k) = \sum_{k=0}^{t-1} r^a(k) G(k)$$

Wenn man also glaubt, dass man mit partiell zensierten Daten die Survivorfunktion G und die Rate r^a (oder, damit äquivalent, die Pseudo-Survivorfunktion $\tilde{G}^a$) einigermaßen korrekt schätzen kann, dann kann man sie sogleich auch verwenden, um die Anteilsfunktion H^a zu schätzen.

VI. Durchschnittliche Verweildauern

An mehreren Stellen wurde bereits über durchschnittliche Verweildauern gesprochen. Es ist nützlich, sich noch einmal die unterschiedlichen Begriffsbildungen zu verdeutlichen. In jedem Fall muss betont werden, dass man nur aus einer retrospektiven Perspektive Feststellungen über Verweildauern machen kann. Natürlich kann man Vermutungen darüber anstellen, wie lange eine noch nicht abgeschlossene Episode dauern wird; aber erst nach ihrer Beendigung kann man wissen, wie lange sie tatsächlich gedauert hat.

Dies gilt auch dann, wenn man sich zunächst nur für die durchschnittliche Verweildauer im Ausgangszustand interessiert. Zu berechnen ist in diesem Fall der Mittelwert $M(T)$ der Verweildauervariablen T, und das setzt eine Kenntnis der gesamten Verteilung voraus. Verwendet man z.B. die Survivorfunktion, muss

$$M(T) = \sum_{t=0}^{\infty} t(G(t) - G(t+1))$$

berechnet werden. Da man bei vielen praktischen Anwendungen nicht die gesamte Verteilung kennt (oder schätzen kann), werden anstelle des Mittelwerts oft Mediane oder andere Quantile angegeben. Dabei sollte jedoch bedacht werden, dass sich der Median einer Verteilung von ihrem Mittelwert erheblich unterscheiden kann. Zum Beispiel liegt bei Bezugsdauern der Sozialhilfe der Median unterhalb eines Jahres, die durchschnittliche Bezugsdauer ist jedoch wesentlich länger als ein Jahr.

Der begriffliche Unterschied wird deutlich, wenn man bedenkt, dass sich Quantile einer Verweildauerverteilung gar nicht auf die Gesamtheit aller Episoden beziehen, sondern nur auf Teilgesamtheiten mit den jeweils kürzesten Episodendauern. Zum Beispiel gibt der Median an, in welcher Zeit die Hälfte der Personen mit den kürzesten Verweildauern den Ausgangszustand verlassen hat; dabei bleiben die Verweildauern der übrigen Personen vollständig unberücksichtigt.

Es erscheint uns deshalb sinnvoll, nicht nur Quantile, sondern auch Mittelwerte zu berechnen. Wenn man nicht die gesamte Verteilung kennt, ist es dennoch oft möglich, bedingte Mittelwerte zu berechnen, die sich auf die Gesamtheit der Episoden bis zu einer maximalen Verweildauer t_0 beziehen, also

$$M(T \mid T < t_0) = \sum_{t=0}^{t_0-1} t \frac{G(t) - G(t+1)}{1 - G(t_0)}$$

Zum Beispiel könnte man auf diese Weise die durchschnittliche Lebensdauer von Menschen berechnen, die maximal 90 oder 100 Jahre alt geworden sind.

Eine scheinbare begriffliche Unklarheit tritt auf, wenn man sich nicht auf Verweildauern im Ausgangszustand, sondern auf Verweildauern bis zum Erreichen eines bestimmten Folgezustands a interessiert, also z.B. bei Arbeitslosigkeitsepisoden für die durchschnittliche Dauer bis zum Wechsel in eine neue Beschäftigung. Es wurde bereits darauf hingewiesen, dass es irreführend wäre, hierfür die Pseudo-Survivorfunktion $\tilde{G}^a$ zu verwenden, denn bereits die Begriffsbildung impliziert, dass man sich nur auf Episoden beziehen kann, die tatsächlich im Zustand a enden. Zu berechnen ist deshalb der Mittelwert der Verweildauervariablen T^a, die für die Teilgesamtheit derjenigen Epi-

soden definiert ist, die im Zustand *a* enden. (Alternativ könnten bedingte Mittelwerte der Form $M(T^u | T^u \geq x)$ betrachtet werden; ihnen entspricht im Kontext unterschiedlicher Mortalitätsrisiken eine „durchschnittliche todesursachenspezifische Lebenserwartung" bzw. „die restliche Lebenserwartung einer Person im Alter *x*, falls diese Person an der spezifischen Ursache *j* [in unserer Notation: *a*] sterben wird" (Mueller 1993: 207).

Auch wenn man nicht weiß, welcher Anteil der Episoden schließlich im Zustand *a* enden wird, ist es oft möglich und sinnvoll, wenigstens bedingte Mittelwerte zu berechnen, die sich auf eine Teilgesamtheit von Episoden beziehen, die bis zu einer bestimmten Dauer t_0 den Zustand *a*, z.B. eine neue Beschäftigung, erreicht haben. Solche bedingten Mittelwerte können nämlich bereits aus der Anteilsfunktion H^a berechnet werden:

$$M(T^a | T^a < t_0) = \sum_{t=0}^{t_0-1} t \frac{G^a(t) - G^a(t+1)}{1 - G^a(t_0)} = \sum_{t=0}^{t_0-1} t \frac{H^a(t+1) - H^a(t)}{H^a(t_0)} = \frac{\sum_{t=0}^{t_0-1} t\, w_t^a}{\sum_{t=0}^{t_0-1} w_t^a}$$

Auf diese Weise wurden z.B. in einer Publikation des Statistischen Bundesamts (1990: 129f.) „durchschnittliche Ehedauern bis zu einer Scheidung" berechnet. Unter Verwendung von Periodendaten, die sich auf das Jahr 1988 beziehen, und mit $t_0 = 26$ wurde so ein Wert von 12.1 errechnet, d.h. die durchschnittliche Ehedauer derjenigen Ehen, die 1988 geschieden wurden und die maximal 25 Jahre dauerten, lag bei 12.1 Jahren. Das ist zugleich eine untere Grenze für die durchschnittliche Dauer aller Ehen bis zu einer Scheidung.

VII. Retrospektiv- und Periodendaten

In Abschnitt IV wurde gezeigt, dass der Pseudo-Survivorfunktion $\tilde{G}^a$, die sich auf den Übergang in einen bestimmten Folgezustand *a* bezieht und bei deren Berechnung alle nicht in diesem Zustand endenden Episoden als rechts zensiert angenommen werden, nicht ohne weiteres eine sinnvolle Bedeutung gegeben werden kann. Das Problem tritt natürlich nur auf, wenn man sich für einen bestimmten Folgezustand interessiert und nicht nur für die undifferenzierte Verweildauer im Ausgangszustand. Aber das ist in den meisten praktischen Anwendungen der Fall. Eine Ausnahme bilden Untersuchungen der Lebensdauer, bei denen der Tod der einzige mögliche Endzustand ist. Aber in allen anderen Fällen ist mindestens der Tod einer Person eine reale Alternative, in der eine Episode enden kann.

Retrospektivdaten. Ob bzw. wie alternative Folgezustände zu berücksichtigen sind, hängt auch von der Art der Datengewinnung ab. In zahlreichen Studien werden retrospektiv erhobene Daten verwendet, bei denen von vornherein nur Personen betrachtet werden können, die mindestens bis zum Interviewzeitpunkt überlebt haben. Infolgedessen liegt

die Vermutung nahe, dass man die Tatsache, dass Episoden auch durch Tod enden können – meistens (abgesehen von Datenerhebungen über bereits sehr alte Menschen) –, ignorieren kann. Das ist jedoch in vielen Fällen zumindest fragwürdig.

Als Beispiel beziehen wir uns auf die Untersuchung von Ehen, die in einer Scheidung enden. Dabei ist offenbar zu berücksichtigen, dass Ehen entweder durch eine Scheidung (Folgezustand *a*) oder durch den Tod eines Ehepartners (Folgezustand *b*) enden können. Eine Retrospektivbefragung von Personen, die in der Vergangenheit geheiratet haben, kann drei Arten von Beobachtungen liefern:

a) Ehen, bei denen bis zum Interviewzeitpunkt eine Scheidung stattgefunden hat;
b) Ehen, bei denen bis zum Interviewzeitpunkt einer der Ehepartner gestorben ist, und
c) Ehen, die zum Interviewzeitpunkt noch bestehen (rechts zensierte Beobachtungen).

In allen Fällen muss auch mindestens einer der (ehemaligen) Ehepartner mindestens bis zum Interviewzeitpunkt leben.

Mit Daten dieser Art kann offenbar die Survivorfunktion G^a für die Ehedauern bis zu einer Scheidung nicht berechnet werden, da man nicht weiß, wie viele Ehen schließlich in einer Scheidung enden. Deshalb wird in den meisten empirischen Untersuchungen (man vgl. z.B. einige der Beiträge in Klein und Kopp 1999) eine Pseudo-Survivorfunktion $\tilde{G}^a$ berechnet. Diese Funktion liefert jedoch nicht die Anteile der Ehen, die in einer Scheidung enden, wie bei Interpretationen oft unterstellt wird. Denn in ihrer Berechnung wird angenommen, dass auch diejenigen Ehen, von denen man weiß oder wissen könnte, dass sie durch den Tod eines Ehepartners enden, schließlich durch eine Scheidung enden könnten.

Wenn man sich für die Anteile der Ehen interessiert, die in einer Scheidung enden, müsste man stattdessen die in Abschnitt III definierte Anteilsfunktion H^a berechnen. Wie in Abschnitt V gezeigt wurde, ist dies grundsätzlich auch mit Retrospektivdaten möglich, bei denen die ermittelbaren Episoden teilweise zensiert sind. Allerdings muss bei der Interpretation beachtet werden, dass man bei Retrospektiverhebungen nur Informationen von Personen gewinnen kann, die mindestens bis zum Interviewzeitpunkt überleben.

Periodendaten. Im Unterschied zu retrospektiv erhobenen Daten liefert die amtliche Statistik in erster Linie Periodendaten, die sich auf einzelne Kalenderjahre beziehen. Um bei unserem Beispiel zu bleiben, betrachten wir ehedauerspezifische Scheidungsraten und -ziffern. Zur Definition werden folgende Größen verwendet: $n_{j,t}$:= Anzahl der Ehen, die im Kalenderjahr j eine Dauer von t Jahren hatten; $w^a_{j,t}$ = Anzahl der Ehen, die im Kalenderjahr j eine Dauer von t Jahren hatten und in diesem Kalenderjahr geschieden wurden.

Ehedauerspezifische Scheidungs*raten* sind dann durch $q^a_j(t) := w^a_{j,t} \big/ n_{j,t}$ definiert. (Diese Definition entspricht derjenigen von Esenwein-Rothe 1982: 293.) Sie beziehen sich auf den Anteil der Scheidungen im Kalenderjahr j bei denjenigen Ehen, die in diesem Jahr eine Ehedauer von t Jahren hatten. Man kann sie auf zwei unterschiedliche Weisen zur Berechnung von Survivorfunktionen verwenden. In einer Periodenbetrachtung bezieht man sich durchgehend auf die Scheidungsraten eines Kalenderjahrs j und erhält

$$\breve{G}_j^a(t) := 1 - \prod_{k=0}^{t-1}(1 - q_j^a(k))$$

Bei einer Quasi-Kohortenbetrachtung geht man von den Ehen aus, die in einem Kalenderjahr j geschlossen worden sind, und verwendet Scheidungsraten für die jeweilige Ehedauer. (Da die Kohortenkonstruktion auf einer zeitlichen Verkettung ehedauerspezifischer Periodendaten beruht, sprechen wir von Quasi-Kohorten.) So gelangt man zur Survivorfunktion

$$\tilde{G}_j^a(t) := 1 - \prod_{k=0}^{t-1}(1 - q_{j+k}^a(k))$$

In beiden Fällen handelt es sich jedoch um Pseudo-Survivorfunktionen, bei deren Konstruktion unberücksichtigt bleibt, dass Ehen auch durch den Tod eines Ehepartners enden können. Sie weisen also die gleichen Mängel auf wie die Pseudo-Survivorfunktionen, die üblicherweise aus Retrospektivdaten berechnet werden.

Tabelle 3: Ehedauerspezifische Scheidungsziffern für das Jahr 1999, multipliziert mit 1000 (Statistisches Bundesamt 1999: 268)

t	$p_{1999}^a(t)$	t	$p_{1999}^a(t)$	t	$p_{1999}^a(t)$	t	$p_{1999}^a(t)$
0	0.3	7	25.7	14	13.8	21	7.7
1	3.5	8	22.9	15	12.6	22	7.0
2	13.4	9	21.6	16	11.3	23	6.2
3	19.2	10	20.6	17	10.5	24	5.6
4	24.2	11	17.9	18	9.8	25	5.1
5	27.2	12	16.1	19	9.3		
6	27.1	13	15.0	20	8.5		

Eine Alternative liegt in der Verwendung von ehedauerspezifischen Scheidungs*ziffern*, die folgendermaßen definiert sind: $p_j^a(t) := w_{j,t}^a \Big/ n_{j-t,0}$. In diesem Fall wird die Anzahl der Scheidungen, die im Jahr j bei Ehen mit einer bisherigen Dauer von t Jahren aufgetreten sind, auf die Anzahl der Eheschließungen im Jahr $j - t$ bezogen. *Tabelle 3* zeigt diese Scheidungsziffern für das Kalenderjahr j = 1999 und Ehedauern t = 0,...,25. Zum Beispiel wurden in diesem Jahr 11139 Ehen geschieden, die im Jahr 1990 geschlossen wurden. Andererseits kennt man aus früheren Erhebungen die Anzahl der Ehen, die 1990 geschlossen wurden, nämlich 516388. So erhält man $p_{1999}^a(9)$ = 11139/516388 = 0.0216. Ein Problem besteht offenbar darin, dass in einer Gesellschaft mit Zu- und Abwanderung die Ehen, die im Jahr j bei einer Dauer von t geschieden werden, keine Teilmenge derjenigen Ehen sind, die im Jahr $j - t$ geschlossen wurden. (Eine weitere Fehlerquelle für Deutschland ist die Gebietsveränderung seit 1990.) Ist man jedoch bereit, von diesem Problem in erster Näherung zu abstrahieren, liefern ehedauerspezifische Scheidungsziffern eine durchaus sinnvolle Möglichkeit, um zu schätzen, wie viele Ehen eines Eheschließungsjahrgangs in einer Scheidung enden.

Ausgehend von den Ehen, die im Jahr j geschlossen worden sind, ist der Anteil der bis zu einer Ehedauer von t geschiedenen Ehen

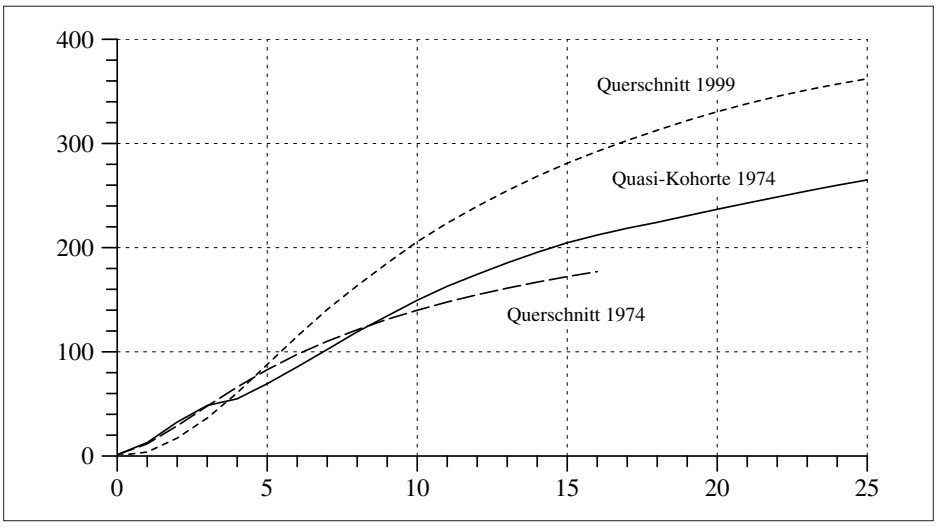
$$H_j^a(t) := \frac{1}{n_{j,0}} \sum_{k=0}^{t-1} w_{j+k,k}^a = \sum_{k=0}^{t-1} p_{j+k}^a(k).$$

Zur Berechnung benötigt man natürlich die ehedauerspezifischen Scheidungsziffern für eine Folge von Jahren. Kennt man sie bis zum Jahr 1999, könnte man z.B. $H_{1974}^a(26)$ berechnen, also – unter der genannten Einschränkung – den Anteil der 1974 geschlossenen Ehen, die bis zum Ende des Jahres 1999 geschieden worden sind. *Abbildung 3* zeigt diese Funktion H_{1974}^a als durchgezogene Linie. (Die Daten entstammen der Tabelle VIII B/181 des Statistischen Bundesamts, die uns freundlicherweise von Herrn H.-P. Bosse zur Verfügung gestellt wurde.) Andererseits könnte man auch eine hypothetische Betrachtung anstellen, bei der man von den ehedauerspezifischen Scheidungsziffern eines bestimmten Kalenderjahrs ausgeht, also

$$\breve{H}_j^a(t) := \sum_{k=0}^{t-1} p_j^a(k)$$

berechnet. Mit den Angaben aus *Tabelle 3* findet man z.B. $\breve{H}_{1974}^a(26) = 0.362$. Unter Zugrundelegung der Scheidungsziffern des Jahres 1999 würden also bei einem hypothetischen Eheschließungsjahrgang bis zu einer Ehedauer von einschließlich 25 Jahren 36.2 Prozent der Ehen geschieden werden. *Abbildung 3* zeigt die entsprechende Funktion als gestrichelte Linie.

Abbildung 3: Kumulierte ehedauerspezifische Scheidungsziffern (multipliziert mit 1000)



VIII. Scheidungen und Todesfälle

In Abschnitt IV wurde gezeigt, dass es sehr große Differenzen zwischen der Pseudo-Survivorfunktion $\bar{G}^a$ und der Anteilsfunktion H^a geben kann. Wie groß die Unterschiede sind, hängt insbesondere von den Häufigkeiten ab, mit denen die unterschiedlichen Folgezustände auftreten, und von ihrer zeitlichen Lagerung in der Prozesszeitachse. In diesem Abschnitt untersuchen wir das Problem bei Ehedauern, wobei die Ehen entweder durch eine Scheidung oder durch den Tod eines Ehepartners enden können. Zuerst konstruieren wir mit Hilfe von Periodendaten der amtlichen Statistik einen hypothetischen Entwicklungsprozess für Ehen, dann verwenden wir diese Daten, um die unterschiedlichen Funktionen zu vergleichen.

Datenkonstruktion. Wir beziehen uns auf Ehen, die 1999 in Deutschland geschlossen wurden, insgesamt 430674 (Statistisches Bundesamt 1999: 258). In der Quelle findet man auch eine Aufgliederung dieser Ehen nach dem Alter des Mannes und der Frau bei der Heirat. Es sei $n_{\tau,\tau'}$ die Anzahl der Ehen, die 1999 geschlossen wurden, bei denen der Mann bei der Heirat τ Jahre und die Frau τ' Jahre alt war. Im Weiteren beschränken wir uns auf diejenigen Ehen, bei denen die Altersangaben nicht gruppiert, sondern in Jahren ausgewiesen sind:

$$n := \sum_{\tau=18}^{64} \sum_{\tau'=18}^{54} n_{\tau,\tau'} = 414030.$$

Es sei n_t die Anzahl der Ehen, die bei einer Ehedauer von t Jahren noch bestehen. Anfangs ist $n_0 = n$. Ehen können durch Scheidungen oder durch den Tod eines Ehepartners enden. Die Scheidungen simulieren wir mit Hilfe der in *Tabelle 3* angegebenen Scheidungsziffern für 1999. Um den Simulationszeitraum über 25 Jahre hinaus ausdehnen zu können, wird für $t > 25$ eine konstante Ehescheidungs*rate* angenommen, die dem Wert für $t = 25$ entspricht.[1]

Um zu simulieren, dass Ehen auch durch den Tod eines Ehepartners enden können, verwenden wir die alters- und geschlechtsspezifischen Sterberaten des Jahres 1999:

$$\delta_\tau^m := \frac{\text{Anzahl der 1999 gestorbenen Männer im Alter } \tau}{\text{jahresdurchschnittliche Anzahl von Männern im Alter } \tau}$$

$$\delta_\tau^f := \frac{\text{Anzahl der 1999 gestorbenen Frauen im Alter } \tau}{\text{jahresdurchschnittliche Anzahl von Frauen im Alter } \tau}$$

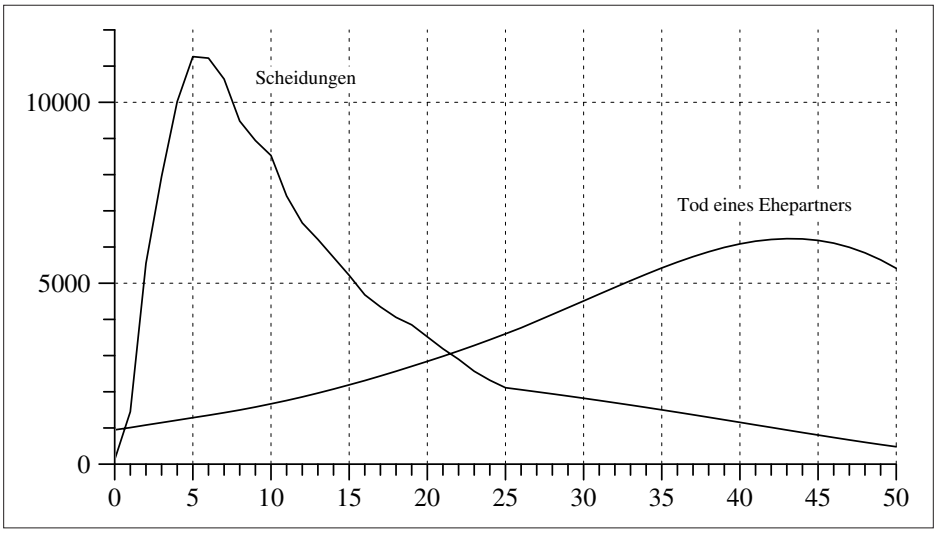
Da uns geeignete Daten nur bis zu einem Alter von 94 Jahren zur Verfügung stehen, nehmen wir an, dass die Sterberaten im Alter 95 den Wert 1 haben. In unserem simulierten Datensatz sterben also alle Personen spätestens im Alter von 95 Jahren. Die erforderlichen Daten haben wir den Segmenten 685 und 1124-26 der STATIS Datenbank des Statistischen Bundesamts entnommen.

1 Für unsere Zwecke wäre es zwar besser, von vornherein ehedauerspezifische Scheidungsraten zu verwenden; solche Scheidungsraten werden jedoch vom Statistischen Bundesamt nicht publiziert, da es für die Gliederung der Ehen nach ihrer Dauer keine Periodendaten gibt.

Tabelle 4: Hypothetischer Entwicklungsprozess von 414030 Ehen, die 1999 geschlossen wurden (zu den Annahmen vgl. man die Erläuterungen im Text)

t	n_t	w_t^a	w_t^b	t	n_t	w_t^a	w_t^b
0	414030	124	947	26	210791	2056	3767
1	412959	1449	1010	27	204968	1999	3952
2	410500	5548	1080	28	199017	1941	4137
3	403872	7949	1147	29	192940	1882	4323
4	394775	10020	1215	30	186735	1821	4510
5	383540	11262	1283	31	180403	1759	4698
6	370995	11220	1350	32	173946	1696	4885
7	358425	10641	1421	33	167365	1632	5069
8	346364	9481	1496	34	160664	1567	5248
9	335387	8943	1578	35	153849	1500	5421
10	324865	8529	1667	36	146928	1433	5583
11	314670	7411	1758	37	139913	1365	5733
12	305501	6666	1857	38	132816	1295	5868
13	296978	6210	1963	39	125652	1225	5989
14	288805	5714	2074	40	118438	1155	6086
15	281017	5217	2191	41	111196	1084	6162
16	273610	4679	2309	42	103950	1014	6209
17	266622	4347	2436	43	96728	943	6231
18	259839	4057	2567	44	89554	873	6224
19	253214	3850	2703	45	82456	804	6183
20	246661	3519	2840	46	75469	736	6108
21	240302	3188	2981	47	68625	669	5993
22	234133	2898	3128	48	61962	604	5840
23	228107	2567	3280	49	55519	541	5646
24	222260	2319	3438	50	49331	481	5411
25	216504	2112	3601	51	43439		

Abbildung 4: Absolute Häufigkeiten von Ehen, die durch eine Scheidung oder durch den Tod eines Ehepartners beendet wurden, entsprechend den Daten in *Tabelle 4*



Es sei nun $n_{\tau,\tau',t}$ die Anzahl der Ehen bei einer Ehedauer von t Jahren, bei denen bei der Heirat der Mann τ und die Frau τ' Jahre alt gewesen sind. Von diesen Ehen enden

$$w^b_{\tau,\tau',t} := n_{\tau,\tau',t}\left(\delta^m_{\tau+t} + \delta^f_{\tau'+t} - \delta^m_{\tau+t}\delta^f_{\tau'+t}\right)$$

durch den Tod mindestens eines Ehepartners. Insgesamt enden also bei einer Ehedauer von t Jahren

$$w^b_t := \sum_{\tau=18}^{64}\sum_{\tau'=18}^{54} w^b_{\tau,\tau',t}$$

Ehen durch den Tod eines Ehepartners. Außerdem treten Scheidungen auf. Wenn $t \leq 25$ ist, gibt es

$$w^a_{\tau,\tau',t} := n_{\tau,\tau',0}\, p^a_{1999}(t)$$

Scheidungen, wobei $p^a_{1999}(t)$ die Scheidungsziffer für 1999 bei einer Ehedauer von t Jahren ist; und wenn $t > 25$ ist, gibt es

$$w^a_{\tau,\tau',t} := n_{\tau,\tau',t}\,\frac{w^a_{\tau,\tau',25}}{n_{\tau,\tau',25}}$$

Scheidungen. Somit kann man den Bestand der Ehen auf folgende Weise fortschreiben:

$$n_{\tau,\tau',t+1} = n_{\tau,\tau',t} - w^a_{\tau,\tau',t} - w^b_{\tau,\tau',t}$$

Abbildung 5: Anteilsfunktion H^a (durchgezogene Linie) und Pseudo-Survivorfunktion $1 - \tilde{G}^a$ (gestrichelte Linie) für die Daten in *Tabelle 4*

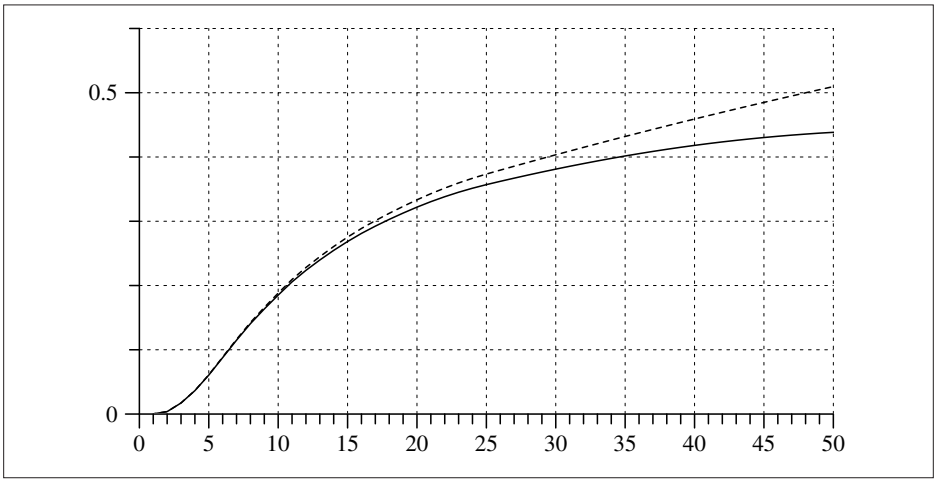


Abbildung 6: Anteilsfunktion H^b (durchgezogene Linie) und Pseudo-Survivorfunktion $1 - \tilde{G}^b$ (gestrichelte Linie) für die Daten in *Tabelle 4*

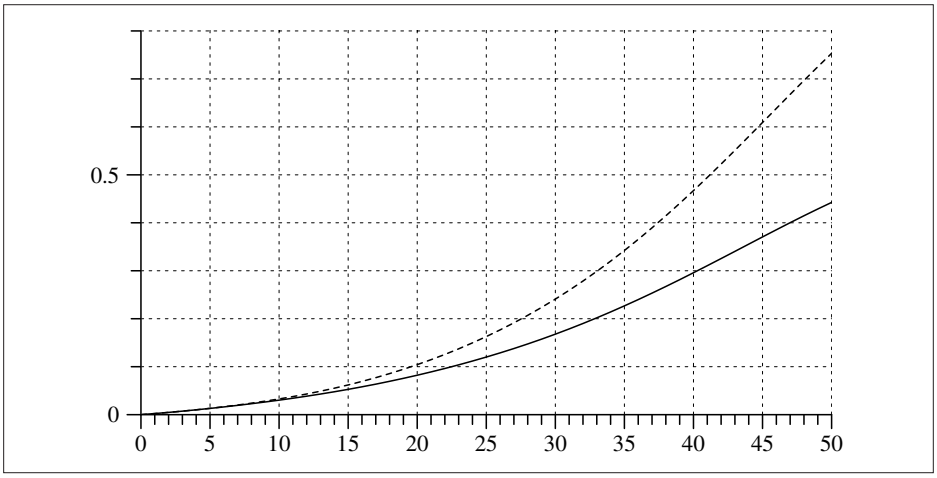
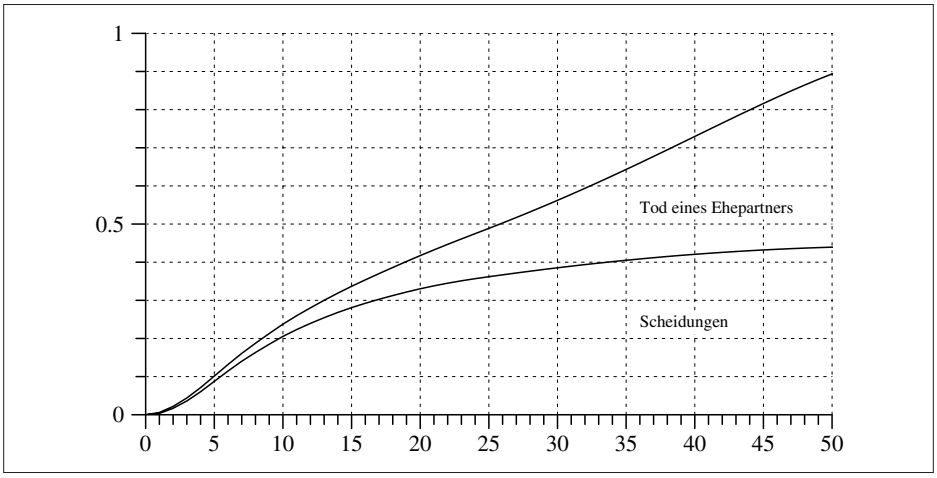


Abbildung 7: Anteilsfunktionen H^a (untere Linie) und $H^a + H^b$ (obere Linie) für die Daten in Tabelle 4



Durch Aggregation gewinnt man die in *Tabelle 4* angegebenen Zahlen. *Abbildung 4* zeigt, wie sich die Häufigkeiten von Ehelösungen durch Scheidung bzw. durch den Tod eines Ehepartners bis zu einer Ehedauer von 50 Jahren verteilen.

Anteils- und Pseudo-Survivorfunktionen. Die Daten in *Tabelle 4* können verwendet werden, um Anteils- und Pseudo-Survivorfunktionen zu berechnen. *Abbildung 5* zeigt als durchgezogene Linie die Funktion H^a und als gestrichelte Linie die Funktion $1 - \tilde{G}^a$. Man erkennt, dass die Unterschiede bei Ehedauern bis zu 20 Jahren nur klein sind,

dann jedoch mit zunehmender Ehedauer immer größer werden. Die Pseudo-Survivorfunktion suggeriert sogar die Vorstellung, dass bis zu einer Ehedauer von etwa 47 Jahren die Hälfte aller Ehen durch eine Scheidung beendet wurde.

Noch deutlicher werden die Diskrepanzen, wenn man sich für diejenigen Ehen interessiert, die nicht durch eine Scheidung sondern durch den Tod eines Ehepartners enden. *Abbildung 6* zeigt wiederum die Anteilsfunktion H^b und die Pseudo-Survivorfunktion $1 - \tilde{G}^b$.

Da Anteilsfunktionen im Unterschied zu Pseudo-Survivorfunktionen additiv sind, können sie sinnvoll zusammengefasst werden. Das wird in *Abbildung 7* illustriert. Die Abbildung zeigt, wie sich in Abhängigkeit von der Ehedauer die Gesamtheit der jeweils bisherigen Ehelösungen auf Scheidungen und den Tod eines Ehepartners verteilt.

IX. Schlussbemerkungen

Der Vorschlag, zur Darstellung von Episoden mit mehreren möglichen Folgezuständen Anteilsfunktionen (für Übergänge in neue Zustände) zu verwenden, setzt als Kontext ein Interesse an deskriptiver Statistik voraus. Dies gilt auch für unsere Kritik an der Verwendung von Pseudo-Survivorfunktionen, wenn diese so interpretiert werden, als ob sie Informationen über Anteile der Wechsel in neue Zustände lieferten.

Ich möchte betonen, dass sich diese Kritik nicht auf Regressionsmodelle für konkurrierende Risiken bezieht. Werden solche Modelle als Übergangsratenmodelle konzipiert, setzen sie bei zustandsspezifischen Übergangsraten an, die als Funktionen von Kovariablen betrachtet werden. Bei dieser Betrachtungsweise können Übergangsraten als zeitlich lokal definierte Sachverhalte angenommen werden, ohne für die begriffliche Explikation einen Verweis auf Verweildauern voraussetzen zu müssen. Schließlich müssen jedoch auch Übergangsratenmodelle interpretiert werden, und dafür ist es insbesondere erforderlich, sich ein Bild der durch die Modelle implizierten Prozesse zu machen. Auch dafür können Anteilsfunktionen sinnvoll verwendet werden.

Ebenfalls unabhängig von der bisher vorgetragenen Kritik kann man über den gelegentlich geäußerten Vorschlag nachdenken, dass man eine Pseudo-Survivorfunktion $\tilde{G}^a$ als eine gewöhnliche Survivorfunktion für eine hypothetische Situation interpretieren kann, in der nur ein Übergang in den Folgezustand a möglich ist. Man sollte sich jedoch bewusst machen, dass es sich um eine spekulative Überlegung handelt, die durch die jeweils vorhandenen Daten und das Kaplan-Meier-Verfahren nicht gerechtfertigt werden kann.

Um das zu verdeutlichen, beziehe ich mich zunächst auf einen in der Literatur über konkurrierende Risiken häufiger verwendeten Modellansatz (man vgl. z.B. Keyfitz 1977: 48ff.). Dieser Modellansatz bezieht sich auf eine Maschine M, die so lange funktioniert, bis zum ersten Mal eine ihrer Komponenten $K_1,\ldots,K_m$ defekt wird. Nun wird angenommen, dass jeder Komponente K_j eine Lebensdauer T_j zugerechnet werden kann, so dass die Lebensdauer T der Maschine M durch das Minimum der Lebensdauern der Komponenten bestimmt wird. Die Fragestellung bezieht sich darauf, ob aus Werten einer Variablen (T,D), wobei D die Nummer der Komponente angibt, die zum Defekt der Maschine geführt hat, Schätzungen der komponentenspezifischen Lebensdauervariablen T_j gewonnen werden können.

Allgemein ist das nicht möglich, da meistens viele unterschiedliche Verteilungen der Variablen $T_1,...,T_m$ mit einer bestimmten Verteilung von (T,D) vereinbar sind. Eine spezielle Situation liegt vor, wenn man annehmen kann, dass $T_1,...,T_m$ stochastisch unabhängige Zufallsvariablen sind. In diesem Fall sind die Pseudo-Survivorfunktionen, die das von Daten für (T,D) ausgehende Kaplan-Meier-Verfahren liefert, näherungsweise mit den Survivorfunktionen für $T_1,...,T_m$ identisch.[2]

Aus zwei Gründen erscheint es mir jedoch fragwürdig, ob dieses Maschinenmodell konkurrierender Risiken auch für sozialwissenschaftliche Anwendungen einen sinnvollen formalen Rahmen liefern kann. Erstens kann bei vielen sozialwissenschaftlichen Anwendungen bezweifelt werden, ob sich die latenten Variablen $T_1,...,T_m$ überhaupt sinnvoll interpretieren lassen. Sie können ja fast immer nicht einmal mehr metaphorisch auf zumindest gedanklich isolierbare Komponenten einer Maschine bezogen werden, sondern verlangen die Vorstellung sozialer Situationen, in denen nur noch jeweils eine der möglichen Alternativen zur Verfügung steht. Zweitens ist bei den meisten sozialwissenschaftlichen Anwendungen unklar, was der für das Maschinenmodell vorausgesetzten Unabhängigkeit der latenten Variablen entsprechen könnte.

Tatsächlich ist bei den meisten sozialwissenschaftlichen Anwendungen nicht nur die Bedeutung dieser Unabhängigkeitsannahme unklar, sondern hat bereits die zugrunde liegende Fragestellung einen vollständig spekulativen Charakter. Zur Illustration beziehe ich mich noch einmal auf das in Abschnitt IV besprochene numerische Beispiel mit den beiden Alternativen (a) Übergang in eine neue Beschäftigung und (b) Verlassen des Arbeitsmarktes (einschließlich Sterbefälle). Mit etwas Phantasie kann man sich vielleicht eine Situation vorstellen, in der nur noch Übergänge in den Zustand (a) möglich sind. Welche Verteilung der Arbeitslosigkeitsdauern bis zum Erreichen einer neuen Beschäftigung ergäbe sich dann? Eine mögliche Antwort liefert die in Tabelle 2 berechnete Pseudo-Survivorfunktion $\tilde{G}^a$. Ebenso gut kann man jedoch beliebig viele andere Antworten konstruieren.

Zum Beispiel kann man auch folgendermaßen vorgehen: Es wird angenommen, dass Personen, die bisher in den Folgezustand (b) gewechselt haben, jetzt – da sie weder den Arbeitsmarkt verlassen noch sterben können – ebenfalls mit der Rate r^a in den Folgezustand a wechseln. Die Anzahl der Personen, die in der Zeitstelle t in den Folgezustand a wechseln, ist dann insgesamt $w_t^* := w_t^a + w_t^b\, r^a(t)$. Dementsprechend ist die Anzahl der Personen, die bis zur Zeitstelle t noch nicht in den Folgezustand a gewechselt haben: $n_t^* = n_{t-1}^* - w_{t-1}^*$ (wobei $n_0^* = n_0 = n$ ist). Also erhält man die Ratenfunktion $r^*(t) := w_t^* / n_t^*$ und die Survivorfunktion $G^*(t) := n_t^* / n$. *Tabelle 5* zeigt die Werte für unser Beispiel, und man erkennt, dass sie sich von den Funktionen r^a und $\tilde{G}^a$, die das Kaplan-Meier-Verfahren liefert, deutlich unterscheiden.

Es erscheint allerdings kaum sinnvoll, jetzt zu fragen, welches der beiden Rechenverfahren „richtiger" wäre. Denn die jeweils berechneten Funktionen beziehen sich (bei der hier vorausgesetzten Interpretation) auf eine fiktive Situation, in der nur noch

2 Ohne weiteres gilt dies nur, wenn man von stetigen Zufallsvariablen ausgeht. Die Komplikationen, die bei diskreten Verweildauervariablen auftreten, sind jedoch für den gegenwärtigen Gedankengang nicht relevant.

Tabelle 5: Berechnung der fiktiven Survivorfunktion $G^*(t)$ aus den Daten in *Tabelle 1*

t	n_t	w_t^a	w_t^b	w_t^*	n_t^*	$r^*(t)$	$G^*(t)$
0	2500	100	50	102.0	2500.0	0.0408	1.0000
1	2350	100	50	102.1	2398.0	0.0426	0.9592
2	2200	100	50	102.8	2295.9	0.0448	0.9183
3	2050	100	50	102.4	2193.1	0.0467	0.8772
4	1900	80	60	82.5	2090.7	0.0395	0.8363
5	1760	80	60	82.7	2008.1	0.0412	0.8032
6	1620	80	60	83.0	1925.4	0.0431	0.7702
7	1480	80	60	83.3	1842.4	0.0452	0.7370
8	1340	60	70	63.1	1759.2	0.0359	0.7037
9	1210	60	70	63.5	1696.0	0.0374	0.6784
10	1080	60	70	63.9	1632.6	0.0391	0.6530
11	950	60	70	64.4	1568.7	0.0411	0.6275
12	820	40	80	43.9	1504.3	0.0292	0.6017
13	700	40	80	44.6	1460.4	0.0305	0.5841
14	580	40	80	45.5	1415.8	0.0322	0.5663
15	460	40	80	47.0	1370.3	0.0343	0.5481
16	340	20	80	24.7	1323.3	0.0187	0.5293
17	240	20	90	27.5	1298.6	0.0212	0.5194
18	130	20	90	33.8	1271.1	0.0266	0.5084
19	20	20	0	20.0	1237.3	0.0162	0.4949
20	0	0	0	0.0	1217.3	0.0000	0.4869

Übergänge in den Folgezustand *a* möglich sind, also arbeitslose Personen nicht mehr aus dem Arbeitsmarkt ausscheiden und nicht mehr sterben können. Wie sich Menschen in solchen fiktiven Situationen verhalten würden, kann schwerlich mithilfe von Daten vorausgesagt werden, die aus ganz andersartigen Situationen stammen.

Literatur

Blossfeld, Hans-Peter, *Alfred Hamerl*e und *Karl Ulrich Mayer*, 1986: Ereignisanalyse. Statistische Theorie und Anwendungen in den Sozialwissenschaften. Frankfurt a.M.: Campus.

Blossfeld, Hans-Peter, und *Götz Rohwer*, 2002: Techniques of Event History Modeling. Mahwah, NJ: Lawrence Erlbaum.

Diekmann, Andreas, und *Thomas Klein*, 1991: Bestimmungsgründe des Ehescheidungsrisikos. Kölner Zeitschrift für Soziologie und Sozialpsychologie 43, 271–290.

Esenwein-Rothe, Ingeborg, 1982: Einführung in die Demographie. Wiesbaden: Franz Steiner.

Keyfitz, Nathan, 1977: Applied Mathematical Demography. Berlin: Springer-Verlag.

Klein, Thomas, und *Johannes Kopp*, 1999. Scheidungsursachen aus soziologischer Sicht. Würzburg: Ergon.

Mueller, Ulrich, 1993: Bevölkerungsstatistik und Bevölkerungsdynamik. Berlin: de Gruyter.

Namboodiri, Krishnan, und *C.M. Suchindran*, 1987: Life Table Techniques and their Applications. New York: Academic Press.

Rohwer, Götz, und *Ulrich Pötter*, 2001: Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Smith, David P., 1992: Formal Demography. New York: Plenum Press.

Statistisches Bundesamt, 1990: Familien heute. Strukturen, Verläufe und Einstellungen. Ausgabe 1990. Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt, 1999: Fachserie 1, Reihe 1. Gebiet und Bevölkerung 1999. Stuttgart: Metzler-Poeschel.

KAUSALANALYSEN MIT SEPARATEN UND ZUSAMMENGEFASSTEN ZEITREIHEN*

Henriette Engelhardt

Zusammenfassung: Zeitreihen gewinnen als Datenmaterial für empirische Studien im Bereich der Sozialwissenschaften eine immer größere Bedeutung. Häufig werden nicht nur Zeitreihen verschiedener Merkmale zusammengeführt, sondern auch Zeitreihen verschiedener Untersuchungseinheiten. Die Datenstruktur dieser zusammengefassten Zeitreihen ähnelt der Struktur von Paneldaten. Während sich die klassischen Paneldaten durch viele Beobachtungseinheiten und wenige Zeitpunkte auszeichnen, sind zusammengefasste Zeitreihen durch relativ wenige Fälle und viele Zeitpunkte gekennzeichnet. In diesem Übersichtsartikel werden multiple Regressionsmodelle für separate und zusammengefasste Zeitreihen vorgestellt und verschiedene Modellvarianten betrachtet. Anschließend werden spezifische Schätzprobleme diskutiert, die sich primär aus spezifischen Zeitreiheneigenschaften ergeben. Als Illustration zu den theoretischen Ausführungen der Analyse zusammengefasster Zeitreihen dient die Analyse des Effektes von Frauenerwerbstätigkeitsraten auf die Fertilität in 22 OECD-Ländern im Zeitraum 1960–2000.

I. Einführung

Zeitreihen gewinnen als Datenmaterial für empirische Studien im Bereich der Sozialwissenschaften eine immer größere Bedeutung. Zeitreihendaten zeichnen sich dadurch aus, dass für eine Einheit (Person, Firma, Stadt, Land etc.) Informationen zu einem Merkmal über die Zeit zur Verfügung stehen. Für Kausalanalysen mit Zeitreihendaten werden zumindest zwei Zeitreihen von zwei verschiedenen Merkmalen miteinander in Beziehung gesetzt. Lange Zeit wurde für Kausalanalysen mit Zeitreihen das traditionelle Regressionsmodell verwendet. Dabei wurden die dynamische Struktur der Zeitreihen und die aus der univariaten Zeitreihen-Literatur bekannten Probleme systematisch ausgeblendet. So wurde etwa angenommen, dass die Eigenschaft der Nicht-Stationarität, welche bei den meisten ökonomischen Zeitreihen gegeben ist (da sie über die Zeit wachsen und keinen konstanten „stationären" Mittelwert haben), die empirischen Analysen nicht tangieren. In der jüngeren Zeitreihenliteratur werden die mit Nicht-Stationarität verbundenen Probleme der Scheinkorrelation intensiv diskutiert und Modelle für so genannte kointegrierte Zeitreihen entwickelt, die in einer langfristigen Gleichgewichtsbeziehung stehen.

Insbesondere im Rahmen international vergleichender Sozialforschung werden aber nicht nur Zeitreihen verschiedener Merkmale zusammengeführt, sondern darüber hin-

* Für wertvolle Hinweise und Kommentare danke ich Andreas Diekmann, Ben Jann und Maria Winkler-Dworak.

aus auch Zeitreihen verschiedener Einheiten. Die Datenstruktur dieser zusammengefassten Zeitreihen ähnelt der Struktur von Paneldaten. Während sich die klassischen Paneldaten durch viele Beobachtungseinheiten und wenige Zeitpunkte auszeichnen, sind zusammengefasste Zeitreihen durch relativ wenige Fälle und viele Zeitpunkte gekennzeichnet.[1] Entsprechend dieser Unterscheidung in der Datenstruktur haben sich in der Panelökonometrie in den letzten Jahren zwei größere Teilgebiete herauskristallisiert. Zum einen gibt es die Literatur, die sich auf Paneldaten mit einer großen Anzahl von (Querschnitts-)Einheiten und einer eher kleinen (konstanten) Anzahl von Zeitperioden konzentriert. Beim zweiten und jüngeren Strang der Panelökonometrie geht man zumeist von einer festen Anzahl Querschnittsbeobachtungen und einer wachsenden Zeitdimension aus. Die Art der Asymptotik, die zur Approximation der stochastischen Eigenschaften der Schätzverfahren verwendet wird, aber auch die Modellierung ist bei zusammengefassten Zeitreihen dem spezifischen Datentyp angepasst und greift Konzepte aus der separaten Zeitreihenanalyse auf.

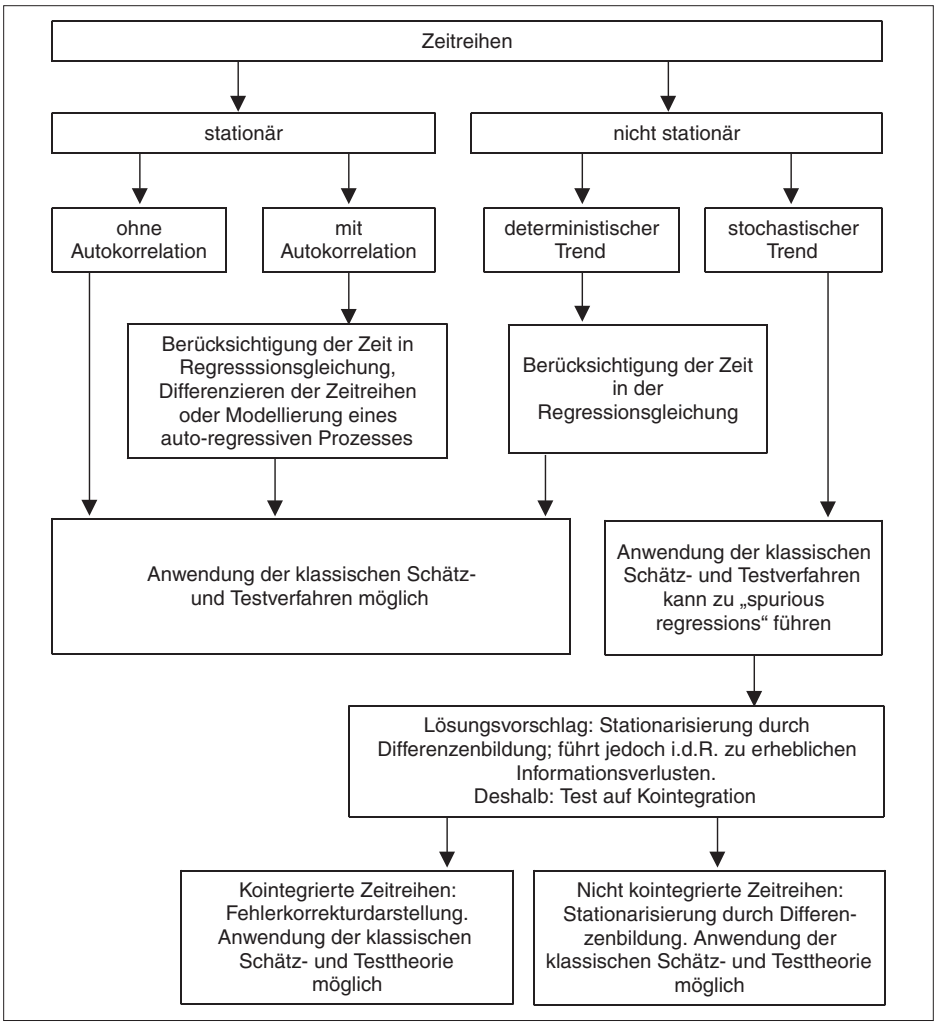
Während speziell in der Soziologie zu den Analysemethoden zusammengefasster Zeitreihen nur Publikationen älteren Datums vorliegen und damit diese Methoden offensichtlich derzeit nicht weiter verfolgt werden (Hibbs 1973; Hannan und Young 1977; Sayrs 1989), finden sich in der Politologie und neuerdings auch in der Kriminologie methodische Diskussionen (Stimson 1985; Hicks 1994; Beck und Katz 1995, 1996; Beck et al. 1998; Beck 2001a, 2001b; Kittel 1999; Podestà 2002, 2003; Kittel und Winner 2004; Plümper et al. 2004; Wilson und Butler 2004; Worall und Pratt 2004). Auf technisch anspruchsvollerem Niveau finden sich auch in der Ökonomik und Statistik einige wenige Übersichtsartikel (Mundlak 1978; Nuamah 1986; Maddala 1991, 1999; Hsiao und Sun 2000) sowie ein Monograph (Dielmann 1989). Der vorliegende Beitrag stellt nicht nur einen weiteren nicht-technischen Übersichtsartikel in diesem Gebiet dar, sondern ergänzt die zitierten Werke insofern, als hier auch neuere Entwicklungen aus dem Gebiet der separaten Zeitreihenanalyse aufgegriffen werden, die bei der Analyse zusammengefasster Zeitreihen zu beachten sind.

Ein Großteil der hier vorgestellten Methoden sind als Routinen in dem unter Soziologen weit verbreiteten Statistikprogrammpaket Stata enthalten oder werden von der Stata-Benutzergemeinde online zur Verfügung gestellt und sind ohne weitere Programmierkenntnisse anwendbar. Eine kurze Einführung in die Schätzung von zusammengefassten Zeitreihenmodellen mit Stata gibt Podestà (2002).

In diesem Artikel werden zunächst multiple Regressionsmodelle für separate und zusammengefasste Zeitreihen vorgestellt (Abschnitt II) und verschiedene Modellvarianten betrachtet (Abschnitt III). Anschließend werden spezifische Schätzprobleme diskutiert, die sich primär aus spezifischen Zeitreiheneigenschaften ergeben (Abschnitt IV). Die komplexe Diskussion der Zeitreiheneigenschaften lässt sich anhand des Überblicks in *Abbildung 1* verfolgen. Als Illustration zu den theoretischen Ausführungen der Analyse zusammengefasster Zeitreihen dient die Analyse des Effektes von Frauenerwerbstätigkeitsraten auf die Fertilität in 22 OECD-Ländern im Zeitraum 1960–2000 (Abschnitt V). Eine kurze Zusammenfassung mit einer graphischen Übersicht über die

1 In der Literatur finden sich für diesen Datentyp die synonymen Bezeichnungen „pooled time series“, „time-series cross-sectional data“, und „panel data“.

Abbildung 1: Zeitreiheneigenschaften und adäquate Modellspezifikationen



Zusammenhänge der diskutierten Zeitreiheneigenschaften und den Modellspezifikationen schließt diesen Beitrag (Abschnitt VI).

II. Multiple Regressionsmodelle

1. Separate Zeitreihen

In den Sozialwissenschaften werden theoretisch postulierte kausale Strukturen zumeist im Rahmen der linearen Regressionsmodelle untersucht; dies geschieht unabhängig davon, ob es sich bei den entsprechenden Merkmalen um zeitabhängige oder zeitunab-

hängige Variablen, um Mikro- oder Makrovariablen handelt.[2] Ausgangspunkt unserer Betrachtungen zur Kausalanalyse mit Zeitreihen ist deshalb – analog der Kausalanalyse mit Querschnittdaten – ein *statisches* lineares Regressionsmodell, das durch folgende Gleichung gegeben ist:

$$y_t = \alpha + \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_t, \quad t = 1,...,T. \tag{1}$$

Hierbei bezeichnet y_t die abhängige Variable und x_{kt} die k'te nicht stochastische unabhängige Variable, deren Werte vorgegeben sind, zum Zeitpunkt t. α und β_k sind zu schätzende Koeffizienten, und ε ist ein zeitlich variierender Störterm, für den folgende Annahmen gelten:

$$\begin{aligned} &E(\varepsilon_t) = 0, \\ &V(\varepsilon_t) = \sigma^2, \\ &Cov(\varepsilon_t, \varepsilon_s) = 0 \quad \text{für } t \neq s. \end{aligned} \tag{2}$$

D.h. im Mittel ist der Fehler Null, die Fehlervarianz ist konstant (Homoskedastizität), und die Fehler in zwei unterschiedlichen Zeitperioden sind unkorreliert (keine serielle Korrelation oder Autokorrelation). Sind diese Standardannahmen erfüllt, können die Parameter mit der Kleinstquadratemethode (OLS) konsistent, effizient und unverzerrt geschätzt werden.[3]

Zeitreihenmodelle leiden nun aber gelegentlich unter Heteroskedastizität (d.h. unterschiedliche Varianzen der Störgrößen) und in der Regel – im Gegensatz zu Modellen mit Querschnittsdaten – unter Autokorrelation. In diesen Fällen ist die einfache OLS-Schätzung nicht effizient. Die verschiedenen Strategien zum Aufdecken von und zum Umgang mit Heteroskedastizität sowie zur Beseitigung von Autokorrelation in Zeitreihen werden in Abschnitt III diskutiert.

2. Zusammengefasste Zeitreihen

Kausalanalyse mit zusammengefassten Zeitreihen ist ungleich komplizierter als mit einfachen Zeitreihen. Bei der Anwendung von klassischen Regressionsmodellen ist die kritische Annahme, dass alle Untersuchungseinheiten durch eine identische Regressionsgleichung zu allen Zeitpunkten beschrieben werden können. Diese Annahme stellt eine besondere Herausforderung an die richtige Spezifikation des Modells.

Unter der Annahme, dass alle Koeffizienten konstant sind und dass die Störterme Unterschiede über die Zeit und über die Untersuchungseinheiten erfassen, können zu-

2 Es ist natürlich nicht möglich, im vorliegenden Beitrag in die Grundlagen der Zeitreihenanalyse einzuführen. Einführungen in die univariate Zeitreihenanalyse bieten z.B. Chatfield (2004) und Enders (2004); zur multivariaten Zeitreihenanalyse siehe z.B. Ostrom (1990) und Harvey (1995).

3 Bei unverzerrten Schätzern entspricht der erwartete Wert des Koeffizienten dem wahren Wert. Ein Schätzer ist effizient, wenn seine Varianz kleiner ist als die jedes anderen Schätzers, und konsistent, wenn die Varianz mit steigender Stichprobengröße gegen Null konvergiert.

sammengefasste Zeitreihen analog zum Regressionsmodell für separate Zeitreihen beschrieben werden:

$$y_{it} = \alpha + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}, \tag{3}$$

wobei y_{it} den Wert der abhängigen Variablen für die $i$te Untersuchungseinheit zum Zeitpunkt t bezeichnet, und x_{kit} die entsprechenden Beobachtungswerte der k unabhängigen Variablen darstellen. α und β_k sind die zu schätzenden Regressionskoeffizienten, und ε_{it} bezeichnet einen Störterm. Gelten analog zum Modell mit separaten Zeitreihen (1) die Annahmen Homoskedastizität und keine Autokorrelation:

$$\begin{aligned} &E(\varepsilon_{it}) = 0, \\ &V(\varepsilon_{it}) = \sigma^2, \\ &Cov(\varepsilon_{it}, \varepsilon_{js}) = 0 \quad \text{für } i \neq j,\ t \neq s, \end{aligned} \tag{4}$$

so sind die Annahmen für das allgemeine lineare Modell erfüllt und das Modell kann mit OLS geschätzt werden. Diese so genannten Gauss-Markov-Annahmen sind in der Regel bei zusammengefassten Zeitreihen nicht erfüllt. Wie bereits bei den separaten Zeitreihenmodellen besteht insbesondere die Gefahr der Autokorrelation. So sind häufig entweder a) die Störgrößen der Untersuchungseinheiten (z.B. Regionen) zum gleichen Zeitpunkt miteinander korreliert, oder b) die Störgrößen einer Untersuchungseinheit sind zu verschiedenen Zeiten miteinander korreliert oder c) beides. In Abschnitt IV werden Verfahren zur Aufdeckung und Beseitigung serieller und räumlicher Korrelation vorgestellt.

Abgesehen von den Annahmen über die Störterme stellt sich natürlich auch die Frage, ob das Modell (3) die Daten angemessen beschreibt. In der Literatur wird eine Vielzahl verschiedener Regressionsmodelle diskutiert, deren Annahmen bezüglich der Konstanz der Parameter weniger restriktiv sind. Judge et al. (1985) diskutieren vier weitere Modelle, die sich danach unterscheiden, ob die Konstante und/oder der Steigungskoeffizient als konstant oder über Raum und Zeit variierend angenommen werden.

Das Modell mit konstanten Steigungskoeffizienten und regional variierender Konstante ergibt sich:

$$y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}. \tag{5}$$

Variiert die Konstante nicht nur regional, sondern auch zeitlich, resultiert:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_k x_{kit} + \varepsilon_{it}. \tag{6}$$

Unterscheiden sich darüber hinaus auch die Koeffizienten zwischen den Regionen, dann ist folgendes Modell von Interesse:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_{ki} x_{kit} + \varepsilon_{it}. \tag{7}$$

Variieren alle Parameter über Zeit und Raum, kann das zusammengefasste Zeitreihenmodell ganz allgemein notiert werden:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_{kit} x_{kit} + \varepsilon_{it} \quad t = 1,\ldots T, \quad i = 1,\ldots,N. \tag{8}$$

Die Modelle (5) bis (8) können weiter danach klassifiziert werden, ob sich die Konstante durch fixe Effekte (fixed effects) oder durch zufällige Effekte (random effects) in Raum und Zeit unterscheidet. Die Annahme fixer Effekte führt zu Modellen mit Dummy-Variablen und zu SURE-Modellen (seemingly unrelated regression models), während die Annahme zufälliger Effekte zu Fehlerkomponentenmodellen (error component models) und zu Swamy's Zufallskoeffizientenmodell (random coefficient model) führt. Da im Rahmen der vorliegenden Arbeit nicht alle Modellvarianten diskutiert werden können, verweisen wir an dieser Stelle auf Judge et al. (1985), Dielmann (1989) und Sayrs (1989). Auf das Konzept zufälliger und fixer Effekte kommen wir im folgenden Abschnitt zurück.

III. Modellvarianten

1. Zeitliche und räumliche Heterogenität

Zur Erfassung von räumlicher Heterogenität in Modellen mit zusammengefassten Zeitreihen werden häufig feste Regionaleffekte angenommen. Formell ergibt sich das Modell mit so genannten fixen Effekten (*fixed effects model*) durch:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_k x_{kit} + \upsilon_i + \varepsilon_{it}, \tag{9}$$

wobei angenommen wird, dass υ_i feste Parameter sind, die eventuell mit x_{it} korreliert sind. Solch ein Modell fokussiert auf die Variation innerhalb einer Region, und die Koeffizienten repräsentieren durchschnittliche Regionseffekte über die Zeit. Dagegen erfassen Zeiteffekte g_t Veränderungen über die Zeit, die allen Regionen gemein sind. Werden die Regions- und Zeitkonstanten kombiniert, resultiert ein Modell, das räumliche und zeitliche Heterogenität explizit einschließt:

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_k x_{kit} + \upsilon_i + \gamma_t + \varepsilon_{it}. \tag{10}$$

Kann angenommen werden, dass die unbeobachtete räumliche und zeitliche Heterogenität das Resultat eines Zufallsprozesses und unkorreliert mit den berücksichtigten unabhängigen Variablen ist, dann ergibt sich das Modell mit zufälligen Effekten (*random effects model*). Die zentrale Unterscheidung zwischen dem Fixed-Effects- und dem Ran-

dom-Effects-Modell ist danach, ob der Fehlerterm eine Komponente enthält, die mit den Regressoren korreliert sein könnte (fixed effects) oder nicht (random effects).

Ob in der empirischen Analyse ein Fixed-Effects- oder ein Random-Effects-Modell verwendet werden sollte, ist sowohl eine substantielle als auch eine statistische Frage. Gibt es keinen substantiellen Grund, eine Korrelation zwischen den unbeobachteten regionsspezifischen Zufallseffekten und den Regressoren anzunehmen, dann ist das Random-Effects-Modell mächtiger und sparsamer in der Modellierung. Gibt es jedoch eine solche Korrelation, produziert das Random-Effects-Modell inkonsistente Schätzer, und das Fixed-Effects-Modell wäre die richtige Wahl. Der klassische Test für eine statistische Modellselektion ist der Spezifikationstest von Hausman (1978).

2. Prozessdynamik

Regressionsmodelle für separate und zusammengefasste Zeitreihen, die die dynamische Entwicklung der abhängigen oder unabhängigen Variablen in Form von Zeitverzögerungen explizit berücksichtigen, werden *dynamisch* genannt.[4] Die Berücksichtigung zeitverzögerter Variablen kann unter Umständen aus psychologischen Gründen (Personen ändern z.B. nicht so schnell ihr Verhalten), aus technologischen Gründen (z.B. Anpassungen von Firmen an Preisänderungen brauchen Zeit) oder institutionellen Gründen (rechtliche Grundlagen verhindern etwa eine unmittelbare Anpassung der Arbeitskräfte an sich verändernde Nachfragebedingungen) theoretisch gerechtfertigt sein (Gujarati 1995). Die gängigsten Beispiele für dynamische Modelle sind das unrestringierte *Distributed-Lag*-Modell mit verzögerter unabhängiger Variable (hier für zusammengefasste Zeitreihen):

$$y_{it} = \alpha + \beta_0 x_{it} + \beta_1 x_{it-1} + \beta_2 x_{it-2} + \ldots + \beta_q x_{it-q} + \varepsilon_{it}, \tag{11}$$

oder das *autoregressive Modell* mit einer zeitverzögerten abhängigen Variablen:

$$y_{it} = \alpha + \beta x_{it} + \gamma_1 y_{it-1} + \gamma_2 y_{it-2} + \ldots + \gamma_p y_{it-p} + \varepsilon_{it}, \tag{12}$$

wobei p und q die Anzahl der Verzögerungen bezeichnen, die in der Regel empirisch bestimmt werden. Dem angelsächsischen Sprachgebrauch folgend werden die zeitverzögerten Variablen auch *Lag-Variablen* genannt. Schließlich können das Distributed-Lag-Modell und das Modell mit verzögerter unabhängiger Variablen zu einem *autoregressiven Distributed-Lag*-Modell kombiniert werden:

$$y_{it} = \alpha + \beta_0 x_{it} + \beta_1 x_{it-1} + \beta_q x_{it-q} + \gamma_1 y_{it-1} + \ldots + \gamma_p y_{it-p} + \varepsilon_{it}. \tag{13}$$

Generell können die Koeffizienten dynamischer Modelle als kurzfristige und langfristige Effekte der jeweiligen Variablen auf die abhängige Variable interpretiert werden. Zum Beispiel wird β_0 (in den Modellen (11) und (13)) als kurzfristiger Effekt bezeichnet, da er die Veränderung von y durch die Änderung von x in der gleichen Periode angibt. $\beta_0 + \beta_1$ gibt die Veränderung von y in der nächsten Periode an, $\beta_0 + \beta_1 + \beta_2$

4 Einige Autoren (z.B. Greene 2003; Gujarati 1995) bezeichnen nur Modelle mit verzögerten abhängigen Variablen als dynamisch.

in der folgenden usw. Der langfristige Effekte von x auf y ergibt sich durch $\beta_0 + \beta_1 + \beta_2 + ... + \beta_q$. Eine grafische Veranschaulichung zu der Interpretation von zeitverzögerten Regressoren findet sich z.B. in Gujarati (1995).

Eine OLS-Schätzung dynamischer Modelle stößt allerdings auf eine wesentliche Schwierigkeit: Da es sich bei den Lag-Variablen letztlich um ein und dieselbe Variable handelt und diese oftmals einen bestimmten Trend aufweist (z.B. stetig ansteigend), kann es in dynamischen Modellen leicht zu hoher *Multikollinearität* kommen. Dieses Problem lässt sich am einfachsten durch das Weglassen einzelner Lag-Variablen verringern. Unklar ist jedoch, welche das sein sollten. Ausgehend von Modell (11) hat Koyck (1954) eine elegante Lösungsstrategie zur Reduzierung der Anzahl zu schätzender Parameter vorgeschlagen, ohne dabei Variablen zu eliminieren. Koyck unterstellt mit der geometrischen Verteilung einen funktionalen Zusammenhang zwischen den Parametern β_q. Zur Schätzung des Koyck-Modells sowie verwandter Modelle wie das partielle Anpassungsmodell und das Modell adaptiver Erwartungen siehe z.B. Gujarati (1995).

IV. Schätzprobleme

1. Heteroskedastizität

Heteroskedastizität in Regressionsmodellen mit Zeitreihen wird traditionellerweise im Zusammenhang mit zusammengefassten Zeitreihenmodellen unter dem Stichwort *Panel-Heteroskedastizät* diskutiert, und erst in jüngster Zeit auch im Rahmen separater Zeitreihenmodelle. Im letzteren Fall interessiert vor allem die dynamische Form der Heteroskedastizität, zu deren Modellierung Engle das Modell der bedingen autoregressiven Heteroskedastizität (autoregressive conditional heteroscedasticity, ARCH) und Bollerslev das Modell der generalisierten bedingten autoregressiven Heteroskedastizität (generalized ARCH, GARCH) entwickelt haben. Im vorliegenden Aufsatz kann aus Platzgründen nicht auf diese Ansätze eingegangen werden; wir verweisen hier z.B. auf Engle (2001) und Enders (2004).

Generell liefert OLS im Vorliegen von (Panel-)Heteroskedastizität zwar unverzerrte Schätzer von β, diese sind aber nicht effizient. Damit werden auch die gewöhnlichen Standardfehler, die t-Werte und F-Statistiken verzerrt geschätzt. In der ökonometrischen Literatur finden sich jedoch Schätzverfahren, die eine größere Verlässlichkeit aufweisen als OLS, d.h. effizienter sind und unverzerrte Standardfehler liefern (z.B. Baltagi 2001; Greene 2003; Hsiao 2003).

Zur Aufdeckung von Heteroskedastizität in Zeitreihen gibt es verschiedene Methoden. Die einfachste Diagnosemöglichkeit besteht darin, die Punktwolke zu betrachten. In einem Modell mit mehr als zwei unabhängigen Variablen ist eine so einfache graphische Analyse kaum mehr praktikabel. Neben der Betrachtung von Residuenplots ist es dann zweckmäßig, analytische Verfahren einzusetzen, mit deren Hilfe Heteroskedastizität diagnostiziert werden kann. Die wohl gebräuchlichsten Verfahren für gewöhnliche OLS-Regressionen sind der Breusch-Pagan-Test, der Goldfeld-Quandt-Test oder der White-Test (siehe z.B. Greene 2003).

Die Methoden zum Umgang mit Heteroskedastizität im Fall von zusammengefassten Zeitreihen sind weniger klar (für einen Überblick siehe Beck und Katz 1995, 1996). Die am meisten verbreitete Methode ist wohl, die Daten mit der Standardabweichung σ jener Variablen z zu gewichten, die für die Heteroskedastizität verantwortlich gemacht wird:

$$\frac{y_{it}}{\sigma_{z_{it}}} = \alpha \frac{1}{\sigma_{z_{it}}} + \beta \frac{x_{it}}{\sigma_{z_{it}}} + \frac{\varepsilon_{it}}{\sigma_{z_{it}}} \; . \tag{14}$$

Diese Methode zur Korrektur von Heteroskedastizität wird *weighted least squares* (WLS) genannt. Problematisch an dieser Methode ist, dass man nicht nur eine genaue Idee darüber braucht, welche Faktoren für den Prozess der Korrelation der Störterme verantwortlich ist, sondern auch über die entsprechenden Daten verfügen muss.

Anstelle einer Gewichtung der Daten schlagen Beck und Katz (1995) deshalb vor, die Standardfehler der Koeffizienten zu korrigieren. Diese *panel-corrected standard errors* (PCSEs) werden dem Prinzip nach kalkuliert wie die Heteroskedasitzitäts-korrigierten Standardfehler mit Querschnittdaten nach White (1980). In Simulationsstudien zeigen Beck und Katz (1995), dass panelkorrigierte Standardfehler für T > 15 sehr präzise sind. Das ist auch dann der Fall, wenn der Störterm in (4) die Gauss-Markov-Annahmen erfüllt. Damit kann durch die Korrektur der Standardfehler auch im Falle von Homoskedastizität zumindest kein Schaden angerichtet werden.

2. Serielle Autokorrelation

Die Annahme, dass die Fehler über die Zeit nicht mit sich selber korrelieren, ist spezifisch für Regressionen mit separaten und zusammengefassten Zeitreihen. Bei Modellen für Querschnittsdaten wird üblicherweise angenommen, dass die in der Regressionsanalyse inkludierten Einheiten auf einer Zufallsauswahl beruhen, und somit unabhängig voneinander sind. Eine beobachtete Zeitreihe wird dagegen als Realisation x_t einer Folge von Zufallsvariablen X_t (synonym: eines stochastischen Prozesses) betrachtet. Die zeitlich geordneten Messungen von X sind in der Regel nicht Realisationen unabhängiger Zufallsexperimente, sondern korrelieren miteinander. Diese Autokorrelationen der einzelnen Zeitreihen beeinflussen die Korrelation zwischen den unterschiedlichen Zeitreihen. Darüber hinaus beinhalten die Residuen Messfehler und nicht berücksichtigte Einflussgrößen, die ebenfalls seriell korrelieren, so dass sie in der Regel nicht als unabhängig voneinander gelten können. Bei Vorliegen von autokorrelierten Störtermen muss in irgendeiner Form dafür Rechnung getragen werden, dass die OLS-Schätzer der Koeffizienten ineffizient sind.

Um Autokorrelation in Zeitreihendaten zu diagnostizieren, kann neben grafischen Analysemethoden ein Durbin-Watson-Test verwendet werden. Dieser Test basiert auf einer Teststatistik, welche sich aus den Residuen $\hat{\varepsilon}_t = y_t - \hat{y}_t$ der OLS-Schätzung wie folgt ergibt:

$$d = \frac{\sum_{t=s+1}^{T} \left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-s}\right)^2}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2} . \tag{15}$$

Der Wert von d ist ungefähr gleich 2 minus zwei Mal die Korrelation von ε_t und ε_{t-s} (Greene 2003). Sind die Fehler unkorreliert, ist der Wert von d nahe dem Wert 2. Ist d wesentlich kleiner (größer) als 2, ist dies ein Zeichen von positiver (negativer) Autokorrelation. Tabellen der Verteilung von d finden sich in vielen ökonometrischen Lehrbüchern, z.B. in Pindyck und Rubinfeld (1981). In der praktischen Anwendung sollte beachtet werden, dass die d-Statistik ein Korrelationsmaß darstellt, das von der Wahl der unabhängigen Variablen im Regressionsmodell abhängig ist.

Andere Tests basieren auf der so genannten Autokorrelationsfunktion ρ. Der geschätzte Wert der Funktion am „Lag" s ergibt sich gemäß der Formel für den Produkt-Moment-Korrelationskoeffizienten durch:

$$\hat{\rho}_s = \frac{Cov\left(\varepsilon_t, \varepsilon_{t-s}\right)}{V\left(\varepsilon_t\right)^{1/2} V\left(\varepsilon_{t-s}\right)^{1/2}} , \quad s = 1,2,... \tag{16}$$

Die Hypothese, dass alle $\hat{\rho}_s$ gemeinsam gleich Null sind, kann mittels der Q-Statistik von Box und Pierce,

$$Q = n\sum_{k=1}^{m} \hat{\rho}_k^2 \tag{17}$$

oder der Ljung-Box (LB) Statistik

$$LB = n(n+2)\sum_{k=1}^{m} \left(\frac{\hat{\rho}_k^2}{n-k}\right) \tag{18}$$

getestet werden, wobei n die Stichprobengröße und m die Anzahl der Lags bezeichnet. Beide Test-Statistiken sind in großen Stichproben approximativ χ^2-verteilt mit m Freiheitsgraden. Bei kleineren Stichproben ist die LB-Statistik mächtiger als die Q-Statistik (Gujarati 1995: 714).

Der Durbin-Watson-Test für Autokorrelation erster Ordnung in den Residuen wurde von Bhargava et al. (1982) für die Anwendung auf balancierte Panel-Daten modifiziert, bei denen für alle Einheiten für alle Zeitpunkte Beobachtungen vorliegen. Baltagi und Wu (1999) haben diesen Test auf den Fall von unbalancierten Panels, bei denen nicht für alle Einheiten und alle Zeitpunkte Beobachtungen vorliegen, und nicht äquidistanten Daten angepasst.

Die Ursachen der Autokorrelation sind vielfältig und resultieren in der Regel aus einer Fehlspezifikation des Modells. Zur Beseitigung der Autokorrelation werden in der Literatur für *stationäre* Zeitreihen verschiedene Lösungsansätze diskutiert. Stationäre Zeitreihen haben im Gegensatz zu nicht-stationären Reihen (siehe hierzu weiter unten) einen zeitunabhängigen Erwartungswert ($E(X_t) = k$), eine zeitunabhängige Varianz

($V(X_t) < \infty$) und eine nur vom Zeitabstand der betrachteten Variablen abhängende Kovarianz ($Cov(\varepsilon_t,\varepsilon_{t+s}) = Cov(\varepsilon_t,\varepsilon_{t-s})$).[5]

Gegen Autokorrelation hilft gelegentlich das Einführen eines *deterministischen Trends* (z.B. mit Hilfe einer Polynomfunktion, Exponentialfunktion, oder logistischen Funktion) als zusätzliche „Kontrollvariable" in der Regressionsgleichung (hier am Beispiel für zusammengefassten Zeitreihen):

$$y_{it} = \alpha + \beta x_{it} + f(t) + \varepsilon_{it}. \tag{19}$$

Die Trendfunktion kann mit OLS bestimmt werden, indem t = 1,2,...,T als Regressor in Gleichung (1) aufgenommen wird. Ein Plot der geschätzten Koeffizienten kann dann Aufschluss über den funktionalen Verlauf des Trends geben.

Zur Beseitigung eines linearen Trends kann alternativ der Zusammenhang nicht für die ursprünglichen Niveau-Variablen, sondern für die Änderungsbeträge $\Delta y_{it} = y_{it} - y_{it-1}$ (t = 2,3,...,T,i= 1,...,N), die sich von Zeitpunkt zu Zeitpunkt ergeben, modelliert werden:

$$\Delta y_{it} = \alpha + \beta \Delta x_{it} + \Delta \varepsilon_{it}. \tag{20}$$

Das Differenzieren der Zeitreihen wird häufig als Allheilmittel gegen Autokorrelation gesehen. Es ist jedoch zu beachten, dass sich die Interpretation der Koeffizienten durch das Differenzieren ändert, was unter Umständen theoretisch nicht erwünscht ist.

Insbesondere bei ökonomischen Zeitreihen ist der Fall eines *stochastischen Trends* häufiger als der eines deterministischen Trends. Zur Illustration eines stochastischen Trends wird eine Zeitreihe als Folge von Zufallsvariablen aufgefasst. Die häufigste Korrektur bei Vorliegen eines solchen Trends findet statt, indem für die Residuen ein *autoregressiver Prozess* erster Ordnung (AR(1)-Prozess) modelliert wird:

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t. \tag{21}$$

Hierbei bezeichnet v_t unabhängige, identisch verteilte Störterme mit Mittelwert Null und ρ den so genannten Autokorrelationsparameter, der in absoluten Werten kleiner als 1 ist.

Autoregressive Prozesse höherer Ordnung (AR(p)-Prozesse, wobei p für das höchste Niveau von autoregressiven Einflüssen steht) sind direkte Generalisierungen des AR(1)-Modells. Z.B. ergibt sich das autoregressive Modell zweiter Ordnung (AR(2)) durch:[6]

$$\varepsilon_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + v_t. \tag{22}$$

Im Gegensatz dazu wird im *Moving-Average-Prozess* erster Ordnung (MA(1)-Modell) der aktuelle Fehler als Funktion eines „random shocks" der aktuellen und vorangehenden Periode modelliert:

$$\varepsilon_t = \psi v_{t-1} + v_t. \tag{23}$$

---

5 In der Zeitreihenliteratur heißen stochastische Prozesse mit diesen drei Eigenschaften *schwach stationär*. Wir bezeichnen in dieser Arbeit schwach stationäre Prozesse als stationär.

6 Greene (2003: 257) betrachtet die Autoregression erster Ordnung als ein vernünftiges Modell für „underlying processes that probaly, in truth, are impenetrably complex".

MA(q)-Prozesse höherer Ordnung sind ähnlich definiert (q steht für die Anzahl moving average Parameter). Schließlich können AR und MA-Terme in einem *autoregressiven Moving-Average*-Prozess (ARMA(p,q)) kombiniert werden. ARMA(1,1)-Fehler folgen etwa dem Prozess:[7]

$$\varepsilon_t = \rho\varepsilon_{t-1} + \psi v_{t-1} + v_t . \tag{24}$$

Die komplexeren Prozesse (22) bis (24) kommen meines Wissens bislang nur bei separaten Zeitreihen mit stochastischen Trends zum Einsatz. Bei zusammengefassten Zeitreihen ist seit dem einflussreichen Aufsatz von Beck und Katz (1995, 1996) die so genannte Prais-Winsten-Methode das dominierende Verfahren zur Parameterschätzung bei Vorliegen von serieller Autokorrelation. Dieses Verfahren hat die bis anhin vorherrschende Parks-Kmenta-Methode abgelöst (Parks 1967; Kmenta 1986). Zu näheren Einzelheiten sei auf die angegebene Literatur verwiesen. Ausgangspunkt der Prais-Winsten-Methode ist ein AR(1)-Panelmodell:

$$y_{it} = \alpha + \beta x_{it} + \rho\varepsilon_{it-1} + v_{it}. \tag{25}$$

Definiert man nach Prais und Winsten (1954):

$$y^*_{it} = \begin{bmatrix} \sqrt{1-\hat{\rho}^2}\, y_{i1} \\ y_{i2} - \hat{\rho} y_{i2} \\ y_{i3} - \hat{\rho} y_{i3} \\ \vdots \\ y_{iT} - \hat{\rho} y_{iT} \end{bmatrix}, \quad x^*_{it} = \begin{bmatrix} \sqrt{1-\hat{\rho}^2}\, x_{i1} \\ x_{i2} - \hat{\rho} x_{i2} \\ x_{i3} - \hat{\rho} x_{i3} \\ \vdots \\ x_{iT} - \hat{\rho} x_{iT} \end{bmatrix}, \tag{26}$$

kann Modell (25) geschrieben werden als:

$$y^*_{it} = \alpha + \beta x^*_{it} + \varepsilon_{it} . \tag{27}$$

Gleichung (27) kann nun mittels OLS konsistent, effizient und erwartungstreu geschätzt werden. Da der zur Transformation von y und x verwendete Parameter ρ allerdings unbekannt ist, wird in einem *Feasible-Generalised-Least-Squares*-Ansatz (FGLS) wie folgt vorgegangen. Zunächst wird Modell (3) mittels OLS geschätzt und die resultierenden Residuen zur Schätzung von ρ verwendet. Diese Schätzung von ρ wird dann zur Transformation der Daten gemäß (26) verwendet, und eine Schätzung von (27) liefert die gewünschten Koeffizienten. Im Vergleich zur Prais-Winsten-Methode vernachlässigt das Verfahren von Cochrane und Orcutt (1949) jeweils die erste Beobachtung in (26) (Greene 2003).

3. Räumliche Autokorrelation

Bei der Analyse zusammengefasster Zeitreihen liegt gelegentlich räumliche Autokorrelation vor, insbesondere dann, wenn die Analyseeinheiten benachbart sind. Handelt es sich bei den Untersuchungseinheiten um geographische Aggregate (Staaten, Länder

7 Zur Schätzung von MA und ARMA-Modellen siehe z.B. Maddala (2001).

oder Gemeinden), ist das Potential für eine Fehlerkorrelation zwischen den Einheiten besonders groß. Räumliche Autokorrelation liegt dann vor, wenn der Wert für eine Variable X am Ort j mit dem Wert der Variablen X am Ort i assoziiert ist.

Räumliche Autokorrelation ist generell dann kein Problem, wenn die Einheiten zufällig ausgewählt werden. Da aber häufig benachbarte Gebiete oder Regionen analysiert werden, muss das (potentielle) Problem räumlicher Autokorrelation bedacht werden.

Zur Aufdeckung räumlicher Autokorrelation sind verschiedene Techniken verfügbar. Die bekannteste Technik dürfte die Kalkulation von Moran's Korrelationskoeffizienten I sein, der wie folgt definiert ist:

$$I = \frac{\sum_i \sum_j w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2}, \tag{28}$$

wobei $\bar{x}$ den Mittelwert der Variable x bezeichnet und w_{ij} die Elemente einer räumlichen Gewichtungsmatrix bezeichnen (vgl. Getis und Ord 1992; Ord und Getis 1995).

Alternativ zu Moran's I schlagen Getis und Ord (1992) das G-Maß vor:

$$G = \frac{\sum_i \sum_j w_{ij}(d) x_i x_j}{\sum_i \sum_j x_i x_j}. \tag{29}$$

Das G-Maß unterscheidet sich von Moran's I durch die Berücksichtigung eines Distanz-Bandes d, ein Gebiet geographischen Interesses, in welchem räumliche Autokorrelation problematisch ist. Die räumliche Gewichtungs-Matrix w_{ij} ist eine binäre Matrix mit Einsen und Nullen, wobei eine Eins Einheiten innerhalb des Distanz-Bandes d indiziert.

Zum Umgang mit räumlicher Autokorrelation existieren zwei relativ einfache Methoden: das *Spatial-Disturbances*-Modell und das *Spatial-Effects*-Modell (Doreian 1980, 1981, 1982).[8] Im Modell mit räumlichen Störtermen sind die Effekte der räumlichen Autokorrelation eingebunden in ein Modell ähnlich dem autoregressiven Modell erster Ordnung. Anstelle der um eine Periode verzögerten Störterme wird der Störterm als Funktion des Produktes der Störterme mit einer binären N × N-Matrix mit räumlichen Gewichten beschrieben, wobei die Gewichte gleich Eins sind, wenn Region j mit Region i eine gemeinsame Grenze hat, 0 sonst. Die Gewichtungsmatrix soll dabei das Interaktionsmuster zwischen den Störtermen in Region i und j abbilden. Im Falle von zusammengefassten Zeitreihen ist diese Methode wegen der zusätzlichen Dimension Zeit sehr rechenaufwendig. So benötigen etwa 30 Regionen zu 10 Zeitpunkten eine räumliche Gewichtungsmatrix von 300 × 300. Meines Wissens gibt es bislang auch keine Software zur Schätzung des Modells mit räumlichen Störtermen für zusammengefasste Zeitreihen.

Im Modell mit räumlichen Effekten sind die Effekte der Autokorrelation in der abhängigen Variablen integriert. Anstelle einer Multiplikation der Störterme mit einer räumlichen Gewichtungsmatrix werden die Werte der abhängigen Variablen für alle benachbarten Einheiten gemittelt und in das Modell als zusätzliche Variable eingeführt.

8 Für weitere Modelle siehe z.B. Haining (2003).

Zur expliziten Berücksichtigung von verzögerter Autokorrelation können diese Werte auch zeitlich verzögert in das Modell eingeführt werden (*Spatial-Lag*-Modell).

Die Wahl zwischen dem Spatial-Disturbances-Modell und dem Spatial-Effects-Modell ist generell eher eine substantielle als eine empirische Frage. Der Forscher muss entscheiden, ob die abhängige Variable oder die Störterme autokorreliert sind. Sind die Werte der abhängigen Variablen korreliert, impliziert dies, dass zwischen den unabhängigen Variablen keine räumliche Beziehung besteht. Räumliche Autokorrelation in den Störtermen unterstellt, dass auch anderen Faktoren neben der abhängigen Variablen über die räumlichen Einheiten korreliert sind.

4. Nicht-Stationarität, Scheinregression und Einheitswurzeltests

Ein besonderer Fall der Autokorrelationsproblematik liegt vor, wenn Zeitreihen variable Trendverläufe aufweisen. Unter variablen Trends versteht man, dass Phasen mit starkem Wachstum von Phasen mit schwachem Wachstum oder gar Rückgängen abgelöst werden und diese Phasen unterschiedlich lang sind. Beispiele hierfür liefern Fertilitätsdaten, Heiratsraten, Scheidungsraten, Arbeitslosenraten, Einkommensdaten etc. Zeitreihen mit einem solchen Trend heißen *nicht-stationär*. Werden bei der Analyse struktureller Zusammenhänge mit verschiedenen Zeitreihen diese Trendkomponenten nicht adäquat berücksichtigt, erhält man in der Regel irreführende Resultate in den Regressionsstatistiken. Bereits Yule (1926) hat auf das Problem hingewiesen, dass in der (linearen) Regression einer Zeitreihenvariablen Y_t auf eine andere Zeitreihenvariable X_t eine statistisch signifikante Beziehung entdeckt werden kann, die sachlogisch nicht zu begründen ist. Ein illustratives Beispiel ist die Fertilitätsrate und die Anzahl Störche in einer Region. Beide Variablen haben einen abwärts gerichteten Trend. Werden nun die Anzahl Störche auf die Fertilitätsrate regressiert, so liefern einfache OLS-Schätzungen aufgrund der gemeinsamen Trends einen signifikanten, positiven Effekt. Granger und Newbold (1974) zeigen denn auch mittels Simulationsstudien, dass im Falle nicht-stationärer Zeitreihen häufig signifikante Koeffizienten und sehr hohe R^2-Werte resultieren (obwohl die Variablen in der Realität in keiner Beziehung stehen), einhergehend mit niedrigen Durbin-Watson-Statistiken. Dieses Problem der Scheinregression wurde von Granger und Newbold (1974) unter dem Begriff *spurious regression* in die Literatur eingeführt.

Nach den Resultaten von Granger und Newbold (1974) können also hohe R^2-Werte und geringe Werte des Durbin-Watson-Tests ein Hinweis auf Scheinregression geben. Philipps (1986) erklärt die Resultate von Granger und Newbold analytisch. Er zeigt, dass im Falle stationärer Zeitreihen die t-Werte der Koeffizienten von Gleichung (1) nicht der t-Verteilung folgen und dass die Werte mit steigenden T gegen unendlich gehen. Dies impliziert, dass für jedes beliebiges Signifikanzniveau die Ablehnungswahrscheinlichkeit der Nullhypothese H_0: $\beta = 0$ mit T steigt. Phillips (1986) zeigt auch, dass die Durbin-Watson-Statistik gegen Null konvergiert, wenn T gegen unendlich geht, während sie zu einem Wert ungleich Null konvergiert, wenn die Variablen in einer Beziehung stehen. In diesem Fall hilft die Durbin-Watson-Statistik, „wahre“ von Scheinregressionen zu unterscheiden. Zusammenfassend sind Scheinregressionen in der

Nicht-Stationarität der Variablen begründet. Das Problem kann nicht durch eine Vergrößerung von T, d.h. durch längere Zeitreihen, gelöst werden; längere Zeitreihen verschlimmern es im Gegenteil sogar.

Für die durch nicht-stationäre Variablen verursachten Probleme haben Ökonometriker drei Lösungsstrategien vorgeschlagen: Im ersten Verfahren wird ein Trend-Term als zusätzliche „Kontrollvariable" in die Regressionsgleichung eingeführt (vgl. Gleichung (19)). Üblich ist die Verwendung einer linearen Funktion der Zeit, aber auch quadratische Polynome oder stückweise lineare Funktionen werden häufig angewandt. Wird ein Prozess durch die Berücksichtigung eines Trends stationär, spricht man von einem *trend-stationären* Prozess.

Beim zweiten Verfahren geht man davon aus, dass die Zeitreihe erst nach d-maligem Differenzieren stationär wird. Wenn eine Zeitreihe durch d-fache Differenzenbildung stationär wird, spricht man – in Anspielung auf das Summieren der Differenzbeträge – auch von *integrierten, unit root* oder *differenzen-stationären* Prozessen d-ter Ordnung, oder kurz von $I(d)$-Prozessen. Für integrierte Prozesse wird der (statische oder dynamische) Zusammenhang mit den transformierten Werten Δy_t und/oder Δx_t modelliert (vgl. Gleichung (20)).[9]

In jüngerer Zeit ist mit den Kointegrations- und Fehlerkorrekturmodellen eine dritte Strategie entwickelt worden, mit der sich das Problem der Nicht-Stationarität von Zeitreihen unter bestimmten Umständen lösen lässt. Diese Strategie eröffnet darüber hinaus die Möglichkeit, eine eventuell vorhandene langfristige Gleichgewichtsbeziehung zwischen zwei oder mehr Variablen und die kurzfristige Anpassungsdynamik, mit der eine Störung des Gleichgewichts vom „System" bearbeitet wird, simultan zu modellieren. Der nachfolgende Abschnitt IV.5 widmet sich diesem Verfahren.

Wie wird nun auf Stationarität getestet? Ein erster Indikator für Stationarität ist ein einfacher Plot einer separaten Zeitreihe. Alternativ kann zum Testen auf Stationarität ein so genannter *Einheitswurzeltes*t (*unit root test*) durchgeführt werden. Betrachten wir für illustrative Zwecke den einfachen Fall einer I(1) Variablen und einen *Random-Walk*-Prozess, bei dem sich der Wert der Variablen zum Zeitpunkt t aus dem Wert in $t - 1$ plus einem Zufallsfehler ergibt: $y_t = y_{t-1} + \varepsilon_t$. Der Fehler ε ist in diesem Fall I(0) und y ist I(1) da $\Delta y_t = \varepsilon_t$. Diese Beziehung kann allgemeiner notiert werden als $y_t = \alpha y_{t-1} + \varepsilon_t$. Ist $|\alpha| < 1$, dann ist y stationär, d.h. I(0). Ist $\alpha = 1$, dann ist y nicht-stationär, d.h. I(1). Die formalen Tests auf Stationarität sind Tests für $\alpha = 1$ und werden deshalb auch Einheitswurzeltests genannt.

Die in vielen Software-Paketen implementierten Standardtests sind der einfache Dickey-Fuller-Test für unabhängige Modell-Residuen, und der erweiterte „augmented" Dickey-Fuller-Test für autokorrelierte Residuen (Dickey und Fuller 1981). In beiden Fällen wird die Nullhypothese, wonach der vorliegende Prozess differenzenstationär sei, gegen die Alternativhypothese, er sei stationär oder trend-stationär, getestet. Der einfache Dickey-Fuller-Test basiert in seiner allgemeinsten Form auf einem Test-Regressionsmodell mit einer Konstanten und einem deterministischen Trend:

$$\Delta y_t = \alpha + \rho y_{t-1} + \gamma t + \varepsilon_t . \tag{30}$$

9 Generell weisen die Zeitreihen bei trend-stationären Prozessen einen deterministischen Trend auf und bei differenzen-stationären Prozessen einen stochastischen Trend.

Die Parameter dieses Modells werden mittels OLS geschätzt, wobei angenommen wird, dass die Residuen unabhängig und identisch normalverteilt sind. Die Alternativhypothese impliziert u.a. die Annahme $\rho < 0$, die Nullhypothese die Annahme $\rho = 0$. Unter der Nullhypothese sind die OLS-Schätzer $\hat{\rho}$ aber nicht normalverteilt. Ihre asymptotischen Eigenschaften sind davon abhängig, ob die Schätzgleichung eine Konstante oder einen Zeitindex als Regressor enthält und ob der unterstellte wahre Prozess einen *drift term* enthält oder nicht. Dickey und Fuller (1979) haben in Monte-Carlo-Experimenten die Verteilungen von $\hat{\rho}$ unter verschiedenen Nullhypothesen ermittelt; die kritischen Werte für unterschiedliche Signifikanzniveaus sind in Tabellenform in verschiedenen Lehrbüchern der Ökonometrie repliziert worden (z.B. Enders 2004).

Alle Variationen des einfachen Dickey-Fuller-Tests implizieren unabhängige Modell-Residuen, oder *white noise errors*. Bei Vorliegen von autokorrelierten Residuen verliert der Test jedoch beträchtlich an Macht. In der Praxis wird deshalb in der Regel der augmented Dickey-Fuller-Test (ADF-Test) verwendet, in welchem die Anzahl der Verzögerungen der abhängigen Variablen in das Regressionsmodell mit aufgenommen werden:

$$\Delta y_t = \alpha + \rho y_{t-1} + \vartheta_1 \Delta y_{t-1} + \vartheta_2 \Delta y_{t-2} + \ldots + \gamma t + \varepsilon_t . \tag{31}$$

In dieser Formulierung dient ein AR(p)-Modell als Basismodell und nicht wie im Standardtest ein AR(1)-Modell. Die Wahl der angemessenen Lag-Zahl dürfte in der Regel von den Daten abhängen. Zur Wahl der besten Anzahl Verzögerungen kann etwa ein Informationskriterium (z.B. AIC oder BIC) herangezogen werden.

In der jüngeren Literatur findet sich eine Vielzahl weiterer Einheitswurzeltests für Paneldaten, von denen die meisten wie die Dickey-Fuller-Tests die Verwendung spezieller kritischer Werte erfordern, z.B. der Test von Levin et al. (2002) sowie Im et al. (1997) für Einheitswurzeln in dynamischen Paneldaten (für einen Überblick über verschiedene weitere Tests siehe Banerjee 1999; Maddala und Wu 1999). Das Hauptproblem all dieser Tests ist jedoch, dass keiner der Tests besonders mächtig ist. Insbesondere bei kurzen Zeitreihen (Faustregel: $t < 100$) ist die Trennschärfe (die „Macht" eines Tests) gegenüber Abweichungen von der Nullhypothese so gering, dass Kritiker die Einheitswurzeltests in diesen Fällen grundsätzlich für nicht sinnvoll einsetzbar halten (z.B. DeJong et al. 1992).

5. Kointegrierte Prozesse und Fehlerkorrekturmodelle

Die Probleme der Scheinregression können bei integrierten Prozessen wie beschrieben mit Hilfe der Differenzenbildung umgangen werden. Dennoch wird in der Zeitreihenliteratur davon abgeraten, dieses Verfahren routinemäßig anzuwenden. Selbst wenn die Zeitreihen, deren struktureller Zusammenhang modelliert werden soll, tatsächlich integrierte Prozesse sind, kann die Differenzenbildung zu einer Fehlspezifikation des Regressionsmodells führen, nämlich dann, wenn zwischen den beiden integrierten Prozessen ein gemeinsamer Trend besteht. Dementsprechend können sie sich langfristig nicht beliebig weit voneinander entfernen. Mit anderen Worten, es besteht eine langfristige ökonomische Gleichgewichtsbeziehung, und die Variablen sind *kointegriert* (Granger

1981; Engle und Granger 1987). In diesem Fall, in dem zwei Variablen X und Y in ihren kurzfristigen Fluktuationen nicht oder nur sehr schwach kovariieren ($\Delta y_t \neq f(\Delta x_t)$), aber in ihren langfristigen Verlaufskomponenten einen strukturellen Zusammenhang bilden ($y_t = f(x_t)$), gerät man durch die Differenzenbildung in die Falle der scheinbaren Nicht-Kausalität.

Der Begriff der Kointegration zweier oder mehrerer nicht-stationärer Variablen setzt drei Bedingungen voraus: 1. die nicht-stationären Trendbewegungen in den beteiligten Zeitreihen sind stochastischer Natur; 2. die beteiligten Zeitreihen sind nicht nur integriert, sondern integriert mit demselben Ordnungsgrad d und können jeweils durch d-fache Differenzenbildung in stationäre Zeitreihen transformiert werden; 3. es existiert mindestens eine Linearkombination dieser Zeitreihen, die stationär ist (Thome 1997). Existiert eine solche Linearkombination, kann in einem Regressionsmodell ein entsprechender Gewichtungsvektor (Kointegrationsvektor) mittels OLS-Regression konsistent geschätzt werden.

Kointegrations-Tests zielen nun darauf ab, synchrone Bewegungen in den Abweichungen vom stochastischen Trend beider Variablen aufzuspüren. Formaler ausgedrückt, Kointegrations-Tests versuchen herauszufinden, ob es eine lineare Kombination von zwei I(1)-Variablen gibt, die stationär ist. In Fällen, wo eine stationäre lineare Kombination existiert, heißen die Variablen kointegriert und die Möglichkeit von Scheinregression kann ausgeschlossen werden.

Im Falle der Kointegration liefert eine OLS-Schätzung von Gleichung (1) zwar konsistente, aber ineffiziente Schätzer, die in kleinen Stichproben auch nicht erwartungstreu sind. Wenn die beiden Variablen nicht kointegriert sind, entstehen die oben abgehandelten Probleme der Scheinregression, und die OLS-Schätzer produzieren nicht-stationäre Residuen. Dieser Tatbestand kann nun nach den klassischen Verfahren von Engle und Granger (1987) für Testzwecke genutzt werden, indem man mit einem Einheitswurzeltest prüft, ob die Residuen der Schätzung der *Kointegrationsregression* stationär sind oder nicht:

$$y_t = \alpha + \beta x_t + \varepsilon_t . \tag{32}$$

Der Test der Nullhypothese über Einheitswurzeln in den Residuen der Kointegrationsregression (32) impliziert also einen Test der Nullhypothese, dass *keine* Kointegration vorliegt. Zusammenfassend umfasst der Engle-Granger-Kointegrationstest zwei Stufen: „In der ersten Stufe geht es um den positiven Nachweis, dass beide Zeitreihen die gleiche Menge von Einheitswurzeln aufweisen (zum gleichen Grad integriert sind); in der zweiten Stufe geht es darum nachzuweisen, dass die Residualreihe keine Einheitswurzel (mehr) aufweist“ (Thome 1997: 215). Die Einheitswurzeltests werden technisch in der gleichen Weise durchgeführt wie mit den Orginalreihen (siehe Abschnitt III.6). Allerdings gelten andere kritische Werte für die Signifikanzniveaus; sie variieren mit der Zahl der Regressoren, dem Vorhandensein einer Konstanten sowie der Stichprobengröße.

Besteht die Kointegrationsgleichung aus mehr als zwei Variablen, resultiert mit dem klassischen Engle-Granger Zwei-Stufen-Verfahren allerdings ein Problem. Der Test kann aufzeigen, ob Kointegration vorliegt, aber er ist nicht in der Lage, eventuelle multiple Kointegrationsbeziehungen aufzudecken. Besteht die Kointegrationsgleichung

etwa aus drei Variablen, so besteht die Möglichkeit, dass alle Variablen durch einen gemeinsamen Faktor bestimmt sind (einen stochastischen Trend), oder durch zwei gemeinsame Faktoren, was auf zwei Kointegrationsbeziehungen schließen lässt. Der Engle-Granger-Ansatz ist nicht in der Lage, mehr als eine Kointegrationsbeziehung aufzudecken; in diesem Fall müssen andere Verfahren wie der Ansatz von Johansen (1991) angewendet werden. Ein Überblick über verschiedene Kointegrations-Tests gibt Banerjee (1999).

Wurde bis jetzt nur die Existenz einer Kointegrationsbeziehung zwischen zwei Variablen oder mehreren untersucht, soll im Folgenden der dynamische Zusammenhang zwischen kointegrierten Variablen dargestellt werden. In diesem Zusammenhang spielt das *Repräsentationstherorem* von Granger eine zentrale Rolle (Engle und Granger 1987), wonach (stark vereinfacht) kointegrierte Variablen immer eine Fehlerkorrekturdarstellung besitzen und umgekehrt ein Fehlerkorrekturmodell aus kointegrierten Variablen besteht.

Die dynamische Spezifikation erfolgt mit Hilfe so genannter Fehlerkorrekturmodelle (*error correction models*), die im bivariaten Fall folgende Gestalt haben:

$$\Delta y_t = \alpha + \sum_{j=1}^{m} \delta_j \left(\Delta y_{t-j} \right) + \sum_{j=1}^{m} \beta_j \left(\Delta x_{t-j} \right) + \gamma z_{t-1} + \varepsilon_t , \tag{33}$$

wobei α, β, δ und γ Konstanten bezeichnen und m und n die Anzahl Verzögerungsintervalle, die entweder theoretisch bestimmt oder datenanalytisch in der Weise ermittelt werden, dass die Residuen ε_t nicht mehr seriell korrelieren. Auf der linken Seite der Gleichung steht die Variable in Differenzen, und auf der rechten Seite sind die verzögerten Differenzen der abhängigen und unabhängigen Variablen. Hinter dem Regressor z_{t-1} verbergen sich die geschätzten Residuen der Kointegrationsgleichung, d.h. die Abweichung vom Gleichgewichtspfad. Enthält die Kointegrationsgleichung nur zwei Variablen, so gilt: $z_{t-1} = (y_t - \beta x_t - \alpha)$. Der Anpassungskoeffizient γ gibt an, wie stark das System zum Gleichgewichtszustand zurückstrebt; darauf bezieht sich dann auch der Begriff „Fehlerkorrektur".

Wie oben bereits erwähnt, besteht in einem multivariaten Modell (mit mehr als zwei Variablen) die Möglichkeit der Existenz mehrerer Kointegrationsbeziehungen. Mit dem Engle-Granger-Verfahren kann aber nur eine einzige Kointegrationsbeziehung ermittelt werden. Einen Ausweg zur Schätzung von multiplen Kointegrationsbeziehungen bieten die so genannten *Vector-Error-Correction*-Modelle nach Johansen (1988) und Johansen und Juselius (1990), auf die hier nicht weiter eingegangen wird.

Während es über Kointegration bei Paneldaten bereits einige Forschungsarbeiten gibt (für Überblicke siehe Banerjee 1999; Kao und Chiang 2000; Baltagi und Kao 2001), ist die Kointegration bei zusammengefassten Zeitreihen ein neues Thema in der ökonometrischen Literatur. Die Politologen Beck und Katz (1996) übertragen einfach die Panel-Kointegrationsmethodologie auf zusammengefasste Zeitreihen. Wie bereits in der Einleitung beschrieben, muss die Art der Asymptotik, die zur Approximation der stochastischen Eigenschaften der Schätzverfahren verwendet wird, dem spezifischen Datentyp angepasst sein. An dieser Stelle sei deshalb vor einem allzu intuitiven Vorgehen bei dieser komplexen Materie gewarnt.

V. Beispiel: Fertilität und Frauenerwerbstätigkeit

Im Folgenden demonstrieren wir einige der oben vorgestellten Methoden am Beispiel des Zusammenhangs der Gesamtfertilitätsrate (TFR) und der Frauenerwerbstätigkeitsquote (FLP) in Raum und Zeit.[10] Der Zusammenhang zwischen diesen beiden Variablen hat seit einiger Zeit verstärkte Aufmerksamkeit erfahren, da sich die Korrelation zwischen FLP und TFR in den OECD-Ländern Mitte der 80er Jahre von einem negativen auf ein positives Vorzeichen gedreht hat (Engelhardt und Prskawetz 2005). Auf Grund des begrenzten Umfangs des vorliegenden Beitrages konzentrieren wir uns an dieser Stelle auf die Schätzung zusammengefasster Zeitreihenmodelle.[11] Dazu vergleichen wir alternative Panelschätzverfahren zur Kontrolle von räumlicher Heterogenität (between effects, fixed effects, random effects models), Prais-Winsten-Schätzer mit autoregressiven Störtermen erster Ordnung zur Berücksichtigung serieller Autokorrelation, alternative Transformationen (Absolutwerte, Logarithmen, erste Differenzen) der Variablen TFR und FLP sowie unterschiedlichen Trendterme zur Abbildung zeitlicher Heterogenität in der Beziehung zwischen TFR und FLP und zur Erfassung eventueller Nicht-Stationarität der separaten Zeitreihen.

Tabelle 1: Einheitswurzeltests für Panel-Daten, p-Werte

	Absolutwerte	Logarithmen	Differenzen
TFR			
Levin-Lin-Chu	0.121	0.437	0.000
Im-Pesaran-Shin	0.196	0.415	0.000
FLP			
Levin-Lin-Chu	0.386	0.813	0.000
Im-Pesaran-Shin	0.970	0.989	0.009

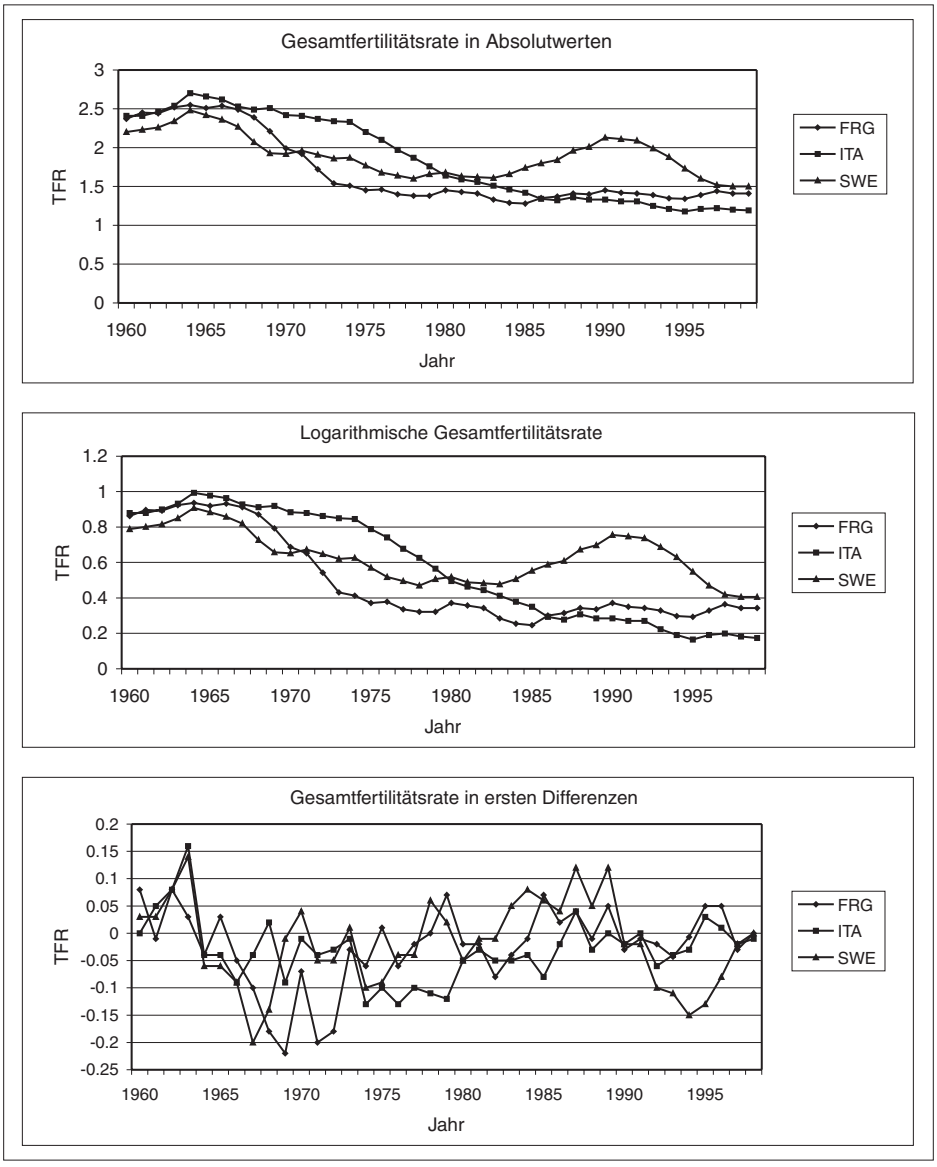
Anmerkungen: Alle Tests wurden auf ein balanciertes Panel angewendet. H_0 für beide Tests: alle Zeitreihen im Panel sind nicht-stationäre Prozesse. Während der Levin-Lin-Chu-Test von der Alternativhypothese ausgeht, dass alle Reihen stationär sind, folgt der Im-Peasaran-Shin-Test der Alternative, dass nur ein Teil der Serien stationär sind.

Tabelle 1 zeigt die Resultate der Einheitswurzeltests nach Levin-Lin-Chu (Levin et al. 2002) und Im-Pesaran-Shin (Im et al. 1997) für balancierte Paneldaten. Ausgewiesen sind die p-Werte für die Nullhypothese, dass alle Zeitreihen im Panel nicht-stationäre Prozesse sind. Die erste Spalte bezieht sich auf die Absolutwerte, die zweite auf die Logarithmen, und die dritte auf die ersten Differenzen der Variablen. Alle Teststatistiken weisen darauf hin, dass TFR und FLP weder in Absolutwerten noch in Logarithmen, allerdings in ersten Differenzen stationär sind. Diese Resultate weisen darauf hin, dass ohne Berücksichtigung der Nicht-Stationarität der Variablen in Absolutwerten und in Logarithmen das Problem der Scheinregression auftreten kann. Wie beschrieben ändert die Verwendung von ersten Differenzen aber die Interpretation der Resultate. Die ge-

10 Die Gesamtfertilitätsrate gibt die nach Alter standardisierte Anzahl Geburten je 1000 Frauen im gebärfähigen Alter zwischen 14 und 49 Jahren an.

11 Zur Schätzung des Zusammenhangs zwischen TFR und FLP anhand separater Zeitreihen siehe z.B. Cheng (1996) und Engelhardt et al. (2004).

Abbildung 2: Gesamtfertilitätsrate in Absolutwerten, Logarithmen und ersten Differenzen für Deutschland, Schweden und Italien, 1960–2000



schätzten Koeffizienten der unabhängige Variablen FLP entsprechen dann nicht mehr der langfristigen Beziehung zwischen den beiden Variablen, sondern reflektieren die Kurzfristbeziehung.

Abbildung 2 zeigt die totale Fertilitätsrate in Absolutwerten, Logarithmen und ersten Differenzen für drei ausgewählte Länder, die repräsentativ für Länder mit geringer, mittlerer und hoher Frauenerwerbsbeteiligung stehen (Italien, Deutschland und Schwe-

Tabelle 2: Geschätzte Koeffizienten vom between effect, random effects und fixed country effects-Modell; t-Werte in Klammern

$TFR_t = \alpha + \beta FLP_t$	BE	FE	RE
β	–0.009+ (–2.079)	–0.044*** (–26.813)	–0.039*** (–24.967)
α	2.459*** (11.103)	4.209*** (51.002)	3.968*** (39.438)
R^2	0.222	0.222	0.222
BIC	18.953	687.001	820.641
F(df, n)	4.32	718.92***	
Chow		38.58***	
Breusch Pagan			456.16***
Hausman	57.99***	109.08***	
ρ		0.980	0.980
Baltagi-Wu LBI		0.144	0.144
Wooldridge		37.097***	

Anmerkungen: BE zeigt OLS-Schätzungen für das between-effects-Modell; RE zeigt GLS-Schätzungen für das Modell mit zufälligen Effekten; FE zeigt OLS-Schätzungen für das Modell mit fixen Länder-Effekten. BIC = –2 log-likelihood + log(n) p, wobei p die Anzahl Parameter im Modell und n die Anzahl Beobachtungen bezeichnet (für RE geschätzt mit MLE). F(df, n)-Test auf „poolability" der Daten, $H_0 : \beta_i = \beta$. Chow-Test auf Absenz von fixen Ländereffekten, $H_0 = \upsilon_i = 0$. Breusch-Pagan-Test auf Absenz von zufälligen Effekten: $H_0 : var(\upsilon_i) = 0$. Hausman-Test: H_0 : Unterschied in den Koeffizienten zwischen aktuellem Modell und RE ist nicht systematisch. ρ ist der geschätzte Autokorrelationskoeffizient von einem AR(1)-Modell. LBI ist die Baltagi-Wu (1999) Locally-best-invariant-Teststatistik von einem AR(1)-Modell, $H_0 : \rho = 0$. Ist LBI wesentlich geringer als 1.5, liegt positive serielle Korrelation vor. Wooldridge-Test für serielle Korrelation von Regression mit Variablen in ersten Differenzen, H_0 : keine Autokorrelation erster Ordnung. *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$, + $p \leq .10$.

den). Diese Reihen veranschaulichen deutlich, dass weder die TFR in Absolutwerten, noch ihre logarithmische Transformation stationär sind, obwohl sich im letzteren Fall die Varianz reduziert. Im Folgenden betrachten wir Panelschätzungen für alternative Transformationen der Variablen TFR und FLP.

Tabelle 2 fasst alternative Panelschätzungen angewendet auf die Variablen in Absolutwerten zusammen. Die erste Spalte weist die Resultate der „between group"-OLS-Schätzung aus. In dieser Schätzung repräsentieren die Daten länderspezifische Zeitdurchschnitte; für die Schätzung wird somit nur die Querschnittsinformation der Daten verwendet und die Zeitreiheninformation wird vernachlässigt. Die zweite Spalte zeigt die OLS-Schätzung der zusammengefassten Daten mit fixen Ländereffekten (approximiert durch eine Dummy-Variable für jedes Land). Diese Schätzung ist identisch mit einer so genannten „within group"-Schätzung, d.h. einer gepoolten OLS-Regression auf die Abweichung jeder Variablen von ihrem Zeitreihenmittel. Im Gegensatz zum „between group"-Schätzer repräsentiert der „fixed effects"-Schätzer die Zeitreihenassoziation zwischen TFR und FLP innerhalb jedes Landes. Die dritte Spalte zeigt die Resultate einer generalisierten Kleinstquadrate-Schätzung mit zufälligen Ländereffekten.

Die Resultate in *Tabelle 2* zeigen, dass unabhängig von der Modellspezifikation die Beziehung zwischen TFR und FLP in dem beobachteten Zeitraum negativ war. Nur im Fall der „between group"-Schätzung erweist sich die negative Beziehung als nicht signifikant. Da der „fixed effects"-Schätzer die erwartete Änderung der TFR innerhalb jeden Landes bei einer Änderung der FLP um eine Einheit, und der „between group"-Schätzer die Änderung der TFR zwischen zwei Ländern bei einer Einheitsänderung der FLP angibt, demonstriert dieses Resultat, dass die Zeitreihenbeziehung zwischen TFR und FLP nicht ihr Vorzeichen wechselt, während sich das Vorzeichen in der länderspezifischen Korrelation zwischen FLP und TFR in der Zeit ändert.

Ferner zeigt *Tabelle 2* die Teststatistik des Hausman-Tests mit der Nullhypothese H_0: Die Differenzen zwischen den geschätzten Koeffizienten des Modells mit fixen Effekten und des Modells mit zufälligen Effekten sind nicht systematisch. Da der Test die Nullhypothese eindeutig ablehnt und darüber hinaus die Modellanpassung (gemessen am Bayesianischen Informationskriteriums BIC) des Modells mit fixen Effekten besser ist als die des Modells mit zufälligen Effekten, scheint die Schätzung mit fixen Effekten geeigneter zu sein als die mit zufälligen Effekten. Ferner weist der F-Test auf „poolability" der Daten darauf hin, dass die Annahme eines konstanten Effektes für FLP im Falle des Modells mit fixen Effekten nicht erfüllt ist. Weiter kann die Nullhypothese auf das Fehlen von Ländereffekten bei der Schätzung mit fixen Effekten (Chow-Test) und mit zufälligen Effekten (Breusch-Pagan-Test) zurückgewiesen werden.

Unabhängig vom Schätzverfahren weisen der Autokorrelationskoeffizient ρ und die Baltagi-Wu-Teststatistik eines AR(1)-Modells und der Wooldridge-Test eines Modells in ersten Differenzen, die im unteren Drittel von *Tabelle 2* aufgeführt sind, auf hohe serielle Korrelation in unseren Daten hin. Aus diesem Grund wählen wir für unsere weiteren Schätzungen den Prais-Winsten-Schätzer mit AR(1)-Störtermen (siehe Abschnitt IV.2).

Tabelle 3: Prais-Winsten-Schätzungen mit fixen Effekten und panel-korrigierten Standardfehlern und AR(1)-Störtermen für unterschiedliche Modell-Spezifikationen; t-Werte in Klammern

Modell	(1) $TFR_t = \alpha + \beta FLP_t$	(2) $\ln TFR_t = \alpha + \beta \ln FLP_t$	(3) $\Delta TFR_t = \alpha + \beta \Delta FLP_t$
β	–0.028*** (–6.800)	–0.577*** (–7.180)	–0.003 (–1.053)
α	3.853*** (14.743)	3.104*** (9.748)	–0.039 (–1.509)
R^2	0.735	0.629	0.009
Wald	251.391***	270.897***	21.256

Anmerkungen: Wald-χ^2-Test für $H_0 : \beta = 0$. *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$, + $p \leq .10$.

Die Resultate der Prais-Winsten-Schätzungen für alternative Transformationen (Absolutwerte, Logarithmen, erste Differenzen) der Variablen FLP und TFR sind in *Tabelle 3* aufgeführt. Differenziert werden Variablen häufig zur Beseitigung serieller Autokorrelation. Gelegentlich finden aber auch logarithmisch transformierte Variablen Verwendung: 1. Sofern die abhängige Variable größer als Null ist, sind die Standardannahmen

von OLS mit einer größeren Wahrscheinlichkeit erfüllt. 2. Die Spannbreite der Werte wird durch die logarithmische Transformation reduziert. Folglich ist die Schätzung weniger anfällig für Ausreißer. 3. Der geschätzte Koeffizient der unabhängigen Variablen repräsentiert die Elastizität der abhängigen Variablen im Hinblick auf die unabhängige Variable und kann somit prozentual interpretiert werden.

Auch wenn die ersten beiden Argumente für die Modellselektion von Relevanz sind, zeigen die Resultate in *Tabelle 3*, dass im Falle logarithmisch transformierter Variablen das Problem serieller Autorkorrelation nicht vermindert wird. Der Autokorrelationskoeffzient ρ bleibt nahezu identisch zu dem Wert der Schätzung mit den Variablen in Absolutwerten in Spalte (1). Die Modellanpassung gemessen an R^2 ist auch geringfügig schlechter im Modell mit Logarithmen. Damit gibt es zumindest von der statistischen Perspektive keinen Grund zur Verwendung logarithmisch transformierter Variablen. Inhaltlich besagt der geschätzte Koeffizient von ln(FLP), dass mit einem Anstieg von FLP um ein Prozent sich die TFR um 0.577 Prozent reduziert.

Spalte (3) zeigt das Resultat der Schätzung des Effekts von FLP für die Variablen in ersten Differenzen. Das wichtigste Resultat ist, dass der Koeffizient von FLP seine Signifikanz verliert. Wie bereits erwähnt, ändert sich durch die Verwendung der Variablen in Differenzen auch die Interpretation des Koeffizienten. Der geschätzte Wert von β gibt nun nicht mehr Auskunft über die langfristige Beziehung zwischen FLP und TFR, sondern reflektiert die kurzfristige Assoziation zwischen den beiden Variablen. Damit ist der Verlust der Signifikanz des Koeffizienten auch nicht überraschend, da er die Absenz einer kurzfristigen Beziehung zwischen FLP und TFR reflektiert. Obwohl sich die serielle Korrelation ρ wesentlich reduziert im Vergleich zu den Resultaten in den Spalten (1) und (2), ist die Transformation der Variablen in erste Differenzen nicht gerechtfertigt, wenn das inhaltliche Interesse auf der Langfristbeziehung zwischen TFR und FLP liegt.

Die wichtigste Botschaft von *Tabelle 3* ist aber, dass sich durch die Anwendung der Prais-Winsten-Schätzung mit AR(1)-Störtermen die Modellanpassung leicht verbessert. Darüber hinaus reduziert sich der Koeffizient von FLP leicht von –0.044 zu –0.028. Die unterschiedlichen Transformationen der Variablen haben das Modell nicht weiter verbessert gemessen an unseren Anforderungen.

Zur Reduktion der Autokorrelation in einem AR(0)-Modell werden in *Tabelle 4* alternativ zum Prais-Winsten-Verfahren unterschiedliche Trend-Terme getestet, die unter Umständen einen Teil der Zeitreihenentwicklung von TFR auffangen können. Im Falle des linearen (Spalte 1) und stückweisen (Spalte 3) Zeittrends verliert die Assoziation zwischen TFR und FLP ihre Signifikanz und die Autokorrelation ist im Vergleich zu dem Modell ohne Trendterm (Spalte 2 in *Tabelle 2*) leicht reduziert, aber dennoch signifikant präsent. Eine logarithmische Spezifikation des Zeittrends in Spalte (2) reduziert die Autokorrelation auch nicht wesentlich im Vergleich zu dem Modell ohne Trendterm.

Von dem Vergleich der Resultate in den *Tabellen 2* bis *4* können wir schließen, dass zur Analyse der Beziehung zwischen TFR und FLP die Prais-Winsten-Schätzung mit einem AR(1)-Störterm angewendet auf die Absolutwerte von TFR und FLP am geeignetsten ist (erste Spalte in *Tabelle 3*). Inhaltlich bedeuten die gefundenen Resultate, dass sich die postulierte kausale Beziehung zwischen TFR und FLP nicht gewandelt

Tabelle 4: Geschätzte Koeffizienten vom AR(0)-Modell mit fixen Effekten mit verschiedenen Trend-Termen; t-Werte in Klammern

Modell	(1) $TFR_t = \alpha + \beta FLP_t + \gamma t$	(2) $TFR_t = \alpha + \beta FLP_t + \gamma \ln t$	(3) $TFR_t = \alpha + \beta FLP_t + \gamma_j t_j$
β	–0.001 (–0.524)	–0.011*** (–5.810)	–0.004+ (–1.944)
γ	–0.036*** (–20.465)	–0.426*** (–23.708)	a)
α	2.833*** (29.865)	3.757*** (56.792)	2.983*** (29.474)
R^2	0.516	0.570	0.619
BIC	346.585	254.900	255.068
F(df,n)	755.68***	891.35***	64.71***
Chow	30.44***	33.68***	43.68***
Wooldridge	29.632***	35.713***	19.610***

Anmerkungen: a) Koeffizienten sind nicht in der Tabelle aufgeführt. BIC = –2 log-likelihood + log(n) p, wobei p die Anzahl Parameter im Modell und n die Anzahl Beobachtungen bezeichnet. F(df, n)-Test auf „poolability" der Daten, $H_0 : \beta_i = \beta$. Chow-Test auf Absenz von fixen Ländereffekten, $H_0 : \upsilon_i = 0$. Wooldridge-Test für serielle Korrelation von Regression mit Variablen in ersten Differenzen, H_0 : keine Autokorrelation erster Ordnung. *** $p \leq .001$, ** $p \leq .01$, * $p \leq .05$, + $p \leq .10$.

hat: Die negative Beziehung von FLP und TFR bleibt über die Zeit erhalten (für detailliertere Analysen siehe Engelhardt und Prskawetz 2005).

VI. Zusammenfassung

Kausalanalysen mit separaten und zusammengefassten Zeitreihen sind nicht trivial und in vielerlei Hinsicht sogar komplexer als Analysen mit Paneldaten. Insbesondere bei zusammengefassten Zeitreihen sollte die Modellierung der kausalen Beziehung nicht nur der Möglichkeit räumlich und zeitlich variierender Konstanten und Parameter Rechnung tragen, sondern auch sonstiger räumlicher und zeitlicher Heterogenität gerecht werden. Gleichermaßen separate wie zusammengefasste Zeitreihen betreffen die Möglichkeit der Modellierung der Prozessdynamik. Obwohl dynamische Modelle auf den ersten Blick durch die explizite Berücksichtigung von zeitverzögerten Rechthandvariablen intuitiv nahe liegend sind, bringt diese Modellierungsstrategie häufig das Problem der Multikollinearität mit sich. Schätzprobleme bei einfachen und separaten Zeitreihen entstehen darüber hinaus durch Heteroskedastizität sowie durch serielle und zeitliche Autokorrelation.

Weniger beachtet bei der Kausalanalyse mit zusammengefassten Zeitreihen wird, dass Zeitreihen häufig variable Trendverläufe aufweisen (siehe *Abbildung 1*). Unter variablen Trends versteht man, dass Phasen mit starkem Wachstum von Phasen mit schwachem Wachstum oder gar Rückgängen abgelöst werden und diese Phasen unterschiedlich lang sind. Solche Daten bezeichnen Ökonometriker als nicht stationär. Wird

diese Eigenschaft der Daten ignoriert, besteht die Gefahr, dass der Forscher aus seiner Analyse auf einen hoch signifikanten Zusammenhang schließt, obwohl die beiden analysierten Zeitreihen voneinander unabhängig sind.

Als man sich in den siebziger Jahren der Problematik annahm, wurde ihr zunächst damit zu Leibe gerückt, dass man mit den Veränderungsraten von Zeitreihendaten arbeitete, da sie der Anforderung der Stationarität genügen. Dieses rein statistisch motivierte Vorgehen vermag aus theoretischer Sicht allerdings nicht zu überzeugen, weil damit theoretisch zu erwartende Zusammenhänge zwischen dem Niveau der Variablen ausgeblendet werden. Stellt man nur auf die Veränderungsraten der Variablen, geht dieser Zusammenhang, wie an unserem Beispiel von Fertilität und Frauenerwerbstätigkeit gezeigt, verloren. Das Dilemma zwischen den Anforderungen der sozialwissenschaftlichen und denjenigen der statistischen Theorie haben die beiden Nobelpreisträger Ganger und Engle durch die Einführung des Begriffes der Kointegration und statistischer Verfahren zur Analyse dieses Phänomens aufgelöst. Von Kointegration sprechen Ökonometriker dann, wenn verschiedene nichtstationäre Zeitreihen einen gemeinsamen variablen Trend besitzen.

Bei Vorliegen von Kointegration lassen sich durch eine dynamische Spezifikation mit Fehlerkorrekturmodellen zumindest in der einfachen Zeitreihenanalyse kausale Zusammenhänge zuverlässig schätzen. Die Übertragung dieser Modelle auf zusammengefasste Zeitreihen steckt allerdings noch in ihren Kinderschuhen. Es ist zu erwarten, dass auf diesem Gebiet der Zeitreihenanalyse in Zukunft neue Forschungsergebnisse vorliegen werden.

Literatur

Baltagi, Badi H., 2001: Econometric Analysis of Panel Data. New York: Wiley.

Baltagi, Badi H., und *Chihwa Kao*, 2001: Nonstationary Panels, Cointegration in Panels and Dynamic Panels: A Survey. Advances in Econometrics 15: 7–52.

Baltagi, Badi H., und *Ping X. Wu*, 1999: Unequally Spaced Panel Data Regression with AR(1) Disturbances. Econometric Theory 15: 814–823.

Bhargava, A., *L. Franzini* und *W. Narendranathan*, 1982: Serial Correlation and the Fixed Effects Model. Review of Economic Studies 49: 533–549.

Banerjee, Anindya, 1999: Panel Data Unit Roots and Cointegration: An Overview. Oxford Bulletin of Economics and Statistics, Special Issue 61: 607–629.

Beck, Nathaniel, 2001a: Time-Series Cross-Section Data. Statistica Neerlandica 55: 110–132.

Beck, Nathaniel, 2001b: Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years? Annual Review of Political Science 4: 271–293.

Beck, Nathaniel, und *Jonathan N. Katz*, 1995: What To Do (and Not To Do) with Time-Series Cross-Section Data. American Political Science Review 89: 634–647.

Beck, Nathaniel, und *Jonathan N. Katz*, 1996: Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Models. Political Analysis 6: 1–36.

Beck, Nathaniel, *Jonathan N. Katz* und *Richard Tucker*, 1998: Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable. American Journal of Political Science 42: 1260–1288.

Chatfield, Christopher, 2004: The Analysis of Time Series, London: Chapman and Hall.

Cheng, Benjamin S., 1996: An Investigation of Cointegration and Causality between Fertility and Female Labour Force Participation. Applied Economics Letters 3: 29–32.

Cochrane, D., und *G. Orcutt*, 1949: Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. Journal of the American Statistical Association 44: 32–61.

DeJong, David N., *John C. Nankervis*, *N.E. Savin* und *Charles H. Whiteman*, 1992: The Power Problem of Unit Root Tests in Time Series with Autoregressive Errors. Journal of Econometrics 53: 31–50.

Dickey, David A., und *Wayne A. Fuller*, 1979: Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association 74: 427–441.

Dickey, David A., und *Wayne A. Fuller*, 1981: Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. Econometrica 49: 1057–1072.

Dielman, Terry E., 1989: Pooled Cross-Sectional and Time Series Data Analysis. New York: Dekker.

Doreian, Patrick, 1980: Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects? Sociological Methods and Research 9: 26–60.

Doreian, Patrick, 1981: Estimating Linear Models With Spatially Distributed Data. Sociological Methodology 13: 359–385.

Doreian, Patrick, 1982: Maximum Likelihood Methods for Linear Models: Spatial Effect and Spatial Disturbance Terms. Sociological Methods and Research 10: 243–269.

Enders, Walter, 2004: Applied Econometric Time Series. New York: Wiley.

Engelhardt, Henriette, *Tomas Kögel* und *Alexia Prskawetz*, 2004: Fertility and Women's Employment: A Macro-Level Time Series Analysis for Developed Countries, 1960–2000. Population Studies 58: 109–120.

Engelhardt, Henriette, und *Alexia Prskawetz*. 2005. A Pooled Time-Series Analysis on the Relation Between Fertility and Female Employment. European Demographic Research Papers.

Engle, Robert, 2001: GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. Journal of Economic Perspectives 15: 157–168.

Engle, Robert F., und *Clive W. J. Granger*, 1987: Cointegration and Error-Correction: Representation, Estimation, and Testing. Econometrica 55: 251–276.

Getis, Arthur, und *J. K. Ord*, 1992: The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis 24(3): 189–206.

Granger, Clive W. J., 1981: Some Properties of Time Series Data and Their Use in Econometrics Model Specification.

Granger, Clive W. J., und *Paul Newbold*, 1974: Spurious Regression in Econometrics. Journal of Econometrics 26: 1045–1066.

Greene, William H., 2003: Econometric Analysis. Englewood Cliffs, N.J.: Prentice Hall.

Gujarati, Damodar N., 1995: Basic Econometrics. New York: McGraw-Hill.

Haining, Robert P., 2003. Spatial Data Analysis: Theory and Practice. Cambridge: Cambridge University Press.

Harvey, Andrew C., 1995: Zeitreihenmodelle, München/Wien: Oldenbourg.

Hannan, Michael T., und *Alice A. Young*, 1977: Estimation in Panel Models: Results on Pooling Cross-Sections and Time Series. Sociological Methodology 8: 52–83.

Hausman, J. A., 1978: Sepcification Tests in Econometrics. Econometrica 46: 1251–1271.

Hibbs, Jr., Douglas A., 1973–74: Problems of Statistical Estimation and Causal Inference in Time-Series Regression Models. Sociological Methodology 5: 252–308.

Hicks, Alexander A., 1994: Introduction to Pooling. In: *T. Janoski* und *A. M. Hicks* (Hg.), The Comparative Political Economy of the Welfare State. Oxford: Cambridge Universtity Press.

Hsiao, Cheng, 2003: Analysis of Panel Data, 2. Aufl. Cambridge: Cambridge University Press.

Hsiao, Cheng, und *Baohong Sun*, 2000: To Pool or Not to Pool Panel Data. In: Panel Data Econometrics: Future Directions, ed. by *P. Balestra*, *J. Krishnakumar* and *E. Ronchetti*. Amsterdam: Elsevier, 181–198.

Im, Kyung So, *M. Hashem Pesaran*, und *Yongcheol Shin*, 1997: Testing for Unit Roots in Heterogeneous Panels. Unpublished Working Paper, Dept. of Applied Economics, University of Cambridge, Dec. 1997.

Johansen, Soren, 1988: Statistical Analysis of Cointegration Vectors. Journal of Economic Dynamics and Control 12: 231–254.

Johansen, Soren, 1991: Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica 59: 1551–1580.

Johansen, Soren, und Katarina Juselius, 1990: Maximum Likelihood Estimation and Inference on Cointegration with Applications to the Demand of Money. Oxford Bulletin of Economics and Statistics 522: 169–210.

Judge, George G., *W. E. Griffith*, *R. Carter Hill* und *Tsoung-Chao Lee*, 1985: The Theory and Practice of Econometrics. New York: Wiley.

Kao, Chihwa, und *Min-Hsien Chiang*, 2000: On the Estimation and Inference of a Cointegrated Regression in Panel Data. Advances in Econometrics 15: 179–222.

Kittel, Bernhard, 1999: Sense and Sensitivity in Pooled Analysis of Political Data. European Journal of Political Research 35: 225–253.

Kittel, Bernard, und *Hannes Winner*, 2004: How Reliable is Pooled Analysis in Political Economy? The Globalization-Welfare State Nexus Revisited. European Journal of Political Research, im Erscheinen.

Kmenta, Jan, 1986: Elements of Econometrics. New York: Macmillan.

Koyck, Leendert M., 1954: Distributed Lags and Investment Analysis. Amsterdam: North Holland.

Levin, Andrew, Chien-Fu Lin und *Chia-Shang J. Chu*, 2002: Unit Root Test in Panel Data: Asymptotic and Finite Sample Properties. Journal of Econometrics 108: 1–24.

Maddala, Gangadharrao S., 1991: To Pool or Not to Pool: That is the Question. Journal of Quantitative Economics 7: 255–264.

Maddala, Gangadharrao S., 1999: On the Use of Panel Data Methods with Cross-Country Data. Annales d'Economie et de Statistique 55/56: 429–449.

Maddala, Gangadharrao S., 2001: Introduction of Econometrics. Chichester: Wiley.

Maddala, Gangadharrao S., und *S. Wu,* 1999: A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test. Oxford Bulletin of Economics and Statistics 61: 631–652.

Mundlak, Yair, 1978: On the Pooling of Time Series and Cross Section Data. Econometrica 46: 69–85.

Nuamah, Nicholas N.N.N., 1986: Pooling Cross Section and Time Series Data. The Statistician 35: 345–351.

Ord, J. K., und *Arthur Getis*, 1995: Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27(4): 286–306.

Ostrom, Charles W., 1990: Time Series Analysis: Regression Techniques. Beverly Hills, CA: Sage.

Parks, Richard W., 1967: Efficient Estimation of a System of Regression Equations when Disturbances are both Serially and Contemporaneously Correlated. Journal of the American Statistical Association 62: 500–509.

Phillips, P. C. B., 1986: Understanding Spurious Regressions in Econometrics. Journal of Econometrics 33: 311–340.

Pindyck, Robert S., und *Daniel L. Rubinfeld*, 1981: Econometric Models and Economic Forecasts. New York: McGraw-Hill.

Podestà, Frederico, 2002: Recent Developments in Quantitative Comparative Methodology: The Case of Pooled Time Series Cross-Section Analysis. DSS Papers Soc 3-02.

Podestà, Frederico, 2003: Econometric Solutions vs. Substantive Results: A Crucial Trade-off in the Time-Series-Cross-Section Analysis. ISER Working Papers 2003-34.

Plümper, Thomas, *Vera Troeger* und *Philip Manow*, 2004: Panel Data Analysis in Comparative Politics: Linking Method to Theory. European Journal of Political Research, im Erscheinen.

Prais, S., und *C. Winsten*, 1954: Trend Estimation and Serial Correlation. Cowls Commission Discussion Paper No. 383. Chicago.

Sayrs, Lois W., 1989: Pooled Time Series Analysis. Beverly Hills: Sage.

Stimson, James A., 1985: Regression in Space and Time: A Statistical Essay. American Journal of Political Sciences 29: 914–947.

Thome, Helmut, 1997: Scheinregression, kointegrierte Prozesse und Fehlerkorrekturmodelle. Zeitschrift für Soziologie 26: 202–221.

Yule, G. Udny, 1926: Why Do We Sometimes Get Nonsense Correlation Between Time Series? Journal of the Royal Statistical Society 89: 1–64.

White, Halbert, 1980: A Heteroscedasticity-Consistent Covariance Matrix and a Direct Test for Heteroscedasticity. Econometrica 48: 817–838.

Wilson, Sven E., und *Daniel M. Butler*, 2004: A Lot More to Do: The Promise and Peril of Panel Data in Political Science. Manuskript, Brigham Young University.

Worrall, John L., und *Travis C. Pratt*, 2004: Estimations Issues Associated with Time-Series-Cross-Section Analysis in Criminology. Western Criminology Review 5: 35–49.

VI. Spezielle Probleme der Datenanalyse

KAUSALANALYSE DURCH MATCHINGVERFAHREN*

Markus Gangl und Thomas A. DiPrete

Zusammenfassung: Auf Grund ihrer Nähe zum Konzept kontrafaktischer Kausalität haben nichtparametrische Matchingverfahren in der neueren statistischen und ökonometrischen Literatur zur Kausalanalyse an Bedeutung gewonnen. Vor diesem Hintergrund führt der Beitrag das Rubin Causal Model (RCM) in die soziologische Methodendiskussion ein und diskutiert seine empirische Umsetzung im Rahmen des Propensity Score Matchings. Der Beitrag verdeutlicht die Relevanz dieser Verfahren für soziologische Fragestellungen sowie die ihnen im Vergleich zu üblichen Regressionsverfahren zu Grunde liegenden Annahmen. Wir illustrieren die Anwendung von Matchingverfahren anhand einer Analyse des kausalen Effekts von Arbeitslosigkeit auf den weiteren Erwerbsverlauf.

I. Einleitung

Die Analyse kausaler Zusammenhänge ist ohne Zweifel das Herzstück einer theoriegeleiteten, hypothesenprüfenden und kumulativen Sozialwissenschaft, in der sich theoretische Modellbildung und deren empirische Überprüfung verbinden: Aus einem Modell abgeleitete Hypothesen oder gar konkurrierende Erklärungen können letztlich nur dadurch auf ihren Realitätsgehalt getestet werden, dass sie mit empirischen Daten konfrontiert werden, die potenziell falsifizierende Aussagen über eine unterstellte Beziehung zwischen Ursache und Wirkung ermöglichen. Dass Kausalschlüssen eine zentrale Rolle für die Verzahnung empirischer und theoretischer Forschung zukommt, ist eine alles andere als neue Erkenntnis.

In den Sozialwissenschaften stellt sich dabei allerdings das zentrale Problem, dass experimentell erhobene Daten, die eine unmittelbar kausale Interpretation empirischer Befunde erlauben, in der Forschungspraxis bestenfalls eine marginale Rolle zukommt.

* Der Beitrag entstand im Rahmen des ESF/DFG-Projekts „Human Capital Effects of the Welfare State". Ein Gastaufenthalt von Thomas DiPrete am Wissenschaftszentrum Berlin ermöglichte uns den Beginn der Arbeiten. Die im Beitrag verwendeten Daten des Sozio-ökonomischen Panels wurden freundlicherweise durch das Deutsche Institut für Wirtschaftsforschung (DIW), Berlin, zur Verfügung gestellt. Selbstverständlich liegt die Verantwortung für die hier durchgeführten Analysen allein bei den Autoren. Wir bedanken uns bei Ulrich Kohler und einem anonymen Gutachter für hilfreiche Anregungen und Kommentare.

Abgesehen von speziellen Anwendungen wie etwa im Bereich der experimentellen Spieltheorie (z.B. Fehr und Gächter 2000; Falk und Fehr 2003), stoßen experimentelle Ansätze in aller Regel sehr schnell an ethische, forschungspraktische und methodologische Grenzen (vgl. Manski 1995; Heckman 1992). Nichtexperimentelle Daten haben ihrerseits den wichtigen Vorteil, dass mit ihnen das reale Verhalten von Personen sowie deren wichtigste Merkmale (repräsentativ) erfasst werden. Andererseits führt gerade die Beobachtung realen Verhaltens dazu, dass kausale Schlussfolgerungen über Ursache-Wirkungs-Zusammenhänge nicht uneingeschränkt möglich sind. In Beobachtungsdaten resultieren in aller Regel eben nicht nur die abhängigen, sondern vor allem *auch die unabhängigen Variablen* der Analyse aus intentionalem Handeln von Akteuren, und können damit eben nicht umstandslos als exogen gegeben angesehen werden.

Man kann vor diesem Hintergrund geneigt sein, den Schluss zu ziehen, dass ein eher naturwissenschaftlich inspiriertes Verständnis von Kausalität als Ursache-Wirkungs-Zusammenhang dem Gegenstandsbereich der Soziologie – außer in Randbereichen wie etwa der Evaluationsforschung – grundsätzlich nur wenig angemessen sei, und die empirische Sozialforschung sich daher auf im Wesentlichen beschreibende Analysen zu beschränken habe, die gegebenenfalls durch akteurszentrierte „causal narratives" unterfüttert werden können (vgl. Goldthorpe 2001). Wir werden im Folgenden dagegen die Position einnehmen, dass – neben dem grundsätzlich immer gebotenen deskriptiven Verständnis sozialer Prozesse (Abbott 1998) – Kausalschlüsse auch in den Sozialwissenschaften nicht nur möglich und angemessen sind, sondern letzten Endes das wesentliche Ziel hypothesen- und mechanismenprüfender empirischer Sozialforschung darstellen. In unserem Verständnis beziehen sich Kausalschlüsse immer auf die empirische Überprüfung theoretisch erwarteter Ursache-Wirkungs-Zusammenhänge, bilden also mithin das zentrale Verbindungsglied zwischen theoretischer Modellbildung und theoriegeleiteter, d.h. generalisierender empirischer Sozialforschung.

Dass Kausalschlüsse innerhalb der Sozialwissenschaften auf besondere Schwierigkeiten stoßen, liegt dagegen angesichts ihres Gegenstandsbereichs auf der Hand: Sozialwissenschaftliche Kausalanalysen können ausschließlich vor dem Hintergrund intentionalen Handelns von Akteuren durchgeführt werden, und beziehen sich auf die *empirisch beobachtbaren Reaktionen der handelnden und interagierenden Akteure auf exogen variierende Rahmenbedingungen*. Da letztere durch Beobachtungsdaten nur bedingt erfasst werden (können), beruhen kausale Schlussfolgerungen notwendigerweise immer auf einer (geglückten) Kombination von Untersuchungsdesign und statistischer Methodologie, die sich einer Experimentalsituation möglichst weit annähert. Innerhalb der statistischen und ökonometrischen Literatur hat sich in den vergangenen 15 Jahren ein Konsens zu den daraus für die statistische Analyse folgenden Implikationen herausgebildet, das so genannte *Rubin Causal Model* (RCM). Wir werden im Folgenden dieses *kontrafaktische Verständnis* von Kausalität ausführlicher darstellen und verschiedene potenziell interessante und empirisch schätzbare kausale Parameter definieren.

Da sie eine direkte statistische Umsetzung des RCM darstellen, stellen wir im Anschluss daran die wesentlichen Grundzüge der *Propensity score matching*-Verfahren zur empirischen Analyse kausaler Effekte mit Hilfe von Beobachtungsdaten vor. Matchingverfahren ermöglichen die direkte Schätzung der wichtigsten kausalen Parameter, und haben entsprechend der zunehmenden Akzeptanz des RCM in den letzten Jahren in

der angewandten Statistik und in ökonometrischen Arbeiten deutlich an Bedeutung gewonnen. Im Unterschied zu üblichen Regressionsverfahren handelt es sich dabei zudem um nichtparametrische Verfahren, deren Anwendung keine oder nur vergleichsweise wenig restriktive statistische Annahmen voraussetzt. Neben einer Präzisierung der Voraussetzungen kausaler Schlussfolgerungen auf der Basis sozialwissenschaftlicher Beobachtungsdaten ist damit in den letzten Jahren auch eine alternative statistische Tradition entwickelt worden, die in der angewandten Forschung zumindest Sensitivitätsanalysen der mit Regressionsverfahren gewonnenen Ergebnisse ermöglicht. Nach der Darstellung ihrer wesentlichen Grundlagen werden wir die empirische Anwendung von Matchingverfahren am Beispiel einer Analyse der Karrierefolgen von Arbeitslosigkeit illustrieren.

II. Kontrafaktische Inferenz im Rubin Causal Model

Das *kontrafaktische* Verständnis von Kausalität wie es im Rubin Causal Model (RCM) zum Ausdruck kommt, ist die gemeinsame methodologische Grundlage aller Matchingverfahren (vgl. Rubin 1974, 1978; Holland 1986; Pratt und Schlaifer 1988; Manski 1995; Heckman 1997; Rosenbaum 2002; aus spezifischer soziologischem Blickwinkel Sobel 1995, 1996; Smith 1997; Winship und Morgan 1999; Winship und Sobel 2004). Gedanklicher Ausgangspunkt des RCM ist die Modellvorstellung potenzieller Ereignisse Y, die jeweils infolge eines kausal wirksamen Faktors T eintreten.[1] Im einfachsten Beispiel eines binären Kausalfaktors (*treatment*) $T=0,1$ besteht die Menge der potenziell möglichen Ereignisse damit aus dem Ereignis Y_1, das unter der Bedingung $T=1$ realisiert wird, sowie dem alternativen Ereignis Y_0 für $T=0$. Konzeptionell ist der kausale Effekt von T dann auf der Ebene individueller Akteure bzw. für einzelne Beobachtungen als *Einheitseffekt* (*unit effect*)

$$\delta_i \equiv Y_{1i} - Y_{0i} \tag{1}$$

definiert.[2] Im unten ausführlicher dargestellten empirischen Beispiel entspräche Y_{1i} dem individuellen Erwerbsverlauf nach Eintritt einer Arbeitslosigkeit zu einem be-

1 Dementsprechend firmiert das RCM in der statistischen Literatur auch als *potential outcome model*. In der ökonometrischen Literatur sind auch die Bezeichnungen *Roy model* oder *Switching (regression) model* üblich (Manski 1995; Heckman et al. 1997, 1998), um denselben Sachverhalt anzudeuten. In der Soziologie wird häufig die Bezeichnung des *interventionistischen* bzw. *manipulativen* Kausalmodells gewählt, um die gedankliche Nähe zur experimentellen Forschung zu betonen (etwa Goldthorpe 2001). Leider entsteht dadurch oft auch der (falsche) Eindruck, der konzeptionelle Rahmen des RCM sei ausschließlich im experimentellen Bereich bzw. in der Evaluation expliziter Interventionen relevant.

2 Wenn auch ursprünglich aus der experimentellen Forschung stammend, ist das RCM konzeptionell nicht auf die Wirkung binärer Faktoren beschränkt (vgl. die ausführlichere Darstellung zur Verbindung mit üblichen Regressionsverfahren bei Wooldridge 2002 oder Berk 2004 bzw. Strukturgleichungsmodellen bei Pearl 1998, 2000). Zur vereinfachenden Illustration verwenden wir im Beitrag jedoch durchgehend das binäre Ereignis Arbeitslosigkeit als potenziell kausal wirksamen Faktor. Ordinale Kausalfaktoren werden im Rahmen des Standardansatzes in multiple Paarvergleiche der Wirkung jeweils binärer Faktoren aufgelöst.

stimmten Zeitpunkt T_0, Y_{0i} dagegen dem individuellen Erwerbsverlauf, in welchem zum Zeitpunkt T_0 keine Arbeitslosigkeit aufgetreten war. Der kausale Effekt der Arbeitslosigkeit besteht in der Differenz zwischen Y_{1i} und Y_{0i}, also dem auf die Arbeitslosigkeit zurückführbaren *Unterschied* im individuellen Erwerbsverlauf – gemessen etwa anhand des realisierten Erwerbseinkommens, der beruflichen Stellung oder der individuellen Arbeitsplatzsicherheit. Ob die kausale Wirkung des Faktors T tatsächlich individuell verschieden ausfällt, ist im Rahmen des kontrafaktischen Modells dann eine lediglich empirische Frage.

1. Das Fundamentalproblem der Kausalanalyse

Die Definition des Einheitseffektes von T als Differenz potenzieller Ereignisse Y ist einerseits unmittelbar intuitiv einsichtig und entspricht andererseits auch gängigen Kausalitätskriterien (etwa Marini und Singer 1988; Pearl 2000): Das Auftreten von T ist dem Ereignis Y zeitlich vorgelagert, es besteht eine Kovariation zwischen T und Y, die nicht auf die Wirkung alternativer (vorgelagerter) Kovariaten X zurückführbar ist, und schließlich, im Zentrum der kontrafaktischen Interpretation, entspricht der oben definierte Einheitseffekt der Wirkung einer bewussten Manipulation von bzw. der Intervention T. Kausale Effekte im Sinne des RCM beschreiben damit Reaktionsfunktionen (*response schedules* in der Terminologie von Pearl 2000; Berk 2004) im Sinne der hypothetisch (kontrafaktisch) zu erwartenden Ereignisse Y infolge einer (wie auch immer praktisch durchführbaren) *Setzung* der Umstände $T=t$.

Aus der Definition kausaler Effekte über den Vergleich potenzieller Ereignisse folgt allerdings auch unmittelbar, dass kausale Wirkungen nicht direkt beobachtbar, sondern immer nur unter bestimmten Annahmen empirisch *erschlossen* werden können. Das fundamentale Problem jedweden Kausalschlusses besteht in der offensichtlichen Unmöglichkeit, dieselbe Person unter sonst gleichen Bedingungen *gleichzeitig* alternativen Rahmenbedingungen ausgesetzt zu sehen und damit ihre empirische Reaktionen auf unterschiedliche Handlungskontexte direkt erfassen zu können. Einheitseffekte könnten nur dann empirisch ermittelt werden, wenn die Erwerbsverläufe Y von Personen i tatsächlich sowohl nach einer Arbeitslosigkeit T_i zu einem bestimmten Zeitpunkt T_0 *als auch* ohne eine Arbeitslosigkeit T_i zum Zeitpunkt T_0 beobachtbar wären.

Abbildung 1: Das Fundamentalproblem der Kausalanalyse

		Potenziell endogene Zuweisung	
		Kausaler Faktor T = 1 (Testbedingung)	Kausaler Faktor T = 0 (Kontrollbedingung)
Potenziell fehlende Vergleichbarkeit	Intervention T=1, Experimentalstichprobe E	(a) $Y_{1i,i\in E} \mid T_i=1$	(b) $\{Y_{0i,i\in E} \mid T_i=0\}$
	Intervention T=0, Kontrollstichprobe C	(c) $\{Y_{1i,i\in C} \mid T_i=1\}$	(d) $Y_{0i,i\in C} \mid T_i=0$

Anmerkung: Latente Ereignisse in geschweiften Klammern.

Kausalschlüsse beziehen sich damit zwar konzeptionell immer auf die Menge potenzieller Ereignisse (Y_{1i}, Y_{0i}), tatsächlich beobachtbar ist aber natürlich jeweils nur das spezifische Ereignis

$$Y_i^* \equiv T_i Y_{1i} + (1 - T_i) Y_{0i} \tag{2}$$

sowie die jeweils realisierte Kausalbedingung $T_i=0,1$ – also entweder der tatsächliche Erwerbsverlauf Y_{1i} einer zum Zeitpunkt T_0 tatsächlich arbeitslos ($T_i=1$) gewordenen Person i *oder* der Erwerbsverlauf Y_{0i} der Person i, die zum Zeitpunkt T_0 nicht arbeitslos geworden ist ($T_i=0$). Wie *Abbildung 1* illustriert, besteht dementsprechend die empirische Datengrundlage kausaler Schlussfolgerungen *prinzipiell* nur aus der partiellen Beobachtung einer Experimentalgruppe $Y_{1i,i\in E}$ (Zelle a), die empirisch der Wirkung des Kausalfaktors T ausgesetzt war, sowie einer Kontrollgruppe $Y_{0i,i\in C}$ (Zelle d), für welche T tatsächlich nicht zutraf. Da Kausalschlüsse definitionsgemäß auf den Vergleich eines empirisch beobachteten Ereignisses $Y_{1i,i\in E}$ (bzw. äquivalent $Y_{0i,i\in C}$) mit einem latenten Ereignis $\{Y_{0i,i\in E}\}$ (bzw. entsprechend $\{Y_{1i,i\in C}\}$) abzielen, haben alle Methoden empirischer Kausalanalyse Annahmen darüber zu treffen, auf welche Weise aus den empirisch vorliegenden Beobachtungen $Y_{0i,i\in C}$ in Zelle (d) eine *kontrafaktische* Abschätzung von $\{Y_{0i,i\in E}\}$, also des *alternativen* Erwerbsverlaufs von i, der ohne Arbeitslosigkeit beobachtet worden wäre (Zelle b), *konstruiert* werden kann, um daraus den kausalen Effekt δ_i empirisch zu bestimmen.[3]

2. Grundlegende Annahmen in der empirischen Analyse

Auch wenn sie die Grundlage des kontrafaktischen Verständnisses von Kausalität darstellen, können Einheitseffekte weder plausibel geschätzt werden, noch stehen sie im Zentrum des sozialwissenschaftlichen Interesses. In aller Regel ist die empirische Forschung ohnehin an der Abschätzung des typischen, d.h. *durchschnittlichen kausalen Effekts* von T auf Y in einer Population oder auch für bestimmte Subgruppen interessiert – d.h. die interessante Frage für eine generalisierende Sozialwissenschaft ist, ob Arbeitslosigkeit im Durchschnitt oder in aller Regel zu verschlechterten Karriereaussichten führt, und eben nicht, ob es nicht vielleicht doch auch Einzelfälle gibt, für die sich eine Arbeitslosigkeit im Nachhinein als Glücksfall herausgestellt hat. Gleichzeitig erlaubt gerade die *Aggregation* der Fragestellung die Abschätzung kausaler Effekte, weil der beobachtbare Gruppenunterschied $E(Y_1^*) - E(Y_0^*)$ unter bestimmten Umständen die Identifikation des durchschnittlichen kausalen Effekts $\bar{\delta}$ ermöglicht.

Um die Voraussetzungen kausalen Schließens zu illustrieren, kann das Verhältnis von empirisch beobachtbaren Unterschieden der Ereignisse Y in der Experimental- und Kontrollgruppe (Zellen a und d in *Abbildung 1*) zum durchschnittlichen kausalen Effekt $\bar{\delta}$ durch die Komponentenzerlegung

3 Die Überlegungen gelten selbstverständlich äquivalent für die Abschätzung des latenten Ereignisses der Zelle (c) durch die empirischen Beobachtungen der Zelle (a).

$$
\begin{aligned}
E\left(Y_{1i}\middle|T_i=1\right)-E\left(Y_{0i}\middle|T_i=0\right) &= E\left(Y_{1i}\right)-E\left(Y_{0i}\right)+E\left(Y_{0i,i\in E}\right)-E\left(Y_{0i,i\in C}\right) \\
&\quad +\left(1-\overline{T}_i\right)\times\left[E\left(Y_{1i,i\in E}-Y_{0i,i\in E}\right)-E\left(Y_{1i,i\in C}-Y_{0i,i\in C}\right)\right] \qquad (3)\\
&= \overline{\delta}+E\left(Y_{0i,i\in E}-Y_{0i,i\in C}\right)+\left(1-\overline{T}_i\right)\left(\overline{\delta}_{i\in E}-\overline{\delta}_{i\in C}\right)
\end{aligned}
$$

Unterschied zwischen Experimental- und Kontrollgruppe = „wahrer" durchschnittlicher kausaler Effekt δ des Faktors T + Unterschied im Basisereignis Y ohne Wirkung von T (Heterogenität) + Unterschied in der Wirkung von T zwischen Experimental- und Kontrollgruppe (Selbstselektion)

beschrieben werden (vgl. Winship und Sobel 2004). Bei dieser Komponentenzerlegung handelt es sich um nichts weiter als eine Formalisierung der Intuition, dass Kausalschlüsse immer dann gerechtfertigt sind, wenn Experimental- und Kontrollgruppe in ihren sonstigen Merkmalen vergleichbar sind (oder genauer, wenn zwischen beiden Gruppen keine sonstige Heterogenität, d.h. keine Unterschiede in den ansonsten zu erwartenden Ereignissen Y vorliegen) und wenn keine Selbstselektion in den Experimentalstatus vorliegt (also der empirisch beobachtete Experimentalstatus T nicht von den Ereignissen Y abhängig ist, und nicht etwa ungemessenes privates Wissen der Akteure ausdrückt).[4] Kurz gefasst sind Gruppenvergleiche also immer dann im Hinblick auf kausale Schlussfolgerungen informativ, wenn

$$\left(Y_1, Y_0\right) \perp T \,|\, X \qquad (4)$$

gilt, d.h. wenn der beobachtete Experimentalstatus T unter Kontrolle sonstiger Kovariaten X unabhängig von den potenziellen Ereignissen Y ist. Nur unter Gültigkeit dieser konditionalen Unabhängigkeitsannahme (*conditional independence assumption* (CIA), nach Rosenbaum und Rubin 1983 auch als *strict ignorability assumption* bezeichnet) rechtfertigt der Gruppenvergleich den kausalen Schluss, dass sich das Ereignis Y im Durchschnitt um den Faktor δ verändert, wenn die Rahmenbedingung T gesetzt wird bzw. eintritt.

Diese Unabhängigkeitsbedingung macht sich die klassische Experimentalforschung zunutze, indem für eine ansonsten homogene Stichprobe von Personen eine zufällige Zuweisung des Experimentalstatus T_i erfolgt. Dadurch ist Bedingung (4) erfüllt und die interne Validität der kausalen Schlussfolgerungen gesichert. In der Analyse von Beobachtungsdaten ist die randomisierte Zuweisung des Kausalfaktors naturgemäß nicht möglich, so dass die Homogenität der Vergleichsgruppen und die konditional unabhängige Zuweisung des Experimentalstatus durch den Einschluss geeigneter Kovariaten X, über eine geeignete statistische Modellierung oder durch ein entsprechendes Untersuchungsdesign erreicht werden muss. Ein wichtiger Unterschied zur gängigen Regressionspraxis besteht darin, dass mit der CIA-Bedingung unmittelbar gefordert wird, im Kovariatenvektor eben nicht nur die relevanten Prädiktoren von Y, sondern explizit auch die Determinanten von T zu berücksichtigen, um dadurch konzeptionell eine Experimentalsituation anzunähern.

4 In der älteren ökonometrischen Literatur firmieren beide Komponenten etwa unter den Begrifflichkeiten der „selection on observables" bzw. „selection on unobservables" (Heckman und Robb 1985). Auch wenn die Unterscheidung im konkreten Einzelfall sicher nicht trennscharf ist, verweisen die verschiedenen Begriffe heuristisch auf den Unterschied zwischen prinzipiell beobachtbaren Hintergrundkovariaten und prinzipiell unbeobachtbarem privatem Wissen der Akteure.

Selbst wenn im konkreten Fall die CIA-Annahme plausiblerweise als erfüllt angesehen werden kann, ist schließlich noch eine weitere, und womöglich noch grundsätzlichere Annahme zu treffen. Kausalschlüsse nach dem Modell des RCM sind auf der Basis von Beobachtungsdaten immer nur dann möglich, wenn die Existenz eines individuell stabilen „wahren" kausalen Effekts δ_i vorausgesetzt werden kann, wenn also die so genannte *stable unit treatment value assumption* (SUTVA, d.h. $(Y_{1i}, Y_{0i}) \perp T_j \; \forall \, i \neq j$) erfüllt ist. Die Annahme ist weitaus weniger trivial als sie zunächst erscheinen mag, da sie die theoretische und gleichzeitig inhärent nicht überprüfbare Annahme beinhaltet, dass die individuelle Kausalwirkung von T nicht von der Kausalwirkung von T auf andere Personen oder auch nicht von der Häufigkeit oder Verteilung der Intervention in der Stichprobe abhängig ist. Die SUTVA-Annahme ist dementsprechend immer dann nicht erfüllt, wenn vermutet wird, dass Interdependenzen zwischen Experimental- und Kontrollgruppe, *Spillover*- oder allgemeine Gleichgewichtseffekte vorliegen, so dass die für die Kontrollgruppe beobachteten Ereignisse nicht unbeeinflusst von den sozialen Prozessen in der Experimentalgruppe sind (vgl. ausführlicher Garfinkel et al. 1992; Heckman, Lochner und Taber 1998). In diesen Fällen wären kausale Effekte grundsätzlich nicht allein mit Daten auf derselben Beobachtungs- und Wirkungsebene, sondern allenfalls durch ein Mehrebenendesign identifizierbar.[5] Kausalanalysen auf der Basis von sowohl Regressions- wie auch Matchingverfahren müssen also in aller Regel die SUTVA-Annahme voraussetzen, die dementsprechend natürlich auch inhaltlich zumindest näherungsweise plausibel sein sollte.

3. Durchschnittliche kausale Effekte als empirische Schätzgrößen

Wir nehmen im Folgenden ebenfalls an, dass die SUTVA-Annahme gerechtfertigt ist, so dass wir den durchschnittlichen kausalen Effekt von T prinzipiell empirisch ermitteln können. Grundsätzlich bestehen unterschiedlichste Möglichkeiten, aus den Einheitseffekten einen durchschnittlichen Kausaleffekt zu definieren; entsprechend der neueren ökonometrischen Literatur (vgl. Heckman und Vytlacil 2002; Heckman et al. 2003) stehen jedoch in der Regel zwei Parameter im Zentrum des kausalanalytischen Interesses: einmal der durchschnittliche kausale Effekt von T in der untersuchten Gesamtpopulation, der so genannte *average treatment effect* (ATE), und zum zweiten der durchschnittliche kausale Effekt für die Personen, die empirisch der Wirkung des kausalen Faktors ausgesetzt waren, der so genannte *average treatment effect on the treated* (ATT).[6]

5 Auch dieser Fall ist in einem experimentellen Design von vornherein ausgeschlossen, da die Interaktion der Probanden unter der Kontrolle des Versuchsleiters steht und in den überwiegenden Fällen bewusst ausgeschlossen wird. Damit soll nun nicht gesagt sein, dass Prozesse, in denen die SUTVA-Annahme nicht erfüllt ist (z.B. Selbstverstärkungsdynamiken oder die Auswirkungen sozialer Präferenzen), nicht untersucht werden könnten. Diese soziologisch hochinteressanten Prozesse erfordern jedoch entweder eine spezifische experimentelle Kontrolle (vgl. etwa Fehr und Gächter 2000), oder, im Falle von Beobachtungsdaten, ein über das RCM-Standarddesign hinausgehendes bzw. es erweiterndes Mehrebenendesign.

6 Ein weiterer zentraler Parameter in der ökonometrischen Diskussion ist der so genannte *local average treatment effect* (LATE, vgl. Imbens und Angrist 1994; Angrist, Imbens und Rubin

Der ATE entspricht dabei konzeptionell dem erwarteten Effekt von T für Y einer zufällig aus der Gesamtpopulation (d.h. Stichproben E und C in *Abbildung 1*) gezogenen Person, der ATT dagegen dem im Durchschnitt beobachteten Effekt von T für Y für eine zufällig aus der Experimentalstichprobe E gezogenen Person. Der ATE erfasst konzeptionell damit die typischen Folgen von T in der untersuchten Population, der ATT die typischen Folgen von T für die (möglicherweise selektive) Gruppe der tatsächlich von T Betroffenen (d.h. für die $T_i=1$ gilt). Die aus der experimentellen Forschung abgeleitete Terminologie des kontrafaktischen Ansatzes sollte insgesamt aber nicht darüber hinwegtäuschen, dass „treatment" hier nicht primär im Sinne einer expliziten Manipulation, sondern allgemeiner als kausal wirksamer Faktor zu verstehen ist.[7] Zugunsten der Konsistenz mit der Literatur folgen wir hier durchgängig den inzwischen eingebürgerten Begrifflichkeiten.

ATE- und ATT-Parameter sind direkt aus den Grundannahmen des RCM definierbar (vgl. zum Folgenden Heckman 1997). Ausgangspunkt ist das folgende Modell potenzieller Ergebnisse

$$\begin{aligned} Y_{0i} &= \mu_0(X_i) + U_{0i} \\ Y_{1i} &= \mu_1(X_i) + U_{1i} \end{aligned} \tag{5}$$

das die potenziellen Ergebnisse Y_0 und Y_1 in Abhängigkeit von beobachteten Variablen X und unbeobachteten Größen U beschreibt. Empirisch beobachtet wird jeweils nur eines der beiden potenziellen Ereignisse Y, Y_{0i} im Fall, dass i dem kausalen Faktor T nicht ausgesetzt war (T_i=0), und Y_{1i} für den Fall, dass i der Wirkung von T ausgesetzt war (T_i=1). Die Wirkung der Kovariaten X wirkt sich vermittelt über die nichtparametrische Funktion $\mu(.)$ auf Y aus, wobei die Wirkung sowohl der beobachteten Kovariaten X als auch der unbeobachteten Einflussfaktoren U prinzipiell von T abhängig sein kann (d.h. im Allgemeinen gilt $\mu_0(.) \neq \mu_1(.)$ und $U_0 \neq U_1$).

1996), der jeweils im Hinblick auf die Wirkung einer Instrumentalvariable definiert ist. Im Zusammenhang mit Matchingverfahren spielt dieser allerdings eine bestenfalls marginale Rolle und wird von uns im Folgenden vernachlässigt. Wir gehen an dieser Stelle ebenfalls nicht weiter auf den *average treatment effect on the untreated* (ATU) ein, der äquivalent zum ATT für die Kontrollstichprobe definiert ist und den durchschnittlichen kausalen Effekt von T unter den empirisch nicht von T betroffenen ($T_i=0$) Personen beschreibt.

7 Aus Hollands (1986) Position, dass Kausalfaktoren prinzipiell manipulierbar sein müssten, hat sich in der soziologischen Rezeption der Konsens gebildet, das kontrafaktische Konzept sei sozialwissenschaftlichen Fragestellungen prinzipiell unangemessen, da in den Sozialwissenschaften die Frage nach den Auswirkungen von unveränderlichen Attributen – etwa Geschlecht, soziale Herkunft oder ethnische Zugehörigkeit – im Unterschied zu den Naturwissenschaften konstituierend sei. Wir teilen weder Hollands spezifische Position noch deren Rezeption in der Soziologie, sondern verstehen Kausaleffekte vielmehr in einem weiteren Sinne als Auswirkung *exogen variierender Randbedingungen* auf soziales Handeln. In diesem Sinn ist es vollkommen legitim, von einem kausalen Effekt des (individuell nicht beeinflussbaren) Geschlechts etwa auf individuelle Lebensverläufe zu sprechen und diese empirisch zu ermitteln (vgl. etwa Sobel 1998). In substanziellen Beiträgen steht bei genauerer Betrachtung in aller Regel jedoch gar nicht die Ermittlung (oft unbestrittener) kausaler Geschlechtereffekte per se, sondern vielmehr der Nachweis der zugrundeliegenden Mechanismen im Mittelpunkt. Aus dem Verständnis des RCM folgt unmittelbar, dass dieser Nachweis nur über ein Untersuchungsdesign zu führen ist, dass in einer nachvollziehbaren Weise exogene Variation in den vermuteten Mechanismen erfasst und dadurch die Identifikation der (jeweils spezifischen) kausalen Effekte ermöglicht.

Unter diesen Annahmen können die beiden Parameter wie folgt definiert werden: zunächst der *bedingte average treatment effect* (ATE_X) als Differenz der alternativen Ereignisse Y an der Kovariatenstelle X als

$$ATE_X = E(\delta_i | X) = \mu_1(X) - \mu_0(X) \tag{6}$$

und daraus abgeleitet der *average treatment effect* (ATE) als Durchschnitt der bedingten average treatment effects ATE_X über die gesamte Verteilung der Kovariaten X als

$$ATE = E(\delta_i) = \int_X \left[\mu_1(X) - \mu_0(X)\right] dF(X) \approx \sum_X ATE_X \; f(X) \tag{7}$$

Äquivalent dazu ist der *bedingte average treatment effect on the treated* (ATT_X) an der Kovariatenstelle X definiert als

$$ATT_X = E(\delta_i | X, T=1) = \mu_1(X) - \mu_0(X) + E(U_1 - U_0 \,|\, X, T=1) \tag{8}$$

und daraus abgeleitet der *average treatment effect on the treated* (ATT) wiederum als Durchschnitt aller bedingten ATT_X über die gesamte Verteilung der Kovariaten X als

$$\begin{aligned} ATT_X = E(\delta_i \,|\, T=1) &= \int_X \left[\mu_1(X) - \mu_0(X)\, E(U_1 - U_0 \,|\, X, T=1)\right] df(X) \\ &\approx \sum_X ATT_X \, f(X). \end{aligned} \tag{9}$$

Danach unterscheiden sich ATE und ATT konzeptionell allein durch den Term $E(U_1 - U_0 \mid X, T=1)$, der für die Experimentalgruppe E die Auswirkung ungemessener Faktoren auf die Ergebnisvariable Y angibt. Der Term umfasst beispielsweise die mögliche Wirkung privaten Wissens der Akteure, aber auch die Wirkung sonstiger, in der Analyse nicht beobachteter Faktoren, welche die empirische Zuweisung der Akteure i in die Experimentalgruppe E bestimmen. Immer wenn angenommen werden muss, dass solche unbeobachteten Einflussfaktoren relevant sind, liegt eine endogene Zuweisung zur Testgruppe vor (das so genannte Selbstselektionsproblem bzw. „selection on unobservables" nach Heckman und Robb 1985; „hidden bias" in der Terminologie von Rosenbaum 2002). Da in diesem Fall $ATT_X \neq ATE_X$, ist die empirische Abschätzung des kausalen Effekts von T in aller Regel – d.h. ohne zusätzliche Informationen – nicht möglich.

Ganz allgemein ergibt sich aus diesen Überlegungen, dass sich der Parameter ATT_X empirisch immer dann identifizieren lässt, wenn das Problem des Vergleichs homogener Einheiten lokal (d.h. für einen spezifischen Kovariatenvektor X) gelöst, und zusätzlich angenommen werden kann, dass die Zuweisung in die Kontrollgruppe C konditional unabhängig vom Ereignis Y ist (also $E(U_0)=0$ gilt; vgl. Heckman, Smith und Todd 1998, Heckman et al. 2003). Der Populationsparameter ATT kann im Anschluss daran immer dann identifiziert werden, wenn ATT_X über die gesamte Verteilung von X schätzbar ist. Auf die entsprechenden (und inhaltlich oft eigentlich interessierenden) ATE-Parameter kann schließlich nur dann generalisiert werden, wenn das Selbstselektionsproblem sowohl für die Experimentalgruppe E als auch die Kontrollgruppe C als gelöst betrachtet werden kann.

Durch die randomisierte Zuweisung von Test- und Kontrollbedingung sind beide Bedingungen in einer Experimentalsituation unmittelbar erfüllt (vgl. Rosenbaum 2002): die Zufallsauswahl der Testgruppe aus der Stichprobe der Versuchspersonen garantiert zum einen im Aggregat die Vergleichbarkeit von Experimental- und Kontrollgruppe anhand ihrer Merkmale *X*, und zum anderen ist eine endogene Zuweisung zu einer der beiden Vergleichsgruppen von vorneherein ausgeschlossen. Die experimentelle Versuchsanordnung identifiziert dann durch einen einfachen Mittelwertvergleich sowohl ATT und – da bei exogener Zuweisung ATT = ATE gilt – auch den ATE für den manipulierten Faktor, und erlaubt, je nach Stichprobengröße, auch die Schätzung subgruppenspezifischer ATEs bzw. ATTs.

Dem partiellen Regressionskoeffizient β_T kommt eine vergleichbare Rolle im Rahmen parametrischer Regressionsmodelle zu. ATE und ATT-Parameter sind jedoch nur dann durch β_T identifiziert, wenn die zugrunde liegenden statistischen Annahmen des Regressionsmodells erfüllt sind. Für die ATE-Parameter trifft dies insbesondere nur dann zu, wenn der Faktor *T* unter Kontrolle der Kovariaten *X* exogen, d.h. konditional unabhängig vom Fehlerterm des Regressionsmodells ist, das Selbstselektionsproblem also durch geeignete Kovariatenauswahl gelöst werden kann (Heckman und Robb 1985).[8] Zudem hängt die Schätzung sowohl der ATT als auch der ATE-Parameter von der korrekten Modellspezifikation ab, also insbesondere davon, dass die Annahmen über die funktionale Form des Modells, mit deren Hilfe v.a. in sparsam besetzten Datenbereichen kontrafaktisch extrapoliert wird, den datengenerierenden Mechanismus zuverlässig abbilden (Wooldridge 2002; Berk 2004).[9] Soweit diese relativ weitreichenden Annahmen als erfüllt angesehen werden, können die bedingten ATT_X- bzw. ATE_X-Parameter durch geeignete Interaktionseffekte geschätzt werden.

Nichtparametrische Matchingverfahren sind vor diesem Hintergrund von Interesse, weil sie zur Schätzung kausaler Effekte nicht auf die Spezifikation eines mathematischen Modells rekurrieren, und durch entsprechende Spezifikationsfehler verzerrte Schätzungen mithin vermeiden. Der Verzicht auf parametrische Annahmen wird allerdings dadurch erkauft, dass Matchingverfahren kausale Wirkungen eines Faktors *T* nur innerhalb des Kovariatenbereichs *S* abschätzen können, in welchem sowohl Personen der Experimental- wie auch der Kontrollgruppe empirisch beobachtet werden. Für diesen Bereich des *common support* $X \in S$ identifizieren Matchingverfahren den bedingten ATT'_X innerhalb von *S* entsprechend Gleichung (6), und den ATT' als

8 Wir gehen hier implizit von einem Eingleichungsmodell, d.h. einer üblichen OLS- oder auch logistischen Regression aus. Für explizite parametrische und nichtparametrische Modellierungen des Selbstselektionsproblems vgl. etwa Manski (1989), Heckman und Robb (1985), Vella und Verbeek (1992), Brüderl (2001) sowie Pötter (2004).

9 Annahmen über die funktionale Form schließen aber auch inhaltlich weitergehende Annahmen wie etwa die Annahme unkorrelierter unbeobachteter Heterogenität (unkorrelierter Fehlervarianzen) über die Ebenen eines Mehrebenenmodells ein (vgl. die ausführlichere Darstellung in Engel 2004).

$$ATT' = E(\delta_i \mid X \in S, T=1) = \frac{\int_{X \in S} \left[\mu_1(X) - \mu_1(X) + E(U_1 - U_0 \mid X, T=1)\right] dF(X \mid T=1)}{\int_{X \in S} dF(X \mid T=1)} \quad (10)$$

(vgl. Heckman, Smith und Todd 1997, 1998).

Da sie sich auf konzeptionell unterschiedliche Parameter beziehen, werden regressions- bzw. matchingbasierte Schätzwerte des ATT(') typischerweise unterschiedlich ausfallen. Eine Konvergenz von regressionsbasiertem ATT und matchingbasiertem ATT' ist immer nur dann zu erwarten, wenn der Bereich des *common support* praktisch die gesamte Matrix der empirischen Kovariatenkombinationen umfasst und die parametrische Spezifikation des Regressionsmodells empirisch (näherungsweise) korrekt ist, oder wenn die „wahre" parametrische Modellspezifikation so ist, dass sie aus den innerhalb des *common support* zur Verfügung stehenden Daten für die Gesamtpopulation korrekt geschätzt werden kann. Zur Identifikation der entsprechenden ATE(')-Parameter muss wie in parametrischen Verfahren die (unter Kontrolle von Kovariaten) konditional unabhängige Zuweisung von T angenommen werden (vgl. Rubin 1974, 1978; Rosenbaum und Rubin 1983; Rosenbaum 2002; Heckman, Smith und Todd 1998).

III. Kausalanalyse durch Propensity Score Matching

Der Verzicht auf parametrische Modellannahmen und die dadurch vergleichsweise größere Robustheit der empirischen Schätzergebnisse ist eine wichtige Eigenschaft von Matchingverfahren, die ihre Anwendung anstelle von oder zumindest zusätzlich zu üblichen Regressionsverfahren rechtfertigt. Ein zweiter Vorzug der Matchingverfahren liegt in ihrer konzeptionellen Klarheit auf Grund der methodologischen Nähe zum Rubin Causal Model, wodurch sich die empirischen Ergebnisse einer Matchinganalyse unmittelbar zu den oben definierten zentralen Parametern des RCM in Beziehung gesetzt werden können. Zudem sind dadurch auch die weitergehenden inhaltlichen Annahmen klar spezifiziert, die jeweils zur Identifikation bestimmter kausaler Parameter erfüllt sein müssen. Ein weiterer praktischer Vorzug der Matchingverfahren ist, dass sie als nichtparametrisches Verfahren konstruktionsbedingt kausale Effekte nur im überlappenden Kovariatenbereich S abschätzen können, und dadurch in der empirischen Anwendung auch dafür sensibilisieren, in welchem Ausmaß regressionsbasierte Schlussfolgerungen von den zugrunde liegenden Annahmen über die funktionale Form des Modells abhängig sind.

Substanziell betrachtet erzwingt die Verwendung von Matchingverfahren schließlich geradezu, sich in der Analyse auf einen (zentralen) kausal wirksamen Faktor zu konzentrieren, dessen Wirkung mit einer bestimmten Datenbasis empirisch getestet werden soll. Quasi als Nebenprodukt der statistischen Strategie wird dazu eine klare Spezifikation der Ausgangshypothese sowie der empirischen Testbedingungen notwendig. Im Unterschied zur gängigen Regressionspraxis ergibt sich dadurch zum Beispiel eine klar hierarchische Struktur der unabhängigen Variablen und ihrer Funktion in der Daten-

analyse: Matchingverfahren verdeutlichen unmittelbar, dass nicht jede unabhängige Variable denselben kausalen Status besitzt. Vielmehr erfordert das Verfahren, den relevanten kausalen Faktor T und die Kovariaten X explizit zu benennen. Da letztere (lediglich) dazu dienen, die Annahmen der *unit homogeneity* und *strict ignorability* zu rechtfertigen, kommt ihnen in der Interpretation erkennbar *kein* kausaler Status zu. Matchingverfahren führen einen Hypothesentest allein für die Wirkung von T auf Y durch und ermitteln damit kausale Effekte im Sinne der „effects of causes" (vgl. Goldthorpe 2001); Matchingverfahren sind wenig geeignet, Varianzerklärungen im Sinne des Konzepts der „causes of effects" zu erfassen.

1. Vorgehensweise

Allen Matchingverfahren gemeinsam ist die Grundidee, den kausalanalytischen Vergleich vergleichbarer Einheiten nur für diejenigen Merkmalskombinationen vorzunehmen, für die empirisch sowohl Fälle aus der Experimental- als auch der Kontrollstichprobe (matches, im Folgenden Vergleichsbeobachtungen) vorliegen. Infolgedessen kann die statistische Analyse *nichtparametrisch*, d.h. unter Verzicht auf weitergehende Annahmen zur funktionalen Form des Modells, erfolgen. Die empirische Schätzung der ATT- und ATE-Parameter erfolgt in aller Regel durch einfache deskriptive Verfahren, etwa durch Mittelwertvergleiche oder weiterführende univariate Verteilungsanalysen in den beiden Vergleichsstichproben.

Tatsächlich können alle Matchingverfahren grundsätzlich einfach als Gewichtungsverfahren aufgefasst werden, die durch jeweils unterschiedliche Algorithmen für eine Anpassung der Verteilung der Hintergrundfaktoren X in der Kontrollstichprobe an die entsprechende Verteilung in der Experimentalstichprobe sorgen. Die Schätzfunktion für den ATT' lässt sich beispielsweise allgemein angeben mit

$$ATT' = \sum_{i \in E} w_i \left[Y_{1i} - \sum_{j \in C} W_{i,j} Y_{0j} \right], \tag{11}$$

d.h. als (mit w_i gewichteter) Durchschnitt der Differenzen zwischen den Ereignissen Y_{1i} der Experimentalfälle i und der mit $W_{i,j}$ gewichteten Ereignisse Y_{0j} der Kontrollbeobachtungen j (Heckman, Smith und Todd 1997, 1998). Einzelne *Matchingalgorithmen* unterscheiden sich in der Konstruktion der Vergleichsgewichte $W_{i,j}$, insbesondere darin, ob ein reales Matching von spezifischen Beobachtungen i und j aus beiden Stichproben vorgenommen wird, oder ob die Vergleichsgruppe als neu gewichteter Durchschnitt aller Kontrollbeobachtungen konstruiert wird. Grundlage der Konstruktion der Vergleichsgruppe wiederum ist das so genannte *Zuweisungsmodell*, d.h. die empirische Beschreibung des Zusammenhangs zwischen der Zugehörigkeit zur Experimentalgruppe E und den beobachteten Kovariaten X. Wir betrachten im Folgenden die einzelnen Komponenten einer Matchinganalyse im Detail.

2. Zuweisungsmodell und Propensity scores

Die Spezifikation eines Zuweisungsmodells (sog. *assignment model*) bildet die zentrale Grundlage jeden Matchingverfahrens. Dem Zuweisungsmodell kommt im Rahmen einer Matchinganalyse dieselbe Funktion zu wie dem Einschluss von Kovariaten in einer Regressionsanalyse: Das Zuweisungsmodell stellt die Vergleichbarkeit der Beobachtungseinheiten her, indem es die Ähnlichkeit bzw. Unähnlichkeit der beiden Stichproben beschreibt, und dadurch anschließend das Matching vergleichbarer Einheiten ermöglicht. Zugleich sollte ein sozialwissenschaftliches Zuweisungsmodell idealerweise auch für die potenziell endogene Zuweisung (Selbstselektion) von Akteuren in den Experimentalstatus kontrollieren, um dadurch ATT- und ATE-Parameter zu identifizieren. In das Zuweisungsmodell sollten deshalb alle beobachteten Kovariaten X eingehen, die sowohl den Experimentalstatus T als auch (direkt oder indirekt) das Ereignis Y beeinflussen. Die Kovariaten X müssen dabei ökonometrisch nicht strikt exogen von Y sein, dürfen aber natürlich nicht endogen von T, also nicht selbst eine *Folge* des Experimentalstatus T sein.[10]

Das einfachste Zuweisungsmodell, das zudem selbst ebenfalls wieder nichtparametrisch ist, besteht darin, direkt den multidimensionalen Kovariatenvektor X als Grundlage eines Matchingalgorithmus zu betrachten, und dementsprechend ein *exaktes Matching* von Experimental- und Kontrollpersonen über alle Kovariaten X vorzunehmen. In der Praxis, und zudem insbesondere dann, wenn viele Kovariaten vorliegen, so dass das Zuweisungsmodell eigentlich inhaltlich zufriedenstellend geschätzt werden kann, wird diese Strategie aber selbst mit großen Datensätzen offensichtlich schnell an ihre Grenzen stoßen.

Von zentraler Bedeutung für die praktische Anwendung von Matchingverfahren ist deshalb das theoretische Ergebnis von Rosenbaum und Rubin (1983), dass es zur Anpassung von Kovariatendifferenzen zwischen beiden Stichproben ausreicht, ein Matching lediglich über einen Ähnlichkeitsindex, den so genannten *Propensity score*, durchzuführen. Der *Propensity score* ist dabei definiert als die bedingte Wahrscheinlichkeit einer Beobachtungsperson i mit Merkmalen X_i, der Wirkung des kausalen Faktors T ausgesetzt zu sein (= als Mitglied der Experimentalgruppe E beobachtet zu werden). Er wird empirisch üblicherweise als Vorhersagewahrscheinlichkeit $P(X)$ z.B. aus einer logistischen oder einer Probit-Regression mit T als abhängiger und X als unabhängigen Variablen bestimmt[11] und im Anschluss daran als Ähnlichkeitsmetrik für das Matching von Experimental- und Kontrollgruppe verwendet. Zur Identifizierung des kausalen Ef-

10 Wenn beispielsweise der kausale Effekt einer Arbeitslosigkeit auf den weiteren Erwerbsverlauf ermittelt werden soll, ist es weder im Rahmen einer Matching- noch innerhalb einer Regressionsanalyse sinnvoll, die Zahl der Stellenwechsel als Kovariate aufzunehmen. Eine im Vergleich zu durchgehend Beschäftigten höhere Zahl von Stellenwechseln unter Arbeitslosen dürfte gerade *infolge* der Arbeitslosigkeit hervorgerufen sein, ist also eine Komponente der untersuchten Wirkung und stellt keine Anpassung der Hintergrundfaktoren *vor* der Arbeitslosigkeit dar.

11 In Simulationsstudien haben sich parametrische Modellannahmen in der Konstruktion der propensity scores als wenig problematisch erwiesen (vgl. Deheija und Wahba 2002; Smith und Todd 2002).

fekts von T wird im Propensity score matching dann angenommen, dass die CIA-Annahme

$$(Y_1, Y_0) \perp T \,\big|\, \Pr(T=1 \,|\, X) \tag{12}$$

empirisch gerechtfertigt ist, d.h. dass der Experimentalstatus T unter Kontrolle des Propensity scores $P(X)$ unabhängig von den potenziellen Ereignissen Y gesetzt ist.

3. Matchingalgorithmen

Ausgehend von einer empirischen Schätzung eines gültigen Zuweisungsmodells und den daraus generierten Propensity scores kann dann mithilfe unterschiedlicher Matchingalgorithmen versucht werden, eine Experimentalsituation dadurch zu imitieren, dass die Verteilung der beobachteten Hintergrundfaktoren X in der Kontrollstichprobe C an die entsprechende Verteilung in der Experimentalstichprobe E angepasst (balanciert) wird.[12] In der angewandten Forschung lassen sich im wesentlichen drei Matchingalgorithmen unterscheiden: Stratifizierung und Nearest-neighbor-Verfahren (Rubin 1978; Rubin und Rosenbaum 1983, 1985; Rosenbaum 2002; Deheija und Wahba 1999, 2002), die vor allem aus dem Bereich der angewandten Statistik stammen, sowie andererseits Kernel Matching und verwandte Verfahren, die in den letzten Jahren in der Ökonometrie entwickelt wurden (v.a. Heckman, Smith und Todd 1997, 1998; Lechner 1999, 2002; Smith und Todd 2002).

Ein wichtiger Unterschied zwischen diesen Strategien besteht darin, dass beim stratifiziertem Matching, aber auch beim Nearest-neighbor Matching nur ein Teil der Kontrollbeobachtungen C zur Konstruktion des kontrafaktischen Ereignisses $\{Y_{0i}\}$ herangezogen werden, im Kernel Matching und verwandten Verfahren (z.B. Local Linear oder Mahalanobis Matching) wird dagegen ein gewichteter Durchschnitt aus der gesamten Kontrollstichprobe gebildet. Ein zweiter Unterschied liegt darin, dass Nearest-neighbor- und verwandte Verfahren zu einem spezifischen Matching zwischen Beobachtungen i der Experimentalgruppe und Beobachtungen j der Kontrollgruppe führen, während Stratifizierung und Kernel Matching die kontrafaktische Beobachtung $\{Y_{0i}\}$ jeweils als gewichtete Durchschnitte aus einer Population von Kontrollbeobachtungen bestimmen. *Tabelle 1* fasst die wesentlichen Merkmale einiger gängiger Algorithmen zusammen.

Wie unser eigenes empirisches Beispiel unten zeigen wird, können die Vorzüge einzelner Ansätze zum Teil auch kombiniert werden, etwa indem innerhalb von Schichten ein kernelbasiertes Matching oder ein Nearest-neighbor Matching durchgeführt wird. Ebenso ist es möglich, durch Angabe eines *calipers*, d.h. einer maximal zulässigen Distanz c, auch innerhalb von Nearest-neighbor-Verfahren wenig adäquate Paarvergleiche mit hohen Differenzen im Propensity score auszuschließen, oder durch multiples Mat-

12 Durch die Balancierung der Beobachtungsdaten wird also gewissermaßen eine Implikation experimenteller Randomisierung imitiert. Wenn das unterstellte Zuweisungsmodell gültig (d.h. die CIA-Annahme erfüllt) ist, führt eine erfolgreiche Balancierung der Kovariatenverteilung zur Identifikation des kausalen Effekts; eine erfolgreiche Balancierung allein ist allerdings naturgemäß kein Nachweis über die Gültigkeit der CIA-Annahme.

Tabelle 1: Zentrale Matchingalgorithmen

	Beschreibung	Gewichtungsfunktion $W_{i,j}$
Stratifizierung	Schichtung der Stichprobe über *P(X)*; Mittelwertvergleich innerhalb der Schichten *k*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in k_i \\ 0 & \text{sonst} \end{cases}$
Nearest-neighbor matching	Direkte Paarbildung aus der Experimental- und Kontrollstichprobe; Matching anhand größter Ähnlichkeit auf *P(X)*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in \arg min_k \left\{ \left\| P_i(X) - P_k(X) \right\| \right\} \\ 0 & \text{sonst} \end{cases}$
Caliper matching	Direkte Paarbildung aus der Experimental- und Kontrollstichprobe; Matching innerhalb eines Ähnlichkeitsradius *c* über *P(X)*	$W_{i,j} = \begin{cases} 1/N_j & \text{für } j \in \left\| P_i(X) - P_j(X) \right\| \leq c \\ 0 & \text{sonst} \end{cases}$
Mahalanobis matching	Kontrollbeobachtung als gewichteter Durchschnitt aller Fälle der Kontrollstichprobe; Gewichtungsfaktoren definiert durch Mahalanobisdistanzen über *X*	$W_{i,j} = \begin{cases} \frac{\left[(X_j - X_i) C_X (X_j - X_i)\right]^{-1}}{\sum_{k \in C} \left[(X_k - X_i) C_X (X_k - X_i)\right]^{-1}} \end{cases}$ mit C_x-Kovarianzmatrix des Kovariatenvektors X
Kernel matching	Kontrollbeobachtung als gewichteter Durchschnitt aller Fälle der Kontrollstichprobe; Gewichtungsfaktoren ermittelt durch Distanzfunktion über *P(X)*	$W_{i,j} = \begin{cases} \frac{K\left[\left(P_j(X) - P_i(X)\right)/h\right]}{\sum_{k \in C} K\left[\left(P_k(X) - P_i(X)\right)/h\right]} \end{cases}$ mit K – Kernelfunktion, z.B. $\phi(.)$ und h – Bandbreitenparameter

ching auch im Rahmen von Nearest-neighbor-Algorithmen die kontrafaktische Beobachtung als gewichteten Durchschnitt mehrerer „ähnlich guter" Vergleichsbeobachtungen zu bilden (z.B. Smith 1997). Welche Kombination dabei auch immer verwendet wird, zentral bleibt das Ziel, eine Kontrollstichprobe mit empirisch der Experimentalstichprobe entsprechenden Hintergrundfaktoren *X* zu erzeugen. Die Güte eines spezifischen Zuweisungsmodells und des darauf aufbauenden Matchingalgorithmus sollte deshalb durch so genannte *balancing tests* auf Mittelwertunterschiede (*bias*) auf *P(X)* oder einzelnen (zentralen) Kovariaten *X* für die Analysepopulation oder auch spezifische Subgruppen ermittelt werden (Rosenbaum und Rubin 1983, 1985; Rosenbaum 2002). Bei festgestellten Verstößen kann dann sowohl der Matchingalgorithmus wie auch das Zuweisungsmodell (etwa durch Aufnahme von Interaktionseffekten) jeweils spezifisch angepasst werden.[13]

Alle Unterschiede in der empirischen Vorgehensweise der verschiedenen Matchingalgorithmen sollten allerdings nicht darüber hinwegtäuschen, dass alle Verfahren asymptotisch äquivalent sind, da der kausale Vergleich bei großem N in jedem Verfah-

13 Balancing tests sind dabei ausschließlich als Tests des durch einen spezifischen Algorithmus erzielten Matchings für ein gegebenes Zuweisungsmodell zu verstehen. Sie stellen keinen statistischen Test für die Validität des unterstellten Zuweisungsmodells dar.

ren ausschließlich auf exakten Vergleichsbeobachtungen basiert. In der Praxis ergeben sich natürlich Unterschiede zwischen den Verfahren, so dass es sich in der angewandten Literatur eingebürgert hat, vor allem bei kleineren Stichproben oder sehr spezifischen Kausaleffekten Sensitivitätsanalysen mit jeweils variablem Analyseaufbau durchzuführen. Als Faustregeln für die empirische Analyse lässt sich sagen, dass Single Nearest-neighbor- Verfahren, evtl. ergänzt um einen *caliper*, in aller Regel sehr gut geeignet sind, vergleichbare Kontrollgruppen mit minimalem *bias* zu erzeugen. Wenn jedoch entweder nur sehr wenige sehr gut geeignete Vergleichsbeobachtungen oder auch sehr viele relativ gut bis sehr gute Vergleichsbeobachtungen zur Verfügung stehen, dann können Multiple Nearest-neighbor-Verfahren oder auch Kernel Matching dazu beitragen, (wenigstens bzw. zusätzlich) die Varianz der Parameterschätzungen zu verringern (z.B. Smith 1997; Smith und Todd 2001). Wie für nichtparametrische Verfahren üblich, werden beide Ziele vor allem in kleinen Stichproben mit wenigen Kontrollbeobachtungen nicht gleichzeitig zu erreichen sein.

4. Punktschätzungen und statistische Inferenz

Da der eigentliche Kern der Matchingverfahren in der Konstruktion der kontrafaktischen Vergleichsgruppe besteht, gestaltet sich die anschließende Schätzung kausaler Effekte sehr unaufwendig. Unter Gültigkeit der Annahme konditionaler Unabhängigkeit des Experimentalstatus lässt sich beispielsweise der *average treatment effect on the treated* (ATT') nach Gleichung (9) als einfacher Mittelwertvergleich für die Experimentalgruppe E und der gematchten bzw. neu gewichteten Kontrollbeobachtungen C berechnen. Entsprechend einfach können subgruppenspezifische (konditionale) Parameter als Mittelwertvergleich innerhalb von Subgruppen ermittelt werden (z.B. Lechner 1999). Und ebenso einfach können durch Vergleiche der univariaten Ergebnisverteilungen auch kausale Effekte auf andere Merkmale, etwa den Median oder die Quartile der Verteilung, bestimmt werden (vgl. Imbens 2004; Gangl und DiPrete 2004). Da zudem bei konditional unabhängiger Zuweisung des Experimentalstatus der konditionale ATT_X dem konditionalen ATE_X entspricht, ergibt sich der ATE'-Parameter aus Gleichung (5) als (der Struktur der Gesamtpopulation angepasster) gewichteter Mittelwert der konditionalen ATT_X-Parameter.

Da die gematchten Stichproben gleichzeitig eine Verteilung individueller ATT- bzw. ATE-Parameter erzeugen, können Signifikanztests für die ermittelten Effekte im Prinzip mithilfe gewöhnlicher t-Tests bzw. äquivalenten nichtparametrischen Rangsummentests durchgeführt werden. Die Verwendung dieser Teststatistiken unterstellt implizit allerdings individuell gegebene Propensity scores. Verteilungstheoretische Modelle, die zusätzlich berücksichtigen, dass empirisch lediglich *geschätzte* Propensity scores vorliegen, existieren bislang lediglich für einzelne (Klassen von) Matchingalgorithmen (vgl. etwa Heckman, Smith und Todd 1998 für Kernel und Local Linear Matching-Verfahren). In der Praxis dominieren deshalb Bootstrapverfahren zur Ermittlung der Standardfehler (vgl. Efron und Tibshirani 1993; Mooney und Duval 1993).

IV. Ein Anwendungsbeispiel: Kausaleffekte von Arbeitslosigkeit

Wir wollen die empirische Anwendung von Matchingverfahren im Folgenden an einer Analyse der Karriereeffekte von Arbeitslosigkeit mit Daten des Sozio-ökonomischen Panels beispielhaft illustrieren. Wir erwarten theoretisch, dass Arbeitslosigkeit kurz- und eventuell auch längerfristige Karrierenachteile auslöst, weil mit dem Verlust des Arbeitsplatzes auch eine Entwertung firmenspezifischen Humankapitals sowie unter Umständen die Notwendigkeit einer beruflichen Neuorientierung verbunden ist. Diese Forschungsfrage zielt eindeutig auf den (kurz- bzw. längerfristigen) kausalen Effekt einer Arbeitslosigkeit, d.h. auf die kontrafaktische Frage nach den *„effects of causes"* ab, wie stark sich die Erwerbsverläufe von empirisch arbeitslos Gewordenen allein durch die Erfahrung der Arbeitslosigkeit *verändert* haben. Welcher Anteil etwa der Einkommensunterschiede in der Bundesrepublik auf Arbeitslosigkeitserfahrungen zurückzuführen ist, oder welche Auswirkung alternative Determinanten von Arbeitsmarkterfolg – etwa Bildung oder Geschlecht – besitzen, ist für die Antwort auf diese Frage nicht von Belang.

Wir betrachten stattdessen im Folgenden also nur die Frage, welche Auswirkung ein Arbeitsplatzverlust auf die nachfolgenden individuellen Beschäftigungschancen, das erzielte Erwerbseinkommen sowie die individuelle Arbeitsplatzsicherheit hat. Zur Vereinfachung der Analyse verwenden wir lediglich Arbeitsplatzverluste aus den Rezessionsjahren 1993 und 1994, d.h. alle Ereignisse, die zwischen den Befragungswellen J und K des Sozio-ökonomischen Panels beobachtet wurden (vgl. zum Datensatz SOEP-Group 2001). Wir beobachten die anschließenden Erwerbsverläufe sowohl unmittelbar (T_0+*1*) als auch bis zu insgesamt fünf Jahre (T_0+*5*) nach einem Arbeitsplatzverlust. Für jeden der bis zu fünf Befragungszeitpunkte nach dem ursprünglichen Arbeitsplatzverlust erfassen wir die Beschäftigungschance als Wahrscheinlichkeit, (abhängig oder selbständig) erwerbstätig zu sein, das logarithmierte reale Bruttomonatseinkommen in dieser Beschäftigung sowie prospektiv aus den Folgewellen die tatsächliche weitere Beschäftigungsdauer beim gegenwärtigen Arbeitgeber als Maß der Arbeitsplatzsicherheit. Da wir uns für den kausalen Effekt eines Arbeitsplatzverlustes für die tatsächlich von Arbeitslosigkeit Betroffenen interessieren, wollen wir im Folgenden kausale Effekte im Sinne der ATT-Parameter ermitteln.

Aufgrund der methodologischen Trennung von Kovariatenkontrolle und Effektschätzung im Matchingverfahren können wir alle drei abhängigen Variablen auf der Basis jeweils desselben Zuweisungsmodells und Matchingalgorithmus bearbeiten und lediglich im letzten Schritt der Punktschätzung dem Skalenniveau jeweils angemessene Verfahren (Mittelwert- bzw. Anteilsvergleich für Einkommen und Beschäftigungschancen, Kaplan-Meier-Schätzer für Ereignisdaten) einsetzen. Da der Median der weiteren Beschäftigungsdauer in aller Regel nicht beobachtet werden kann, beschreiben wir die Beschäftigungsstabilität durch den Wert der Überlebensfunktion *G(36)* für jeweils den Zeitpunkt 3 Jahre nach dem betreffenden Wellenинterview T_0+*k, k=1...5*.

Wir stellen für jede der drei abhängigen Variablen drei unterschiedliche Analysen vor, die jeweils unterschiedliche Matchingalgorithmen verwenden: Zum einen einen 1×1 Nearest-neighbor-Algorithmus, der für jede Experimentalbeobachtung *i* innerhalb des common support *S* und des Calipers *c* ein Matching mit einer (der nächstliegen-

den) Beobachtung *j* aus der Kontrollgruppe erzeugt, sowie eine zweite (1×10) Nearest-neighbor-Variante, in der (soweit vorhanden) bis zu 10 verfügbare Kontrollbeobachtungen im common support und innerhalb des Calipers als Paarvergleiche für eine Experimentalbeobachtung herangezogen werden. Wir verwenden in beiden Analysen einen sehr engen Caliper von c=0.05*Standardabweichung des Propensity scores, um nur sehr gute Paarvergleiche in der Analyse zu berücksichtigen und dadurch eine sehr hohe Biasreduktion zu erreichen.[14] Wir kontrastieren die Nearest-neighbor-Verfahren mit einer Analyse auf der Basis eines Kernel Matching, in welchem wir einen Epanechnikov-Kernel mit einer Bandbreite von ebenfalls h=0.05*Standardabweichung des Propensity scores verwenden. Diese drei Analysen werden zusätzlich jeweils stratifiziert für die alten bzw. neuen Bundesländer durchgeführt, so dass Vergleichsbeobachtungen jeweils nur innerhalb der Teilregionen gesucht werden.[15]

Allen Analysen liegt zudem ein gemeinsames Zuweisungsmodell zugrunde, welches das Risiko eines Arbeitsplatzverlustes zwischen den SOEP-Wellen J und K durch Bildung, Berufserfahrung, Geschlecht, ethnische Zugehörigkeit und Migrationsstatus, bisheriges Erwerbseinkommen, detaillierten Beruf und Branche der gegenwärtigen Erwerbstätigkeit, sowie die Firmengröße des gegenwärtigen Arbeitgebers vorhersagt. Wir schätzen dazu – schichtspezifisch – eine logistische Regression, mit der wir ein Pseudo-R^2 von insgesamt 7,6 Prozent (West: 7,8 Prozent, Ost 7,4 Prozent) erreichen. Da wir im SOEP über weiterreichende Längsschnittinformationen verfügen, würde es sich in einer inhaltlichen Analyse aber in jedem Fall anbieten, zusätzlich etwa noch für die Anzahl der Arbeitslosigkeitsspells innerhalb der letzten fünf Jahre oder für ähnliche Längsschnittkovariaten zu kontrollieren und dadurch die Plausibilität der *conditional independence*-Annahme zu erhöhen.

Wie die in *Tabelle 2* dargestellten *Balancing tests* zeigen, gelingt uns in allen drei Matchingalgorithmen und für jeweils alle drei abhängigen Variablen eine sehr weitreichende Angleichung der Verteilung der Propensity scores in der Kontrollstichprobe an die entsprechende Verteilung in der Experimentalgruppe. Die verbleibenden Mittelwertunterschiede sind minimal, ebenso wie (dank der großen Stichprobe des SOEP) die Unterschiede zwischen den einzelnen Algorithmen. In keiner Analyse ist der ver-

14 Diese Strategie hängt offensichtlich von der empirischen Datenlage ab, und dabei insbesondere von der Verfügbarkeit sehr guter (sehr naher) Kontrollbeobachtungen möglichst über den gesamten common support. Diese Bedingung ist tendenziell natürlich in großen Stichproben, aber auch bei wenig speziellen Experimentalgruppen eher erfüllt. In unseren eigenen Analysen gelingt es uns mit jedem der gewählten Verfahren, zu jeweils allen Experimentalbeobachtungen innerhalb des common support adäquate Kontrollbeobachtungen zu finden. Zusätzlich umfasst der Bereich des common support immer über 95 Prozent aller Experimentalfälle, so dass eine hinreichende Generalisierbarkeit der Ergebnisse gegeben erscheint. Vor diesem Hintergrund ist zu erwarten, dass sich das 1×1 Nearest-neighbor-Verfahren als relativ ineffizient erweisen wird, da sowohl das multiple Nearest-neighbor als auch das Kernel Matching die reichlich vorhandenen Kontrollbeobachtungen vollständig(er) in die Parameterschätzung einbeziehen.

15 Die regionale Stratifizierung entspricht einer fixed-effects-Strategie im Rahmen von Regressionsverfahren, mit der in diesem Fall alle beobachteten und unbeobachteten Effekte auf der regionalen Ebene konstant gehalten werden (vgl. Heckman et al. (1997) für eine ausführliche Sensitivitätsanalyse der Konsequenzen von regional unstratifizierten Matchings im Falle von amerikanischen Arbeitsmarktdaten).

Tabelle 2: Balancing Tests der Modellschätzungen

Zeitpunkt		AV: Beschäftigung (1) Nearest neighbor 1×1	AV: Beschäftigung (2) Nearest neighbor 1×10	AV: Beschäftigung (3) Kernel matching	AV: Logarithmiertes Realeinkommen (1) Nearest neighbor 1×1	AV: Logarithmiertes Realeinkommen (2) Nearest neighbor 1×10	AV: Logarithmiertes Realeinkommen (3) Kernel matching	AV: Beschäftigungsstabilität (1) Nearest neighbor 1×1	AV: Beschäftigungsstabilität (2) Nearest neighbor 1×10	AV: Beschäftigungsstabilität (3) Kernel matching
T+1	(1)	–0,0001	–0,0001	<0,0001	–0,0001	–0,0001	≤0,0001	–0,0001	–0,0001	<0,0001
(N_i = 490, 344,	(2)	0,002	0,002	<0,001	0,002	0,002	<0,001	0,002	0,002	<0,001
393)	(3)	–99,8%	–99,8%	–100,0%	–99,8%	–99,8%	–100,0%	–99,8%	–99,8%	–100,0%
T+2	(1)	–0,0001	–0,0001	<0,0001	–0,0002	–0,0002	<0,0001	–0,0002	–0,0002	<0,0001
(N_i = 467, 295,	(2)	0,002	0,002	0,001	0,002	0,002	0,001	0,002	0,002	0,001
334)	(3)	–99,8%	–99,8%	–99,9%	–99,8%	–99,8%	–99,9%	–99,8%	–99,8%	–99,9%
T+3	(1)	–0,0001	–0,0001	<0,0001	–0,0002	–0,0002	<0,0001	–0,0002	–0,0002	<0,0001
(N_i = 451, 274,	(2)	0,002	0,002	0,001	0,002	0,002	<0,001	0,002	0,002	<0,001
296)	(3)	–99,8%	–99,8%	–99,9%	–99,8%	–99,8%	–100,0%	–99,8%	–99,8%	–100,0%
T+4	(1)	–0,0002	–0,0002	<0,0001	–0,0002	–0,0002	<0,0001	–0,0002	–0,0002	<0,0001
(N_i = 420, 234,	(2)	0,002	0,002	<0,001	0,003	0,003	<0,001	0,002	0,002	0,001
251)	(3)	–99,8%	–99,8%	–100,0%	–99,7%	–99,7%	–100,0%	–99,7%	–99,7%	–99,9%
T+5	(1)	–0,0002	–0,0002	<0,0001	–0,0002	–0,0002	≤0,0001	–0,0002	–0,0002	<0,0001
(N_i = 391, 197,	(2)	0,002	0,002	0,001	0,003	0,003	<0,001	0,002	0,002	<0,001
216)	(3)	–99,8%	–99,8%	–99,9%	–99,7%	–99,7%	–100,0%	–99,7%	–99,7%	–100,0%

Anmerkung: Stratifiziertes Nearest-neighbor Matching mit caliper $c = 0{,}05 \cdot \mathrm{sd}(P(X))$ und stratifiziertes Kernel matching mit Epanechnikov-Kernel, Bandbreite $h = 0{,}05 \cdot \mathrm{sd}(P(X))$; Stratifizierung jeweils für die alten bzw. neuen Bundesländer. Angaben in den Tabellenzellen jeweils für (1) Mittelwertdifferenz $P_E(X)–P_C(X)$ im Propensity score zwischen Experimental- und gematchter Kontrollgruppe, statistische Signifikanzangaben für * p < .05 (Bootstrap-Standardfehler, N=100 Replikationen); (2) Standardisierter Bias des Propensity score in der gematchten Stichprobe und (3) Relative Biasreduktion gegenüber den Ausgangsstichproben.

Datenquelle: Sozio-ökonomisches Panel, Wellen J-R, eigene Berechnungen.

Tabelle 3: Der kausale Effekt eines Arbeitsplatzverlustes 1993/94 auf den weiteren Erwerbsverlauf (ATT')

	AV: Beschäftigung			AV: Logarithmiertes Realeinkommen			AV: Beschäftigungsstabilität (G_{36})		
Zeitpunkt	(1) Nearest neighbor 1×1	(2) Nearest neighbor 1×10	(3) Kernel matching	(1) Nearest neighbor 1×1	(2) Nearest neighbor 1×10	(3) Kernel matching	(1) Nearest neighbor 1×1	(2) Nearest neighbor 1×10	(3) Kernel matching
T+1	–0,553*	–0,563*	–0,563*	–0,107	–0,120*	–0,121*	–0,208*	–0,194*	–0,191*
	(0,030)	(0,024)	(0,027)	(0,057)	(0,042)	(0,042)	(0,045)	(0,035)	(0,033)
T+2	–0,261*	–0,289*	–0,289*	–0,153*	–0,094	–0,089	–0,138*	–0,124*	–0,125*
	(0,032)	(0,027)	(0,028)	(0,073)	(0,055)	(0,048)	(0,052)	(0,033)	(0,038)
T+3	–0,266*	–0,235*	–0,241*	–0,168*	–0,125*	–0,127*	–0,052	–0,089*	–0,083*
	(0,037)	(0,031)	(0,027)	(0,064)	(0,045)	(0,051)	(0,052)	(0,040)	(0,032)
T+4	–0,224*	–0,224*	–0,232*	–0,188*	–0,153*	–0,145*	–0,033	–0,066	–0,089*
	(0,034)	(0,025)	(0,026)	(0,064)	(0,050)	(0,052)	(0,049)	(0,039)	(0,043)
T+5	–0,184*	–0,168*	–0,172*	0,003	–0,089	–0,088	0,012	–0,033	–0,040
	(0,041)	(0,031)	(0,028)	(0,071)	(0,057)	(0,047)	(0,058)	(0,045)	(0,042)

Anmerkung: Stratifiziertes Nearest-neighbor Matching mit caliper $c = 0{,}05 \cdot \mathrm{sd}(P(X))$ und stratifiziertes Kernel matching mit Epanechnikov-Kernel, Bandbreite $h = 0{,}05 \cdot \mathrm{sd}(P(X))$; Stratifizierung jeweils für die alten bzw. neuen Bundesländer. Angaben in den Tabellenzellen jeweils für ATT' = Mittelwertdifferenz $Y_E(P(X)) - Y_C(P(X))$ in der abhängigen Variable zwischen Experimental- und gematchter Kontrollgruppe. Bootstrap-Standardfehler des ATT' in Klammern (N=100 Replikationen; statistische Signifikanzangaben für * $p < .05$).

Datenquelle: Sozio-ökonomisches Panel, Wellen J-R, eigene Berechnungen.

bleibende Bias (Zeile (1)) auch nur annähernd statistisch signifikant. Wir gehen also davon aus, dass das Matching der Stichproben jeweils erfolgreich war und eine Abschätzung der ATT-Parameter möglich ist.[16] Es sollte an dieser Stelle aber noch einmal explizit betont werden, dass der Nachweis des erfolgreichen Matchings nicht gleichbedeutend mit einem Modelltest für das Zuweisungsmodell zu verstehen ist. Ob und inwieweit in einer konkreten empirischen Analyse tatsächlich die kausalen ATT- und ATE-Parameter identifiziert sind, hängt allein von der *inhaltlichen* Plausibilität des Zuweisungsmodells ab – die, wie oben angedeutet, in der konkreten Analyse durchaus noch verbesserungsfähig wäre.

Tabelle 3 schließlich gibt die durch die Matchingverfahren ermittelten ATT'-Parameter für den kausalen Effekt eines Arbeitsplatzverlustes 1993/94 auf den weiteren Erwerbsverlauf. Die in der Tabelle ausgewiesenen Koeffizienten sind jeweils als Mittelwertdifferenz in der gematchten Stichprobe bzw. für die Beschäftigungsstabilität als Differenz der Überlebensfunktion *G(36)* ermittelt. Alle Matchingalgorithmen sind konsistent dahingehend, dass ein Arbeitsplatzverlust unmittelbar, aber auch mittelfristig geringere Beschäftigungschancen, Einkommensverluste sowie eine geringere Beschäftigungsstabilität nach sich zieht. Für alle drei abhängigen Variablen fallen die kurzfristigen Effekte dabei deutlich negativer aus als im mittelfristigen Vergleich nach bis zu fünf weiteren (Berufs-)Jahren. Zum Teil gelingt es den Betroffenen also, entstandene negative Karrierefolgen wieder auszugleichen, insbesondere im Hinblick auf die mittelfristige Beschäftigungssicherheit.

Im Vergleich der Ergebnisse der verschiedenen Matchingalgorithmen zeigen sich im Detail – und trotz nahezu äquivalenter Biasreduktion – zudem die erwarteten Unterschiede in der Performanz in vergleichsweise großen Stichproben. Während das 1×10 Nearest-neighbor und das Kernel Matching zu fast deckungsgleichen Ergebnissen gelangen, weichen die durch das 1×1 Nearest-neighbor-Verfahren ermittelten Schätzwerte zum Teil deutlich ab. Dies betrifft einmal die Parameterschätzungen des ATT' selbst, die vor allem in der Einkommensanalyse, aber zum Teil auch in der Analyse der Beschäftigungsstabilität offenbar auch in dieser großen Stichprobe noch relativ sensitiv sind.[17] Zudem liegen die Standardfehler der Koeffizienten im 1×1 Nearest-neighbor-Modell in allen Analysen deutlich über den entsprechenden Werten der alternativen Algorithmen, in welchen die Kontrollstichprobe umfassender genutzt wird.

16 In einer inhaltlichen Analyse würde man u.U. zusätzliche Balancing tests für Subgruppen oder auch für spezifische Kovariaten *X* durchführen, um die Qualität des Matchings dezidiert jeweils innerhalb wichtiger Subgruppen bzw. für zentrale Kovariaten zu überprüfen.

17 Beide Nearest-neighbor-Analysen wurden „mit Zurücklegen", d.h. mit potenziell mehrmaliger Verwendung derselben Kontrollbeobachtung durchgeführt, wodurch sich die Sensitivität der Ergebnisse bekanntermaßen erhöht. Wie das 1×10 Matching zeigt, kann diesem Nachteil der Nearest-neighbor-Verfahren aber durch multiples Matching abgeholfen werden (vgl. auch Smith 1997).

V. Ausblick und Perspektiven

Aus vielerlei Gründen stellen Matchingverfahren unserer Meinung nach eine attraktive und sinnvolle Erweiterung des Methodenspektrums der empirischen Sozialforschung dar. Matchingverfahren sind intuitiv einleuchtend, leicht handhabbar und relativ voraussetzungslos in den ihnen zugrundeliegenden statistischen und mathematischen Annahmen. Zudem besitzen Matchingverfahren eine solide konzeptionelle Fundierung durch das Rubin Causal Model und identifizieren direkt die relevanten Kausalparameter des Modells. Damit sind Matchinganalysen in aller Regel sowohl technisch relativ einfach durchführbar als auch inhaltlich leicht kommunizierbar.

Gleichzeitig ist die Anwendung von Matchingverfahren auf sozialwissenschaftliche Fragestellungen noch ein relativ junges statistisches Feld, das sich in den letzten Jahren sehr schnell entwickelt hat. Es ist deshalb zu erwarten, dass der weiteren Verbreitung der Verfahren auch eine stärkere Standardisierung folgt, in welche zunehmend vielfältigere Erfahrungen in der Anwendung unterschiedlicher Algorithmen in bestimmten Daten- und Fragekontexten einfließen. Zudem entwickelt sich das Feld auch insofern dynamisch weiter, als die Überschneidungen und Unterschiede zu traditionellen ökonometrischen Verfahren zunehmend deutlich werden (Heckman und Navarro-Lozano 2004), Matchingschätzer z.B. als bedingte Differenzen-in-Differenzen-Schätzer in der Analyse von Längsschnittdaten eingesetzt werden (Smith und Todd 2002) oder sensitivitätsanalytische Methoden entwickelt werden, um Verzerrungen durch Effekte ungemessener Variablen abschätzen zu können (Rosenbaum 2002; DiPrete und Gangl 2004). Zudem sind vielfältige neue Anwendungsmöglichkeiten denkbar, etwa auf makrosoziologische oder international vergleichende Anwendungen, während bislang in der statistischen und ökonometrischen Literatur ausschließlich mikroanalytische Fragestellungen betrachtet wurden.

Dennoch, auch Matchingverfahren sind natürlich kein Wundermittel der empirischen Sozialforschung, sondern immer nur mehr oder minder geeignetes statistisches Hilfsmittel bei der – notwendigerweise annahmebehafteten – empirischen Überprüfung sozialwissenschaftlicher Hypothesen. Die Validität matchingbasierter kausaler Schlussfolgerungen hängt dabei von einer relativ weitgehenden Annahme ab: dass es gelungen ist, durch die beobachteten Kovariaten sowohl vergleichbare Kontextbedingungen in der Experimental- und Kontrollgruppe herzustellen als auch die potenziell endogene Zuweisung des Experimentalstatus zu kontrollieren. Die Plausibilität dieser Annahme ist letzten Endes nicht mit statistischen Methoden zu beantworten, sondern verlangt eine sorgfältige theoretische Spezifikation der sich in den Beobachtungsdaten widerspiegelnden sozialen Mechanismen und Prozessen und die inhaltlich begründete Abschätzung, inwieweit dieselben durch die beobachteten Kovariaten hinreichend erfasst werden.

Aus dieser Perspektive ist vielleicht die verfahrenslogisch notwendige Transparenz der Analysen und der ihnen zugrunde liegenden inhaltlichen Annahmen das stärkste Argument für die Verwendung von Matchingverfahren in der Soziologie: um überhaupt empirische Ergebnisse erhalten zu können, müssen die methodologischen Grundlagen – die Frage nach der Identifizierbarkeit eines bestimmten kausalen Effekts mit den zur Verfügung stehenden Daten, der kausale Status der Kovariaten sowie die

konkrete Operationalisierung der Testbedingung und des Zuweisungsmodells – notwendigerweise sehr viel expliziter geklärt werden, als dies bei (mangelhafter Anwendung) manch anderer Methoden der Fall ist. Insofern steht zu hoffen, dass die Beschäftigung mit kausalanalytischen Matchingverfahren zu einem verbesserten Verständnis der theoretischen und empirischen Grundlagen von Kausalschlüssen in den Sozialwissenschaften führt, so dass in empirischen Analysen die den Schlussfolgerungen jeweils zugrundeliegenden Annahmen unmittelbar offen gelegt und damit diskutabel werden.

Literatur

Abbott, *Andrew*, 1998: The Causal Devolution. Sociological Methods & Research 27: 148–181.

Angrist, *Joshua*, *Guido W. Imbens* und *Donald Rubin*, 1996: Identification of Causal Effects Using Instrumental Variables (mit Diskussion). Journal of the American Statistical Association 91: 444–472.

Berk, *Richard A.*, 2004: Regression Analysis. A Constructive Critique. Thousand Oaks: Sage.

Brüderl, *Josef*, 2000: Regressionsverfahren in der Bevölkerungswissenschaft. S. 589–642 in: *Ulrich Müller*, *Bernhard Nauck* und *Andreas Diekmann* (Hg.), Handbuch der Demographie, Band I. Berlin: Springer.

Dehejia, *Rajeev H.*, und *Sadek Wahba*, 1999: Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94: 1053–1062.

Dehejia, *Rajeev H.*, und *Sadek Wahba*, 2002: Propensity Score-Matching Methods for Nonexperimental Causal Studies. Review of Economics and Statistics 84: 151–161.

DiPrete, *Thomas A.*, und *Markus Gangl*, 2004: Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments. Sociological Methodology 34: 271–310.

Efron, *Bradley*, und *Robert J. Tibshirani*, 1993: An Introduction to the Bootstrap. New York: Chapman & Hall.

Engel, *Uwe*, 2004: Sozialer Kontext in der Mehrebenenanalyse. In: *Andreas Diekmann* (Hg.), Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Opladen: Westdeutscher Verlag.

Falk, *Armin*, und *Ernst Fehr*, 2003: Why Labour Market Experiments? Labour Economics 10: 399–406.

Fehr, *Ernst*, und *Simon Gächter*, 2000: Cooperation and Punishment in Public Goods Experiments. American Economic Review 90: 980–994.

Gangl, *Markus*, und *Thomas A. DiPrete*, 2004: Unemployment Insurance and Post-unemployment Wages: Evidence from Matching Estimators. Berlin/Durham: mimeo.

Garfinkel, *Irwin*, *Charles F. Manski* und *Charles Michalopolous*, 1992: Micro Experiments and Macro Effects. S. 253–273 in: *Charles F. Manski* und *Irwin Garfinkel* (Hg.), Evaluating Welfare and Training Programs. Cambridge, MA: Harvard University Press.

Goldthorpe, *John H.*, 2001: Causation, Statistics, and Sociology. European Sociological Review 17: 1–20.

Heckman, *James J.*, 1992: Randomization and Social Policy Evaluation. S. 201–230 in: *Charles F. Manski* und *Irwin Garfinkel* (Hg.), Evaluating Welfare and Training Programs. Cambridge, MA: Harvard University Press.

Heckman, *James J.*, 1997: Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. Journal of Human Resources 32: 441–462.

Heckman, *James J.*, 2001: Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programmes. Economic Journal 111: F654–F699.

Heckman, *James J.*, und *Salvador Navarro-Lozano*, 2004: Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models. Review of Economics and Statistics 86: 30–57.

Heckman, *James J.*, und *Richard Robb*, 1985: Alternative Methods for Evaluating the Impact of Interventions. S. 156–245 in: *James J. Heckman* and *Burton Singer* (Hg.), Longitudinal Analysis of Labor Market Data. New York: Wiley.

Heckman, *James J.*, *Hidehiko Ichimura* und *Petra E. Todd*, 1997: Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. Review of Economic Studies 64: 605–654.

Heckman, *James J.*, *Hidehiko Ichimura* und *Petra E. Todd*, 1998: Matching as an Econometric Evaluation Estimator. Review of Economic Studies 65: 261–294.

Heckman, *James J.*, *Robert J. LaLonde* und *Jeffrey A. Smith*, 1999: The Economics and Econometrics of Active Labor Market Programs. S. 1865–2097 in: *Orley Ashenfelter* und *David Card* (Hg.), Handbook of Labor Economics, Band III. Amsterdam: North Holland.

Heckman, *James J.*, *Lance Lochner* und *Christopher Taber*, 1998: General-Equilibrium Treatment Effects: A Study of Tuition Policy. American Economic Review 88: 381–386.

Heckman, *James J.*, *Justin L. Tobias* und *Edward Vytlacil*, 2003: Simple Estimators for Treatment Parameters in a Latent Variable Framework. Review of Economics and Statistics 85: 748–755.

Holland, *Paul W.*, 1986: Statistics and Causal Inference. Journal of the American Statistical Association 81: 945–960.

Imbens, *Guido W.*, und *Joshua D. Angrist*, 1994: Identification and Estimation of Local Average Treatment Effects. Econometrica 62: 467–75.

Imbens, *Guido W.*, 2004: Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. Review of Economics and Statistics 86: 4–29.

Lechner, *Michael*, 1999: Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. Journal of Business & Economic Statistics 17: 74–90.

Lechner, *Michael*, 2002: Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies. Review of Economics and Statistics 84: 205–220.

Manski, *Charles F.*, 1989: Anatomy of the Selection Problem. Journal of Human Resources 24: 343–360.

Manski, *Charles F.*, 1995: Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press.

Marini, *Margaret M.*, und *Burton Singer*, 1988: Causality in the Social Sciences. Sociological Methodology 18: 347–409.

Mooney, *Christopher Z.*, und *Robert D. Duval*, 1993: Bootstrapping. A Nonparametric Approach to Statistical Inference. Newbury Park: Sage.

Pearl, *Judea*, 1998: Graphs, Causality, and Structural Equation Models. Sociological Methods & Research 27: 226–284.

Pearl, *Judea*, 2000: Causality. Models, Reasoning, and Inference. Cambridge: Cambridge University Press.

Pötter, *Ulrich*, 2004: Das Problem der Selektionsverzerrung bei nicht-experimentellen Daten. In: *Andreas Diekmann* (Hg.), Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Opladen: Westdeutscher Verlag.

Pratt, *John W.*, und *Robert Schlaifer*, 1988: On the Interpretation and Observation of Laws. Journal of Econometrics 39: 23–52.

Rosenbaum, *Paul R.*, 2002: Observational Studies. 2. Auflage. New York: Springer.

Rosenbaum, *Paul R.*, und *Donald B. Rubin*, 1983: The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70: 41–55.

Rosenbaum, *Paul R.*, und *Donald B. Rubin*, 1985: Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistician 39: 33–38.

Rubin, *Donald B.*, 1974: Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology 66: 688–701.

Rubin, *Donald B.*, 1978: Bayesian Inference for Causal Effects. Annals of Statistics 6: 34–58.
Rubin, *Donald B.*, 1991: Practical Implications of the Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. Biometrics 47: 1213–1234.
Smith, *Herbert L.*, 1997: Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology 27: 325–353.
Smith, *Jeffrey A.*, und *Petra E. Todd*, 2001: Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. American Economic Review 91: 112–118.
Smith, *Jeffrey A.*, und *Petra E. Todd*, 2002: Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? University of Western Ontario und University of Pennsylvania: mimeo.
Sobel, *Michael E.*, 1995: Causal Inference in the Social and Behavioral Sciences. S. 1–38 in: *Gerhard Arminger*, *Clifford C. Clogg* und *Michael E. Sobel* (Hg.), Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum.
Sobel, *Michael E.*, 1996: An Introduction to Causal Inference. Sociological Methods & Research 24: 353–379.
Sobel, *Michael E.*, 1998: Causal Inference in Statistical Models of the Process of Socioeconomic Achievement. Sociological Methods & Research 27: 318–348.
SOEP Group, 2001: The German Socio-Economic Panel (GSOEP) after more than 15 Years – Overview. Vierteljahreshefte zur Wirtschaftsforschung 70: 7–14.
Winship, *Christopher*, und *Michael E. Sobel*, 2004: Causal Inference in Sociological Studies. S. 481–503 in: *Melissa Hardy* und *Alan Bryman* (Hg.), Handbook of Data Analysis. Thousand Oaks: Sage.
Winship, *Christopher*, und *Stephen L. Morgan*, 1999: The Estimation of Causal Effects from Observational Data. Annual Review of Sociology 25: 659–707.
Wooldridge, *Jeffrey M.*, 2002: Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

DIAGNOSTIK VON REGRESSIONSSCHÄTZUNGEN BEI KLEINEN STICHPROBEN

Ben Jann

> We are usually happier about asserting a regression relation
> if the relation is still apparent after a few observations
> (any ones) have been deleted – that is, we are happier
> if the regression relation seems to permeate all the observations
> and does not derive largely from one or two
> (Anscombe 1973: 18).

> An apparently wild (or otherwise anomalous) observation is a
> signal that says: „Here is something from which we may learn
> a lesson, perhaps of a kind not anticipated beforehand,
> and perhaps more important than the main object of the study"
> (Kruskal 1960: 1).

Zusammenfassung: Wie alle anderen statistischen Verfahren konzentriert sich auch die Methode der Regression nur auf die Analyse ausgewählter Aspekte vorliegenden Datenmaterials. Entsprechend sind zu gegebenen Regressionsergebnissen ganz unterschiedliche Datenkonstellationen denkbar, wovon aber für die Interpretation der Ergebnisse nicht alle unproblematisch sind. So besteht besonders bei kleinen Stichproben die Gefahr, dass die Regressionsschätzung entscheidend von einzelnen Extremwerten abhängt, was die Verlässlichkeit der daraus abgeleiteten Schlussfolgerungen beeinträchtigt. In diesem Beitrag werden deshalb anhand von Beispielen einige einfache grafische und formale Instrumente zur Diagnose einflussreicher Datenpunkte vorgestellt, die im Prozess der Datenanalyse standardmäßig angewendet werden sollten. Weiterhin werden nach Identifikation „atypischer" Datenpunkte zu verfolgende Analysestrategien diskutiert.

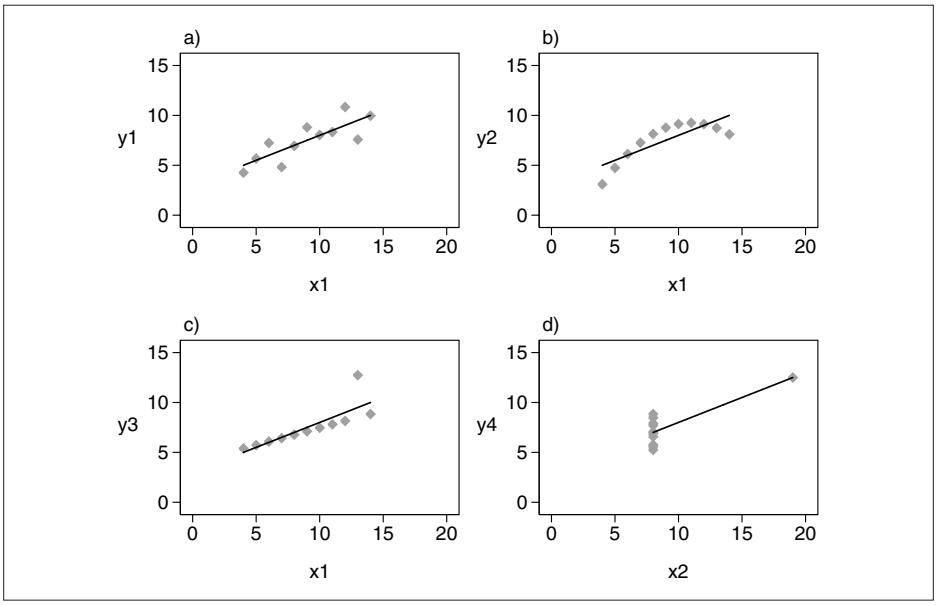
I. Viele Wege führen nach Rom

Die Methode der (linearen) Regression ist zweifelsohne eines der am häufigsten verwendeten Verfahren zur statistischen Behandlung von sozialwissenschaftlichen Fragestellungen. Trotz der sehr günstigen allgemeinen Eigenschaften und der unumstrittenen Nützlichkeit des Verfahrens ist es aber keineswegs so, dass eine Regressionsschätzung vorliegendes Datenmaterial umfassend beschreiben würde. Viele Eigenschaften der Daten bleiben verborgen und können an den üblicherweise im Rahmen einer Regressionsanalyse berechneten Zahlen nicht abgelesen werden. Entsprechend sind zu gegebenen Regressionsergebnissen ganz unterschiedliche Datenkonstellationen denkbar.

Ein eindrückliches Beispiel hierfür liefert Anscombe (1973): Die Daten, die den vier Diagrammen in *Abbildung 1* zu Grunde liegen, führen jeweils zu den gleichen Schätzergebnissen. Die Regressionsgerade lautet in allen Diagrammen $Y = 3.0 + 0.5X$

und eine ganze Reihe weiterer Kennwerte ist jeweils praktisch identisch: Die Mittelwerte und Standardabweichungen der Variablen sind gleich ($\overline{X}$ = 9.0, s_X = 3.32, $\overline{Y}$ = 7.5, s_Y = 2.03), und in allen vier Fällen werden 67 Prozent der Varianz der abhängigen Variablen erklärt. Zudem ergeben sich für die Regressionsparameter immer die gleichen Standardfehler (ca. 1.125 und 0.118), d.h. auch die inferenzstatistischen Schlüsse (Hypothesentests über den Einfluss von *X* auf *Y*), die man aus den Ergebnissen ziehen würde, sind identisch.

Abbildung 1: Anscombes Quartett



Weder an den normalerweise in Publikationen angegebenen Kennzahlen zu einer Regressionsschätzung noch anhand der Standardausgabe eines Statistikprogramms sind die vier Fälle also unterscheidbar. Dennoch würde man wohl nur von den Daten in Diagramm a) vorbehaltlos behaupten, dass sie durch die Regressionsgleichung adäquat beschrieben werden. Die Daten in Diagramm b) beruhen offensichtlich auf einem nichtlinearen Zusammenhang, in Diagramm c) gibt es einen einzelnen Datenpunkt der von einem sonst sehr deutlichen Zusammenhang abweicht und in Diagramm d) wird eine Regressionsschätzung überhaupt erst durch den Datenpunkt am rechten Rand der Abbildung ermöglicht. Ohne diesen Punkt wäre die Varianz der *X*-Variablen gleich null und somit die Steigung der Regressionsgeraden nicht bestimmbar.

Das Beispiel von Anscombe verdeutlicht die Notwendigkeit, die einer Regressionsschätzung zu Grunde liegenden Daten etwas genauer unter die Lupe zu nehmen und somit „Regressionsdiagnostik" zu betreiben. Einerseits werden bei der Schätzung von Regressionsmodellen verschiedene Annahmen getroffen, die mehr oder weniger stark verletzt sein können (in *Abbildung 1b*, zum Beispiel, wird die Annahme eines linearen Zusammenhangs zwischen *Y* und *X* verletzt), andererseits kann sich das Problem stellen, dass einzelne Messwerte in auffälliger Weise vom Rest der Daten abweichen (wie

in *Abbildung 1c*) und unter Umständen einen sehr starken Einfluss auf die Ergebnisse der Regressionsschätzung nehmen (*Abbildung 1d*), was die Gültigkeit der Ergebnisse zumindest in Frage stellt.

Besonders in kleinen bis moderaten Stichproben können die Resultate einer Regression ganz entscheidend von einzelnen Extremwerten abhängen, und die Daten sollten standardmäßig bezüglich des Vorliegens einer solchen Situation evaluiert werden. Ich werde deshalb in diesem Beitrag einige einfache grafische und formale Instrumente zur Diagnose von atypischen Datenkonstellationen vorstellen, wobei das Schwergewicht auf der Identifikation einflussreicher Datenpunkte liegen wird.[1] Nach einer kurzen Übersicht zur Notation werden in Abschnitt II einige Plots zur grafischen Veranschaulichung der Datenstruktur vorgestellt. Abschnitt III befasst sich mit den wichtigsten Maßzahlen zur Diagnose von Ausreißern und einflussreichen Fällen, und in Abschnitt IV wird eine exemplarische Anwendung der vorgestellten Instrumente durchgespielt. Abschnitt V fasst kurz zusammen und diskutiert sinnvolle Analysestrategien.

Grundlegendes Modell und Notation: Im Folgenden wird von einem multiplen linearen Regressionsmodell

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon \text{ mit } E(\varepsilon) = 0,\ E(\varepsilon\varepsilon') = \sigma''\mathrm{I} \tag{1}$$

ausgegangen, wobei y ein $n \times 1$ Vektor der Werte der abhängigen Variablen, X eine $n \times k$ Matrix der Werte der Regressoren (inklusive Konstante), β ein $k \times 1$ Vektor der unbekannten Koeffizienten und ε ein $n \times 1$ Vektor der unbekannten Fehler ist. Es gelten die üblichen Annahmen (vgl. z.B. Cook und Weisberg 1999; Fox 1997; Greene 2003; Wooldridge 2003), insbesondere dass der Erwartungswert der Fehler gleich null ist und die Fehler unabhängig sind und konstante Varianz σ^2 aufweisen. Der Parametervektor β wird auf Grundlage der Kleinste-Quadrate-Methode (OLS) geschätzt als

$$\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \tag{2}$$

und die Residuen (die geschätzten Fehler) ergeben sich als

$$\hat{\varepsilon} = \mathrm{y} - \hat{\mathrm{y}} = \mathrm{y} - \mathrm{X}\hat{\beta} \tag{3}$$

Die Kovarianz-Matrix der OLS-Koeffizienten schließlich wird berechnet als

$$Cov(\hat{\beta}) = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1} \tag{4}$$

wobei

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n - k} \tag{5}$$

ein Schätzer für die Fehlervarianz ist.

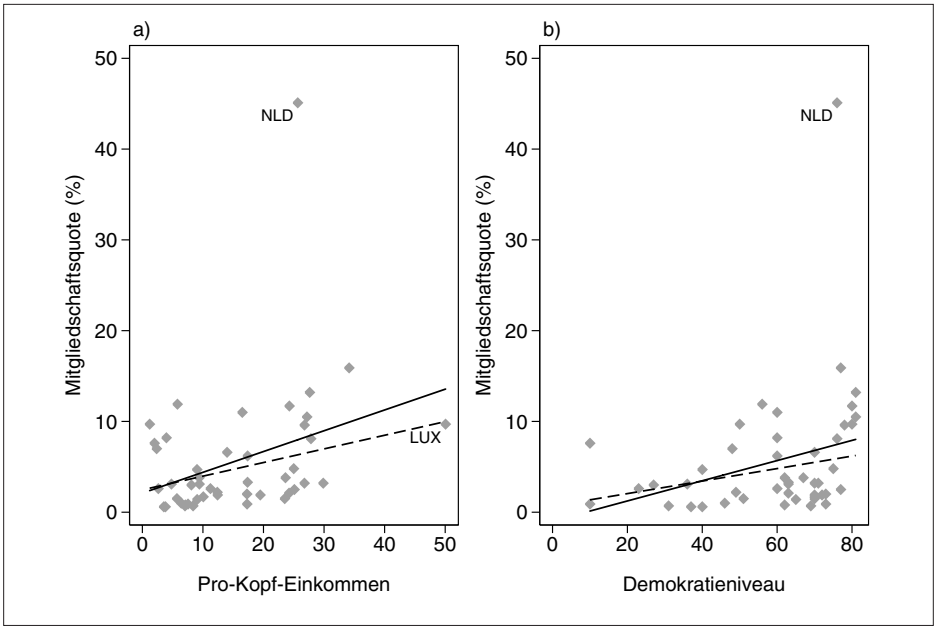
1 Die wichtigsten Grundlagenwerke zu diesem Thema sind Belsley et al. (1980), Cook und Weisberg (1982a) und Chatterjee und Hadi (1988); zugängliche Einführungen liefern Fox (1991) und Bollen und Jackman (1990). In deutscher Sprache finden sich entsprechende Abschnitte z.B. in Brüderl (2000) oder Kohler und Kreuter (2001).

II. Grafische Beurteilung der Daten

Einen guten ersten Überblick über die Beschaffenheit des Datenmaterials kann man sich meistens mit Hilfe von Grafiken verschaffen. Grafiken liefern zwar keine exakten Kennwerte, können aber sehr viel Information verdichten und auf einen Blick erfassbar machen. Im Falle einer Einfachregression kann zum Beispiel bereits ein einfaches Streudiagramm zwischen *X* und *Y* viele Eigenschaften der Daten wie etwa das Vorhandensein von Ausreißern oder auch Nichtlinearität oder Heteroskedastizität offen legen.

Die Streudiagramme in *Abbildung 2* zeigen ein extremes Beispiel für einen Ausreißer. Die Daten stammen aus einer Studie von Dalton (2004), in der es darum geht, die länderspezifischen Unterschiede in der Mitgliedschaftsquote in Umweltorganisationen zu erklären. Bei den Mitgliedschaftsquoten handelt es sich um aggregierte Werte aus dem World Values Survey (1999–2002, je nach Land). Dalton untersuchte unter anderem, inwieweit der Zulauf von Umweltorganisationen vom Wohlstand eines Landes (gemessen am kaufkraftbereinigten Pro-Kopf-Einkommen) und vom Grad der Demokratisierung (gemessen mit einem Indikator für die Pressefreiheit) abhängt. Klar erkennbar ist der Ausreißer Holland mit einer phänomenalen Mitgliederrate von rund 45 Prozent, und es stellt sich die Frage, inwieweit Holland überhaupt mit den übrigen Ländern verglichen werden kann. Es handelt sich offensichtlich um einen Spezialfall (oder um einen „Messfehler" z.B. aufgrund eines nicht vergleichbaren Messinstru-

Abbildung 2: Mitgliedschaft in Umweltgruppen in Abhängigkeit von Wohlstand und Demokratieniveau (die gestrichelte Linie entspricht der Regressionsgerade unter Ausschluss von NLD)



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

ments).[2] Auffällig ist zudem in Diagramm a) der Datenpunkt für Luxemburg, bei dem es sich um einen – allerdings weniger dramatischen – Ausreißer bezüglich der *X*-Achse, also des Pro-Kopf-Einkommens, handelt.

Neben bivariaten Streudiagrammen können viele weitere Grafiken zur Beurteilung von Regressionsschätzungen verwendet werden (eine sehr ausführliche Übersicht gibt Schnell 1994). Beispiele sind Streudiagramme der Residuen gegen die Vorhersagewerte zur Ermittlung von Heteroskedastizität und Histogramme oder Normal-Quantil-Plots der Residuen zur Beurteilung der Normalitätsannahme. Zur Diagnose von Ausreißern und einflussreichen Datenpunkten sind vor allem zwei Plots von Bedeutung und sollen deshalb hier genauer besprochen werden: partielle Regressionsplots und partielle Residuenplots.

1. Partielle Regressionsplots

Für Modelle mit mehreren unabhängigen Variablen können Streudiagramme zwischen *Y* und den unabhängigen Variablen ein irreführendes Bild vermitteln, weil sie nur die bivariaten Zusammenhänge widerspiegeln, nicht jedoch die *partiellen* Einflüsse der Regressoren. Als Ergänzung sollte hier die Verwendung von partiellen Regressionsplots (*partial regression plot*; auch *added variable plot* oder *partial regression leverage plot*) in Betracht gezogen werden.

Ein partieller Regressionsplot für die Variable X_j in einem multiplen Regressionsmodell (1) wird erstellt, indem die Residuen einer Regression von *Y* auf $X_{\ell \neq j}$, d.h. auf alle anderen Regressoren *ausser* X_j, gegen die Residuen einer Regression von X_j auf $X_{\ell \neq j}$ abgetragen werden. Es werden also zwei Modelle

$$y = X_{(-j)}\alpha + v \quad \text{und} \quad x_j = X_{(-j)}\gamma + \nu$$

geschätzt und dann deren Residuen $\hat{v}$ und $\hat{\nu}$ in einem Streudiagramm abgebildet ($X_{(-j)}$ bezeichnet die Matrix der Werte der Regressoren ausschließlich X_j). Dies ist sinnvoll, da der Steigungskoeffizient des Modells

$$\hat{v} = b_0 + b_1\hat{\nu} + e \tag{6}$$

gerade dem Koeffizienten für X_j in der multiplen Regression entspricht, also $\hat{b}_1 = \hat{\beta}_j$. Zudem gilt $\hat{e} = \hat{\varepsilon}$, d.h. die Residuen von (6) sind identisch mit den Residuen aus der multiplen Regression (1). Von der Struktur her ist ein partieller Regressionsplot also immer noch ein einfaches Streudiagramm zwischen *Y* und X_j, außer dass die beiden Variablen um den Teil, der durch die anderen Regressoren $X_{\ell \neq j}$ erklärt wird, bereinigt wurden: Anstatt der ursprünglichen *Y*- und *Yj*-Werte werden die korrigierten Werte

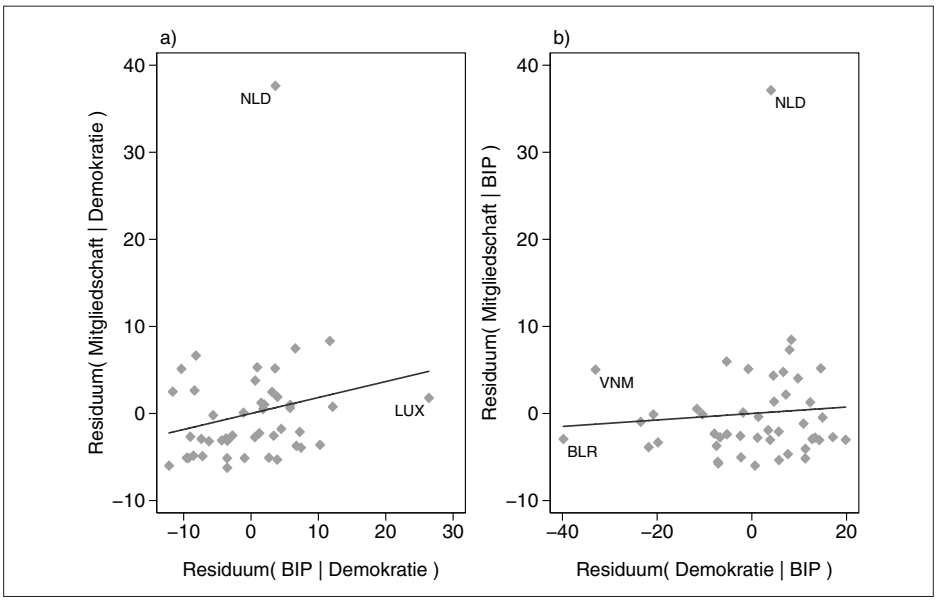
$$\hat{v} = y - X_{(-j)}\hat{\alpha} \quad \text{und} \quad \hat{\nu} = x_j - X_{(-j)}\hat{\gamma}$$

verwendet.

2 Dalton (2004) berichtet denn auch sämtliche Resultate jeweils unter Einschluss wie auch unter Ausschluss von Holland.

Abbildung 3 zeigt die partiellen Regressionsplots für die Daten von Dalton (2004), wenn die Mitgliedschaftsrate in Umweltgruppen simultan auf das Pro-Kopf-Einkommen und das Demokratieniveau regressiert wird. Auffallend ist erstens, dass die Steigung der Geraden im Plot für das Demokratieniveau (Diagramm b) im Vergleich mit dem bivariaten Streudiagramm deutlich abgenommen hat (diese Aussage ist hier allerdings nur möglich, da die Skalierungen der Achsen im Vergleich zum bivariaten Streudiagramm etwa gleich geblieben sind): Unter Kontrolle des Pro-Kopf-Einkommens (BIP) ist der Einfluss auf die Mitgliedschaftsquote praktisch null. Zweitens treten in Diagramm b) zwei Ausreißer bezüglich der *X*-Achse stärker hervor. Es handelt sich um die zwei Länder in der Stichprobe, die die geringste Pressefreiheit aufweisen: Vietnam (VNM) und Weissrussland (BLR).

Abbildung 3: Partielle Regressionsplots für die Mitgliedschaft in Umweltgruppen in Abhängigkeit von Wohlstand und Demokratieniveau



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

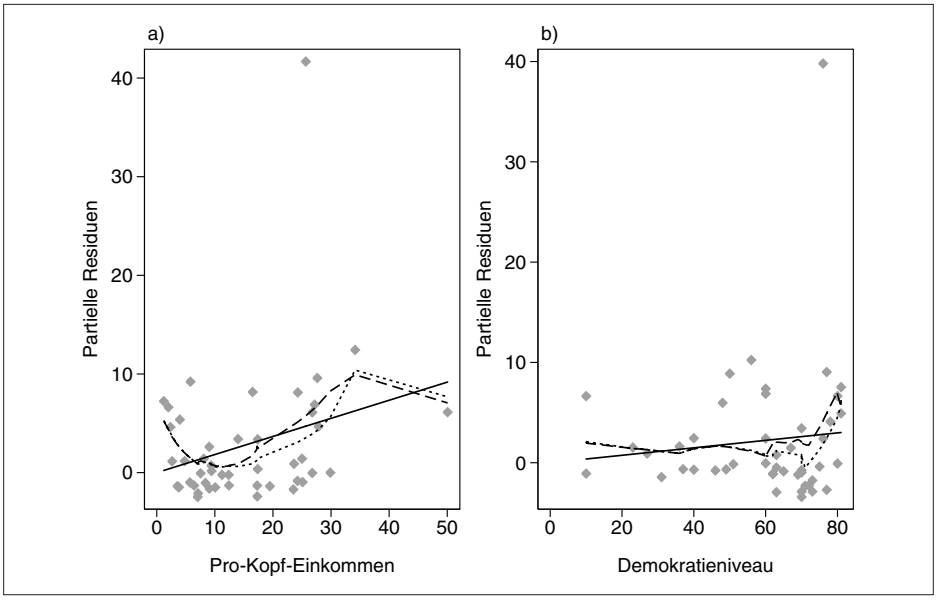
2. Partielle Residuenplots

Als Alternative und Ergänzung zu den partiellen Regressionsplots können auch partielle Residuenplots (*partial residual plot*; auch *component-plus-residual plot*) berechnet werden, wobei sich diese allerdings etwas weniger gut zur Identifikation von einflussreichen Datenpunkten eignen. Ihre Stärke liegt darin, dass sie nicht-lineare Zusammenhänge relativ gut sichtbar machen. Die Bezeichnung *partial residual* bzw. *component-plus-residual* kommt daher, dass den Residuen der durch die Variable X_j erklärte Teil der Differenz $Y_i - \overline{Y}$ wieder beigefügt wird, also

$$\rho_j = \hat{\varepsilon} + \hat{\beta}_j \mathrm{x}_j \tag{7}$$

Diese „partiellen" Residuen werden dann gegen die Werte von X_j abgetragen. Auch hier entspricht die Steigung der Regressionsgeraden in der Grafik gerade wieder dem Koeffizienten $\hat{\beta}_j$ im multiplen Modell. Zur Verdeutlichung der Zusammenhangsform wird häufig eine Lowess-Kurve (*locally weighted scatterplot smoother*, Cleveland 1979; zur Berechnung auch z.B. Schnell 1994: 112f. oder Fox 1991: 85ff.) eingetragen.[3]

Abbildung 4: Partielle Residuenplots für die Mitgliedschaft in Umweltgruppen



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

Abbildung 4 zeigt die partiellen Residuenplots für die Daten von Dalton (2004). Eingezeichnet ist jeweils eine Lowess-Kurve unter Berücksichtigung aller Datenpunkte (gestrichelt) und eine Lowess-Kurve unter Ausschluss von Holland (gepunktet). Bezüglich des Pro-Kopf-Einkommens (Diagramm a) zeichnet sich ein eher U-förmiger Zusam-

3 In einer Variation des partiellen Residuenplots, die nichtlineare Zusammenhänge u.U. noch stärker verdeutlicht, wird für die zu untersuchende Variable zusätzlich ein quadratischer Term eingesetzt (augmented partial residual plot, Mallows 1986). Es wird also ein Modell

$$Y_i = \sum_{\ell \neq j} \beta_\ell X_{i\ell} + \gamma_1 X_{ij} + \gamma_2 X_{ij}^2 + \upsilon_i$$

geschätzt und die (erweiterten) partiellen Residuen berechnen sich als

$$\rho_{ij} = \hat{\upsilon}_i + \hat{\gamma}_1 X_{ij} + \hat{\gamma}_2 X_{ij}^2, \quad i = 1,\ldots,n$$

Man beachte, dass, anders als bei den einfachen partiellen Residuenplots, für jede darzustellende Variable ein eigenes Modell geschätzt werden muss. Ferner spiegelt die Regressionsgerade im erweiterten partiellen Residuenplot nicht mehr den Koeffizienten $\hat{\beta}_j$ wider.

menhang ab, wobei Luxemburg (ganz rechts) nicht ins Bild passt; im Plot für das Demokratieniveau (Diagramm b) verläuft die Lowess-Kurve bis zu Skalenwert 70 praktisch horizontal und steigt danach stark an. Um diesen Ergebnissen gerecht zu werden, könnte das Modell erweitert werden mit einem quadratischen Term für das Pro-Kopf-Einkommen (*BIP*) und einem Strukturbruch im Effekt des Demokratieniveaus (*D*) bei Skalenwert 70, also

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 BIP_i + \beta_2 BIP_i^2 + \beta_3 D_i + \beta_4 D_i^{(70+)} + \varepsilon_i \\ &\text{mit } D_i^{(70+)} = \begin{cases} D_i - 70 & \text{falls } D_i > 70 \\ 0 & \text{sonst} \end{cases} \end{aligned} \tag{8}$$

Tatsächlich verbessert sich die Anpassung des Modells dadurch spürbar. Modell 1a in *Tabelle 1* zeigt das Ausgangsmodell mit den zwei linearen Termen. Der Anteil erklärter Varianz verdoppelt sich beinahe, wenn das Modell gemäß (8) erweitert wird (Modell 2a; die Modellverbesserung ist allerdings bei einem Signifikanzniveau von 5 Prozent knapp nicht signifikant). Eine Betrachtung der Koeffizienten zeigt, dass zwar der Strukturbruch im Effekt des Demokratieniveaus gut abgebildet wird, sich der erwartete U-förmige Einfluss des Pro-Kopf-Einkommens jedoch nicht manifestiert. Dies ist offensichtlich auf den Ausreißer Luxemburg mit dem ungewöhnlich hohen Pro-Kopf-Einkommen zurückzuführen. Die Modelle 1b und 2b in *Tabelle 1* zeigen zum Vergleich die Regressionsergebnisse, wenn Luxemburg und der *Y*-Ausreißer Holland von den Berechnungen ausgeschlossen werden. Während sich das lineare Ausgangsmodell nicht maßgeblich verändert, ist die durch die Erweiterung erzeugte Verbesserung der Modellanpassung nun ziemlich dramatisch und hoch signifikant (Erhöhung von R^2 um ca. 250 Prozent). *Abbildung 5* illustriert die Zusammenhänge in Modell 2b mit Hilfe von modifizierten partiellen Residuenplots.[4] Eingezeichnet sind auch die partiellen Prognosefehler für die beiden ausgeschlossenen Punkte (hervorgehoben).[5]

Aufgrund der guten Modellanpassung und der signifikanten Koeffizienten könnte man nun versucht sein, nur Modell 2b als Forschungsergebnis zu berichten. Das wäre hier aber natürlich höchst unzulässig. Erstens wurden die modellierten Zusammenhangsformen vollkommen ad hoc und unter Abwesenheit jeglicher inhaltlicher Überlegungen ermittelt. Es müssten zumindest eine plausible inhaltliche Erklärung mitgeliefert werden (Warum sollte der Einfluss des Volkseinkommens parabolisch sein? Wieso sollte das Demokratieniveau erst ab Wert 70 einen Einfluss haben?). Zweitens wurden die beiden Datenpunkte, die am schlechtesten zum Modell passen, willkürlich elimi-

4 Das Verfahren zur Erstellung der Plots ist hier ähnlich wie bei den erweiterten partiellen Residuenplots (vgl. Fussnote 3), da jeweils zwei Parameter an einem Zusammenhang beteiligt sind. D.h. die partiellen Residuen werden berechnet als $\rho_{i,BIP} = \hat{\varepsilon}_i + \hat{\beta}_1 BIP_i + \hat{\beta}_2 BIP_i^2$ und $\rho_{i,D} = \hat{\varepsilon}_i + \hat{\beta}_3 D_i + \hat{\beta}_4 D_i^{(70+)}$, und die eingezeichneten Kurven entsprechen $f(BIP) = \hat{\beta}_1 BIP + \hat{\beta}_2 BIP^2$ und $f(D) = \hat{\beta}_3 D + \hat{\beta}_4 D^{(70+)}$.

5 Das Auftreten von Luxemburg in Diagramm 5b als Ausreißer (am unteren Rand) ist auf die schlechte Vorhersage durch den parabolischen Effekt des Pro-Kopf-Einkommens in Abbildunbg 5a zurückzuführen. Dies verdeutlicht, nebenbei bemerkt, die Abhängigkeit der partiellen Residuenplots untereinander. Wird bezüglich eines Regressors X_j eine Veränderung am Modell vorgenommen, ändern sich die partiellen Residuenplots auch für alle anderen Variablen $X_{\ell \neq j}$.

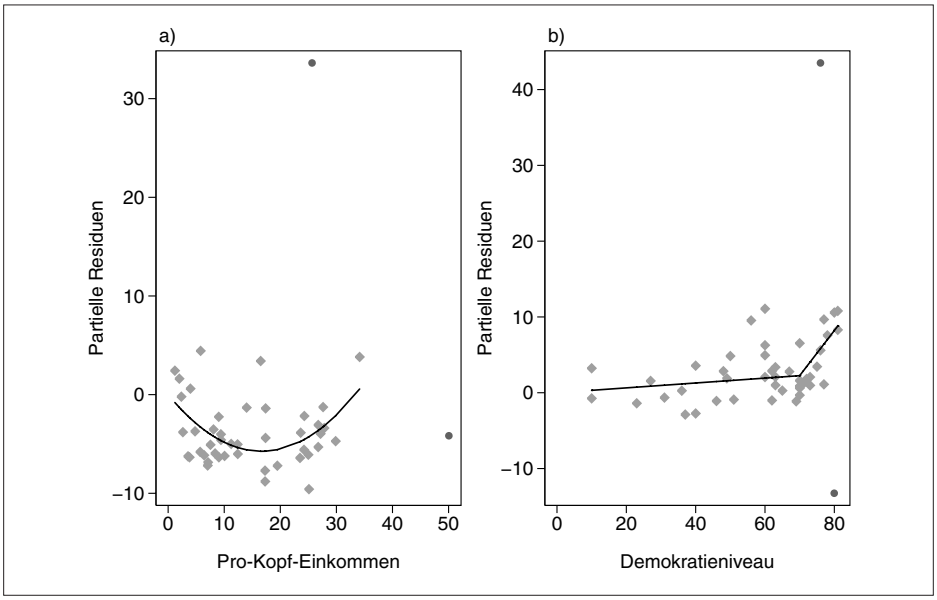
Tabelle 1: Verschiedene Regressionsschätzungen zum länderspezifischen Anteil Mitglieder in Umweltgruppen (Standardfehler in Klammern)

	Alle Fälle		Ohne NLD, LUX	
	Modell 1a	Modell 2a	Modell 1b	Modell 2b
Pro-Kopf-Einkommen	0.183 (0.135)	–0.014 (0.336)	0.128 (0.090)	–0.689** (0.253)
Pro-Kopf-Einkommen²/100		–0.011 (0.702)		2.067** (0.740)
Demokratieniveau	0.037 (0.078)	0.011 (0.082)	0.018 (0.045)	0.032 (0.039)
Demokratieniveau 70+		0.981* (0.422)		0.563* (0.210)
Konstante	0.596 (3.627)	3.277 (3.685)	1.765 (2.031)	5.661** (1.858)
R^2	0.120	0.226	0.129	0.450
F-Test (*p*-Wert)[a]		0.073		0.000
N	46	46	44	44

Quelle: Daten aus Dalton (2004), eigene Berechnungen.

[a] Test für die Verbesserung der Erklärungskraft im Vergleich zu Modell 1.
Signifikanzen: * p < 0.5, ** p < 0.01 (zweiseitig).

Abbildung 5: Partielle Residuenplots zu Modell 2b in Tabelle 1



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

niert. Da erstaunt es kaum, wenn sich die Modellanpassung verbessert. Drittens sind die Fallzahlen eher zu gering, um solche statistischen Kürübungen zu vollführen – das Eis, auf dem man sich bewegt, ist dünn. Relativ geringe Änderungen der Daten (kleinere Messfehler, leicht unterschiedliche Operationalisierung der Variablen etc.) oder zum Beispiel die Aufnahme einer zusätzlichen Variablen können die Ergebnisse unter Umständen radikal verändern. Trotz dieser Einwände ist es aber dennoch eindrucksvoll, wie sehr das Modell durch Ausschluss von nur zwei Datenpunkten „optimiert" werden konnte. Ferner wird verdeutlicht, wie stark ein einzelner Punkt (hier Luxemburg) auf die passende Zusammenhangsform Einfluss nehmen kann.

III. Formale Identifikation einflussreicher Datenpunkte

Ausreißer werden klassischerweise als Datenpunkte definiert, die stark abweichen vom Großteil der anderen Punkte (vgl. z.B. Beckman und Cook 1983; Barnett und Lewis 1995). Nicht jeder Ausreißer ist aber unbedingt ein Problem bei der Schätzung eines Regressionsmodells. Es geht also nicht einfach darum zu ermitteln, ob gewisse Fälle beispielsweise unüblich große oder kleine Y-Werte aufweisen. Vielmehr ist das Ziel, diejenigen Punkte zu identifizieren, die einen besonderen Einfluss auf die Regressionsergebnisse ausüben. Belsley et al. (1980: 11) definieren solche Fälle wie folgt: „An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of the various estimates (coefficients, standard errors, *t*-values, etc.) than is the case for most of the other observations."

Ein offensichtlicher Ansatz zur Bestimmung des Einflusses von Beobachtungen, der diese Definition sehr direkt umsetzt, liegt in der Messung der Veränderung der Schätzresultate, wenn ein Punkt (oder eine Gruppe von Punkten) aus der Datenmatrix gelöscht wird. Der Ausschluss eines Datenpunktes kann sich dabei auf unterschiedliche Kennzahlen der Regressionsschätzung auswirken: „The question ‚Influence on what?' is, therefore, an important one" (Chatterjee und Hadi 1986: 380). Entsprechend findet sich in der Literatur eine ganze Reihe von Maßzahlen, die den Einfluss auf unterschiedliche Zielgrößen operationalisieren oder unterschiedlich standardisiert sind. Die meisten dieser Maßzahlen lassen sich jedoch als Funktion zweier zentraler Elemente verstehen: der Abweichung eines Punktes bezüglich des Vorhersagewertes der abhängigen Variablen sowie der relativen Lage des Punktes bezüglich der Werte der unabhängigen Variablen und somit der „Hebelwirkung" auf die Regressionsfunktion. Für die Bestimmung des Einflusses eines Datenpunktes ist also wohl von Bedeutung, inwieweit es sich um einen Ausreißer bezüglich der Y-Werte handelt (*fit-outlier*), jedoch auch, ob der Datenpunkt extreme X-Werte aufweist (*factor-outlier*).

1. Leverage

Die Leverage h_i („Hebelwirkung") erfasst den potentiellen Einfluss eines Datenpunktes (Y_i, x_i) auf die Vorhersagewerte $\hat{Y}$. Gegeben sei wiederum ein lineares Regressionsmodell

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

Durch Einsetzen des OLS-Parameterschätzer $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ erhält man die Vorhersagewerte für *Y:*

$$\hat{\mathrm{y}} = \mathrm{X}\hat{\beta} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

Man erkennt, dass die Vorhersagewerte von *Y* als Funktion der beobachteten *Y*-Werte verstanden werden können, was durch die folgende Formulierung verdeutlicht wird:

$$\hat{\mathrm{y}} = \mathrm{H}\mathrm{y} \quad \text{mit} \quad \mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \tag{9}$$

Die $n \times n$ Matrix H wird als „hat matrix" bezeichnet weil sie die *Y*-Werte in die $\hat{Y}$-Werte überführt (Hoaglin und Welsch 1978; Chatterjee und Hadi 1986 hingegen sprechen von einer „Prädiktionsmatrix"). Als Projektionsmatrix ist H symmetrisch und idempotent, d.h. $\mathrm{H}' = \mathrm{H}$ und $\mathrm{H}^2 = \mathrm{H}$, und es gilt

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 = \sum_{j=1}^{n} h_{ij}^2 \quad \text{mit } h_{ii} \in [0,1]$$

Ein Diagonalelement $h_i = h_{ii}$ lässt sich somit als ein zusammenfassendes Maß für den potentiellen Einfluss von (Y_i, x_i) auf die Vorhersagewerte interpretieren und wird Leverage (Hebelwirkung) oder „Hut-Wert" (*hat value*) genannt.

Wie man an (9) erkennt, ist h_i lediglich eine Funktion der beobachteten *X*-Werte und hat mit den Werten der abhängigen Variablen nichts zu tun. Tatsächlich kann h_i als relative (quadratische) Distanz zwischen x_i und $\bar{\mathrm{x}}$, dem Vektor der Mittelwerte der Regressoren, verstanden werden, was durch die Berechnungsformel für h_i im Falle eines einfachen Regressionsmodells mit nur einem Regressor verdeutlicht wird:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}$$

Für Modelle mit mehreren Regressoren lässt sich dies allgemein formulieren als

$$h_i = \frac{1}{n} + \frac{MD_i^2}{n-1}$$

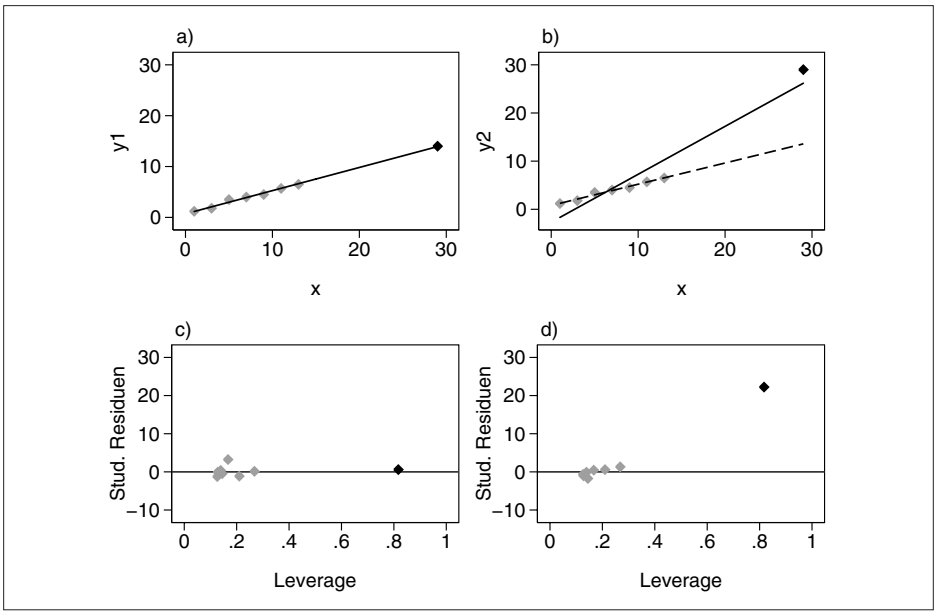
mit MD_i als der Mahalanobis-Distanz der X_i-Werte vom Vektor der Mittelwerte, also

$$MD_i^2 = (\mathrm{x}_i - \bar{\mathrm{x}})' C(X)^{-1} (\mathrm{x}_i - \bar{\mathrm{x}})$$

(vgl. z.B. Weisberg 1985: 112; $C(X)$ ist die Kovarianz-Matrix der X-Variablen). Extremwerte auf den unabhängigen Variablen führen somit zu hoher Leverage.

Zur Identifikation von Datenpunkten mit beachtenswerter Leverage wird i.d.R. der Schwellenwert $h_i > 2k/n$ verwendet (die Schranke entspricht dem zweifachen Durchschnitt der h_i, da $\sum_{i=1}^{n} h_i = k$, vgl. Hoaglin und Welsch 1978; siehe auch Belsley et al. 1980: 17). Hohe Leverage wirkt sich indes aber nur dann auf die Schätzer für die Regressionsparameter aus, wenn zusätzlich Y_i selbst „aus der Reihe tanzt", d.h. bei der Leverage handelt es sich nur um ein Maß für den *potentiellen* Einfluss. Dies wird in *Abbildung 6* veranschaulicht. Für den markierten Punkt in den Diagrammen a) und b) besteht jeweils eine sehr hohe Leverage ($h_i = 0.82$), nur in Diagramm b) hat der Punkt aber auch einen starken Einfluss auf die Lage der Regressionsgeraden. Wird der Punkt bei der Berechnung weggelassen, verläuft die Regressionsgerade viel flacher (gestrichelte Linie).

Abbildung 6: Leverage und Einfluss



2. Residuen

Um Datenpunkte zu identifizieren, die einen starken Einfluss auf die Lage der Regressionsgeraden ausüben, muss also neben der Leverage auch noch berücksichtigt werden, inwieweit es sich bei den korrespondierenden Y-Werten um „Ausreißer" handelt, wofür die Residuen

$$\hat{\varepsilon} = y - \hat{y} = [I - H]y$$

herangezogen werden können. Man muss dabei aber bedenken, dass hohe Leverage zu tendenziell kleineren Residuen führt. Große Hebelwirkung hat ja gerade zur Folge, dass die Regressionsgerade näher an den entsprechenden Datenpunkt „herangezogen" wird. Auch wenn für die Fehler ε_i Homoskedastizität besteht, also $Var(\varepsilon_i) = \sigma^2$, lässt sich zeigen, dass die Residuen $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ unterschiedliche Varianzen aufweisen, nämlich $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_i)$ (vgl. z.B. Cook und Weisberg 1982a: 10ff.). Hohe Leverage führt also aufgrund der Hebelwirkung zu geringerer Residualvarianz, und es liegt nahe, die Residuen mit Hilfe ihrer geschätzten Varianz zu standardisieren, also

$$\hat{\varepsilon}'_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}} \tag{10}$$

Allerdings sind $\hat{\sigma}$ und $\hat{\varepsilon}_i$ nicht unabhängig voneinander, so dass die so genannten studentisierten Residuen i.d.R. vorgezogen werden.[6] Diese sind definiert als

$$\hat{\varepsilon}^*_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}} \tag{11}$$

wobei $\hat{\sigma}_{(-i)}$ den Schätzer für σ unter Ausschluss von Fall i bezeichnet.[7] Für $\hat{\sigma}_{(-i)}$ kann die folgende Berechnungsformel hergeleitet werden

$$\hat{\sigma}^2_{(-i)} = \frac{(n-k)\hat{\sigma}^2}{n-k-1} - \frac{\hat{\varepsilon}_i^2}{(n-k-1)(1-h_i)} \tag{12}$$

Die studentisierten Residuen $\hat{\varepsilon}^*_i$ entsprechen somit lediglich einer monotonen Transformation der standardisierten Residuen $\hat{\varepsilon}'_i$ (Atkinson 1981). Sie haben jedoch den Vorteil, dass sie große Abweichungen stärker hervorheben und unter den Standardannahmen der linearen Regression einer t-Verteilung mit $n - k - 1$ Freiheitsgraden folgen. Als Schwelle zur Identifikation von beachtenswerten Residuen wird i.d.R. $|\hat{\varepsilon}^*_i| > 2$ empfohlen. Unter Normalbedingungen befinden sich ca. 5 Prozent der Datenpunkte in diesem Bereich.

6 Das Attribut „studentisiert" (studentized) geht zurück auf William Sealey Gosset (1867–1937), der unter dem Pseudonym „Student" publizierte.

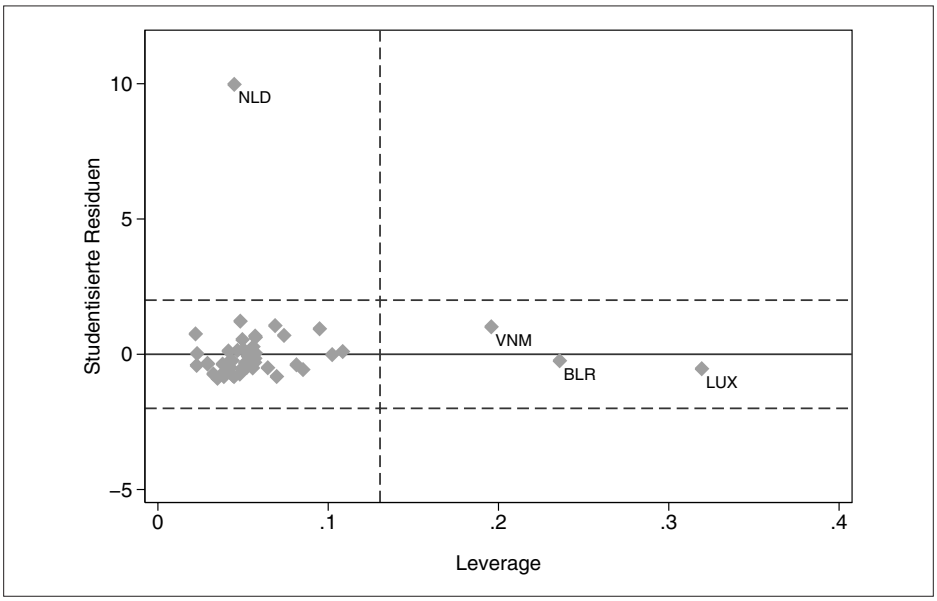
7 Bezüglich der Benennung von $\hat{\varepsilon}'_i$ und $\hat{\varepsilon}^*_i$ herrscht ein ziemliches Durcheinander. Während einige Autoren im ersten Fall von den „standardisierten" Residuen sprechen, wenden Cook und andere auch hier den Begriff „Studentized residuals" oder genauer „internally Studentized residuals" an (z.B. Cook und Weisberg 1982a; Beckman und Cook 1983; Weisberg 1985). Für $\hat{\varepsilon}^*_i$ findet man neben dem Begriff „studentized residual" (Belsley et al. 1980; Velleman und Welsch 1981; Fox 1991) auch „externally Studentized residual" (Cook und Weisberg 1982a), „crossvalidatory residual" (Atkinson 1981) oder „jackknifed residual" (Rousseeuw und Leroy 1987). In SPSS wird $\hat{\varepsilon}'_i$ als „Studentized residual" (SRESID) und $\hat{\varepsilon}^*_i$ als „Studentized deleted residual" (SDRESID) bezeichnet. Der Begriff „Standardized residual" (ZRESID) bezieht sich in SPSS auf den Ausdruck $\hat{\varepsilon}_i/\hat{\sigma}$.

3. Einfluss auf die Koeffizienten und Vorhersagewerte

Erst durch die gemeinsame Betrachtung der Leverage h_i und der studentisierten Residuen $\hat{\varepsilon}_i^*$, wenn beispielsweise die beiden Größen in einem Streudiagramm gegeneinander abgetragen werden, können Datenpunkte identifiziert werden, die einen großen Einfluss auf die Regressionsschätzung ausüben.[8] In *Abbildung 6* wird dies exemplarisch veranschaulicht: Nur in Diagramm b) hat der markierte Punkt einen starken Einfluss auf die Lage der Regressionsgeraden, da er im Plot mit den Residuen und der Leverage auf beiden Achsen hohe Werte aufweist (Diagramm d). Am Vergleich der Diagramme b) und d) erkennt man zudem sehr schön die Beziehung zwischen den unstandardisierten und den studentisierten oder standardisierten Residuen: Der markierte Punkt liegt aufgrund der hohen Leverage kaum weiter weg von der Regressionsgerade als andere Punkte, das studentisierte Residuum, bei dem um den Einfluss der Leverage korrigiert wird, ist jedoch sehr groß.

Abbildung 7 zeigt ferner ein Streudiagramm der Leverage und studentisierten Residuen für die in Abbildung 2 vorgestellten Daten (es handelt sich um die h_i- und $\hat{\varepsilon}_i^*$-Werte, die man bei einer Regression der Mitgliedschaftsquoten in Umweltgruppen auf das Pro-Kopf-Einkommen und das Demokratieniveau erhält). Zusätzlich sind als gestrichelte Linien die oben angegebenen Schwellenwerte eingetragen ($\hat{\varepsilon}^* = \pm 2$ und

Abbildung 7: Studentisierte Residuen und Leverage für das Modell zur Erklärung länderspezifischen Mitgliederrate in Umweltgruppen



Quelle: Daten aus Dalton (2004), eigene Berechnungen.

8 Eine andere Darstellungsform mit ähnlichem Informationsgehalt ist der L-R-Plot (leverage residual plot), bei dem h_i gegen $\hat{\varepsilon}_i^2 / (n-k)\hat{\sigma}^2$ abgetragen wird (Gray 1986, 1989a).

$h = 2k/n = 2 \cdot 3/46 = 0.13$). Man kann erkennen, dass der massive *Y*-Ausreißer Holland (NLD) für die Lage der Regressionsparameter gar nicht unbedingt ein so gravierendes Problem darstellt, da er eine relativ geringe Leverage hat. Weiterhin ist das im bivariaten Streudiagramm für den Regressor „Pro-Kopf-Einkommen" identifizierte Luxemburg (LUX) zwar ein Punkt mit hoher Leverage, der tatsächliche Einfluss ist aber ebenfalls gering aufgrund des kleinen Residuums. Weitere, allerdings ebenfalls eher problemlose Extremwerte bezüglich der Leverage sind Vietnam (VNM) und Weißrussland (BLR).

a) Gesamteinfluss. Man kann nun versuchen, das Zusammenspiel von Leverage und studentisierten Residuen mit Hilfe einer einzigen Maßzahl für den Gesamteinfluss eines Datenpunktes auf die Regressionsschätzung auszudrücken. Die am häufigsten verwendeten derartigen Statistiken sind Cooks *D* (Cook 1977) und die Maßzahl *DFFITS* (Belsley et al. 1980). Cooks *D* wurde entwickelt als Maß für die Veränderung im geschätzten Parametervektor, wenn eine Beobachtung *i* aus dem Datensatz eliminiert wird, und ist definiert als

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})'\mathrm{X}'\mathrm{X}(\hat{\beta}_{(-\mathrm{i})} - \hat{\beta})}{k\sigma^2}, \quad i = 1,...,n$$

wobei $\hat{\beta}_{(-i)}$ den Parameterschätzer unter Ausschluss des Datenpunktes *i* symbolisiert. Alternativ lässt sich D_i auch als Veränderung in den Vorhersagewerten interpretieren, wenn Punkt *i* bei der Schätzung weggelassen wird, also

$$D_i = \frac{(\hat{\mathrm{y}}_{(-i)} - \hat{\mathrm{y}})'(\hat{\mathrm{y}}_{(-i)} - \hat{\mathrm{y}})}{k\sigma^2}$$

mit $\hat{\mathrm{y}}_{(-i)}$ als Vektor der Vorhersagewerte bei Verwendung der Schätzers $\hat{\beta}_{(-i)}$, also $\hat{\mathrm{y}}_{(-i)} = \mathrm{X}\hat{\beta}_{(-i)}$. Die Maßzahl kann aber auch einfach als Funktion der Leverage h_i und des standardisierten Residuums $\hat{\varepsilon}'_i$ formuliert werden:

$$D_i = \frac{\hat{\varepsilon}'^2_i}{k} \cdot \frac{h_i}{1-h_i} = \frac{1}{k} \cdot \frac{\hat{\varepsilon}^2_i}{\hat{\sigma}^2(1-h_i)} \cdot \frac{h_i}{1-h_i} \tag{13}$$

Die Maßzahl *DFFITS* hat einen leicht anderen Fokus. Sie misst die (standardisierte) Veränderung des Vorhersagewertes Y_i, wenn Punkt *i* für die Schätzung des Parametervektors weggelassen wird, und ist definiert als

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_i}} = \frac{\mathrm{x}_i(\hat{\beta} - \hat{\beta}_{(-i)})}{\hat{\sigma}_{(-i)}\sqrt{h_i}}, \quad i = 1,...,n$$

Auch diese Statistik lässt sich mit Hilfe der Leverage und der Residuen ausdrücken, nämlich

$$DFFITS_i = \hat{\varepsilon}^*_i\sqrt{\frac{h_i}{1-h_i}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}\sqrt{\frac{h_i}{1-h_i}}. \tag{14}$$

Aus (13) und (14) folgt zudem

$$D_i = \frac{1}{k}\left(DFFITS_i \frac{\hat{\sigma}_{(-i)}}{\hat{\sigma}}\right)^2$$

Die beiden Maße sind also trotz der unterschiedlichen Ausgangslage sehr ähnlich. Es sollte folglich nicht eine bedeutende Rolle spielen, welche der Statistiken man verwendet. Der einzige substantielle Unterschied ist, dass bei *DFFITS* das studentisierte Residuum verwendet wird, bei Cooks *D* jedoch das standardisierte, was *DFFITS* etwas sinnvoller erscheinen lässt (vgl. auch Bollen und Jackman 1990: 266). Als kritische Bereiche werden i.d.R. $|DFFITS| > 2\sqrt{k/n}$ und folglich $D > 4/n$ vorgeschlagen.

Tabelle 2 zeigt die Einflussstatistiken für einige Fälle aus dem Beispiel von Dalton (2004). Von den im Plot mit den studentisierten Residuen und der Leverage identifizierten Punkten (*Abbildung 7*) übersteigt nur Holland die kritischen Werte für Cooks *D* und *DFFITS* aufgrund des enormen Residuums. Die Leverage-Extremwerte Luxemburg, Vietnam und Weißrussland jedoch liegen relativ gut im linearen Trend der übrigen Daten, haben also kleine Residuen und werden durch die beiden Maße nicht als einflussreiche Datenpunkte identifiziert (wobei allerdings Vietnam den kritischen Wert fast erreicht).

Tabelle 2: Ausreißer im Modell zur länderspezifischen Mitgliederrate in Umweltgruppen (Modell 1a in Tabelle 1)

Land	Stud. Residuen	Leverage	Cooks *D*	*DFFITS*	*COVR.*	*DFBETAS*	
						BIP	*D*
ZWE	0.101	0.109	0.000	0.035	1.203	0.004	–0.025
BLR	–0.239	0.236	0.006	–0.133	1.399	–0.065	0.123
LUX	–0.534	0.319	0.045	–0.366	1.545	–0.336	0.159
VNM	1.016	0.196	0.084	0.501	1.241	0.147	–0.423
NLD	9.977	0.045	0.473	2.161	0.029	0.725	0.466
Schwelle	±2.000	0.130	0.087	±0.511	±0.196	±0.295	±0.295

Quelle: Daten aus Dalton (2004), eigene Berechnungen.

Anmerkungen: In der Tabelle sind alle Punkte enthalten, die für mindestens eine der aufgeführten Statistiken im kritischen Bereich liegen. Die nach den im Text angegebenen Regeln berechneten kritischen Werte sind im Fuß der Tabelle aufgeführt. BIP: Pro-Kopf-Einkommen; D: Demokratieniveau.

b) Einfluss auf individuelle Koeffizienten. Die Maße Cooks *D* und *DFFITS* drücken den Einfluss eines Datenpunktes auf den Schätzer des *gesamten* Parametervektors aus. Häufig möchte man jedoch wissen, wie stark sich die individuellen Koeffizienten in Abhängigkeit von einzelnen Datenpunkten verändern. Zur Bestimmungen des Einflusses eines Datenpunktes i auf den Schätzer für einen Koeffizienten β_j ist die Differenz $\hat{\beta}_j - \hat{\beta}_{j(-i)}$ maßgebend. Belsley et al. (1980) schlagen die folgende Statistik vor:

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE_{(-i)}(\hat{\beta}_j)} \qquad (15)$$

wobei mit $SE_{(-i)}(\hat{\beta}_j)$ der Standardfehler von $\hat{\beta}_j$ unter Ausschluss von Punkt i gemeint ist (zu den genauen Berechnungsformeln siehe Belsley et al. 1980 oder Chatterjee und Hadi 1986). Die Ergebnisse, die mit der *DFBETAS*-Statistik gewonnen werden, sind relativ unübersichtlich, da für jeden Datenpunkt k Werte zu berechnen sind. Zur Betrachtung der Werte können deshalb Streudiagramme sehr nützlich sein. Als Kriterium für beachtenswerte Punkte geben Belsley et al. (1980) den Bereich $|DFBETAS| > 2/\sqrt{n}$ an. Bollen und Jackman (1990) schlagen alternativ den Bereich $|DFBETAS| > 1$ vor: So werden Punkte identifiziert, die den Parameterschätzer um mehr als einen Standardfehler verändern.

Wie oben angesprochen, werden Vietnam, Luxemburg und Weißrussland im Beispiel von Dalton trotz ihrer großen Leverage-Werte durch die Statistiken für den Gesamteinfluss nicht als beachtenswerte Punkte identifiziert. Die Betrachtung der *DFBETAS*-Werte gibt hier ein differenzierteres Bild (*Tabelle 2*). Zwar erscheint Weißrussland auch bezüglich der einzelnen Koeffizienten als nicht besonders bedeutend, Luxemburg und Vietnam jedoch nehmen je auf einen Koeffizienten spürbar vermindernden Einfluss. Wird Luxemburg weggelassen, erhöht sich der Effekt des Pro-Kopf-Einkommens; bei Ausschluss von Vietnam steigt der Einfluss des Demokratieniveaus. Den stärksten Einfluss auf beide Koeffizienten hat jedoch, kaum überraschend, der Ausreißer Holland.

4. Einfluss auf die Präzision der Schätzung

Die bisher besprochenen Maße beziehen sich auf den Einfluss auf den Koeffizientenvektor der Regressionsschätzung bzw. die Vorhersagewerte. Einzelne Datenpunkte können aber auch einen bedeutenden Einfluss auf die Größe der Standardfehler der Parameterschätzer, d.h. auf die Präzision der Modellschätzung haben. Eine Maßzahl, die die relative Veränderung des (quadrierten) Volumens des gemeinsamen Konfidenzraumes aller Parameter bei Ausschluss der Beobachtung i approximiert und somit als Maß für den Einfluss auf die Schätzgenauigkeit dienen kann, wurde von Belsley et al. (1980) definiert als

$$COVRATIO_i = \frac{\det\left[\hat{\sigma}^2_{(-i)}(\mathrm{X}'_{(-i)}\mathrm{X}_{(-i)})^{-1}\right]}{\det\left[\hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}\right]}$$

d.h. als der Quotient aus der Determinanten der Kovarianz-Matrix des Modells ohne Punkt i und des vollen Modells. Mit Hilfe von (11) und (12) lässt sich auch hier zeigen, dass die Maßzahl als eine Funktion der Leverage und Residuen dargestellt werden kann, nämlich

$$COVRATIO_i = \left(\frac{\hat{\sigma}^2_{(-i)}}{\hat{\sigma}^2}\right)^k \cdot \frac{1}{1-h_i} = 1 \Big/ \left(\frac{n-k-1+\hat{\varepsilon}^{*2}_i}{n-k}\right)^k (1-h_i) \qquad (16)$$

Wert 1 für $COVRATIO_i$ bedeutet, dass kein Einfluss auf die Standardfehler besteht. Werte größer 1 weisen darauf hin, dass sich der Datenpunkt positiv auf die Schätzgenauigkeit auswirkt (Verkleinerung des Konfidenzraumes im Vergleich mit dem Modell ohne den Datenpunkt). Dies ist zum Beispiel für Punkte mit hoher Leverage h_i und kleinem studentisiertem Residuum $\hat{\varepsilon}_i^*$ der Fall. Werte kleiner 1 bedeuten, dass sich der Konfidenzraum durch Hinzunahme des Datenpunktes vergrößert, d.h. die Schätzung wird ungenauer (z.B. bei kleinem h_i und großem $\hat{\varepsilon}_i^*$). Gemäß Belsley et al. (1980) sind Punkte mit

$$|COVRATIO_i - 1| > 3k/n$$

beachtenswert.

Tabelle 2 zeigt die fünf Fälle im Beispiel von Dalton, die diese Schwelle überschreiten. Es handelt sich einerseits um die vier bereits bezüglich der anderen Maßzahlen besprochenen Punkte: Das große Residuum von Holland führt zu einer massiven Verschlechterung der Präzision der Modellschätzung (durch Ausschluss von Holland fällt der Standardfehler des Modells, also $\hat{\sigma}$, um fast 50 Prozent von 6.9 auf 3.8; entsprechend halbieren sich auch die Standardfehler der einzelnen Koeffizienten beinahe). Für Luxemburg, Weißrussland und Vietnam ist die Situation gerade entgegengesetzt. Es handelt sich um Punkte mit eher kleinen Residuen dafür aber hoher Leverage. Sie wirken sich somit positiv auf die Schätzgenauigkeit aus. Der fünfte identifizierte Punkt ist Zimbabwe, ein Punkt, der praktisch auf der Regressionsfläche liegt. Das geringe Residuum in Kombination mit einer relativ hohen Leverage (Zimbabwe hat nach Luxemburg, Weißrussland und Vietnam die vierthöchste Leverage) führt hier ebenfalls zu dem günstigen Einfluss auf die Präzision der Regressionsschätzung.

5. Übersicht und weitere Ansätze

Tabelle 3 fasst die besprochenen Maßzahlen nochmals zusammen und gibt zudem Hinweise zur Verfügbarkeit der Maßzahlen in den häufig verwendeten Statistikprogrammen SPSS und Stata. In der Literatur werden noch weitere Einflussmaße und Varianten vorgeschlagen (z.B. Andrews und Pregibon 1978; Draper und John 1981; Welsch 1982; vgl. auch etwa die Übersichten in Hocking 1983 sowie Chatterjee und Hadi 1986), eine Analyse der hier besprochenen Größen in Verbindung mit den grafischen Methoden sollte aber normalerweise ausreichen. Wenn es jedoch darum geht, die optimale Zusammenhangsform mit Hilfe von Transformationen der Variablen zu finden, können zusätzliche Statistiken von Nutzen sein, die den Einfluss einzelner Datenpunkte auf die Transformation quantifizieren (vgl. z.B. Atkinson 1982, 1986; Cook und Wang 1983; Carroll und Ruppert 1987). Ferner beschränken sich die hier besprochenen Methoden auf die lineare Regression. Für entsprechende Instrumente im Rahmen der logistischen Regression siehe z.B. die Einführungen in Hosmer und Lemeshow (2000) und Menard (2002) sowie den Grundlagentext von Pregibon (1981) oder für multinomiale Modelle Lesaffre und Albert (1989). Zur Regressionsdiagnostik in der Ereignisanalyse siehe z.B. Hosmer und Lemeshow (1999).

Tabelle 3: Übersicht über die Maßzahlen zur Identifikation einflussreicher Daten

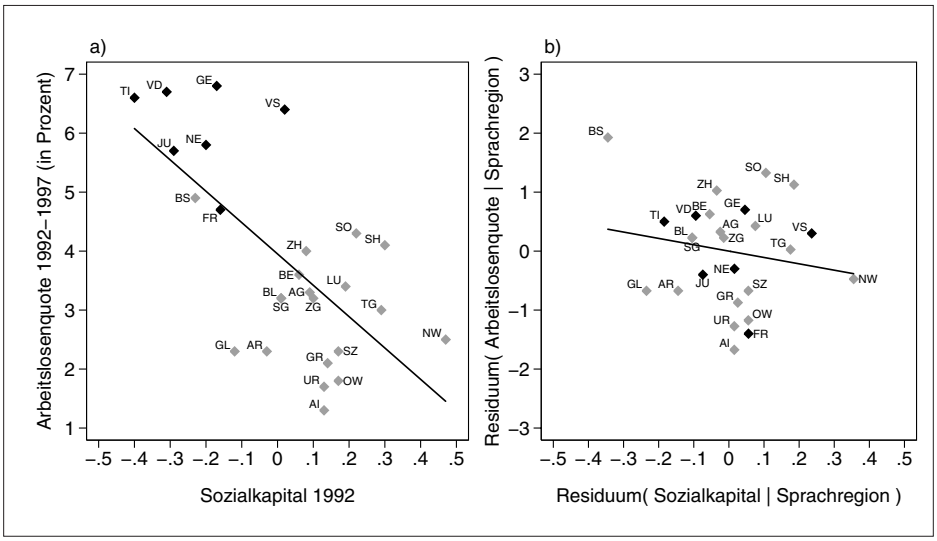
Maß	Formel	kritischer Bereich	Verfügbarkeit[a]
Potentieller Einfluss (Leverage)	$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$	$h_i > 2k/n$	SPSS: LEVER (Leverage values) Stata: leverage, hat (Leverage)
Standardisierte Residuen	$\hat{\varepsilon}_i' = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_i}}$	$\lvert\hat{\varepsilon}_i'\rvert > 2$	SPSS: SRESID (Studentized residuals) Stata: rstandard (Standardized residuals)
Studentisierte Residuen	$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$	$\lvert\hat{\varepsilon}_i^*\rvert > 2$	SPSS: SDRESID (Studentized deleted residuals) Stata: rstudent (Studentized residuals)
Gesamteinfluss auf Koeffizienten	$D_i = \frac{\hat{\varepsilon}_i'^2}{\kappa} \cdot \frac{h_i}{1-h_i}$	$D_i > 4/n$	SPSS: COOK (Cook's Distance) Stata: cooksd (Cook's Distance)
	$DFFITS_i = \hat{\varepsilon}_i^* \sqrt{\frac{h_i}{1-h_i}}$	$\lvert DFFITS_i\rvert > 2\sqrt{k/n}$	SPSS: SDFIT (Standardized DfFit) Stata: dfits (DFITS)
Einfl. auf einzelne Koeffizienten	$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE_{(-i)}(\hat{\beta}_j)}$	$\lvert DFBETAS_{ij}\rvert > 2/\sqrt{n}$	SPSS: SDBETA (Standardized DfBeta(s)) Stata: dfbeta() (DFBETA)
Einfluss auf die Präzision	$COVRATIO_i = \left(\frac{\hat{\sigma}^2_{(-i)}}{\hat{\sigma}^2}\right)^k \cdot \frac{1}{1-h_i}$	$\lvert COVRATIO_i - 1\rvert > 3k/n$	SPSS: COVRATIO (Covariance Ratio) Stata: covratio (COVRATIO)

[a] Angegeben ist die jeweilige Bezeichnung in der Befehlssyntax sowie in Klammern die Beschriftung in den Dialogfeldern. Die Maßzahlen können in SPSS mit Hilfe der SAVE-Option im REGRESSION-Befehl gespeichert werden. In Stata werden die Maßzahlen mit dem predict-Befehl berechnet.

Ein weiterer Aspekt, der bisher ausgeklammert wurde, betrifft das Problem, dass sich einflussreiche Datenpunkte gegenseitig „verdecken" können. Die besprochenen Maße beziehen sich alle auf den Effekt auf die Regressionsergebnisse, wenn ein *einzelner* Datenpunkt eliminiert wird. Es kann aber auch sein, dass multiple Ausreißer in Gruppen auftreten und durch die klassischen Diagnosemaße aufgrund des so genannten Maskierungseffektes nicht erkannt werden (die Eliminierung eines einzelnen Punktes der Gruppe ändert nicht viel an den Regressionsergebnissen, weil die restlichen Punkte ihren Einfluss weiterhin ausüben). Die besprochenen Maße wie z.B. Cooks *D* lassen sich zwar relativ problemlos generalisieren für die Betrachtung des simultanen Einflusses mehrerer Beobachtungen (vgl. z.B. Cook und Weisberg 1982b: 347ff.). Dies ist allerdings meistens relativ nutzlos für die Identifikation von einflussreichen Gruppen, weil es sehr viele Möglichkeiten zur Gruppierung der Daten gibt. Verschiedene Ansätze, mit dem Problem umzugehen, wurden vorgeschlagen (vgl. etwa Gray und Ling 1984; Rousseeuw und van Zomeren 1990; Ritschard und Antille 1992; Crawford et al. 1995; Peña und Yohai 1995; Peña und Prieto 2001), können aber hier nicht erörtert werden. Glücklicherweise erkennt man multiple Ausreißer meistens auch relativ gut anhand von Scatterplots und partiellen Regressionsplots.

Ein Beispiel für multiple Ausreißer findet sich in einer Arbeit von Freitag (2000). Freitag untersuchte den Zusammenhang zwischen Sozialkapital und Arbeitslosenquoten in Schweizer Kantonen. Die empirischen Ergebnisse deuten auf einen – gemäß Theorie erwarteten – starken negativen Zusammenhang: Je höher das Sozialkapital in einem Kanton, desto geringer die Arbeitslosenquote (*Abbildung 8a*).

Abbildung 8: Streudiagramm zwischen Arbeitslosenquote und Sozialkapital (a) und partieller Regressionsplot unter Kontrolle der Sprachregion (b)



Quelle: Daten aus Freitag (2000), eigene Berechnungen.

Bei genauerer Betrachtung des Streudiagramms erkennt man jedoch, dass der Zusammenhang hauptsächlich durch den Unterschied zwischen zwei Clustern von Datenpunkten zustande kommt. Die beiden Cluster repräsentieren nun gerade die Kantone der (überwiegend) „lateinischen" Schweiz (FR, GE, JU, NE, TI, VD, VS) und der Deutschschweiz (einzige Ausnahme ist Basel-Stadt). Die Annahme liegt also nahe, dass ein klassisches Drittvariablenproblem vorliegt und nicht das Sozialkapital eine Auswirkung auf die Arbeitslosenquote hat, sondern beide Variablen durch gemeinsame, mit der Sprachregion zusammenhängende Faktoren bedingt sind. Wird das Modell durch Aufnahme eines Indikators für die Sprachregion erweitert, verschwindet denn auch der Zusammenhang zwischen Sozialkapital und Arbeitslosenquote fast vollständig, wie man am partiellen Regressionsplot unter Kontrolle der Sprachregion (*Abbildung 8b*) erkennen kann (bzw. in beiden Sprachregionen ist der Effekt des Sozialkapitals praktisch null). Zur Verteidigung von Freitag (2000) muss allerdings gesagt werden, dass mit dem Sprach-Dummy eigentlich nicht viel erklärt wird. Es ist lediglich ein Hinweis darauf, dass weitere, mit dem sprachkulturellen Hintergrund zusammenhängende Faktoren von Bedeutung sind.

IV. Reputationseffekte in Internet-Auktionen: Ein Anwendungsbeispiel

Im Folgenden sollen die vorgestellten Methoden der Regressionsdiagnostik an den Daten einer Studie von Diekmann und Wyder (2002) nochmals veranschaulicht werden. Diekmann und Wyder untersuchten die Auswirkungen von Reputation in Transaktionen zwischen anonymen Akteuren anhand von Internet-Auktionen. Die Daten stammen aus der Beobachtung der Auktionen von neuen Nokia-Mobiltelefonen (Nokia 8310) bei Ricardo.ch im Zeitraum Oktober 2001 bis Januar 2002. Unter anderem interessierten sich Diekmann und Wyder für den Effekt der Reputation auf den realisierten Verkaufspreis und schätzten entsprechende Regressionsmodelle. Die Hypothese war, dass eine gute Reputation den Verkaufspreis im Sinne einer Risikoprämie erhöht. Die Hypothese konnte in der Studie weitgehend bestätigt werden.

Für die folgende Reanalyse des Reputationseffektes wurden die Daten zuerst einer Bereinigung anhand des Rohdatenmaterials (gespeicherte Screenshots der Auktionen) unterzogen. Neben der Korrektur einiger kleinerer Übertragungsfehler musste eine Reihe von Fällen ausgeschlossen werden, weil es sich a) um doppelt erfasste Auktionen handelte, b) die Rohdaten unvollständig oder c) die Auktionen nicht vergleichbar waren („neuwertige" anstatt „neue" Produkte, Mehrfachangebote etc.). Zur Analyse der Verkaufpreise verbleiben nach der Bereinigung 84 Fälle.

Diekmann und Wyder verwendeten als Regressoren zur Erklärung des Verkaufspreises (in Franken) die Reputation (operationalisiert als die Anzahl Bewertungen oder die logarithmierte Anzahl Bewertungen; es wurden zudem nur Fälle mit durchschnittlich positiver Reputation berücksichtigt), den Mindestpreis (das vom Anbieter festgelegte Startgebot), die veranschlagten Versandkosten, die Dauer der Auktion, die vom Verkäufer festgelegte Mindesterhöhung (Mindestschrittweite der Gebote) und die Anzahl der Gebote. Für die Reputation und die Mindesterhöhung wurde je ein signifikanter

positiver Effekt festgestellt, für die Versandkosten ein stark negativer Effekt. Der Mindestpreis und die Anzahl Gebote hatten keinen Effekt.

Aus Gründen der Übersichtlichkeit werde ich mich hier mit einem vereinfachten Modell ohne die Variablen „Dauer der Auktion" und „Anzahl der Gebote" befassen und als abhängige Variable den Bruttopreis (Verkaufspreis plus Versandkosten) verwenden. Als zusätzliche Variable wird jedoch das Startdatum der Auktion (Kalenderzeit) aufgenommen, da es sich um einen sehr wichtigen Prädiktor für den Verkaufspreis handelt. Die Resultate der Modellschätzung sind in *Tabelle 4* (Spalte 1) abgebildet (die Bootstrap-Standardfehler in eckigen Klammern sollen vorerst außer Acht gelassen werden). Wie bei Diekmann und Wyder (2002) zeigt sich auch in dem modifizierten Modell ein signifikanter Reputationseffekt (mit einem *t*-Wert von 2.83) und ein überproportionaler Effekt der Mindesterhöhung (mehr als 3 Franken Zusatzerlös bei einer Erhöhung der Schrittweite der Gebote um einen Franken; $t = 5.90$).[9] Im Vergleich mit den Resultaten von Diekmann und Wyder (2002: Tabelle 3) fällt die deutlich verbesserte Erklärungskraft des Modells auf, was zum Teil mit der oben angesprochenen Be-

Tabelle 4: Regressionsmodelle zur Erklärung des Verkaufspreises bei Internet-Auktionen

	OLS1	OLS2	OLS3	RREG
Reputation	0.615	0.611	0.836	0.892
	(0.217)	(0.188)	(0.215)	(0.180)
	[0.296]	[0.270]	[0.260]	[0.357]
Startpreis	–0.019	0.000	–0.001	0.009
	(0.024)	(0.021)	(0.020)	(0.020)
	[0.027]	[0.022]	[0.023]	[0.023]
Mindesterhöhung	3.401	1.492	2.282	1.997
	(0.577)	(0.621)	(0.579)	(0.593)
	[1.339]	[0.669]	[0.580]	[0.761]
Kalenderzeit	–0.885	–0.941	–0.947	–0.991
	(0.152)	(0.133)	(0.118)	(0.127)
	[0.155]	[0.136]	[0.119]	[0.151]
Konstante	518.609	524.102	516.510	512.363
	(9.558)	(8.356)	(7.501)	(7.988)
	[10.280]	[9.710]	[8.360]	[9.874]
R^2	0.510	0.526	0.622	
N	84	83	79	83

Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

Anmerkungen: OLS-Standardfehler in runden Klammern; Bootstrap-Standardfehler in eckigen Klammern (1000 Replikationen; vgl. z.B. Mooney und Duval 1993).

Modelle: OLS1: Lineare Regression mit allen Fällen; OLS2: Lineare Regression ohne Fall 71; OLS3: Lineare Regression ohne die Ausreißer 9, 18, 32, 71 und 73; RREG: Robuste Regression nach Hamilton (1991). Die abhängige Variable ist in allen Modellen der Bruttoverkaufspreis in Franken.

9 Bei der Datenbereinigung wurde festgestellt, dass einige Anbieter mit mehreren Auktionen in der Stichprobe vertreten sind. Dies hat zur Folge, dass die Beobachtungen genau genommen nicht unabhängig sind, was bei der Berechnung der Standardfehler und somit der *t*-Werte zu berücksichtigen wäre. Ich verzichte jedoch hier auf diese Korrektur, weil sie sich kaum auf die Resultate auswirkt.

reinigung des Datenmaterials zusammenhängt, in der Hauptsache jedoch auf den starken Effekt der Kalenderzeit zurückzuführen ist.

Es stellt sich nun die Frage, ob die Resultate einem genaueren diagnostischen Check standhalten. Bei der Betrachtung der bivariaten Scatterplots zwischen dem Verkaufspreis und den Regressoren (*Abbildung 9*) sticht ein Ausreißer mit einem sehr hohen Verkaufspreis von rund 800 Franken sofort ins Auge (Nr. 71).[10] In allen vier Diagrammen liegt der Punkt weit weg vom Rest der Daten. Bezüglich der Variable „Mindesterhöhung" scheint der Ausreißer besonders problematisch zu sein, weil er auch in der Richtung der *X*-Achse deutlich von der Punktewolke abweicht und somit sehr starken Einfluss auf die Steigung der Regressionsgeraden nehmen kann. Im Diagramm, das sich auf den Zusammenhang zwischen dem Verkaufspreis und der Reputation bezieht, fällt weiterhin Punkt 73 als extremer *X*-Wert mit einer relativ großen Abweichung von der Regressionsgerade auf. Zudem gruppiert sich der Großteil der Punkte in einem Bereich von 0 bis ca. 15 Bewertungen (ca. 85 Prozent der Punkte liegen in diesem Bereich; in 32 Prozent der Fälle ist die Anzahl Bewertungen gleich null) und es besteht die Gefahr, dass der Zusammenhang von den wenigen, lose verteilten Punkten im Bereich von 20 bis 100 Bewertungen dominiert wird.[11]

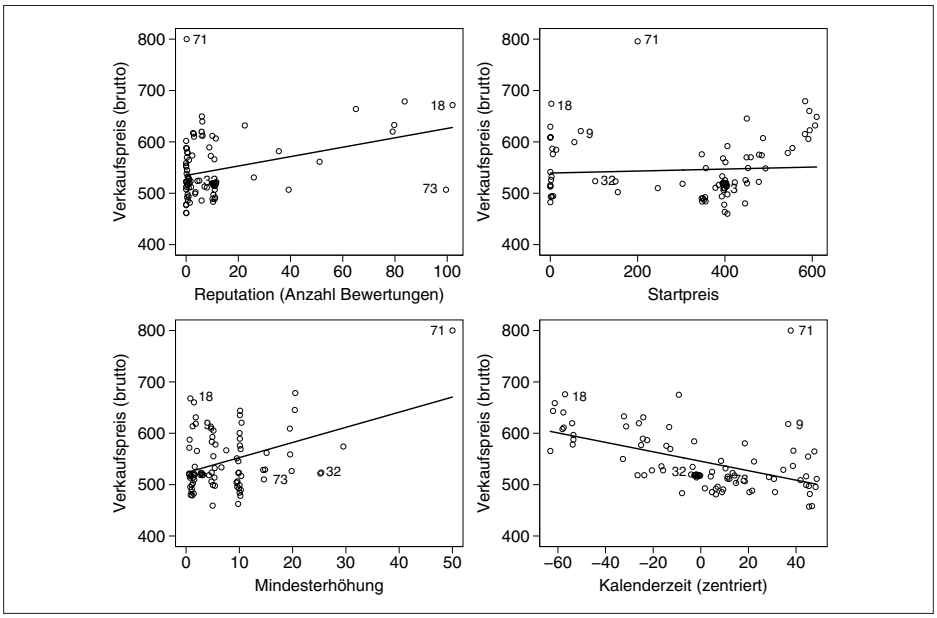
Das Ergebnis von Punkt 71 als starkem Ausreißer wird durch die partiellen Regressionsplots (*Abbildung 10*) bestätigt, die im Falle einer multiplen Regression ein verlässlicheres Bild ergeben. Im Plot für die Mindesterhöhung sticht der Punkt nun sogar noch stärker hervor. Ebenfalls erscheint der bereits identifizierte Punkt 73 im Plot für die Reputation deutlich als einflussreicher Ausreißer. Weiterhin gibt es offensichtlich einige bedeutsame Fälle, die mit den bivariaten Streudiagrammen nicht identifiziert werden konnten. Zum Beispiel hebt sich Punkt 9 in allen vier partiellen Regressionsplots mit einem relativ großen Residuum von der Punktewolke ab. Zudem tritt Punkt 73 nun auch in den Plots bezüglich des Startpreises, der Mindesterhöhung und der Kalenderzeit in Erscheinung. Weitere bedeutsame Punkte sind Nr. 32 mit einem eher großen Residuum in allen vier Diagrammen und Nr. 18 als Ausreißer bezüglich der *X*-Achse in den Plots für die Reputation und den Startpreis.

In *Abbildung 11* sind die Werte einiger Maßzahlen für den Einfluss einzelner Datenpunkte auf die Regressionsergebnisse mit Hilfe von Indexplots grafisch dargestellt. Die Maße Cooks *D* und *DFFITS* beziehen sich auf den Einfluss auf die Koeffizienten. Die bisherigen Ergebnisse werden bestätigt: Es gibt fünf Punkte (Nr. 9, 18, 32, 71 und 73), die die jeweilige Schranke zur Identifikation problematischer Fälle spürbar über-

10 Bei den markierten Punkten Nr. 9, 18, 32, 71 und 73 handelt es sich um diejenigen Fälle, die nach Cooks *D* und *DFFITS* den größten Gesamteinfluss auf die Regressionsschätzung nehmen (mehr dazu weiter unten).

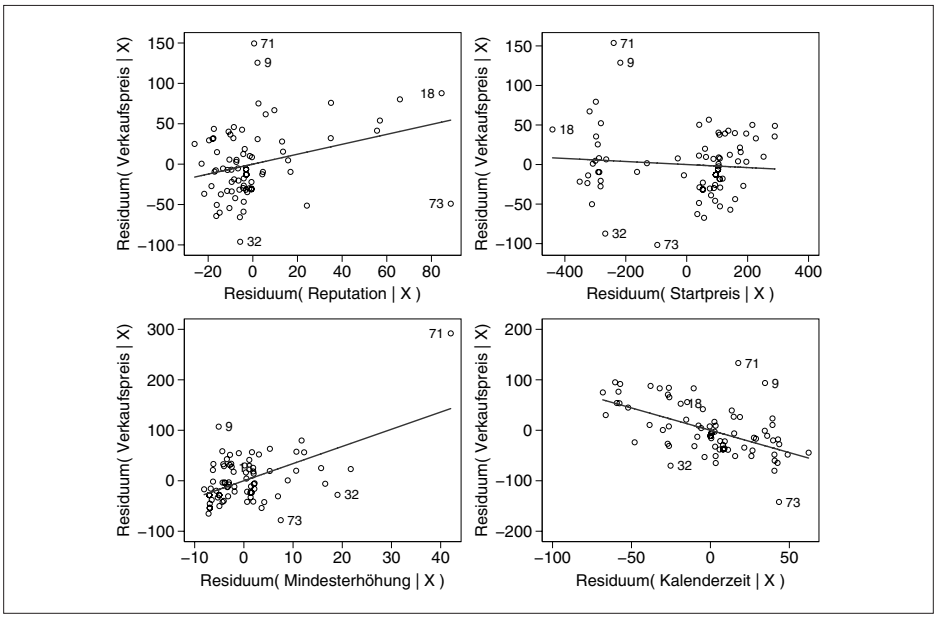
11 Weiterhin erkennt man im Diagramm zum Startpreis, dass der Verkaufspreis von unten her zensiert ist: Es können nur Gebote abgegeben werden, die über dem Startpreis liegen. Sämtliche Punkte im Plot befinden sich folglich oberhalb der Diagonale (man beachte, dass die *Y*-Achse erst bei Wert 400 beginnt). Dieses Zensierungsproblem, das auch von Diekmann und Wyder (2002) erkannt und behandelt wurde, kann mit einem zensierten Regressionsmodell angegangen werden (siehe dazu z.B. Maddala 1983 oder Breen 1996). Zu vermerken ist ferner, dass die Verkaufspreise zusätzlich von oben her begrenzt sind, da einige Anbieter einen „Sofortpreis" angegeben haben. Bietet ein Käufer den Sofortpreis, wird die Auktion gestoppt und das Produkt zu diesem Preis verkauft.

Abbildung 9: Bivariate Scatterplots für das Reputationsmodell



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

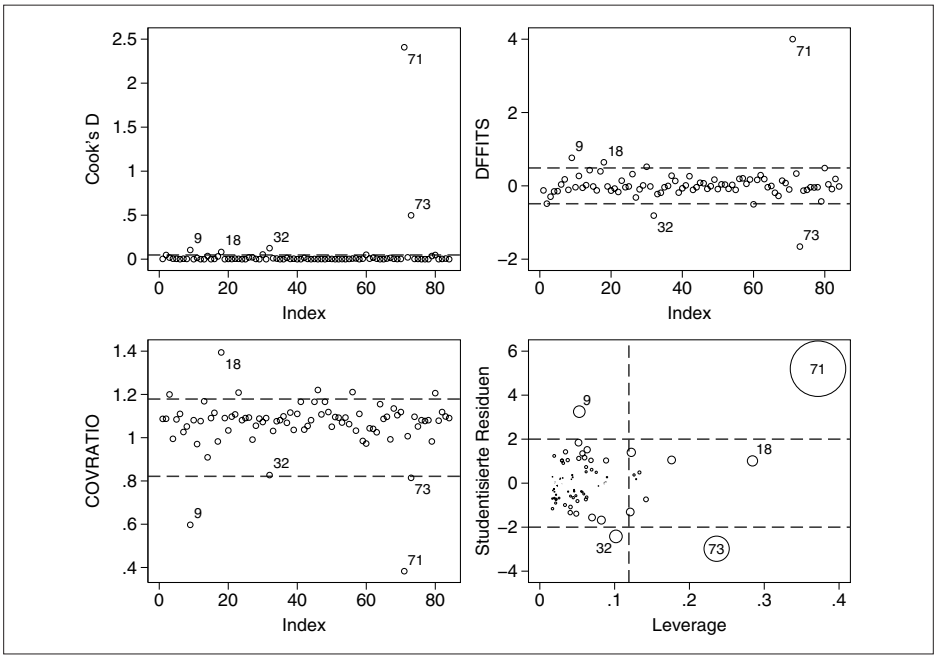
Abbildung 10: Partielle Regressionsplots für das Reputationsmodell



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

schreiten. Mit einer Ausnahme haben diese Punkte aufgrund ihrer großen studentisierten Residuen alle einen negativen Effekt auf die Präzision der Schätzung (*COVRATIO* < 1). Punkt 18 hingegen erhöht die Schätzgenauigkeit, weil sein Residuum relativ klein ist, die Leverage aber eher groß (in *Abbildung 10* erkennt man, dass Punkt 18 immer relativ nahe an der Regressionsgerade liegt, jedoch bezüglich der Reputation und des Startpreises Extremwerte annimmt). Recht informativ ist der letzte Plot in *Abbildung 11*, in dem die studentisierten Residuen gegen die Leverage abgetragen werden und die Größe der Symbole vom Wert von Cooks *D* abhängt.[12] Man sieht hier sehr gut, wie die Kombination aus Residuum und Leverage zu den Werten für die Einflussstatistiken führt. Als Ergänzung oder als Ersatz für die Grafiken kann es weiterhin sinnvoll sein, eine Tabelle mit den Ausreißern zu erstellen, die die Werte der verschiedenen Einflussstatistiken und ggf. weiterer Variablen aufführt. Dies erlaubt es, die Extremfälle einzeln bezüglich verschiedener Größen zu untersuchen. Zudem kann man sich mit der Berechnung der *DFBETAS*-Statistiken ein differenziertes Bild der Effekte von Ausreißern auf die einzelnen Koeffizienten im Modell verschaffen. Aus Platzgründen wird hier jedoch auf diese Schritte verzichtet.

Abbildung 11: Verschiedene Einflussstatistiken für das Reputationsmodell



Quelle: Bereinigte Daten aus Diekmann und Wyder (2002), eigene Berechnungen.

12 Ein weiteres Beispiel für einen Plot, der die Information aus verschiedenen Maßzahlen verdichtet darstellt, gibt Gray (1989b).

Interessant ist nun zu betrachten, inwieweit sich die Regressionsergebnisse durch die Eliminierung einzelner Datenpunkte tatsächlich inhaltlich verändern. In *Tabelle 4* ist neben dem Ausgangsmodell mit allen 84 Fällen (OLS1) ein Modell enthalten, bei dessen Schätzung der stärkste Ausreißer (Fall 71) ausgeschlossen wurde (OLS2). Man erkennt, dass der überproportionale Effekt der Mindesterhöhung durch Ausschluss dieser einen Beobachtung stark schrumpft und nicht mehr weit vom theoretisch erwarteten Wert 1 entfernt ist. Interessant ist auch noch ein anderes Ergebnis: Der Standardfehler des Koeffizienten wird im vollen Modell verglichen mit einem mittels des Bootstrap-Verfahrens gewonnenen Standardfehler durch die OLS-Methode stark unterschätz. Dies ist höchst plausibel. Bootstrap-Standardfehler werden mit Hilfe der Varianz der geschätzten Koeffizienten ermittelt, die sich ergibt, wenn die Modellschätzung anhand einer Reihe von Stichproben aus den vorliegenden Daten repetiert wird (die Bootstrap-Stichproben haben jeweils den gleichen Umfang wie die Originalstichprobe, werden jedoch mit Zurücklegen gezogen und unterscheiden sich deshalb voneinander). Die einzelnen Bootstrap-Koeffizienten für die Mindesterhöhung hängen nun stark davon ab, ob sich Fall 71 gerade in der Stichprobe befindet oder nicht, was zu starker Variabilität des Koeffizienten führt. Ein großer Bootstrap-Standardfehler verglichen mit dem gewöhnlichen OLS-Schätzer kann somit ein Indiz für die Präsenz von einflussreichen Ausreißern sein.

Tabelle 4 enthält weiterhin ein Modell, bei dem zusätzlich die Punkte 9, 18, 32, und 73 ausgeschlossen wurden (OLS3), sowie eine „robuste" Regression (RREG; es handelt sich um einen so genannten *M*-Schätzer, bei dem verschiedene Gewichtungsfunktionen zur Anwendung kommen; zu den Details vgl. Hamilton 1991). Die Resultate sind ähnlich, wobei der Effekt der Mindesterhöhung in diesen Modellen wieder etwas stärker ist. Ebenfalls leicht zugenommen hat in diesen Modellen der Effekt der Reputation, wobei zu berücksichtigen ist, dass auch hier der OLS-Schätzer für den Standardfehler verglichen mit dem Bootstrap-Schätzer jeweils etwas tief ausfällt. Dies hat mit der bereits angesprochenen starken Massierung der Datenpunkte im unteren Bereich der Reputationsskala (0–15 Bewertungen) und dem damit verbundenen relativ starken Einfluss der weit gestreuten restlichen Punkte zu tun (der Effekt der Reputation hängt stark davon ab, welche Fälle mit hoher Reputation in einem Bootstrap-Sample enthalten sind). Der Reputationseffekt ist also aufgrund der schiefen Verteilung der *X*-Werte weniger stabil als man auf den ersten Blick annehmen könnte und ist gemessen am Bootstrap-Standardfehler nur noch knapp signifikant.

Zusammenfassend lässt sich als Resultat der Ausreißeranalyse festhalten, dass erstens der Effekt der Mindesterhöhung in dem Modell aufgrund eines einzelnen Extremwerts stark überschätzt wird und zweitens der Reputationseffekt etwas weniger verlässlich ist, als es durch eine einfache OLS-Schätzung suggeriert wird. Trotzdem muss die Schlussfolgerung von Diekmann und Wyder (2002) bezüglich der Hypothese eines positiven Effektes von Reputation auf den Auktionserlös nicht revidiert werden. Der Reputationseffekt ist in allen Modellen signifikant und liegt jeweils in einem Bereich von 0.6 oder höher, d.h. pro zusätzliche (positive) Bewertung ist gemäß den Daten mit einer Erhöhung des Preises um gut 0.6 Franken zu rechnen (bei Diekmann und Wyder 2002 betrug der Effekt 0.67 Franken).

V. Diskussion

Wie alle statistischen Verfahren, die zum Ziel haben, gegebene Daten auf eine überschaubare Menge an Kennzahlen zu reduzieren, erfassen auch Regressionsschätzungen nur einen Teil der Eigenschaften und spiegeln die Daten nicht immer in angemessener Weise wider. Insbesondere in kleinen Stichproben besteht die Gefahr, dass die gewonnenen Resultate durch einige wenige Beobachtungen dominiert werden, was Zweifel an der Verlässlichkeit einer Regressionsschätzung und den daraus abgeleiteten Schlussfolgerungen aufkommen lassen kann.

Die in diesem Beitrag vorgestellten grafischen und formalen Instrumente sind einfach zu berechnen und stehen in den üblichen Statistikprogrammen mehrheitlich zur Verfügung. Sie sollten zumindest bei kleinen bis mittleren Stichproben, in denen einzelne „atypische" Beobachtungen einen besonders starken Einfluss ausüben können, im Prozess der Datenanalyse standardmäßig angewendet werden. Es bleibt jedoch darauf hinzuweisen, dass durchaus auch bei großen Stichproben übermäßig einflussreiche Datenpunkte auftreten können. Diese Gefahr besteht besonders bei komplexen Modellen mit vielen Parametern und z.B. Interaktionseffekten. Auch bei größeren Stichproben lassen sich die Instrumente also unter Umständen gewinnbringend einsetzen.

Trotz der Wichtigkeit der Regressionsdiagnostik muss man bei der Interpretation der diagnostischen Ergebnisse auch eine gewisse Großzügigkeit walten lassen. Bei kleineren Stichproben ist die Wahrscheinlichkeit groß, dass vermeintlich systematische Muster entstehen, die in Wahrheit aber nur auf Zufall beruhen (der Mensch ist bekanntlich anfällig dafür, in Zufallsprozessen eine Systematik zu erkennen; vgl. Schnell 1994: 213f.). Bei kleineren Unregelmäßigkeiten besteht sicherlich kein Anlass zu Besorgnis und man muss aufpassen, dass man die Daten nicht „überinterpretiert". Erst deutliche Abweichungen bedürfen einer genaueren Betrachtung. Im Allgemeinen sind auch die Faustregeln zur Identifikation bedeutender Datenpunkte ziemlich sensitiv, d.h. selbst unter normalen Bedingungen ist zu erwarten, dass ein Teil der Punkte im kritischen Bereich liegt.

Es stellt sich schließlich noch die Frage, was zu tun ist, wenn Ausreißer und einflussreiche Datenpunkte identifiziert worden sind.

1. Die Ausreißer sollten einer detaillierten Betrachtung unterzogen werden. Ratsam ist beispielsweise, für alle Punkte, die auf mindestens einer der Einfluss-Maßzahlen extreme Werte erzielen, die *Y*- und *X*-Werte sowie die Ausreißerstatistiken zu tabellieren. Man sollte sich auch nicht scheuen, das Rohdatenmaterial nochmals zu sichten, da Ausreißer unter Umständen auf Übertragungsfehler zurückzuführen sind oder die Rohdaten zusätzliche Informationen enthalten, die die Ausreißer erklären können. Im Beispiel mit den Internet-Auktionen (Abschnitt IV) konnte so für den kritischen Punkt Nr. 71 ausfindig gemacht werden, dass im Lieferumfang einige Zusatzaccessoires enthalten waren, in der Beschreibung des Produkts ein sehr hoher Neupreis von 900 Franken angegeben wurde (der Ladenpreis des Geräts betrug ca. 700 Franken) und der realisierte Verkaufspreis von 800 Franken dem Sofortpreis entsprach. Einerseits scheint der Wert der gelieferten Ware also tatsächlich etwas höher als üblich gewesen zu sein. Andererseits hatten unter Umständen der angegebene Neupreis sowie die Sofortpreis-

Option eine zusätzliche Signalwirkung die zur Realisierung des hohen Verkaufspreises beitrugen.

2. Ausreißer, bei denen es sich um offensichtliche Messfehler handelt, können getrost eliminiert bzw., falls möglich, korrigiert werden. Bei dem Ausschluss von Messfehlern sollte man allerdings nicht voreilig handeln, da die Klassifizierung als fehlerhafte Beobachtung nicht immer eindeutig ist und der Ausschluss von gültigen Daten unter Umständen zu Fehlschlüssen führen kann. Eine berühmte Kontroverse um den Einbezug von vermeintlich fehlerhaftem Datenmaterial bezieht sich auf eine Studie von Jasso (1985), in der die Entwicklung der ehelichen Koitusfrequenz in Abhängigkeit der Ehedauer untersucht wurde. Kahn und Udry (1986) warfen Jasso vor, vier Fälle mit Wert 88 (!) für die monatliche Koitusfrequenz, bei denen es sich um fehlcodierte fehlende Werte handle (der Code für fehlende Werte war 99), übersehen zu haben und dass dadurch ihre Regressionsschätzungen substantiell verzerrt seien. Zwar mag Jasso (1986) damit Recht haben, dass durch Ausschluss der höchsten Werte und die damit verbundene Stutzung der beobachteten Häufigkeitsverteilung ebenfalls eine Verzerrung der Regressionsresultate herbeigeführt werden könne, ihre verteidigenden Argumente für die Gültigkeit der vier in Frage stehenden Werte (z.B. „a figure of ‚88' represents a mean weakly figure of 22, which, in turn, represents three daily plus a weekly lagniappe ...", S. 741) klingen allerdings nicht sonderlich überzeugend.[13] Jasso hätte zumindest auf das Vorhandensein der Extremwerte hinweisen müssen.

3. Allgemein empfiehlt es sich, einige „robuste" Modelle zu schätzen, bei denen Ausreißer und einflussreiche Punkte ausgeschlossen werden. Ziel dieses Verfahrens kann allerdings nicht sein, einfach die Ausreißer wegzulassen und Modellschätzungen ohne die fraglichen Datenpunkte zu berichten. Oftmals geben ja gerade die unüblichen Datenpunkte wichtige Hinweise über die zu untersuchenden Sachverhalte, und es wäre fahrlässig, diese Punkte zu ignorieren. Überdies kann die Entscheidung, welche Ausreißer im Detail auszuschließen sind, eine höchst subjektive Angelegenheit sein (dies demonstrieren z.B. Collett und Lewis 1976), und es lassen sich durch den gezielten Ausschluss von Punkten, „die nicht ins Bild passen", fast beliebige Resultate erzielen. Oder allgemeiner: Eliminiert man alle Befunde, die gegen eine bestimmte Vorstellung (z.B. eine spezifische Theorie) sprechen, wird man die Vorstellung auch eher bestätigen können. Der endgültige Ausschluss einzelner Fälle ist somit nicht zu empfehlen bzw. sollte ggf. genau dokumentiert und stichhaltig begründet werden.

Der Zweck von Modellschätzung unter Ausschluss von Ausreißern ist vielmehr zu ermitteln, inwieweit sich die Regressionsergebnisse tatsächlich inhaltlich verändern. Bleiben die Schätzergebnisse im Großen und Ganzen stabil, ist man auf der sicheren Seite und kann davon ausgehen, dass von den Ausreißern keine grundlegende Verzerrung ausgeht. Sind die Schätzergebnisse jedoch stark abhängig von geringen Änderungen der Daten, dann sollte dies bei der Interpretation der Resultate verdeutlicht wer-

13 Zudem ist Jassos Behauptung, dass ihre Schätzungen auch bei tatsächlicher Fehlerhaftigkeit der vier Werte erwartungstreu seien, falsch. Jasso macht geltend, dass gemäß statistischer Theorie die Schätzer auch dann unverzerrt sind, wenn die abhängige Variable mit *zufälligen* Messfehlern behaftet ist. Die Definition eines Zufallsfehlers in dieser Theorie ist jedoch nicht kompatibel mit der vorliegenden Situation.

den. Auf Ergebnisse instabiler Modelle ist kein Verlass und sie sollten in Folgestudien mit neuen Daten überprüft werden.

Eine Alternative zur Schätzung von Modellen unter Ausschluss von Ausreißern wird manchmal auch in der Anwendung von Methoden der robusten Regression gesehen. Dazu ist zu bemerken, dass mit der klassischen robusten Regression, die auf der Anwendung von Gewichtungsfunktionen beruht (sog. *M*-Schätzer wie z.B. die Median-Regression; vgl. Huber 1981, 1996; Hampel et al. 1986 sowie einführend Berk 1990), das Ziel verfolgt wird, im Falle von nicht normalverteilten Fehlern eine größere Effizienz zu erreichen als mit der OLS-Regression. Die Verfahren sind somit zwar robust gegen *Y*-Ausreißer, jedoch nicht gegen Datenpunkte mit hoher Leverage (*X*-Ausreißer). Man spricht in diesem Zusammenhang davon, dass diese Schätzer genau gleich wie die OLS-Regression einen tiefen „Breakdown-Point" haben: Durch die willkürliche Platzierung eines einzigen (!) Datenpunktes kann die Regressionsschätzung beliebig stark verändert werden. Zur Begrenzung des Einflusses von Punkten mit hoher Leverage wurden zwar so genannte Bounded-Influence- bzw. Generalized-*M*-Schätzer (z.B. Welsch 1980; Krasker und Welsch 1982) oder High-Breakdown-Schätzer (Rousseeuw 1984; Rousseeuw und Leroy 1987) vorgeschlagen (für Übersichten siehe auch Davies 1993; Chave und Thomson 2003; Lawrence 2003); diese Schätzer sind aber leider bis jetzt noch kaum praxistauglich.[14]

4. Ausreißer und einflussreiche Datenpunkte können schließlich darauf hinweisen, dass ein Modell falsch spezifiziert ist und zum Beispiel wichtige zusätzliche Regressoren oder Interaktionseffekte übersehen oder unpassende Zusammenhangsformen modelliert wurden (ein illustratives Beispiel gibt etwa Gray 1989a). Insbesondere das Vorliegen ganzer Cluster von Ausreißern weist auf das Vorhandensein starker Dritteinflüsse hin (vgl. dazu das Beispiel zum Zusammenhang zwischen Sozialkapital und Arbeitslosenquoten in Abschnitt III.5).

Man kann also versuchen, das Zusammenhangsmodell neu zu spezifizieren, wobei dies möglichst theoriegeleitet geschehen sollte. Eine rein explorative Anpassung des Modells an die Daten ist nicht zu empfehlen, weil man so mit großer Wahrscheinlichkeit Eigenschaften modelliert, die zwar in der spezifischen Stichprobe erkennbar sind, jedoch in der Grundgesamtheit kein Gegenüber finden. So kann beispielsweise die passende Zusammenhangsform ebenfalls hauptsächlich von einigen wenigen einflussreichen Datenpunkten abhängen (vgl. die Literaturhinweise in Abschnitt III.5). Oft besteht ferner das Problem, dass zwar Ausreißer relativ deutliche Hinweise über mögliche weitere Einflussfaktoren geben, entsprechende Daten aber nicht zur Verfügung stehen oder sich die Faktoren auf Phänomene beziehen, die in der Stichprobe zu selten auftreten, um noch statistisch behandelt werden zu können. In beiden Fällen wird man zur Klärung der genauen Zusammenhänge auf Folgestudien angewiesen sein.

14 Zum Beispiel ließen sich für das Reputationsmodell in Abschnitt IV keine einigermaßen stabilen High-Breakdown-Schätzer ermitteln (verwendet wurde für die Berechnungen das lqs-Modul von Brian Ripley für R; vgl. Venables und Ripley 2002: 156ff.).

Literatur

Andrews, David F., und *Daryl Pregibon,* 1978: Finding the Outliers that Matter. Journal of the Royal Statistical Society (Series B) 40: 85–93.

Anscombe, F. J., 1973: Graphs in Statistical Analysis. The American Statistician 27: 17–21.

Atkinson, A. C., 1981: Two Graphical Displays for Outlying and Influential Observations in Regression. Biometrika 68: 13–20.

Atkinson, A. C., 1982: Regression Diagnostics, Transformations and Constructed Variables. Journal of the Royal Statistical Society (Series B) 44: 1–36.

Atkinson, A. C., 1986: Diagnostic Tests for Transformations. Technometrics 28: 29–37.

Barnett, Vic, und *Toby Lewis,* 1995: Outliers in Statistical Data, 3. Aufl. New York: John Wiley & Sons.

Beckman, R. J., und *R. D. Cook,* 1983: Outlier.....s. Technometrics 25: 119–149.

Belsley, David A., Edwin Kuh und *Roy E. Welsch,* 1980: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons.

Berk, Richard A., 1990: A Primer on Robust Regression. S. 292–324 in: *John Fox* und *J. Scott Long* (Hg.), Modern Methods of Data Analysis. Newbury Park, CA: Sage.

Bollen, Kenneth A., und *Robert W. Jackman,* 1990: Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. S. 257–291 in: *John Fox* und *J. Scott Long* (Hg.), Modern Methods of Data Analysis. Newbury Park, CA: Sage.

Brüderl, Josef, 2000: Regressionsverfahren in der Bevölkerungswissenschaft. S. 589–642 in: *Ulrich Mueller, Bernhard Nauck* und *Andreas Diekmann* (Hg.), Handbuch der Demographie, Band 1. Berlin: Springer.

Breen, Richard, 1996: Regression Models. Censored, Sample Selected, or Truncated Data. Newbury Park, CA: Sage.

Carroll, R. J., und *D. Ruppert,* 1987: Diagnostics and Robust Estimation when Transforming the Regression Model and the Response. Technometrics 29: 287–299.

Chatterjee, Samprit, und *Ali S. Hadi,* 1986: Influentual Observations, High Leverage Points, and Outliers in Linear Regression. Statistical Science 1: 379–393.

Chatterjee, Samprit, und *Ali S. Hadi,* 1988: Sensitivity Analysis in Linear Regression. New York: John Wiley & Sons.

Chave, Alan D., und *David J. Thomson,* 2003: A Bounded Influence Regression Estimator Based on the Statistics of the Hat Matrix. Applied Statistics 52: 307–322.

Cleveland, William S., 1979: Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 74: 829–836.

Collett, D., und *T. Lewis,* 1976: The Subjective Nature of Outlier Rejection Procedures. Applied Statistics 25: 228–237.

Cook, R. Dennis, 1977: Detection of Influential Observation in Linear Regression. Technometrics 19: 15–18.

Cook, R. Dennis, und *P. C. Wang,* 1983: Transformations and Influential Cases in Regression. Technometrics 25: 337–343.

Cook, R. Dennis, und *Sanford Weisberg,* 1982: Residuals and Influence in Regression. New York: Chapman and Hall.

Cook, R. Dennis, und *Sanford Weisberg,* 1982: Criticism and Influence Analysis in Regression. Sociological Methodology 13: 313–361.

Cook, R. Dennis, und *Sanford Weisberg,* 1999: Applied Regression Including Computing and Graphics. New York: John Wiley & Sons.

Crawford, Kelly D., Daniel J. Vasicek und *Roger L. Wainwright,* 1995: Detecting Multiple Outliers in Regression Data Using Genetic Algorithms. Proceedings of the 1995 ACM Symposium on Applied Computing, February 26–28, 1995, Nashville, Tennessee, United States.

Dalton, Russell J., 2004: The Greening of the Globe? Cross-national Levels of Environmental Group Membership. University of California.

Davies, P. L., 1993: Aspects of Robust Linear Regression. The Annals of Statistics 21: 1843–1899.

Diekmann, Andreas, und *David Wyder,* 2002: Vertrauen und Reputationseffekte bei Internet-Auktionen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 674–693.
Draper, N. R., und *J. A. John,* 1981: Influential Observations and Outliers in Regression. Technometrics 23: 21–26.
Fox, John, 1991: Regression Diagnostics. Newbury Park, CA: Sage.
Fox, John, 1997: Applied Regression Analysis, Linear Models, and Related Methods. Thousand Oaks, CA: Sage.
Freitag, Markus, 2000: Soziales Kapital und Arbeitslosigkeit. Eine empirische Analyse zu den Schweizer Kantonen. Zeitschrift für Soziologie 29: 186–201.
Gray, J. Brian, 1986: A Simple Graphic for Assessing Influence in Regression. Journal of Statistical Computation and Simulation 24: 121–134.
Gray, J. Brian, 1989: On the Use of Regression Diagnostics. The Statistician 38: 97–105.
Gray, J. Brian, 1989: The Four-Measure Influence Plot. Computational Statistics & Data Analysis 8: 135–225.
Gray, J. Brian, und *Robert F. Ling,* 1984: *K*-Clustering as a Detection Tool for Influential Subsets in Regression. Technometrics 26: 305–318.
Greene, William H., 2003: Econometric Analysis, 5. Aufl. Upper Saddle River, NJ: Pearson Education.
Hamilton, Lawrence C., 1991: How Robust is Robust Regression? Stata Technical Bulletin: 21–26.
Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw und *Werner A. Stahel,* 1986: Robust Statistics. The Approach Based on Influence Functions. New York: John Wiley & Sons.
Hoaglin, David C., und *Roy E. Welsch,* 1978: The Hat Matrix in Regression and ANOVA. The American Statistician 32: 17–22.
Hocking, R. R., 1983: Developments in Linear Regression Methodology: 1959–1982. Technometrics 25: 219–230.
Hosmer, David W., Jr., und *Stanley Lemeshow,* 1999: Applied Survival Analysis. Regression Modeling of Time to Event Data. New York: John Wiley & Sons.
Hosmer, David W., Jr., und *Stanley Lemeshow,* 2000: Applied Logistic Regression, 2. Aufl. New York: John Wiley & Sons.
Huber, Peter J., 1981: Robust Statistics. New York: John Wiley & Sons.
Huber, Peter J., 1996 (1977): Robust Statistical Procedures, 2. Aufl. Philadelphia: Society for Industrial and Applied Mathematics.
Jasso, Guillermina, 1985: Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. American Sociological Review 50: 224–241.
Jasso, Guillermina, 1986: Is it Outlier Deletion or is it Sample Truncation? Notes on Science and Sexuality. American Sociological Review 51: 738–742.
Kahn, Joan R., und *J. Richard Udry,* 1986: Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. American Sociological Review 51: 734–737.
Kohler, Ulrich, und *Frauke Kreuter,* 2001: Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung. München: Oldenbourg.
Krasker, William S., und *Roy E. Welsch,* 1982: Efficient Bounded-Influence Regression Estimation. Journal of the American Statistical Association 77: 595–604.
Kruskal, William H., 1960: Some Remarks on Wild Observations. Technometrics 2: 1–3.
Lawrence, David E., 2003: Cluster-Based Bounded Influence Regression. Blacksburg, Virginia: Virginia Polytechnic Institute and State University.
Lesaffre, E., und *A. Albert,* 1989: Multiple-Group Logistic Regression Diagnostics. Applied Statistics 38: 425–440.
Maddala, G. S., 1983: Limited Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.
Mallows, C. L., 1986: Augmented Partial Residuals. Technometrics 28: 313–319.
Menard, Scott, 2002: Applied Logistic Regression Analysis, 2. Aufl. Newbury Park, CA: Sage.

Mooney, Christopher Z., und *Robert D. Duval,* 1993: Bootstrapping. A Nonparametric Approach to Statistical Inference. Newbury Park, CA: Sage.

Peña, Daniel, und *Francisco J. Prieto,* 2001: Multivariate Outlier Detection and Robust Covariance Matrix Estimation. Technometrics 43: 286–310.

Peña, Daniel, und *Victor J. Yohai,* 1995: The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix. Journal of the Royal Statistical Society (Series B) 57: 145–156.

Pregibon, Daryl, 1981: Logistic Regression Diagnostics. The Annals of Statistics 9: 705–724.

Ritschard, Gilbert, und *Gerard Antille,* 1992: A Robust Look at the Use of Regression Diagnostics. The Statistician 41: 41–53.

Rousseeuw, Peter J., 1984: Least Median of Squares Regression. Journal of the American Statistical Association 79: 871–880.

Rousseeuw, Peter J., und *Annick M. Leroy,* 1987: Robust Regression and Outlier Detection. New York: John Wiley & Sons.

Rousseeuw, Peter J., und *Bert C. van Zomeren,* 1990: Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association 85: 633–639.

Schnell, Rainer, 1994: Graphisch gestützte Datenanalyse. München: Oldenbourg.

Velleman, Paul F., und *Roy E. Welsch,* 1981: Efficient Computing of Regression Diagnostics. The American Statistician 35: 234–242.

Venables, William N., und *Brian D. Ripley,* 2002: Modern Applied Statistics With S, 4. Aufl. New York: Springer.

Weisberg, Sanford, 1985: Applied Linear Regression, 2. Aufl. New York: John Wiley & Sons.

Welsch, Roy E., 1980: Regression Sensitivity Analysis and Bounded-Influence Estimation. S. 153–167 in: *Jan Kmenta* und *James Bernard Ramsey* (Hg.), Evaluation of Econometric Models. New York: Academic Press.

Welsch, Roy E., 1982: Influence Functions and Regression Diagnostics. S. 149–169 in: *Robert L. Launer* und *Andrew F. Siegel* (Hg.), Modern Data Analysis. New York: Academic Press.

Wooldridge, Jeffrey M., 2003: Introductory Econometrics. A Modern Approach, 2. Aufl. Mason, Ohio: Thomson, South-Western.

COMPUTERUNTERSTÜTZTE ANALYSE QUALITATIVER DATEN

Udo Kuckartz

Zusammenfassung: Die Auswertung qualitativer Daten wird zunehmend mit Hilfe von EDV-Programmen (QDA-Software) durchgeführt. Der Beitrag behandelt die grundlegenden Analysetechniken, in deren Mittelpunkt eine kategorienbasierte systematische Erschließung des Textmaterials steht. In den Anfängen leistete QDA-Software primär die Segmentierung und Codierung des Materials mit anschließender Synopse der zu einer Kategorie codierten Segmente, mittlerweile gehört auch die Suche nach komplexen Mustern und die Visualisierung von Zusammenhängen in den Daten zum methodischen Repertoire. Anhand eines Beispiels aus der Forschungspraxis werden der Analyseprozess und seine verschiedenen Phasen skizziert und neuartige Formen der Visualisierung von Codes und ihren Relationen vorgestellt. Abschließend werden die in der Methodendiskussion thematisierten verschiedenen Facetten des Einsatzes von QDA-Methoden dargestellt, insbesondere die Effekte auf die Forschungspraxis und die Qualität qualitativer Forschung. Die neuen Techniken ermöglichen eine methodische Kontrolliertheit und Nachvollziehbarkeit der Analyse, die für die früher eher auf anekdotische Evidenz angewiesenen qualitativen Analyseverfahren einen beträchtlichen Gewinn darstellen kann.

I. Überblick

1. Historie und Anwendungsfelder

Bei der computerunterstützten Analyse qualitativer Daten handelt es sich um eine vergleichsweise neue Entwicklung in der Methodik der empirischen Sozialforschung. Erst Mitte der 1980er Jahre waren mit der Verbreitung des Personal Computers die technischen und organisatorischen Voraussetzungen hierfür geschaffen (vgl. Drass 1980; Seidel und Clark 1984). Die Diskussion um qualitative oder quantitative Sozialforschung, wie sie in den 1970er und 1980er Jahren geführt wurde (vgl. z.B. Schwartz und Jacobs 1979; Hopf und Weingarten 1984), ging implizit noch davon aus, dass es sich beim Computer per se um ein Hilfsmittel der quantitativen Sozialforschung handle. Computer wurden mit *Rechenmaschinen* gleichgesetzt und insofern mit quantitativer Methodik assoziiert (vgl. Brent 1987; Kuckartz 1990; Weitzman 2000). Mittlerweile ist es auch für qualitativ Forschende zur Selbstverständlichkeit geworden, diverse Technologien zu benutzen. Dabei ist der Einsatz von technischen Geräten und Verfahren nicht immer ein bewusster Akt, die Technik hat sich eher unbemerkt in den qualitativen Forschungsprozess eingeschlichen, beispielsweise als Telefon- und E-Mail-Kommunikation oder in Form von Mikrofon und Kassettenrecorder, mit deren Hilfe Interviews üblicherweise aufgezeichnet werden (vgl. Flick 2002: 261f.; 2002: 364; Miles und Huberman 1994: 44). Inzwischen ist auch der Computer für die qualitative Sozialforschung

„destigmatisiert". Die Mehrzahl der qualitativ Forschenden arbeitet mit Textverarbeitungssoftware, und eine wachsende Zahl akademischer und nicht-akademischer Nutzer setzt spezielle, für die qualitative Datenanalyse entwickelte EDV-Programme wie *AQUAD*, *ATLAS.ti, MAXqda/winMAX, NVivo/NUDIST* oder *The Ethnograph* ein.[1] Diese Programme werden meist *QDA-Software* genannt, wobei QDA für Qualitative Daten Analyse steht. Im englischsprachigen Raum findet man auch die Bezeichnung CAQDAS (Computer Assistierte Qualitative Daten Analyse Software; vgl. Fielding und Lee 1998; Flick 2002: 362). In diesem Beitrag werden die grundlegenden Funktionen von QDA-Software dargestellt sowie ein Überblick über den methodischen Diskurs und seine verschiedenen Facetten gegeben. Verzichtet wird auf Leistungsbeschreibungen und -vergleiche der verschiedenen Programme. Solche Vergleiche sind an anderer Stelle zu finden (z.B. Weitzman und Miles 1995; Alexa und Züll 1999; Creswell und Maietta 2002) und insofern problematisch, als sie meist bereits bei Drucklegung veraltet sind.

QDA-Software unterstützt eine Vielzahl von Auswertungsoperationen, die seit langem im qualitativen Forschungsprozess Anwendung finden, wie beispielsweise die Zusammenstellung von Schlüsselpassagen eines Textes, das Codieren von Textpassagen nach bestimmten inhaltlichen Kriterien und die spätere Zusammenstellung von mit dem gleichen Code codierten Textsegmenten.

Computerunterstützte Analyse qualitativer Daten wird heute in einer Vielzahl von Wissenschaftsdisziplinen und Praxisfeldern betrieben, wobei Soziologie, Psychologie, Politikwissenschaft und die Erziehungswissenschaft den Schwerpunkt bilden. Soziologen analysieren beispielsweise die Biographien von Studenten oder Lebensvorstellungen von Jugendlichen. In der Psychologie und Psychoanalyse geht es etwa um Bewältigungsstrategien in kritischen Lebenssituationen. Ein typisches politikwissenschaftliches Anwendungsfeld ist die Analyse von Dokumenten und Protokollen, beispielsweise von Politikerreden oder Parlamentsdebatten. In der Erziehungswissenschaft, Sozialarbeit und Bildungsforschung geht es zum Beispiel um schulische Qualitätszirkel, Berufsbiographien arbeitsloser Lehrer, berufliche Einmündung von Diplompädagogen, Orientierungsmuster straffällig gewordener Jugendlicher oder um die subjektive Bedeutung des Computers für Schülerinnen. Ein stetig wachsendes Einsatzfeld für QDA-Software ist die Kommunikations- und Medienwissenschaft: Die Inhaltsanalyse von Massenmedien ist dabei ebenso zu nennen wie die Analyse von Kommunikationsformen, die mit den neuen Medien, Internet und Multimedia, assoziiert sind (z.B. Kommunikation in User-Groups, Foren, Online-Konferenzen und Chat-Rooms). Auch in der Geschichtswissenschaft und in interdisziplinär ausgerichteten Feldern werden QDA-Programme häufig eingesetzt, u.a. in der Genderforschung, den Gesundheits- und Pflegewissenschaften, den Rehabilitationswissenschaften und Umweltwissenschaften.

1 Aktuelle Informationen über die Programme lassen sich am besten über die jeweiligen Webseiten erhalten: <www.aquad.de>, <www.atlasti.de>, <www.qualisresearch.com> (Ethnograph), <www.qsr-international.de> (NVivo), <www.maxqda.de>.

2. Datenarten

Der Begriff qualitative Daten umfasst eine große Vielfalt unterschiedlicher Arten von Daten: Diese reichen von den Datentypen der klassischen qualitativen Forschung (Feldnotizen, Beobachtungsprotokolle und narrative Interviews) über die Printmedientexte, Dokumente und Materialien der Kommunikations- und Medienforschung bis hin zu den relativ kurzen Antworttexten, wie Sie in teilstandardisierten Surveys als Antworten auf offen gestellte Fragen anfallen. Ein noch weiter gefasster Begriff von qualitativen Daten umfasst auch Audio-, Bild- und Videomaterial. Einige QDA-Programme beinhalten auch Analysefunktionen für Audio- oder Videodaten (zumindest in einfacher Form), dieser Beitrag beschränkt sich allerdings wegen der in diesen Bereichen doch stark verschiedenen Methodik auf die Darstellung der computerunterstützten Analyse von *Text*. Der Begriff qualitative Daten steht im Weiteren für nicht-numerische Daten, genauer gesagt Textdaten, er impliziert jedoch nicht, dass die Daten auch im Kontext eines sich explizit als qualitativ verstehenden Projektes erhoben sein müssen, wenngleich das in der Mehrzahl der Fälle durchaus so ist.

Die Datenarten, die mit QDA-Software analysiert werden können, sind sehr vielfältig, neben den bereits genannten Textformen sind u.a. auch Leitfadeninterviews, fokussierte Interviews, qualitative Online-Befragungen und Focus Groups, aber auch mit Internet und neuen Medien zunehmend leicht verfügbare Daten aus Online- und CD-Rom-Quellen zu nennen.

Der Einsatz computerunterstützter Analyse ist überall dort sinnvoll, wo das Datenmaterial ohnehin in digitalisierter Form vorliegt oder leicht in diese überführt werden kann und wo dieses mit der Intention von *Systematisierung* und *Zusammenfassung* ausgewertet werden soll. Die computerunterstützte Analyse qualitativer Daten ist keine standardisierte Methode, die in immer gleicher Form an das Datenmaterial herangetragen wird. Ihre konkrete Ausgestaltung hängt sowohl von Art und Umfang des Materials als auch vom methodologischen bzw. theoretischen Ansatz ab. Für sequenzanalytische Vorgehensweisen, bei denen Texte Wort für Wort in extensiver Weise interpretiert werden, ist die Unterstützungsleistung von QDA-Software naturgemäß eher gering, wenngleich auch dort bspw. die Möglichkeit, Textstellen über Hyperlinks miteinander zu verknüpfen, für explizierende Analysen durchaus wertvoll sein kann.

II. Was leistet die computergestützte Analyse?

1. Segmentieren und codieren von Text

So wie der innere Kern der klassischen Inhaltsanalyse durch Berelsons Satz: „Die Inhaltsanalyse steht und fällt mit ihren Kategorien" (Berelson 1952: 147), prägnant zum Ausdruck gebracht wird, so lässt sich für die computerunterstützte Analyse qualitativer Daten formulieren, sie stehe und falle mit „Cut-and-Paste" und „Code-and-Retrieve". Unter „Cut-and-Paste" wird die Segmentierung und Kategorisierung von Texten verstanden: Inhaltlich bedeutsame Textpassagen werden identifiziert und ihnen wird ein Code zugeordnet. Der Ausdruck „Cut-and-Paste" nimmt Bezug auf die im Vor-Com-

puter-Zeitalter betriebene handwerkliche Bearbeitung von Texten: Mit Schere, Klebstoff und Karteikarten bearbeitete der Forscher seine Texte; er kopierte das Material und schnitt jene Stellen aus, die zu einem bestimmten Thema relevant erschienen, klebte sie auf eine Karteikarte und vermerkte Stichworte bzw. Kategorien (vgl. Lofland und Lofland 1984: 134). Im zweiten Schritt des elektronischen Analyseprozesses, nach dem Segmentieren und Codieren, steht das Wiederfinden („Retrieve") im Mittelpunkt. Im einfachsten Fall geht es um das Wiederfinden und die Zusammenstellung all der Textpassagen, die mit der gleichen Kategorie oder Subkategorie codiert wurden. Diese beiden Techniken stellen die Schlüsseltechniken der computerunterstützten Analyse qualitativer Daten dar, was allerdings nicht bedeutet, dass die Nutzung von QDA-Software sich nur in Segmentieren, Codieren und Retrieval erschöpfen würde. Basierend auf diesen Grundideen haben sich im letzten Jahrzehnt sehr vielfältige Methoden und Verfahren entwickelt, ganz ähnlich wie in der sozialwissenschaftlichen Statistik, wo die Ideen von Mittelwert und Varianz als Bezugsgrößen und des linearen Zusammenhangs von Variablen zwar zentral sind, aber sich darauf aufbauend ein sehr vielfältiges Bild von unterschiedlichen Verfahren bietet.

Die Kernfunktionen des „Cut-and-Paste" und „Code-and-Retrieve" sind in allen heutigen QDA-Programmen verfügbar, d.h. alle Programme unterstützten auf effektive Weise das Segmentieren und Codieren von Text sowie das Arbeiten mit Kategorien. Technisch funktioniert dies denkbar einfach: Textstellen werden, wie aus Textverarbeitungsprogrammen gewohnt, mit der Maus markiert und es wird ein Code zugewiesen.

Die Codes werden in Form eines Kategoriensystems verwaltet, wobei die verschiedenen Programme dies in unterschiedlicher Form tun, z.B. als lineare Liste von Codes, als hierarchisch organisierten Codebaum oder Code-Netzwerk. Programme unterscheiden sich u.a. dadurch, welche Möglichkeiten zur Strukturierung und Organisation des Kategoriensystems (Ausdifferenzieren von Codes, Fusionieren von Codes, Bilden von Subkategorien, Umorganisation von Beziehungen) sie bieten und wie komfortabel die Nutzung dieser Funktionen ist.

Wie unterschiedlich die Programme im Hinblick auf ihre Leistungen auch sein mögen, so gilt doch für alle, dass sie im Vergleich zum herkömmlichen „Cut-and-Paste" mit Schere, Karteikarte und Klebstoff einen immensen Forschritt darstellen. So wie auch das kleinste und unbequemste Auto noch eine andere Welt von Mobilität und Geschwindigkeit ermöglicht, so eröffnet jedes QDA-Programm dem Forscher eine neue Dimension im Hinblick auf Organisation und Management seiner Texte. Nicht zuletzt ist es wohl diese große Lücke zwischen handwerklichen und computerunterstützten Techniken, die zu einer großen Vielfalt der Programme geführt hat und auch noch lange solche QDA-Software am Leben erhält, die kaum mehr als die Basisfunktionen bietet und technisch vielleicht nicht mehr auf dem neuesten Stand der Softwaretechnik ist.

Was unter einem Code verstanden wird, wie Kategoriensysteme aussehen und wie der Codierungsprozess gehandhabt wird, wird nicht durch die Wahl eines QDA-Programms bestimmt. Es ist Angelegenheit des Forschers, die Eigenschaften seines Kategoriensystems zu bestimmen. Für die Bildung von Codes bzw. die Konstruktion eines Kategoriensystems lässt sich eine Dualität von Verfahrensweisen ausmachen. Auf der einen Seite findet man gänzlich induktiv orientierte Verfahren wie die Grounded Theory

Glaserscher Lesart (Glaser 2002), bei denen die Codes aus der sorgfältigen Analyse der Texte – dem Anspruch nach möglichst ohne theoretisches Vorwissen – entwickelt werden. Diametral entgegen gesetzt hierzu sind deduktiv orientierte Verfahren, bei denen ein vorgegebenes Kategoriensystem auf die Daten angewendet wird. Diese Form der Kategorienbildung findet man häufig im Kontext von leitfadenorientierten Interviews, deren Fragen und Themen mehr oder weniger direkt in thematische Auswertungskategorien übersetzt werden.

Zwischen diesen beiden Extremen – vollständig induktive bzw. vollständig deduktive Kategorienbildung – existieren vielfältige Mischformen, beispielsweise Analyseformen, die mit einem Raster vorgegebener formaler Kategorien beginnen, welche dann im zweiten Schritt auf der Basis des empirischen Materials dimensionalisiert und ausdifferenziert werden (vgl. Kuckartz 1996), oder etwa die Konzeption der Grounded Theory, wie sie durch Strauss und Corbin ausgearbeitet wurde, in der deduktive Elemente durch die zu Grunde liegenden Theorien, Forschungsfragestellungen und Vorerfahrungen der Forscher bei der Konstruktion des Codesystems anerkannt werden.

QDA-Software ermöglicht das Codieren und Arbeiten mit den Codes, nimmt aber keine selbstständigen oder automatischen, nach festen Regeln funktionierende Codierungen vor. Der intellektuelle Codiervorgang wird wirksam unterstützt, wobei gerade die Auswertungsvorgänge rund um das Codieren eben in Analysestilen wie der Grounded Theory oder der qualitativen Inhaltsanalyse eine große Rolle spielen. Codes werden integriert, fusioniert, ausdifferenziert, gruppiert, abstraktere Codes werden entwickelt und bilden in Form von Schlüsselkategorien einen Dreh- und Angelpunkt für die „Story Line".

Fortgeschrittene QDA-Programme machen es möglich, den Entstehungsprozess eines Codes wie auch die Codierungen zu protokollieren und in Form einer Signatur festzuhalten, damit ist zu jedem Zeitpunkt klar, wer welche Kategorie an welchem Datenmaterial entwickelt hat. In Form von Code-Memos oder Kategorienbeschreibungen kann festgehalten werden, wie ein Code definiert ist und wie Ankerbeispiele aussehen. Dies ist insbesondere für die gemeinsame Arbeit in einer Forschergruppe von nicht zu unterschätzendem Wert und eine elementare Hilfe bei der Herstellung der Reliabilität der Analyseergebnisse.

Die Art und Weise, wie nun ein Text inhaltlich erschlossen und kategorisiert wird, ist Angelegenheit des Forschers und des von ihm gewählten methodischen Ansatzes, sie wird nicht von QDA-Programmen vorgeschrieben. Das Spektrum reicht hier von der sehr freien Vorgehensweise der Grounded Theory bis zur stark vorgegebenen hochgradig resümierenden und aggregierenden Deskriptorzuweisung.

2. Visualisierung von Codierungen und Arbeiten mit dem Kategoriensystem

Zu den handwerklichen Techniken des Arbeitens mit verbalen Daten gehörten das Markieren von Text mit farbigen Markierstiften und das Schreiben von Bemerkungen oder Codes an den Rand. Computergestützte Auswertung erlaubt dies ebenfalls, und zwar mit größerer Systematik und höherem Informationsgehalt. Programme wie ATLAS.ti, NVivo oder MAXqda visualisieren in Form von farbigen Codierstreifen, wo

etwas codiert worden ist und zeigen die zugehörige(n) Kategorie(n) gleichzeitig mit an. In so genannten Tooltipps, die erscheinen, sobald die Maus über die Codierung geführt wird, sind weitere Informationen verfügbar, z.B. über den Autor und das Datum der Codierung. So ist ersichtlich, ob eine Textstelle mehrere Codierungen aufweist und welche Codes ggf. zugeordnet sind. Auch die Definition von Codes und Ankerbeispiele sind sofort verfügbar.

Im Allgemeinen ist das Arbeiten mit Codes und Kategoriensystemen in der qualitativen Datenanalyse zentral. Charakteristisch ist, dass die Arbeit mit den Codes einerseits zirkuläre Elemente aufweist, d.h. es wird nicht mit einem vorab fertigen Kategoriensystem gearbeitet, andererseits, dass die benutzten Codes mit dem Fortgang der Analyse theoretischer und abstrakter werden, d.h. sich vom Datenmaterial und der Sprache der Akteure entfernen. Verschiedene Ansätze und Analysestile entwickeln hier ihre je eigenen Regeln: Während die Grounded Theory zum frühzeitigen Theoretisieren ermutigt und das Paraphrasieren und Zusammenfassen von Textinhalten ablehnt, propagieren andere Ansätze wie Mayrings qualitative Inhaltsanalyse gerade eine Arbeitsweise, die mit Paraphrasieren beginnt und über Zusammenfassung und Abstraktion zur Bildung von theoretisch gehaltvollen Kategorien „voranschreitet". Beide Ansätze werden durch QDA-Software wirksam unterstützt, denn diese erlauben den flexiblen Umgang mit dem Kategoriensystem. Codes können leicht umbenannt, umorganisiert und fusioniert werden. Sie lassen sich ggf. unter abstraktere Konzepte subsumieren. Technisch geschieht dies, indem Codes im komplexen Kategoriensystem mit der Maus hin und her geschoben werden, wobei die zugehörigen Codierungen von Textstellen erhalten bleiben. Durchlaufen wird so ein iterativer Prozess der Kategorienbildung, der dem Ideal qualitativer Arbeitsweisen weitaus näher kommt als die früheren handwerklichen Techniken.

3. Systematische Suche nach Zusammenhängen von Codes

Die Schlüsseltechnik „Code-and-Retrieve" bedeutet längst nicht mehr nur das Wiederfinden von mit dem gleichen Code codierten Textpassagen, sondern dies kann sehr komplexe Suchvorgänge beinhalten. Die Nähe verschiedener Codes zueinander innerhalb eines Textes wie auch ihre Abfolge kann ebenso zum Suchkriterium werden wie das gleichzeitige Vorkommen oder Nicht-Vorkommen von Codes. Solche Suchanfragen nach der Beziehung von Codes zueinander werden meist mit Hilfe der Logikregeln der Booleschen Algebra formuliert. In elaborierten Programmen kann gefragt werden nach:

- der Nähe von Codes mit Angabe eines maximalen Abstands gemessen in Absätzen;
- der Sequenz von Codes, im Sinne von Code A folgt auf Code B, ebenfalls mit Festlegung eines maximal erlaubten Abstandes;
- Überschneidungen von Codes oder von Teilmengen eines Code-Sets, wobei die Anzahl von sich überschneidenden Codes festgelegt werden kann;
- Codierungen, die in andere Codierungen mit bestimmten interessierenden Codes eingebettet sind;
- Codierungen, die sich außerhalb von bestimmten interessierenden Codes befinden.

Berücksichtigt man, dass diese komplexen Suchstrategien nicht nur auf einzelne Codes, sondern auch auf eine Struktur von Codes, d.h. beispielsweise eine Baumstruktur, angewendet werden kann, wird deutlich, dass das analytische Potential hier weit höher ist als beim einfachen Wiederfinden von mit dem gleichen Code codierten Textpassagen.

Bei der Auswertung der Beziehungen zwischen einzelnen Codes im bearbeiteten Material entsteht eine Matrix der Codierungen. Diese Organisation des Datenmaterials lässt sich mit dem Erstellen einer Kontingenztafel bei der Auswertung von quantitativen Daten vergleichen. Bei der quantitativen Analyse interessieren aber primär Parameter und Koeffizienten, die wie das Chi-Quadrat im Falle der Kreuztabelle den Tabelleninhalt und die vorhandenen Zusammenhänge in einer einzigen Zahl zusammenfassen. Die qualitative Analyse zielt natürlich nicht auf die Ermittlung eines resümierenden Zahlenwertes, sondern auf präzise und verstehbare Interpretation dessen, was in einer solchen *Themenmatrix* enthalten ist. In den einzelnen Zellen der Matrix befinden sich nicht Zahlen, sondern Textstellen, auf die man jederzeit Zugriff hat, so dass man gewissermaßen immer nur einen Klick von den Originaltexten entfernt ist. So wird es möglich zu selektieren, zu separieren und zu abstrahieren, ohne die Kontextkontrolle aufzugeben.

Memos: Einen weiteren wichtigen Bestandteil computerunterstützter Analyse qualitativer Daten stellt die Arbeit mit Memos dar. Nun gehört es gewiss auch zur Arbeitsweise quantitativer Forscher, im Verlaufe der Datenauswertung Aufzeichnungen über Hypothesen, theoretische Bezüge und Zusammenhänge in den Daten zu erstellen, ohne dass dieser Vorgang methodisiert worden wäre. Im qualitativen Analyseprozess kommt dem Anfertigen von Memos allerdings ein weitaus größerer Stellenwert zu. Vor allem die Grounded Theory hat dem Arbeitsmittel Memos große Aufmerksamkeit gewidmet (vgl. Strauss und Corbin 1996). In den Standardwerken der Grounded Theory findet man eine Ausdifferenzierung von verschiedenen Typen von Memos, z.B. Code-Memos, theoretischen Memos, integrativen Memos. Bei Strauss und Corbin werden zudem sehr ausführliche Anleitungen zum Schreiben von Memos und zum Arbeitsprozess mit Memos gegeben. QDA-Software setzt diese Ideen um und ermöglicht es, Memos nicht nur für beliebige Textstellen und für jede Kategorie oder Subkategorie zu erstellen, sondern auch für Texte, Projekte, Textgruppen, gewissermaßen für fast alle „Einheiten", die Teil eines qualitativen Datenkorpus sind. Elaborierte Programme enthalten ein spezielles Memosystem, in dem die Memos ähnlich wie in einem Karteikasten ein von den Texten und Codierungen unabhängiges Arbeitsmaterial darstellen, in welchem gesucht werden kann, das aggregiert und zusammengefasst und selbst wiederum codiert werden kann. Memos sind schnell geschrieben, aber wenn ihre Anzahl eine bestimmte Größenordnung überschreitet, sind sie nur noch schwer zu überblicken. Ein Zugriff auf die Gesamtheit aller erstellten Memos ist deshalb von nicht zu unterschätzendem Wert: Können Memos nach Autoren, dem Zeitpunkt der Erstellung oder zugeordneten Stichworten sortiert werden und auf das Vorkommen bestimmter Wörter hin durchsucht werden, so ist die jederzeitige Verfügbarkeit der analytischen Zwischenschritte der Forscher gewährleistet.

4. Informationswissenschaftliche Methoden

Mehr und mehr umfasst QDA-Software auch Techniken und Prozeduren, die im Feld der qualitativen Methoden neu sind und im Vor-Computer-Zeitalter wegen des hohen Zeitaufwandes praktisch schlichtweg nicht möglich waren. Solche Verfahren stammen meist aus dem Bereich der Informationswissenschaft. Beispiele hierfür sind die lexikalische Suche und Methoden des Text-Mining und damit verbundene Möglichkeiten zur automatischen Vercodung. Insbesondere die lexikalische Suche nach Wörtern und Wortkombination innerhalb bestimmter genau definierter Textabschnitte besitzt eine große Attraktivität, denn im Unterschied zu den Code-and-Retrieve Methoden ist hier kein aufwändiger Codierungsprozess notwendig, sondern die Interpretation und Analyse des so selektierten Datenmaterials kann quasi ohne Vorarbeit beginnen. Die Möglichkeit, komplexe lexikalische Suche mit automatischen Codierprozessen zu verbinden, ist dabei besonders attraktiv.

Die informationswissenschaftlichen Methoden nutzen die Stärke des Computers, nämlich die Möglichkeit des schnellen Zugriffs auf große Datenmengen, optimal aus. Zur Exploration des Datenmaterials lassen sie sich hervorragend nutzen. Dem qualitativ Forschenden bieten diese Funktionen die Möglichkeit zum „Surfen" im Datenkorpus. Diesen Prozess kann man mit dem aus dem Diskurs über neue Medien stammenden Begriff Serendipity[2] (assoziative Informationsbeschaffung) umschreiben, d.h. man macht Entdeckungen durch Zufälle und Scharfsichtigkeit und findet Zusammenhänge, nach denen man vielleicht gar nicht gesucht hat. Dem Reiz, den die lexikalische Suche und die automatische Vercodung als schneller, voraussetzungsloser Zugriff auf die Daten ausübt, steht allerdings die Gefahr mangelnder Systematik gegenüber, d.h. dass man Exploration mit Analyse verwechselt und sich ein Gefühl von „Lost in Cyberspace" einstellt, man in der Datenmenge und der in ihr enthaltenen Beziehungsvielfalt verloren geht. Dennoch sollte ein Computerwerkzeug zur Qualitativen Datenanalyse hier keine methoden-erzieherische Einschränkung vornehmen, vielmehr ist auch hier der Forscher gefordert, die seinem Forschungsgegenstand und methodischem Ansatz angemessene Entscheidung bei der Nutzung des Werkzeuges zu treffen.

2 In der Soziologie wurde der Begriff „Serendipity" frühzeitig von Robert K. Merton eingeführt, auf den auch zurückgeht, dass der Ausdruck in den USA zum Modewort wurde (vgl. zusammenfassend Merton und Barber 2004). Bei Merton ist Serendipity ein wichtiger Aspekt seiner Wissenschaftssoziologie. Auf einer einschlägigen Webseite (www.polarluft.de) findet sich folgende Begriffserläuterung: „Serendipity *(englisch)*. Die Erfahrung, durch Zufall bei einer Tätigkeit überraschend etwas Schönes zu finden, zum Beispiel beim Suchen im Lexikon (oder im Web) woanders hängen zu bleiben. Die Bezeichnung geht auf das persische Märchen *‚The Three Princes of Serendip'* zurück, auf das sich der englische Schriftsteller Horace Walpole (1717–1797) in einem Brief bezieht. Das Märchen erzählt vom König von Serendip (heute vermutlich Sri Lanka, das frühere Ceylon), der seine drei Söhne ausschickt, um das kostbarste, was es auf der Welt gibt, zu finden. Auf seinen Reisen gerät jeder der drei mehr oder weniger zufällig in Situationen, die ihn dazu bringen, seine Meinung, was das Wertvollste sei, zu ändern. Endlich kommen die Prinzen zu ihrem Vater zurück und erfahren, dass dieser das kostbarste auf der Welt schon gefunden hat – gleich vor den Mauern seines Palastes. Auf eine Kurzformel gebracht steht *Serendipity* für die Entdeckerfreude, wenn man – vielleicht durch eine Fügung des Schicksals – auf etwas Unerwartetes stößt."

5. Inhaltsanalytische Methoden

Techniken der klassischen, auf Quantifizierung zielenden Inhaltsanalyse (vgl. Früh 2001; Merten 1995) sind bislang noch relativ wenig in das Feld der computerunterstützten Analyse qualitativer Daten infiltriert. Das mag zum einen an der traditionellen Distanz zwischen quantitativen und qualitativen Methoden liegen, die auch mit einer Portion gegenseitiger Ignoranz einhergeht. Vielleicht hat aber auch die Existenz von Software zur quantitativen Inhaltsanalyse bewirkt, dass man diese Methoden und Techniken lange Zeit nicht in QDA-Software zu integrieren versucht hat, weil man möglicherweise glaubte, sie seien bereits durch diese Spezialsoftware abgedeckt. Schon seit geraumer Zeit gibt es nämlich Computerprogramme wie Textpack, General Inquirer oder CoAn, die automatische wortbasierte Vercodungen auf der Basis eines Diktionärs durchführen (vgl. Alexa und Züll 1999). Zu den Funktionen dieser Software gehören u.a. die Erstellung von Häufigkeitswortlisten, alphabetischen Wortlisten, bereinigten Wörterbüchern und sowie Keyword-in-context-Analysen. Alle diese Verfahren eignen sich gut für die Textexploration und Textdeskription (vgl. Züll und Mohler 1992, 2001).

Allmählich finden diese Funktionen auch in QDA-Software Eingang: ATLAS.ti bietet einen sogenannten Word-Cruncher an, der für einen oder mehrere Texte Häufigkeitswortlisten erstellt. MAXqda offeriert neben Häufigkeitswortlisten und bereinigten Wörterbüchern auch die komplexe lexikalische Suche mit automatischer Codierung sowie ein Zusatzmodul MAXdictio, das neben den genannten Standardfunktionen auch die Möglichkeit bietet, Diktionäre aufzubauen und automatisch auf der Basis eines Diktionärs zu codieren.

6. Methodenmix

Die computerunterstützte Analyse qualitativer Daten erleichtert die Kombination von qualitativen und quantitativen Auswertungsprozeduren, und zwar auf verschiedenen Ebenen:

Erstens können die vorgenommenen Codierungen als Datenmatrix (Texte*Codierungen) exportiert und statistisch analysiert werden. Diese Matrix enthält also für jeden Text die Information, welche Kategorie bzw. Subkategorie wie häufig zugeordnet wurde. Dies entspricht der Datenmatrix, wie man sie von Verfahren der klassischen Inhaltsanalyse (z.B. der Frequenzanalyse) kennt.

Zweitens steht in manchen Programmen mit den Fallvariablen (Ethnograph, MAXqda) bzw. Attributen (NVivo) ein Hilfsmittel für die Nutzung von quantitativen Informationen in Verbindung mit den Texten zur Verfügung. Hier können zum einen a priori vorhandene Informationen über den Text festgehalten werden, z.B. soziodemographische Daten, Informationen über den Interviewer und den Interviewverlauf. Zum anderen kann dieses Werkzeug auch dazu benutzt werden, um Klassifizierungen und Bewertungen, die auf der Grundlage der Textinterpretation vorgenommen werden, in Form von Variablenwerten festzuhalten. Die Fallvariablen/Attribute können als Selektionskriterien für das Text-Retrieval, d.h. das Wiederfinden codierter Textpassagen he-

rangezogen werden. Wenn man beispielsweise die Variablen „Alter", „Anzahl der Kinder" und „Einkommenshöhe" als Attribute erfasst hat, lässt sich die Frage „Wie sehen die sozialen Wunschprojektionen von Befragten mit Kindern und einem Monatseinkommen über 6000 DM aus?" leicht beantworten. Auch lassen sich mit den Fallvariablen/Attributen, ähnlich wie mit den Codehäufigkeiten, statistische Berechnungen durchführen.

Drittens besteht bei manchen QDA-Programmen die Möglichkeit, eine komplette SPSS-Datenmatrix zu importieren und mit den Textdaten zu verknüpfen. Bedingung ist natürlich, dass für jede Person beide Datentypen – sowohl Text als auch standardisierte Daten – vorliegen. Dies ist beispielsweise bei teilstandardisierten Befragungen der Fall, wo neben Fragen mit Antwortvorgaben auch eine gewisse Anzahl von offenen Fragen gestellt wird. Das gleiche gilt für das problemzentrierte Interview nach Witzel, wo neben dem transkribierten Interviewtext üblicherweise die Fragen, die als Frage-Antwort-Schema aufgebaut sind, in Form eines Kurzfragebogens erhoben werden (Witzel 2000).

Methodenmix erleichtert die Realisierung von Konzepten der Triangulation im Sinne von Daten-Triangulation bzw. Across-method-Triangulation im Sinne von Denzin (1970), d.h. des systematischen Vergleichs und aufeinander Beziehens von Daten, die mit unterschiedlichen Methoden erhoben wurden.

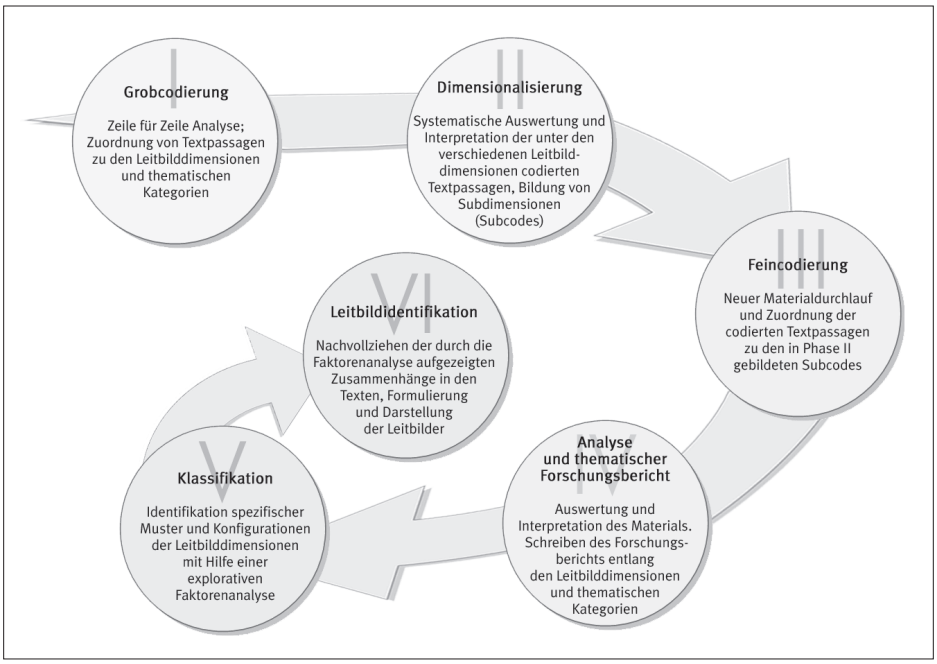
III. Beispiel aus der Forschungspraxis

Das folgende Beispiel aus der Forschungspraxis soll den Analyseprozess mit QDA-Software anschaulich machen. Es stammt aus dem Projekt „Umweltkommunikation und Lokale Agenda 21", einer Feldstudie, in der über einen Zeitraum von 18 Monaten hinweg Lokale Agenda 21-Initiativen in sechs verschiedenen Bezirken von Berlin untersucht wurden.[3] Die hier gewählte analytische Vorgehensweise stellt *eine* der möglichen Varianten des Einsatzes von QDA-Software dar. Die Darstellung der Auswertung der Befragung der Akteure der Agenda-Initiativen will beispielhaft den Gang der computergestützten Analyse aufzeigen und beansprucht nicht, eine Darstellung der inhaltlichen Ergebnisse des Projektes zu geben.[4] Die durch einen Leitfaden strukturierten Interviews (Dauer zwischen 45 und 90 Minuten) wurden auf Tonband aufgenommen und vollständig transkribiert. Der Leitfaden beinhaltete direkt auf die Projektfragestellung bezogene Themen, z.B. nach den persönlichen Motiven für die Agendaarbeit,

3 Das Projekt wurde unter dem Titel „Initiativen für eine nachhaltige Entwicklung: Neue Dialogformen und Kommunikationsstile im Zusammenhang mit der Agenda 21" vom Umweltbundesamt gefördert. Inhalt: Im empirischen Teil der Studie wurden folgende Methoden eingesetzt: a) offene, leitfadenstrukturierte Interviews mit Akteuren der Agenda-Initiativen (mit je 5 Personen pro Initiative wurden Interviews geführt, insgesamt also n=30); b) leitfadenstrukturierte Interviews mit Experten (n=15) aus dem Umfeld der Initiativen (Vertreter von Politik, Wirtschaft, Kommunalverwaltung und Nicht-Regierungsorganisationen); c) Feldnotizen der ethnographisch orientierten Feldstudie, teilweise auch Beobachtungsprotokolle der Sitzungen der Initiativen; d) telefonische Befragung der Bevölkerung mit einem überwiegend standardisierten Fragebogen.

4 Diese ist zu finden in de Haan/Kuckartz/Rheingans (2000).

Abbildung 1: Analysephasen bei der computergestützten Leitbildanalyse



nach Formen und Inhalten der Arbeit sowie den persönlichen Wünschen und den Vorstellungen der Machbarkeit. Der an die sorgfältige Lektüre des transkribierten Materials anschließende Analysegang ist in *Abbildung 1* schematisiert.

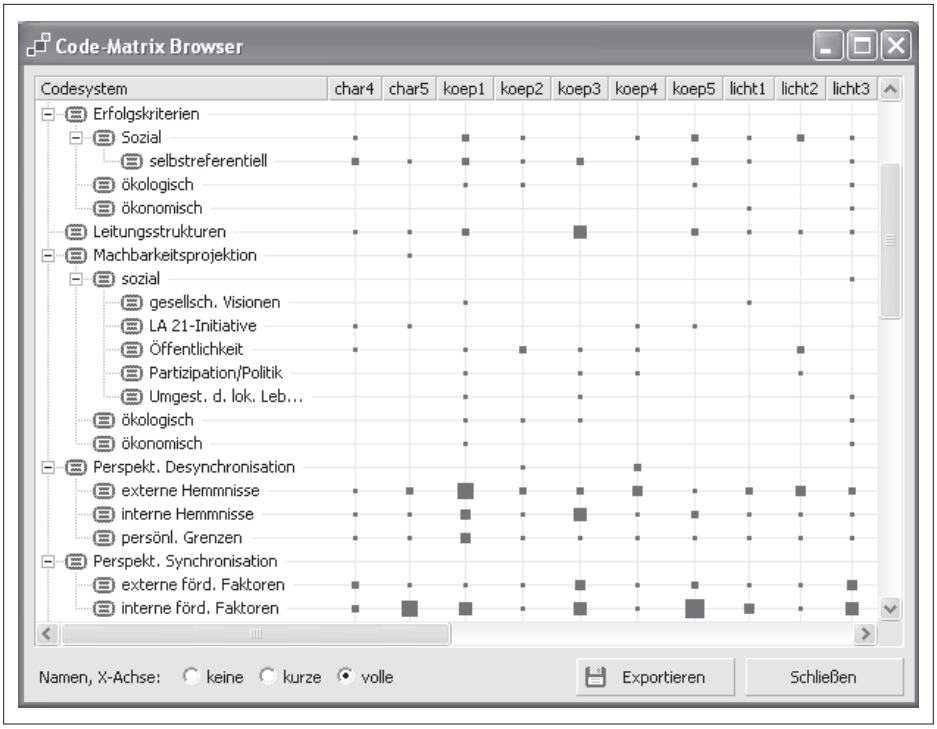
In der *Phase der Grobcodierung* werden die Interviews systematisch durchgearbeitet. Auf der Basis einer Zeile für Zeile-Analyse werden Textpassagen entsprechend dem aus der Technikgeneseforschung stammenden Konzept der Leitbildanalyse (vgl. Marz und Dierkes 1992) verschiedenen Leitbilddimensionen zugeordnet. Ferner werden weitere, direkt mit den Themen des Leitfadens assoziierte Kategorien, wie beispielsweise die Kategorie „Erfolgskriterien", codiert.

In der Phase II, *der Dimensionalisierung,* stehen die Zusammenschau, die Interpretation und die Systematisierung der unter den verschiedenen Leitbilddimensionen codierten Textsegmente im Zentrum. Zu den Leitbilddimensionen werden induktiv, d.h. in den Daten gegründete, Subdimensionen gebildet. Beispielsweise zur Kategorie „Machbarkeitsprojektionen" die Subdimensionen „ökonomisch", „ökologisch" und „sozial", je nachdem, auf welche Bereiche sich die geäußerten Machbarkeitsvorstellungen beziehen. Die soziale Dimension wird dann weiter untergliedert in „Gesellschaftliche Visionen", „Partizipation/Politik", „Öffentlichkeit", „LA 21-Initiative" und „Umgestaltung des lokalen Lebensraums".

Diese neu gebildeten Subcodes werden in der Phase III, *der Feincodierung,* auf das Material angewendet, d.h. alle in Phase I codierten Segmente werden neu zugeordnet. Mit diesem zweistufigen Codierprozess sind alle notwendigen Vorarbeiten geleistet: Gegründet auf den inhaltlich zugeordneten Textsegmenten kann eine Zusammenstellung

und systematische Interpretation des Materials erfolgen und ein Forschungsbericht geschrieben werden, der den Kategorien des Leitbildkonzepts folgt (Phase IV). Dabei können auch quantitative Aspekte eine Rolle spielen, denn es ist durchaus von Belang, ob 20 von 30 Agenda-Akteuren Erfolgskriterien definieren, die sich selbstbezüglich auf die Entwicklung der Agenda Initiativen beziehen, oder ob dies nur in wenigen Fällen zu registrieren ist. Die Differenz zu quantitativ argumentierender Forschung ist aber, dass es nicht auf das Berichten der bloßen Zahlen bzw. von Prozentanteilen ankommt, sondern auf die Wörter und Argumentationen, mit denen die Frage der „Erfolgskriterien" von den Befragten thematisiert wird. QDA-Software kann die fallbezogenen Code-Häufigkeiten der verschiedenen Dimensionen und Subdimensionen direkt zur Verfügung stellen, so dass leicht unterschieden werden kann, ob es sich um ein von vielen Befragten genanntes Kriterium (das dann zu Recht als ein „typischerweise von Akteuren genanntes Kriterium" bezeichnet werden kann) oder lediglich um eine randständige Einzelmeinung handelt. Zur Verdeutlichung stehen jeweils Zitate zur Verfügung, deren „Repräsentativität" kontrollierbar und nachprüfbar ist. Einer bloß episodischen Evidenz, wie sie qualitativer Analyse häufig vorgeworfen wird, kann so wirksam vorgebeugt werden.

Abbildung 2: Visualisierung der codierten Segmente pro Text

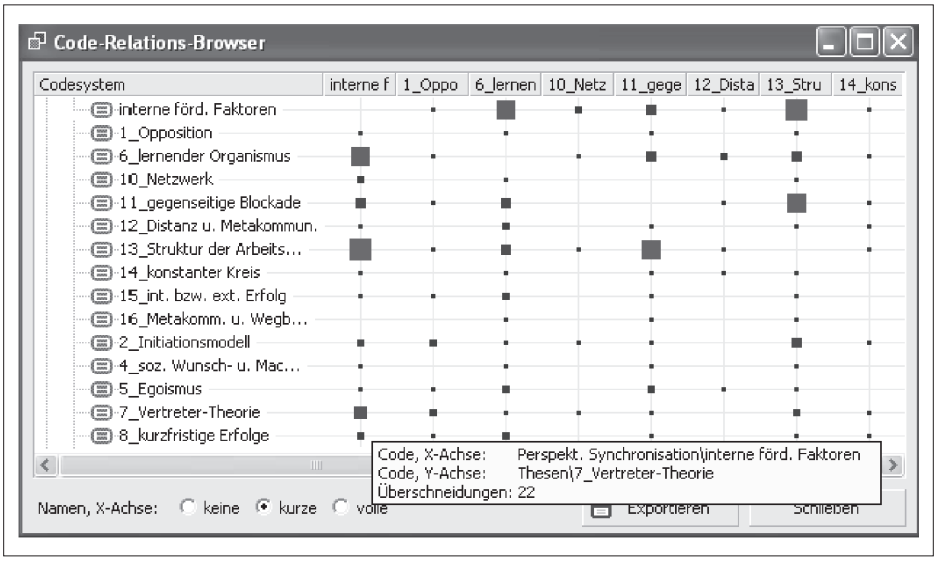


Im Auswertungsprozess erweist sich eine Visualisierung der vorgenommenen Codierungen, wie sie in MAXqda in Form des Code-Matrix-Browser verfügbar ist, als sehr hilfreich. Auf einen Blick lässt sich ersehen, welche Personen zu welchen Themen viele

bzw. nur wenige Codierungen aufweisen. In *Abbildung 2* werden die Spalten durch die Personen und die Zeilen durch die Codes bzw. Subcodes gebildet. Die Größe der Knoten in den einzelnen Zellen der Matrix entspricht der relativen Häufigkeit der entsprechenden Codierungen. Im Text „koep5" ist beispielsweise sehr viel über „interne fördernde Faktoren" die Rede, während darüber in den Texten „koep2" und „koep4" kaum etwas zu finden ist. Wird die Maus über den entsprechenden Knoten der Matrix bewegt, wird die exakte Information über die Anzahl der codierten Segmente angezeigt. Ein Doppelklick bewirkt, dass diese Segmente in einem Ergebnisfenster zusammengestellt werden und für eine weitergehende Inspektion zur Verfügung stehen. Die visuelle Darstellung ist vor allem für die Exploration sehr nützlich und unterstützt eine systematische Analyse des Datenmaterials – so lassen sich recht einfach ausgewählte Personen zum Vergleich nebeneinander anordnen. Für weitergehende statistische Analysen kann eine Matrix Texte*Codes erstellt werden, die die Zahlen der jeweils vorhandenen Textsegmente enthält. Diese Matrix entspricht in ihrem Aufbau einer SPSS-Datei, bei der die Fälle in den Zeilen und die Variablen in den Spalten dargestellt werden. Sie ist Ausgangspunkt für die in *Phase V* vorgenommene *Klassifikation.* In dieser Phase geht es darum, die Strukturen und Zusammenhänge zwischen Kategorien zu ermitteln. Dazu dient einerseits die Interpretation und systematische Suche nach dem gleichzeitigen Vorkommen von Kategorien und die Identifikation spezifischer Muster und Konfigurationen.

Die Darstellung einfacher Überschneidungen von Kategorien und das Auffinden der jeweiligen Textstellen lassen sich computerunterstützt leicht realisieren. *Abbildung 3* zeigt eine Darstellung des Code-Relations-Browsers von MAXqda, in der die Überschneidungen des Codes „intern fördernde Faktoren" mit anderen Codes dargestellt sind. Je größer der Knoten, desto häufiger finden sich Überschneidungen im Datenmaterial.

Abbildung 3: Visualisierung des gleichzeitigen Vorkommens von Codes



In diesem Forschungsbeispiel wurde nicht nur nach einfachen Überschneidungen von Codes gesucht, sondern die Matrix der Codierungen (Texte*Codes) wurde einer formalisierten statistischen Analyse, nämlich einer Faktorenanalyse, unterzogen. Diese Matrix entspricht im Prinzip der Ergebnistabelle einer klassischen quantitativen Inhaltsanalyse, wie sie auch mit einer Inhaltsanalyse-Software (z.B. Textpack) erstellt werden kann. Dort ist die Matrix Resultat eines hoch reliablen, aber möglicherweise wenig validen automatischen Codiervorgangs. Hier basiert die Matrix auf intellektueller Codierung. Die Ergebnisse der Faktorenanalyse ermöglichten uns, die Identifikation von Orientierungsmustern, z.B. „Das Leitbild des starken persönlichen Engagements" oder „Das Leitbild der Lokalen Agenda 21-Initiativen als große, von Fachkräften angeleitete Gemeinschaften" (vgl. de Haan u.a. 2000: 102ff.). Im letzten Auswertungsschritt erfolgt dann erneut ein Zugriff auf die Texte, indem die Textstellen solcher Personen gebündelt werden, die das entsprechende Leitbild in besonders ausgeprägter Weise verkörpern.

In anderen Forschungsprojekten hat es sich als nützlich erwiesen, das formalisierte Verfahren der Clusteranalyse anzuwenden, um eine Gruppierung der Befragten in homogene Subgruppen zu erreichen. Gegenüber der analytischen Vorgehensweise, wie sie etwa in der quantitativen Inhaltsanalyse stattfindet, lassen sich für den hier dargestellten Analysegang drei wesentliche Unterschiede hervorheben:

1. Die Codierung des Materials erfolgt hier stärker im Sinne einer konzeptuellen Durchdringung, wie sie beispielsweise auch in der Grounded Theory angestrebt wird. Die Codierung einer Textstelle macht den Text keineswegs überflüssig.
2. Exploration spielt in jeder Phase der Datenauswertung eine Rolle, QDA-Software bietet die Möglichkeiten, lexikalische Suche in den Texten und den codierten Segmenten durchzuführen und Fundstellen im Kontext zu sehen. Wir stellten beispielsweise fest, dass manche Textstellen darauf hinwiesen, dass bei den Akteuren der Initiative möglicherweise doch ein parteipolitisches Engagement eine Rolle spielt. Nach Eingabe einer Liste von einschlägigen Suchworten (SPD, Grüne, Partei etc.) und anschließender Suche in den Texten ließen sich immerhin 243 Textstellen finden, die zur Überprüfung inspiziert werden konnten.
3. Bis zuletzt werden Textstellen systematisiert und interpretiert, Typen gebildet und mit Primärmaterial „verständlich" gemacht. Das Endresultat stellen keine statistischen Koeffizienten, etwa Korrelationen zwischen Kategorien, oder statistische Zusammenhangsmodelle zwischen Kategorien dar. Auch dort, wo statistisch Verfahren zu Hilfe genommen wurden, geschieht dies mit Hinblick auf eine bessere Gruppierung von Daten und Fällen und nicht als finaler Zweck der Analyse.

IV. Forschungsstand und Probleme

1. Die computerunterstützte qualitative Analyse in der Methodendiskussion

In der Literatur sind Arbeiten über *methodische Aspekte* und Probleme von QDA-Software weitaus seltener anzutreffen als Leistungsbeschreibungen und -vergleiche. In den letzten Jahren hat die Methodendiskussion aber erheblich zugenommen (vgl. z.B. Cof-

fey et al. 1996; Kelle 1997a, 1997b; Lee und Fielding 1996; Gibbs et al. 2002; Glaser 2002; Mruck 2000; Fielding und Lee 2002; Weitzman 2000). Gibbs hat den methodischen Gewinn von QDA-Software in vier kurzen Schlagworten zusammengefasst: „more accurate, reliable, more transparent, easier" (Gibbs 2002: 10). Insgesamt erweist sich die Diskussion über die methodischen Aspekte von QDA-Software als thematisch breit und detailreich. Es lassen sich folgende Hauptthemen der Diskussion unterscheiden:

Effektivierung. Was wird durch den Einsatz von QDA-Software erreicht und welche Vorteile und Nachteile sind damit verbunden? Mehrheitsmeinung ist, dass das qualitative Forschen durch Computereinsatz erleichtert wird: Die Textdaten lassen sich besser organisieren, sind schneller zugänglich und es lassen sich prinzipiell mehr Daten verarbeiten (Flick 2002: 365; Kelle 2000; Gibbs et al. 2002). Es erheben sich auch einige kritische Stimmen (z.B. Mruck 2000; Laucken 2002), die demgegenüber eher den kreativen, einer Kunstlehre ähnlichen Charakter qualitativer Forschung betonen und zu bedenken geben, dass es beispielsweise nicht auf Zahl und Umfang der analysierten Texte, sondern auf die Tiefe der Analyse ankomme.

Transparenz, Dokumentation und Glaubwürdigkeit. Durch computergestützte Verfahren wird eine weit größere Transparenz als bei manuellen Vorgehensweisen erzielt. Es ist ein Zuwachs an interner Validität zu verzeichnen (Kelle und Laurie 1995). Die computergestützte Vorgehensweise lässt sich sehr gut dokumentieren und dadurch steigt auch die Glaubwürdigkeit qualitativer Forschung, die bisher vor allem deshalb von Anhängern einer strikt quantitativen Methodik in Zweifel gezogen wird, weil die Vorgehensweise und die Selektionsmechanismen nicht nachvollziehbar seien und ein gehöriges Maß an subjektiver Willkür des Forschers implizierten. Dieser Diskursstrang um den Gewinn an Konsistenz und Konsequenz (vgl. Seale 1999) geht unmittelbar in die Frage eines durch Einsatz von QDA-Software veränderten Forschungsprozesses über, die weiter unten noch Thema sein wird. Manche Autoren stellen den Prestige- und Reputationsgewinn in den Vordergrund, „entgegen den Vorwürfen des bloß Subjektivistischen und Essayistischen qualitativer Sozialforschung" (Mruck 2000: 29).

Teamarbeit. QDA-Software erleichtert es, im Team zusammenzuarbeiten (Gibbs et al. 2002; Ford et al. 2000). Das Zustandekommen von Kategorien und Codierungen ist leicht nachvollziehbar, der Codierungsprozess kann anders, z.B. als „disjunktive Gruppentechnik" (Kühn und Witzel 2000), organisiert werden. Durch Memos kann der interne Kommunikationsprozess zwischen den Mitgliedern der Forschergruppe intensiviert werden. Die Daten und Analyseergebnisse können zwischen u.U. weit entfernt voneinander arbeitenden Teammitgliedern ausgetauscht werden, damit werden andere Organisationsformen von qualitativen Projekten ermöglicht. Auch dieser Diskussionsstrang beinhaltet Aspekte, die nicht nur die Methoden, sondern den Forschungsprozess und seine Organisation tangieren.

Archivierung und Sekundäranalyse. Anders als in der quantitativen Forschung sind in der qualitativen Forschung Sekundäranalysen bisher weitgehend unbekannt. In den letzten Jahren ist eine rege Diskussion entstanden, inwieweit hier durch QDA-Software Abhilfe geschaffen werden kann. In Deutschland wurde vor allem im Rahmen des Bre-

mer Sonderforschungsbereiches 186 „Statuspassagen und Risikolagen im Lebenslauf" die Frage der Archivierung digitalisierter qualitativer Daten diskutiert. In diesem Kontext sind praktische Vorschläge entwickelt worden (z.B. Corti et al. 2000). Auch international sind entsprechende Diskussionen und Aktivitäten zu verzeichnen, z.B. wurde in Essex das Qualidata-Archiv eingerichtet, das qualitative Daten archiviert und für weitere Lehre und Forschung zur Verfügung stellt (Corti et al. 2000). Die Standards zur Archivierung digitalisierter qualitativer Daten sind allerdings noch Gegenstand der Diskussion (Carmichael 2002; Kuckartz 1997b; Muhr 2000). Nicht abschließend gelöst ist auch das Problem der Anonymisierung, das ungleich komplizierter ist als bei quantitativen Erhebungen.

2. Neue Methoden zur Analyse qualitativer Daten

Bei der Diskussion um neue Methoden sind zwei Richtungen zu unterscheiden: Auf der einen Seite dreht sich die Diskussion um neue Methoden, die eng an spezielle Theorien bzw. theoretische Ansätze gebunden sind wie beispielsweise die „Contextual Analysis" (Svederberg 2001) oder das Gabek Verfahren (vgl. Buber und Zelger 2000), auf der anderen Seite geht es um eher universell verwendbare Verfahren wie etwa die Abbildung von Zusammenhängen in *Netzwerkdarstellungen* (Muhr 1994), das *Testen von formalisierten Hypothesen* (Hesse-Biber et al. 1991; Huber 1992; Kelle 1997a, 1997b), die *logische Minimierung* (Huber 1992, 1997; Ragin 1987), die *Typenbildung* (Kelle und Kluge 1999; Kluge 1999; Kuckartz 1999) oder komplexe Verfahren des Text-Retrievals (z.B. „proximity search" und die Suche nach „co-occurences" vgl. Weitzman 2000). Einige methodische Neuerungen sind Gegenstand heftiger Kontroversen. Dies gilt prototypisch für die Methode des Hypothesentestens, des sogenannten „Theory Building" (vgl. Kelle 1995, 1997a, 1997b; Mangabeira 1995; Richards und Richards 1994). Diese Methode ist u.a. in den Ansätzen von Hesse-Biber und Huber (vgl. Hesse-Biber et al. 1991; Huber 1997) und in den entsprechenden Programmen Hyper-Research und AQUAD zu finden. Im Kern geht es hier um die Überprüfung von Hypothesen, die in formalisierter Form als logische Verknüpfung von Codes formuliert werden. Dieses stark deterministisch geprägte Verfahren hat viele Kritiker auf den Plan gerufen. Kelle (1997a; b) kritisiert etwa, dass der Hypothesenbegriff sehr eigenwillig verwendet werde und ein solches Verfahren selbstverständlich auch an bestimmte Voraussetzungen der Codes geknüpft sein müsse. Die Codes sollten distinkt sein, d.h. sich wechselseitig ausschließen, und die Reliabilität des Codiervorgangs müsse sicher gestellt sein (Kelle 1997a: 4). Diese Anforderungen korrespondieren aber, so Kelle, nicht mit dem Konzept der Grounded Theory, derzufolge es sich bei theoretischen Codes um abstrakte Konzepte handelt. Zudem bleibe es beim Hypothesentesten sensu Hesse-Biber unklar, wie man sich gegen die Zufälligkeit von Ergebnissen absichern könne. Kelle gesteht diesem Verfahren nur einen explorativen Nutzen zu, es könne als „heuristic device" genutzt werden (Kelle 1997a: 5), allenfalls bei Codes, die Fakten in den Texten bezeichnen, sei ein solches „Testen" begründbar. Strikte Regeln des Theorie Testens auf „fuzzy codes" anzuwenden, müsse notwendigerweise zu Artefakten als Resultate führen.

3. Integration qualitativer und quantitativer Methoden

Eine Vielzahl von Arbeiten ist der Integration von quantitativen und qualitativen Methoden und der Triangulation gewidmet (Flick 2002; Kuckartz 1999; Mayring 2001; Kluge 1999; Kelle und Kluge 1999). Das auf die Arbeiten von Denzin (1989) zurückgehende Konzept der Triangulation versucht verschiedene Methoden und Untersuchungssettings miteinander zu kombinieren. Es wird durchaus verschieden ausbuchstabiert (vgl. Flick 1995: 249f.), wobei die Integration von QDA-Software bislang eher eine geringe Rolle spielt. In Bezug auf die Kombination von qualitativen und quantitativen Methoden jenseits des Triangulationskonzeptes ist die Entwicklung allerdings bereits weiter fortgeschritten. Hier werden teilweise sehr detaillierte Ablaufschemata für einen solchen Methodenmix entwickelt (Kühn und Witzel 2000; Mayring 2001; Kluge 1999; Kuckartz 1999; de Haan et al. 2000), die auch bereits praktisch in Projekten erprobt werden. Mayring stellt heraus, dass die Hinzuziehung quantitativer Analyseschritte nicht nur zu einer gesteigerten Systematisierung, sondern auch zu einer größeren Verallgemeinerbarkeit der Ergebnisse führe (Mayring 2001). Eine besondere Bedeutung haben im Rahmen der Integration von quantitativer und qualitativer Methodik Ansätze zur Typenbildung (Kelle und Kluge 1999; Kluge 1999; Kuckartz 1999). Eine beachtliche Zahl von praktischen Arbeiten, in denen mit teilweise sehr unterschiedlicher Kombination von qualitativen und quantitativen Verfahren gearbeitet wurde, liegt vor, nicht nur aus dem Bereich der Sozialwissenschaften, sondern beispielsweise auch aus der Publizistik.

4. Probleme: Die Effekte von QDA-Software

Bei der qualitativen Datenanalyse spielt der Computer bzw. entsprechende Analysesoftware eine Rolle, die sich erheblich von der Rolle bei der quantitativ-statistischen Analyse unterscheidet. Bei letzterer ist es der Computer, der die Analyse, z.B. in Form einer Varianz- oder Faktorenanalyse, durchführt. Die eigentliche Analyse besteht in einem mathematischen Kalkül. Dem Forscher obliegt die Aufgabe, die Resultate, d.h. die errechneten Parameter und Koeffizienten, zu interpretieren. Bei der computergestützten Analyse qualitativer Daten ist es hingegen nicht der Computer, der die Daten automatisch analysiert und Assoziationsmaße liefert, sondern es ist weiterhin der Forscher bzw. die Forscherin. Die QDA-Software erbringt vornehmlich eine *Unterstützungs- und Systematisierungsleistung,* verbunden mit Zeitgewinn. Es hat relativ lange gedauert, bis sich die Computernutzung auch in der qualitativen Forschung etabliert hat. Dies scheint nun aber mehr und mehr der Fall zu sein und so rechnen auch jene Experten, die bisher keineswegs zu den Protagonisten einer Computerisierung qualitativer Forschung zählten, mit einem schwerwiegenden Veränderungsprozess (Flick 2002: 362). Es stellt sich somit die Frage, welche Effekte der Einsatz von QDA-Software auf die Entwicklung der qualitativen Forschung selbst besitzt. Dabei geht es zum einen um die Steigerung der Qualität der qualitativen Forschung durch die Computerunterstützung, zum anderen um die Effekte der QDA-Software auf den Forschungsprozess und auf das Analyseverhalten der Forscher.

5. Effekte auf die Qualität qualitativer Forschung

In den letzten Jahren ist eine generelle Diskussion um Qualität und Qualitätskriterien qualitativer Forschung in Gang gekommen, allerdings vornehmlich im englischsprachigen Raum. Seale und Silverman hatten 1997 mit einem unter dem Titel „Ensuring rigour in qualitative research" publizierten Artikel den Weg zu mehr Systematik und methodischer Strenge vorgezeichnet. Vor allem Seale hat durch seine weiteren Beiträge (insbes. Seale 1999) die Diskussion voran getrieben und mit dem Konzept des „subtilen Realismus" für ein Qualitätskonzept plädiert, das einerseits vom klassischen Objektivitätsbegriff quantitativer Forschung kritisch-rationalistischer Prägung abrückt, andererseits aber auch ein radikal-konstruktivistisches bzw. postmodernes Weltbild zurückweist. Seale plädiert deshalb für eine *systematische* Suche nach Evidenz und Gegenevidenz und für den Fallibilismus, d.h. die Suche nach Falsifizierendem statt nach Bestätigendem. Hier werden pragmatische Gesichtspunkte aufgenommen, die sich bereits bei Hammersley (1992), einem kritischen Betrachter der Qualität ethnographischer Methodik, finden.

Im deutschsprachigen Raum haben Kelle (1995), Flick et al. (2000) und Flick (2002) wichtige Beiträge geliefert. Kelle plädiert für das aus der Grounded Theory entlehnte Konzept der „constant comparative method", d.h. für eine ständige (möglichst maximale) Kontrastierung von Fällen. Dies steht dem Sealeschen Fallibilismus recht nahe. Flick entfaltet ein Konzept der Triangulation und empfiehlt ein am Total Quality Management orientiertes prozessbegleitendes Qualitätsmanagement (Flick 2002).

Im Rahmen der Qualitätsdiskussion werden von vielen Autoren (so Kelle 1995; Richards und Richards 1994; Welsh 2002; Kuckartz 1999; Mruck 2000) auch die potentiellen Qualitätszuwächse durch QDA-Software diskutiert. Mit Hilfe von Methoden computergestützter Analyse seien folgende Qualitätszuwächse zu erzielen:

- Management von größeren Stichproben und damit von größeren Textmengen (Kelle und Laurie 1995; Webb 1999);
- schnellere Bewältigung von redundanten, nicht kreativen Aufgaben (Fielding und Lee 1991; Moseley et al. 1997);
- Steigerung der Transparenz und Nachvollziehbarkeit der Analyse;
- effizientere und einfachere Vornahme von Datenreduktion, Speichern der Daten und Datenmanagement (Kelle 1995, 1997a; b);
- größere Nähe zu den Daten durch die umfangreichen Funktionen des Text-Retrievals und die jederzeitige Möglichkeit zur Re-Kontextualisierung (vgl. Creswell und Maietta 2002; Weitzman und Miles 1995);
- erweiterte komplexe qualitative Analysen, etwa formalisierte Hypothesentests (Tesch 1991; Fielding und Lee 1991).

Bislang nur vereinzelt vorliegende Studien über die Forschungspraxis zeigen allerdings, dass aus potentiellem Qualitätszuwachs nicht unbedingt auch tatsächlicher Qualitätszuwachs wird. Seale analysierte Forschungsarbeiten, die QDA-Software nutzen, und kam zu dem Ergebnis, dass die Programme meistens nur in einfacher Form genutzt würden und die komplexeren Programmfeatures überhaupt nicht zum Einsatz kämen. Grunenberg (2001) kam für deutschsprachige Forschungsarbeiten zu dem Ergebnis, dass nur in weniger als 5 Prozent von qualitativen Publikationen mit QDA-Software gearbeitet

worden war. Diese empirischen Arbeiten geben aber nur erste Hinweise und können nicht ohne weiteres verallgemeinert werden. Es bedarf noch erheblicher Forschungsarbeit um zu klären, ob bzw. unter welchen Bedingungen der Einsatz von QDA-Software tatsächlich zu einem Qualitätszuwachs führt.

6. Effekte auf den qualitativen Forschungsprozess

In Bezug auf die Anwenderseite ist die Forschung bislang noch recht dürftig (z.B. Fielding und Lee 1998). In den USA hatte bereits in den frühen 1980er Jahren eine Diskussion über den Computereinsatz in der qualitativen Sozialforschung und seine Konsequenzen stattgefunden (z.B. in den Zeitschriften „Computer and the humanities", „Qualitative Sociology" und „Human Organizations"). Bereits 1987 führten Brent et al. eine Erhebung durch, die den Grad und die Art der Computernutzung durch qualitativ Forschende klären sollte. Ein wichtiges Ergebnis war, dass es vor allem die jüngere Wissenschaftlergeneration war, die den Computer rege nutzte. Ähnlich sind die Resultate der Studie von Fielding und Lee (2002): Häufig sind es kleinere Projekte und Qualifikationsarbeiten, in deren Rahmen mit QDA-Software gearbeitet wird. In der überwiegenden Zahl der Fälle richten sich die Forscher nach ihren eigenen Angaben nicht nach einem bestimmten methodischen Paradigma, sondern praktizieren eine pragmatische, prozedural wenig fixierte Art der Textauswertung.

Carvajal (2002) weist auf das Missverständnis von Novizen hin, die gegenüber QDA-Software drei unrealistische Erwartung hegen würden: Erstens, dass die Software die Daten mehr oder weniger automatisch auswerten würde. Zweitens, dass sich die Analysezeit erheblich reduzieren würde und drittens, dass QDA-Programme einen Output ähnlich wie Statistik-Programme erzeugen würden.

Fielding und Lee, Initiatoren des an der University of Surrey angesiedelten CAQDAS Projektes[5], führten Gruppendiskussionen mit Anwendern in England durch. Ein bemerkenswertes Resultat ihrer Studie ist die Entdeckung, dass es sich bei den Nutzern von QDA-Software in den meisten Fällen um Neulinge innerhalb der qualitativen Forschung handelt, d.h. um Personen, die zuvor nicht mit herkömmlichen, nicht-elektronischen Mitteln qualitativ geforscht haben. Sie sind Novizen in diesem Feld und haben sich nicht zuletzt durch die Möglichkeit computergestützter Analyse für diese Methodik interessiert. Dies lenkt den Blick darauf, dass die Frage, wie der qualitative Forschungsprozess durch Computereinsatz verändert wird, nicht darauf verengt werden kann, welchen Veränderungen der Forschungsalltag vormals handwerklich arbeitender Wissenschaftler ausgesetzt ist. QDA-Software interessiert offenbar neue Personenkreise für qualitative Methoden und trägt auf diese Weise zu deren wachsender Popularität bei. Mit der Ausweitung des Anwenderspektrums sind auch neue Impulse

5 Das Projekt CAQDAS („Computer Assisted Qualitative Data Analysis Software") ist ein vom UK Economic and Social Research Council gefördertes Projekt, das vor allem durch Workshops und Lehrveranstaltungen die praktischen Fähigkeiten im Umgang mit QDA-Software englandweit fördert. Ferner stellt das Projekt verschiedene Plattformen zum Dialog über die Software und deren Implikationen und Folgen zur Verfügung. Hierzu gehört unter anderem die Mailingliste „Qual-Software".

für die methodologischen Diskurse der Qualitativen Datenerhebung und -analyse zu erwarten.

Die Protagonisten von QDA-Software betonen primär die Qualitätsverbesserung und die neuen analytischen Möglichkeiten (Gibbs et al. 2002; Morison und Moir 1998; Richards und Richards 1994) und verweisen darauf, dass die Software die Nutzer keineswegs auf ein bestimmtes methodisches Paradigma festlege (vgl. Weitzman 2000: 803). QDA-Software habe den Charakter eines Werkzeugs, man könne sie so oder so benutzen, ja sogar als Literaturverwaltungsprogramm zweckentfremden. Die Kritiker bezweifeln den reklamierten Qualitätszuwachs zunächst auch gar nicht und befassen sich nicht etwa mit der Widerlegung des Arguments der Qualitätsverbesserung auf der Mikroebene. Sie wechseln mit ihrer Argumentation auf eine andere Ebene, nämlich auf die einer Art Gesamtsicht qualitativer Forschung und bemängeln eine Verengung oder auf die Dauer drohende Verengung des Methodenspektrums (prototypisch Coffey et al. 1996; vorher bereits Seidel 1991; Agar 1991). Das Argument, dass die Software kein bestimmtes Paradigma vorschreibe, lassen sie nicht gelten. Auch vorhandene Beispiele, QDA-Software für andere Interpretationsansätze jenseits der Grounded Theory einzusetzen (so etwa für die Konversationsanalyse bei Ten Have 1999), überzeugen die Kritiker nicht. Der Charakter eines neutralen Werkzeugs sei nur bei oberflächlicher Betrachtung gegeben, in Wirklichkeit beinhalte QDA-Software eine Art „hidden curriculum", das die Forscher, von ihnen selbst weitgehend unbemerkt, in eine bestimmte Richtung führe (Seidel 1991; Coffey et al. 1996; Brown 2002). Man könne nicht bestreiten, dass die Software manches sehr gut und manches hingegen eher schlecht könne. Vor allem sehen die Kritiker die Gefahr, dass die Software zur quantitativen Analyse, einer der interpretativen Sozialforschung zutiefst fremden Vorgehensweise, ermutige. Auch werde eine Form qualitativer Datenanalyse bevorzugt, die auf Codierung abstellt, während sequenzanalytische Vorgehensweisen weniger gut zur Logik der QDA-Software passen würden. Umgekehrt befördere die Stärke der QDA-Software eine Inflation von Codes, hierarchischen Baumstrukturen und Codierungen (vgl. Seale 1999).

Diese tatsächlichen oder vermeintlichen Gefahren des Arbeitens mit QDA-Software werden relativ häufig diskutiert (Laucken 2002; Glaser 2002; Coffey et al. 1996; Kelle 1997a, b; Lee und Fielding 1996). In besonderer Weise richtet sich dabei das Augenmerk auf die analytische Technik des Codierens. Fielding und Lee (1998: 119) warnen davor, dass die extensive Nutzung von QDA-Software dazu führen kann, dass das Codieren nicht mehr die Analyse unterstütze, sondern diese gewissermaßen ersetze. Auch drohe die Gefahr, dass sich durch die Zwischenschaltung des Computers die Distanz zu den Daten vergrößere. Zudem befördere der Codierungsprozess die Suggestion, dass die Bedeutung gewissermaßen außerhalb des Textes in den Codierungen liege. So könne das Codieren tatsächlich dazu führen, dass man durch die De-Kontextualisierung, die mit dem „Cut-and-Paste" einhergeht, das eigentliche Phänomen aus den Augen verliere. Während solche Warnungen noch auf dem Hintergrund einer eigentlich technikfreundlichen Grundstimmung erfolgen, setzt die Kritik von Glaser (2002) sowie von Roberts und Wilson (2002) fundamentaler an. Roberts und Wilson sehen prinzipielle Gegensätze zwischen Computern und qualitativer Forschung: „Computer techniques of logic and precise rules are not compatible with the unstructured, ambiguos na-

ture of qualitative data and so it may distort or weaken data or stifle creativity" (ebd.: 15). Für Kritik dieses Typs gilt der gesamte Vorgang der computergestützten Analyse als „abwegig", weil er unnötig viel Zeit binden und damit von der eigentlichen qualitativen Analyse abzweigen würde.

Durch diese Argumentation schimmert, wie Fielding und Lee schon 1991 bemerkten, eine Art Frankenstein-Mythos durch: Der Computer als eine Art außer Kontrolle geratenes menschliches Geschöpf, das zum den Erzeuger verschlingenden Monster mutiert. Viele Autoren beschwören eine drohende disziplinübergreifende Homogenität der Methoden herauf (Barry 1998; Hinchcliffe et al. 1997), die sie vor allem in der zunehmenden Beliebtheit des Grounded Theory-Ansatzes zu erkennen glauben. Coffey et al. 1996 befürchten einen zunehmend homogenisierten Ansatz, einen „gold standard" qualitativer Forschung. Fielding und Lee setzen dagegen, dass die Grounded Theory zwar an Beliebtheit gewonnen habe, dass aber nach wie vor eine Mehrheit qualitativ Forschender mit anderen Ansätzen arbeiten würden. Kelle weist zudem darauf hin, dass die Bezugnahme auf die Grounded Theory, wie sie etwa in Forschungsanträgen immer öfter vorzufinden sei, auch Resultat eines vorherrschenden Legitimationszwangs für das in Projekten angewendete methodische Vorgehen sein könne.

V. Perspektiven

Im letzten Jahrzehnt ist die Entwicklung computergestützter Techniken zur Analyse qualitativer Daten mit rasantem Tempo verlaufen. Dies wird vermutlich auch in der näheren Zukunft noch andauern und einerseits zu einer Steigerung von Funktionsumfang und Bedienerfreundlichkeit der einschlägigen Software führen und andererseits zu einem wachsenden Kreis von Nutzern beitragen. Deshalb dürfte Flicks Prognose „Computer werden sich in der qualitativen Forschung – wie sonst auch – durchsetzen" (Flick 2002: 378) wohl realistisch sein. Soweit sich das heute absehen lässt, werden die zukünftigen Entwicklungen der Software wahrscheinlich in folgenden Schwerpunkten stattfinden:

– Steigerung der *Visualisierungsmöglichkeiten* auch hinsichtlich des Zusammenhangs von Kategorien;
– Verbesserung der *Teamworkfunktionen* – heute handelt es sich bei QDA-Software fast immer um Einzelplatzsysteme, die kein simultanes Arbeiten der Projektmitarbeiter an den Texten erlauben (vgl. die Perspektiven bei Weitzman);
– Zunehmende Integration qualitativer und quantitativer Analysetechniken.

Die Nutzung von Computerunterstützung wird zukünftig für qualitativ Forschende vermutlich zum Alltag gehören, so wie heute quantitative Forscher sich kaum mehr einen Analysealltag ohne statistische Softwarepakete vorstellen mögen. Was aber trotz wachsender Computernutzung bleiben wird, ist die Vielfalt qualitativer Methoden und Ansätze (vgl. etwa die Systematisierung bei Tesch 1990), die es unwahrscheinlich erscheinen lässt, dass sich ein „best program" etablieren wird, das für alle unterschiedlichen Ansätze in gleicher Weise optimal ist.

Was vermutlich ebenfalls bleiben wird, sind qualitative Methoden, die sich der Computerisierbarkeit entziehen. Es lässt sich schwerlich bestreiten, dass es vor allem

die zusammenfassenden und systematisierenden Techniken sind, die durch die Computerunterstützung befördert werden, während hermeneutische und sequenzanalytische Verfahren vergleichsweise weniger profitieren. Insgesamt erscheint es so, als wachse mit den computerunterstützten Methoden zur Analyse qualitativer Daten ein neues Arsenal von sozialwissenschaftlichen Methoden zusammen, für das die Denkweise der strikten Trennung von qualitativen und quantitativen Verfahren nicht mehr bestimmend ist. Die neue Form methodischer Kontrolliertheit und Nachvollziehbarkeit kann für die vormals eher auf anekdotische Evidenz angewiesenen qualitativen Analyseverfahren einen nicht zu unterschätzenden Gewinn darstellen.

Literatur

Agar, Michael, 1991: Microcomputers as Field Tools. Computers and the Humanities 17: 19–26.

Barry, Christine A., 1998: Choosing Qualitative Data Analysis Software: Atlas/ti and Nudist Compared. Sociological Research Online [On-line Journal], 3(3). Verfügbar über: http://www.socresonline.org.uk/ [Zugriff: 12.08.2002].

Berelson, Bernard, 1952: Content Analysis in Communication Research. Gencoe, Ill.: The Free Press.

Brent, Edward, James Scott und *John Spencer,* 1987: The Use of Computers among Qualitative Sociologists. Qualitative Sociology 10(3): 309–313.

Brown, David, 2002: Going Digital and Staying Qualitative: Some Alternative Strategies for Digitizing the Qualitative Research Process [29 paragraphs]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Buber, Renate, und *Josef Zelger* (Hg.), 2000: GABEK II. Zur qualitativen Forschung. Innsbruck/Wien/München: Studien Verlag.

Carmichael, Patrick, 2002: Extensible Markup Language and Qualitative Data Analysis [39 paragraphs]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Carvajal, Diógenes, 2002: The Artisan's Tools. Critical Issues When Teaching and Learning CAQDAS [46 paragraphs]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Coffey, Amanda, et al., 1996: Qualitative Data Analysis: Technologies and Representations. Sociological Research Online 1: 1.

Corti, Louise, 2002: QUALIDATA. Qualitative Data Archival Resource Centre. Verfügbar über: http://www.qualidata.essex.ac.uk [Zugriff: 31.07.2002].

Corti, Louise, Susann Kluge, Katja Mruck und *Diane Opitz* (Hg.), 2000: Text. Archiv. Re-Analyse *Forum* Qualitative Sozialforschung/Forum: Qualitative Social Research [On-line Journal] 1: Nr. 3, Dezember 2000.

Corti, Louise, und *Paul Thompson,* 1998: Are You Sitting on Your Qualitative Data? Qualidata's Mission. Int. J. Social Research Methodology 1(1): 85–89.

Creswell, John W., und *Ray C. Maietta*, 2002: Qualitative Research. In: *C. Delbert, Neil J. Miller* und *J. Salkind* (Hg.), Handbook of Research Design and Social Measurement. 6th Edition. Thousand Oaks: Sage Publications.

De Haan, Gerhard, Udo Kuckartz und *Anke Rheingans*, 2000: Bürgerbeteiligung und Lokale Agenda 21-Initiativen. Analysen zu Kommunikations- und Organisationsformen. Opladen: Leske + Budrich.

Denzin, Norman K., 1970: The Research Act in Sociology. A Theoretical Introduction to Sociological Methods. Chicago: Aldine.

Denzin, Norman K., und *Yvonna S. Lincoln*, 2000: Handbook of Qualitative Research. 2nd Edition. Thousand Oaks: Sage Publications.

Drass, Kriss A., 1980: The Analysis of Qualitative Data: A Computer Program. Urban Life: 332–353.

Fielding, Nigel G., und *Raymond M. Lee* (Hg.), 1991: Using Computers in Qualitative Research. London: Sage Publications.

Fielding, Nigel G., und *Raymond M. Lee*, 1998: Computer Analysis and Qualitative Research. Sage Publications: London.

Fielding, Nigel G., und *Raymond M. Lee*, 2002: New Patterns in the Adoption and Use of Qualitative Software. Field Methods 14:2 (5): 197–216.

Flick, Uwe, 1995: Qualitative Forschung. Theorie, Methoden, Anwendungen in Psychologie und Sozialwissenschaften. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe, 2002: Qualitative Sozialforschung. Eine Einführung. Reinbek bei Hamburg: Rowohlt.

Flick, Uwe, Ernst von Kardorff und *Ines Steinke* (Hg.), 2000: Qualitative Forschung. Ein Handbuch. Reinbek bei Hamburg: Rowohlt.

Ford, Kate, Iddo Oberski und *Steve Higgins,* 2000: Computer-Aided Qualitative Analysis of Interview Data: Some Recommendations for Collaborative Working. The Qualitative Report [Online Journal], 4(3/4). Verfügbar über: http://www.nova.edu/ssss/QR/BackIssues/index.html [Zugriff: 12.08.2002].

Früh, Werner, 2001: Inhaltsanalyse. 5. Auflage. Konstanz: UVK Verlagsgesellschaft.

Gibbs, Graham R., 2002: Qualitative Data Analysis: Explorations with NVivo. Buckingham: Open University Press.

Gibbs, Graham R., Susanne Friese und *Wilma C. Mangabeira,* 2002: Technikeinsatz im qualitativen Forschungsprozess. Einführung zu FQS Band 3(2) [34 Absätze]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs.htm [Zugriff: 31.07.2002].

Giegler, Helmut, 1992: Zur computerunterstützten Analyse sozialwissenschaftlicher Textdaten: Quantitative und qualitative Strategien. S. 335–388 in: *Jürgen H.P. Hoffmeyer-Zlotnik* (Hg.), Analyse verbaler Daten. Ueber den Umgang mit qualitativen Daten. Opladen: Westdeutscher Verlag.

Glaser, Barney, 2002: Data Management: Computer Data Collection. Manuskript o.O.

Grunenberg, Heiko, 2001: Die Qualität qualitativer Forschung. Eine Metaanalyse erziehungs- und sozialwissenschaftlicher Forschungsarbeiten. Marburg: Diplomarbeit.

Hammersley, Martyn, 1992: What's Wrong with Ethnography? London: Routledge.

Hesse-Biber, Sharlene, Paul Dupuis und *T. Scott Kinder,* 1991: HyperRESEARCH: A Computer Program for the Analysis of Qualitative Data with an Emphasis on Hypothesis Testing and Multimedia Analysis. Qualitative Sociology 14 (4): 289–306.

Hinchcliffe, Steve, et al., 1997: Software for Qualitative Research: 2. Some Thoughts on 'aiding' Analysis. Environment and Planning A 29: 1109–1124.

Hopf, Christel, und *Elmar Weingarten* (Hg.), 1984: Qualitative Sozialforschung. Stuttgart: Klett-Cotta.

Huber, Günter L. (Hg.), 1992: Qualitative Analyse: Computereinsatz in der Sozialforschung. München: Oldenbourg.

Huber, Günter L., 1997: Analysis of Qualitative Data with Aquad Five for Windows. Schwangau: Ingeborg Huber Verlag.

Jakob, Alexander, 2001: Möglichkeiten und Grenzen der Triangulation quantitativer und qualitativer Daten am Beispiel der (Re-)Konstruktion einer Typologie erwerbsbiographischer Sicherheitskonzepte. Arbeitspapier 2 des SFB 536 Reflexive Modernisierung. München.

Kelle, Udo, 1994: Empirisch begründete Theoriebildung: zur Logik und Methodologie interpretativer Sozialforschung. Weinheim: Deutscher Studien-Verlag.

Kelle, Udo (Hg.), 1995: Computer-aided Qualitative Data Analysis. Theory, Methods and Practice. Thousand Oaks: Sage Publications.

Kelle, Udo, 1997a: Capabilities for Theory Building & Hypothesis Testing. In: Software for Computer Aided Qualitative Data Analysis. Data Archive Bulletin 65.

Kelle, Udo, 1997b: Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. Sociological Research Online 2: 2.

Kelle, Udo, 2000: Computergestützte Analyse qualitativer Daten. S. 485–501 in: *Uwe Flick, Ernst von Kardorff* und *Ines Steinke* (Hg.), Handbuch qualitativer Sozialforschung. Reinbek bei Hamburg: Rowohlt.

Kelle, Udo, und *Heather Laurie*, 1995: Computer Use in Qualitative Research and Issues of Validity. S. 19–28 in: *Udo Kelle* (Hg.), Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice. London: Sage Publications.

Kelle, Udo, und *Susann Kluge*, 1999: Vom Einzelfall zum Typus: Fallvergleich und Fallkontrastierung in der qualitativen Sozialforschung. Opladen: Leske + Budrich.

Kluge, Susann, 1999: Empirisch begründete Typenbildung. Zur Konstruktion von Typen und Typologien in der qualitativen Sozialforschung. Opladen: Leske + Budrich.

Kluge, Susann, und *Diane Opitz,* 1999: Die Archivierung qualitativer Interviewdaten: Forschungsethik und Datenschutz als Barrieren für Sekundäranalysen? Soziologie, Mitteilungsblatt der DGS, Nr. 4: 48–63.

Knorr-Cetina, Karin, 1984: Die Fabrikation von Erkenntnis. Zur Anthropologie der Naturwissenschaft. Frankfurt a.M.: Suhrkamp.

Kuckartz, Udo, 1995: Case-oriented Quantification. In: *Udo Kelle* (Hg.), Computer-Aided Qualitative Data Analysis. Theory, Methods and Practice. London: Sage.

Kuckartz, Udo, 1996: Argumentationen und Leitbilder computergestützt analysieren. Historical Social Research. Historische Sozialforschung 21, 3: 115–136.

Kuckartz, Udo, 1997a: Qualitative Daten computergestützt auswerten: Methode, Techniken, Software. S. 584–589 in: *Barbara Frieberthäuser* und *A. Prengel* (Hg.), Handbuch Qualitative Forschungsmethoden in der Erziehungswissenschaft. München: Juventa Verlag.

Kuckartz, Udo, 1997b: A Standardized Text Interchange Format for Computer Programs for Qualitative Data Analysis. S. 259–266 in: *Frank Faulbaum* (Hg.), SOFTSTAT '97. Advances in Statistical Software, The 9th Conference on the Scientific Use of Statistical Software. Stuttgart/Jena/New York: Gustav Fischer Verlag.

Kuckartz, Udo, 1999: Computergestützte Analyse qualitativer Daten. Eine praktische Einführung in Methoden und Arbeitstechniken. Opladen: Westdeutscher Verlag.

Kühn, Thomas, und *Andreas Witzel,* 2000: Der Gebrauch einer Textdatenbank im Auswertungsprozess problemzentrierter Interviews [115 Absätze]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [Online-Journal], 1(3). Verfügbar über: http://www.qualitative-research.net/fqs/fqs.htm [Zugriff: 31.07.2002].

Laucken, Uwe, 2002: Qualitätskriterien als wissenschaftspolitische Lenkinstrumente [83 Absätze]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research [On-line Journal], 3(1): Verfügbar über: http://www.qualitative-research.net/fqs-texte/1-02/1-02laucken-d.htm [Datum des Zugriffs: 20.02.2003].

Lee, Ray M., und *Nigel G. Fielding,* 1996: Qualitative Data Analysis: Representations of a Technology: A Comment on Coffey. Holbrook and Atkinson' Sociological Research Online 1: 4.

Lofland, John, und *Lynn H. Lofland,* 1984: Analyzing Social Settings: A Guide to Qualitative Observation and Analysis (2nd edition). Belmont, CA: Wadsworth.

Mangabeira, Wilma, 1995: Computer Assistance, Qualitative Analysis and Model Building. In: *M. Raymond Lee* (Hg.), Information Technology for the Social Scientist. London: UCL Press.

Marz, Lutz, und *Meinolf Dierkes,* 2000: Leitbildprägung und Leitbildgestaltung. Zum Beitrag der Technikgenese-Forschung für eine prospektive Technikfolgen-Regulierung. WZB Papers FS II 92-105. Berlin. Wissenschaftszentrum Berlin für Sozialforschung.

MAX, 1990: MAX – ein Paket zur soziologischen Analyse verbaler Daten. S. 203ff. in: *Johannes Gladitz* und *Klaus G. Troitzsch*, Computer Aided Sociological Research. Berlin.

Mayring, Philipp, 1993: Qualitative Inhaltsanalyse. Grundlagen und Techniken. 4. erweiterte Auflage. Weinheim: Deutscher Studienverlag.

Mayring, Philipp, 2001: Kombination und Integration qualitativer und quantitativer Analyse [31 Absätze]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [Online-Journal], 2(1). Verfügbar über: http://www.qualitative-research.net/fqs/fqs.htm [Zugriff: 31.07. 2002].

Merten, Klaus, 1995: Inhaltsanalyse: Einführung in Theorie, Methode und Praxis, 2. verbesserte Auflage. Opladen: Westdeutscher Verlag.

Merton, Robert K., und *Elinor Barber*, 2004: The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science. With an Introduction by James L. Shulman. Princeton, NJ: Princeton University Press.

Miles, Matthew B., und *Huberman A. Michael,* 1994: Qualitative Data Analysis. An Expanded Sourcebook. London: Sage Publications.

Morison, Moya, und *Jim Moir,* 1998: The Role of Computer Software in the Analysis of Qualitative Data: Efficient Clerk, Research Assistant or Trojan Horse? Journal of Advanced Nursing 28(1): 106–116.

Moseley, Laurence G., D.M. Mead und *F. Murphy,* 1997: Applying Lexical and Semantic Analysis to the Exploration of Free-text Data. Nurse Researcher 4(3): 46–68.

Mruck, Katja, 2000: Qualitative Sozialforschung in Deutschland [54 Absätze]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 1(1). Verfügbar über: http://qualitative-research.net/fqs [Zugriff: 31.07.2002].

Muhr, Thomas, 1994: ATLAS/ti. Ein Werkzeug für die Textinterpretation. In: *Andreas Boehm, Andreas Mengel* und *Thomas Muhr* (Hg.), Texte verstehen. Konzepte, Methoden, Werkzeuge. Konstanz: Universitätsverlag.

Muhr, Thomas, 2000: Forum: Increasing the Reusability of Qualitative Data with XML: Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Ragin, Charles C., 1987: The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies. Los Angeles: Berkely.

Richards, Tom, und *Lyn Richards*, 1994: Creativity in Social Sciences: The Computer Enhancement of Qualitative Data Analysis. S. 365–383 in: *Terry Dartnall* (Hg.), Artificial Intelligence and Creativity: An Interdisciplinary Approach. Dordrecht: Kluwer.

Roberts, Kathryn A., und *Richard W. Wilson,* 2002: ICT and the Research Process: Issues Around the Compatibility of Technology with Qualitative Data Analysis [52 paragraphs]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research [On-line Journal], *3*(2). Verfügbar über: http://www.qualitative-research.net/fqs-texte/2-02/2-02robertswilson-e.htm.

Schwartz, Howard, und *Jerry Jacobs,* 1979: Qualitative Sociology. A Method to the Madness. New York: MacMillan.

Seale, Clive, 1999: The Quality of Qualitative Research. London: Sage Publications.

Seale, Clive, und *David Silverman,* 1997: Ensuring Rigour in Qualitative Research. European Journal of Public Health 7: 379–384.

Seidel, John, 1991: Method and Madness in the Application of Computer Technology to Qualitative Data Analysis. In: *Nigel G. Fielding* und *Raymond Lee* (Hg.), Using Computers in Qualitative Research. London: Sage Publications.

Seidel, John, und *J. Clark,* 1984: The Ethnograph: A Computer Programme for the Analysis of Equalitative Data. Qualitative Sociology 7: 110–125.

Strauss, Anselm L., 1994: Grundlagen qualitativer Forschung. Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung. München: Fink.

Strauss, Anselm L., und *Juliet M. Corbin*, 1996: Grounded Theory. Grundlagen qualitativer Sozialforschung. Weinheim: Beltz.

Svederberg, Eva, 2001: Consumers' Views Regarding Health Claims on Food Packages. Contextual Analysis by Means of Computer Support [48 paragraphs]. Forum Qualitative Sozialforschung/ Forum Qualitative Social Research [On-line Journal], 3(1). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Ten Have, Paul, 1999: Doing Conversation Analysis: A Practical Guide. Sage Publications: London.

Tesch, Renate (Hg.), 1991: Computers und Qualitative Data II, Special Issue, Part 1,2. Qualitative Sociology 14: 3, 4.

Webb, Christine, 1999: Analysing Qualitative Data: Computerised and Other Approaches. Journal of Advanced Nursing, 29(2): 323–330.

Weitzman, Eben A., 2000: Software and Qualitative Research. In: *Norman K. Denzin* und *Yvonna S. Lincoln* (Hg.), Handbook of Qualitative Research. Thousand Oaks: Sage Publications.

Weitzman, Eben A., und *Matthew B. Miles*, 1995: A Software Source Book: Computer Programs for Qualitative Data Analysis. Thousand Oaks: Sage Publications.

Welsh, Elaine, 2002: Dealing with Data: Using NVivo in the Qualitative Data Analysis Process [12 paragraphs]. Forum Qualitative Sozialforschung/Forum Qualitative Social Research [On-line Journal], 3(2). Verfügbar über: http://www.qualitative-research.net/fqs/fqs-eng.htm [Zugriff: 31.07.2002].

Witzel, Andreas, 2000: Das problemzentrierte Interview [26 Absätze]. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research [On-line Journal], 1(1). Verfügbar über: http://www.qualitative-research.net/fqs-texte/1-00/1-00witzel-d.htm [Datum des Zugriffs: Tag, Monat, Jahr].

Züll, Cornelia, und *Peter Ph. Mohler* (Hg.), 1992: Textanalyse. Anwendungen der computerunterstützten Inhaltsanalyse. Opladen: Westdeutscher Verlag.

Züll, Cornelia, und *Peter Ph. Mohler, 2001:* Computerunterstützte Inhaltsanalyse: Codierung und Analyse von Antworten auf offene Fragen. ZUMA How-to-Reihe, Nr. 8. Mannheim: ZUMA.

META-ANALYSE ALS METHODE DER SOZIALFORSCHUNG

Michael Wagner und Bernd Weiß

Zusammenfassung: In der empirischen Sozialforschung unterscheidet man gewöhnlich Primär- von Sekundäranalysen. Mittlerweile hat sich ein dritter Typus empirischer Sozialforschung etabliert, der von Glass (1976) Meta-Analyse genannt wurde. Es handelt sich hierbei um eine statistische Analyse empirischer Befunde, die aus voneinander unabhängigen Studien stammen. Meta-Analysen werden mit dem Ziel vorgenommen, publizierte empirische Befunde aus Primär- oder Sekundäranalysen zusammenzufassen und mögliche Unterschiede in den Ergebnissen verschiedener Studien aufzuklären. Der vorliegende Beitrag verfolgt drei Ziele: Erstens soll die Meta-Analyse als ein Forschungsprozess aufgefasst werden, der von der Problemstellung über die Datenerhebung bis hin zur Interpretation der Ergebnisse reicht. Zweitens wird dargestellt und durch Beispiele illustriert, welche verschiedenen Verfahren bei der Synthese von Befunden angewandt werden. Der Beitrag schließt mit einer Diskussion methodologischer Schwierigkeiten und der Frage nach dem Anwendungspotenzial für die Soziologie.

I. Meta-Analyse: Begriff[1]

In der empirischen Sozialforschung unterscheidet man gewöhnlich Primär- von Sekundäranalysen. Primäranalysen sind erstmalige Untersuchungen von Daten, die von den Forschern selbst erhoben wurden, während Sekundäranalysen Datenmaterial heranziehen, das von anderen erhoben und in der Regel bereits ausgewertet wurde. Sekundäranalysen können sich auf die ursprünglichen oder auf neue Forschungsfragen richten (Friedrichs 1973; Schnell et al. 1995: 238). In den letzten zwei bis drei Jahrzehnten hat sich nun ein dritter Typus empirischer Sozialforschung etabliert, der von Glass (1976) Meta-Analyse genannt wurde. Es handelt sich hierbei um eine statistische Analyse empirischer Befunde, die aus voneinander unabhängigen Studien stammen. Meta-Analysen werden in der Regel mit dem Ziel vorgenommen, publizierte empirische Befunde aus Primär- oder Sekundäranalysen zusammenzufassen und mögliche Unterschiede in den Ergebnissen verschiedener Studien aufzuklären. Bedeutsam ist, dass mit einer Meta-Analyse bezweckt wird, zu genaueren und sichereren Schlussfolgerungen zu gelangen als es mit einer einzelnen Studie oder einer qualitativen Zusammenfassung empirischer Befunde möglich wäre.

Die statistische Analyse publizierter Befunde ist nur ein Typ der Meta-Analyse, wenn auch derjenige, der weitaus am häufigsten vorkommt. Von einer Meta-Analyse spricht man aber auch, wenn Originaldaten verschiedener Primärstudien zusammenge-

1 Teile aus der Einleitung dieses Beitrags wurden auch in Wagner und Weiß (2004b) veröffentlicht.

fasst und ausgewertet werden („gepoolte Auswertung", vgl. Blettner et al. 1997). Mit diesem Typ der Meta-Analyse wollen wir uns im Folgenden jedoch nicht genauer befassen.

„Meta-Analyse" ist bereits zu einem Sammelbegriff für diverse Techniken zur Synthese empirischer Befunde geworden, deren Kritik und Weiterentwicklung mittlerweile einen eigenen Zweig in der Methodendiskussion verschiedener Disziplinen ausmacht. Der wohl wichtigste Grund für die rasche Entwicklung dieses Forschungsfelds ist die in zahlreichen Disziplinen beobachtbare rasante Expansion der empirischen Forschung, was einige auch als „increasingly chaotic output of contemporary research" (Hunt 1997: 1) bezeichnen. Selbst Experten eines Forschungsgebiets überblicken nicht mehr den Stand der Forschung. Diese Wissensexplosion ist in der Medizin besonders augenfällig: Heitmann (2001: 41) spricht von über 25000 biomedizinischen Zeitschriften, in denen weltweit jährlich über zwei Millionen Artikel erscheinen. Aus der Unüberschaubarkeit wissenschaftlicher Arbeiten können eine Reihe sehr nachteiliger Konsequenzen erwachsen. Erstens beruhen politische Maßnahmen, die mit einem großen finanziellen Aufwand betrieben werden, häufig auf wissenschaftlichen Erkenntnissen über die Wirksamkeit dieser Maßnahmen. Wird hier der Forschungsstand nicht oder nicht vollständig einbezogen, kann dies zu einer teuren Fehlallokation finanzieller Mittel führen. Ein zweites Beispiel betrifft die Medizin. Sind Ärzte, die ernste Krankheiten behandeln, nicht auf dem neuesten Stand des Wissens, kann dieses gravierende, wenn nicht sogar tödliche Folgen für die Patienten haben. Drittens werden finanzielle Mittel für die Forschung nicht effizient eingesetzt, wenn diese Forschungsvorhaben sich nicht an dem erreichten Forschungsniveau orientieren. Zu den wesentlichen Voraussetzungen wissenschaftlicher Erkenntnisbildung und Tätigkeit gehört die Vorstellung, dass man es hier mit einem kumulativen Prozess zu tun hat. Meta-Analysen tragen dazu bei, dass die Kumulation empirischer Forschung gezielt und systematisch vorgenommen wird (Hedges 1987).

Zwar lässt sich nicht sicher feststellen, ob in der Soziologie von Jahr zu Jahr mehr empirische Untersuchungen publiziert werden. Dennoch nimmt der Bestand an Studien, die zu einem bestimmten Thema durchgeführt wurden, kontinuierlich zu. Diese Entwicklung führt verstärkt zu Fragen nach dem erreichten Forschungsstand. Meta-Analysen sind daher in der Soziologie genauso erforderlich wie in den Wissenschaften, in denen sich dieses Verfahren bereits seit Jahren etabliert hat.

Es wird also immer notwendiger, systematische Reviews zu verfassen. Dafür genügt es auch nicht mehr, narrative oder qualitative Reviews durchzuführen, bei denen keine quantitativen Analysen empirischer Befunde erfolgen. Narrative Reviews weisen nämlich mehrere Nachteile auf: Erstens ist das Vorgehen bei solchen Reviews wenig standardisiert und damit kaum nachprüfbar. Beispielsweise sind Literaturrecherchen häufig unvollständig, vorhandene Informationen werden nicht voll genutzt oder die Gewichtung einzelner Studien wird nach subjektiven Kriterien vorgenommen. Zweitens sind narrative Reviews selten in der Lage, widersprüchliche Forschungsergebnisse zu integrieren. Drittens können sie in der Regel eine große Anzahl von Forschungsbefunden nicht mehr angemessen bearbeiten und interpretieren (Wagner und Weiß 2001; Wolf 1986: 10).

Im Vergleich zwischen qualitativen und quantitativen Reviews eine Entweder-oder-Position einzunehmen, ist jedoch weder notwendig noch produktiv. Light und Pillemer (1984: 105ff.) geben mehrere Beispiele, aus denen hervorgeht, dass quantitative Zusammenfassungen des Forschungsstands von der Hinzunahme nicht quantifizierbarer Studieninformationen profitieren. Qualitative Reviews haben dort ihren Platz, wo es um die Kritik und den systematischen Vergleich von Theorien oder von inhaltlichen Interpretationen empirischer Befunde geht. Meta-Analysen ersetzen also nicht synthetische Leistungen auf dem Gebiet der Theorieentwicklung. Ferner können qualitative Einzelfallstudien Einsichten in ungewöhnliche oder unerwartete soziale Tatbestände vermitteln, die theoretisch durchaus relevant sind. Schließlich kann es vorkommen, dass quantitative Untersuchungen qualitativ begutachtet werden müssen, bevor sie in eine Meta-Analyse einbezogen werden können. So könnte durch ein Expertenrating festgestellt werden, ob Studien bestimmte methodologische Mindeststandards einhalten.

Auch Meta-Analysen tragen zur Theorieentwicklung bei. Wie bei Primär- und Sekundäranalysen können mit Hilfe von Meta-Analysen eher beschreibende oder eher hypothesentestende Fragestellungen verfolgt werden (Light und Pillemer 1984). Die Übergänge zwischen beiden Vorgehensweisen sind fließend. Genau genommen unterliegen jeder Meta-Analyse theoretische Annahmen über die Bedeutsamkeit bestimmter Sachverhalte und Zusammenhänge, andernfalls könnte überhaupt keine systematische Auswahl von Studien und Variablen erfolgen. Aber dies ist keine spezielle Eigenschaft von Meta-Analysen, sondern charakterisiert jede Art empirischer Forschung.

Zahlreiche neuere Einführungen, Lehr- und Handbücher zur Meta-Analyse belegen die Bedeutung, die dieses Forschungsfeld mittlerweile erreicht hat (Cooper und Hedges 1994a; Hunter und Schmidt 1990; Lipsey und Wilson 2001; Rosenthal 1991; Rustenbach 2003; Schulze et al. 2003; Sutton et al. 2000). Eine Auswertung von drei sozialwissenschaftlichen Datenbanken (PsycINFO, ERIC und Social Scisearch) im Zeitraum zwischen 1974 bis 1989 durch White (1994: 43) ergab beträchtliche Zuwachsraten. So wurden in PsycINFO, einer Literaturdatenbank für psychologische Publikationen, in den Jahren von 1986 bis 1989 jährlich zwischen 120 und 160 Meta-Analysen gezählt. Anfang der 90er Jahre wurden mit Meta-Analysen in der Medizin ähnlich hohe Werte erreicht (Bausell et al. 1995). Ein weiterer Beleg für die hohe praktische Bedeutsamkeit der Meta-Analyse ist das reichhaltige Angebot an spezieller Software für die Durchführung von Meta-Analysen.

Gegenwärtig werden Meta-Analysen in der medizinischen und pharmakologischen Forschung, der Erziehungswissenschaft, der Psychologie und der Soziologie durchgeführt. Auch in den Naturwissenschaften sind seit langem Verfahren in Anwendung, die beispielsweise Befunde unterschiedlicher physikalischer oder chemischer Experimente zusammenfassen. So gehört die Agrarwissenschaft zu den Disziplinen, in denen Meta-Analysen am frühesten durchgeführt wurden (Olkin 1996). Wir beschränken uns im Folgenden auf die Medizin und die Sozialwissenschaften, was nicht bedeutet, dass es bei der Synthese empirischer Befunde prinzipielle Differenzen zwischen Sozial- und Naturwissenschaften gibt. Wie Hedges (1987) zeigt, sind die Gemeinsamkeiten zwischen Sozial- und Naturwissenschaften in dieser Hinsicht groß.

In der Medizin und Epidemiologie wurde die Notwendigkeit systematischer Reviews schon Anfang der 70er Jahre von Archibald Cochrane hervorgehoben (Sutton et al. 2000: 3). Altman (2000) datiert eine der frühesten medizinischen Meta-Analysen auf das Jahr 1980, in dem eine Studie über die Behandlung des Stotterns publiziert wurde (Andrews et al. 1980). Im Anschluss an Cochranes Arbeit begann eine Gruppe von Medizinern und Methodologen in Oxford die Ergebnisse randomisierter Studien auf dem Gebiet der Geburtshilfe zu systematisieren. Ziel dieses Projektes war es herauszufinden, für welche medizinischen Maßnahmen die beste empirische Evidenz vorliegt. Mittlerweile hat sich die „evidence based medicine" (Sackett et al. 1997) institutionalisiert. Im Oktober 1992 wurde das erste Cochrane Centre in Oxford eröffnet, ein Jahr später wurde die Cochrane Collaboration gegründet (Sauerland und Galandi 2001). Die große Bedeutung der evidence based medicine ist ein Grund dafür, dass Meta-Analysen mittlerweile zu einem Standardtyp medizinischer Publikationen geworden sind (Altman 2000).

In den Erziehungswissenschaften wurden ebenfalls zahlreiche Meta-Analysen durchgeführt. So behandelt die frühe Meta-Analyse von Glass und Smith (1979) den Zusammenhang zwischen Klassengröße und Schulleistung. Bekannt ist auch die Untersuchung zum „Pygmalioneffekt" von Raudenbush (1984), in der es beispielsweise um die Frage geht, ob Lehrer solche Schüler besonders fördern, von denen die Lehrer glauben, dass sie sich besonders positiv entwickeln werden. Ein Jahr später wurde in einer soziologischen Fachzeitschrift eine Meta-Analyse zu der Frage publiziert, ob die schulischen Leistungen von Kindern schwarzer Hautfarbe mit dem Ausmaß schulischer Segregation zusammenhängen (Wortman und Bryant 1985).

Auch in der Psychologie sind Meta-Analysen mittlerweile unverzichtbar. Bei der großen Anzahl psychologischer Meta-Analysen können wir wiederum nur einige exemplarisch herausgreifen. Eine der ersten Meta-Analysen überhaupt widmete sich der Frage nach der Effektivität von Psychotherapien (Smith und Glass 1977). Meta-Analysen zu diesem Thema wurden später wiederholt (Markus et al. 1990). Schon Anfang der 1980er Jahre wurden Meta-Analysen zur Frage durchgeführt, ob sich Männer und Frauen in kognitiven Fähigkeiten unterscheiden (Becker und Hedges 1984; Hyde 1981). Andere untersuchen Geschlechtsunterschiede in der Persönlichkeitsentwicklung oder im Sozialverhalten (Cohn 1991; Eagly und Wood 1991).

In der Soziologie etabliert sich die Meta-Analyse dagegen in kleinen Schritten. In einem kürzlich erschienenen Übersichtsartikel in der renommierten Reihe Sociological Methodology über die Entwicklung statistischer Verfahren in der Soziologie zwischen den Jahren 1950 und 2000 wird die Meta-Analyse noch mit keinem Wort erwähnt (Raftery 2001). Bislang sind wohl nur in der US-amerikanischen Soziologie eine Reihe von Meta-Analysen erstellt worden. Crain und Mahard (1983) publizierten im American Journal of Sociology eine Meta-Analyse zu den Folgen der Rassentrennung in den Schulen für die Schulleistungen (vgl. auch Wortman und Bryant 1985). Weiterhin findet sich die Meta-Analyse von Gartrell und Gartrell (1985) zum Zusammenhang von Status und technischen Innovationen in der Landwirtschaft. Auffällig ist, dass eine Reihe von Meta-Analysen familiensoziologische Probleme untersucht (Amato 2001; Amato und Gilbreth 1999; Amato und Keith 1991a; Amato und Keith 1991b; Ganong et al. 1990; Twenge et al. 2003; Vemer et al. 1989; Wells und Rankin 1991). Hinzu kom-

men Meta-Analysen von Experimenten zum Entscheidungshandeln beim Gefangenendilemma (Sally 1992) und zum Aushandeln von Kompromissen in Konfliktsituationen (Druckman 1994).

II. Durchführung von Meta-Analysen

Eine Meta-Analyse umfasst sämtliche Elemente des sozialwissenschaftlichen Forschungsprozesses wie sie auch bei einer Primärforschung vollzogen werden (Cooper 1982; Schnell et al. 1995). 1. Am Beginn stehen die Auswahl eines geeigneten Forschungsproblems sowie die genaue Eingrenzung des Forschungsgegenstandes. 2. Die Datenerhebung besteht im Fall publikationsbasierter Meta-Analysen in einer systematischen und möglichst erschöpfenden Literaturrecherche. 3. Im Anschluss daran werden die Informationen in den gesammelten Publikationen vercodet und elektronisch aufbereitet. 4. Die eigentliche (statistische) Datenanalyse besteht in der Regel aus zwei Schritten: der Befundintegration und der anschließenden Heterogenitätsanalyse. 5. Abschließend sind die Ergebnisse angemessen aufzubereiten und mit Bezug auf das Forschungsproblem zu interpretieren.

1. Problemspezifikation

Die Wahl des Forschungsgegenstandes beeinflusst unmittelbar die Form der Literaturrecherche, der Datenerhebung sowie der Datenanalyse (Hedges 1994; Lipsey und Wilson 2001). Im Gegensatz zur sozialwissenschaftlichen Primärforschung, die sich prinzipiell auf jeden (sozialen) Tatbestand beziehen kann, ist eine Meta-Analyse auf solche Themen beschränkt, die bereits Gegenstand empirischer Forschung waren (Cooper und Hedges 1994b).[2]

Jeder Forschungsgegenstand wird durch theoretische Konzepte und ihre Beziehungen zueinander strukturiert. Wie jedes andere empirische Vorhaben auch, erfordert eine Meta-Analyse die Definition dieser theoretischen Konzepte und deren Operationalisierung (Cooper 1982; Schnell et al. 1995). Allerdings werden im Rahmen von Meta-Analysen verschiedene Operationalisierungen desselben theoretischen Konstrukts berücksichtigt. Dies ist im Hinblick auf die Konstruktvalidität vorteilhaft, und Befunde, die unabhängig von der Art der Operationalisierung bestehen, sind besonders robust (Hall et al. 1994: 20; vgl. ausführlich Abschnitt III.1).

Das sichtbare Ergebnis der Problemspezifikation ist ein Regelkatalog, der es in der anschließenden Recherchephase erlaubt, relevante von irrelevanten Publikationen zu unterscheiden; zentral sind vor allem die inhaltliche Fragestellung der Studien sowie ihr methodisches Design (Lipsey und Wilson 2001). In der Meta-Analyse von Wagner und Weiß (2003a; 2004b) über Scheidungsrisiken setzt sich das Regelset wie folgt zusam-

2 Die Frage, wie viel Forschungsmaterial bereits vorliegen muss, beantworten Cooper und Hedges (1994b: 9) mit dem Hinweis, dass „methods of meta-analysis can be applied to literatures containing as few as two hypothesis tests". Inwieweit ein solches Review jedoch den Forschungsprozess befördert, sei dahingestellt (vgl. auch Hall et al. 1994).

men: 1. empirische Studien aus Deutschland zur Erklärung von Ehescheidungs- oder Trennungsrisiken, 2. individuelle Längsschnittstudien, in denen Partnerschafts- und Eheverläufe erhoben wurden, 3. vor dem Jahr 1990 wurden nur westdeutsche, danach ost- und westdeutsche Studien einbezogen. In einer Meta-Analyse von Amato und Keith (1991b), die sich mit den Konsequenzen einer Ehescheidung auf das Wohlbefinden von Kindern auseinandersetzt, nennen die Autoren vier Kriterien, die eine Studie/Publikation erfüllen muss: 1. Stichproben müssen sowohl Kinder von Alleinerziehenden als auch Kinder aus Familien mit beiden Elternteilen enthalten; 2. es muss mindestens eine Statistik zum Wohlbefinden der Kinder („well-being") vorhanden sein; 3. die berichteten Statistiken müssen sich in Effektstärken überführen lassen und 4. Untersuchungsgegenstand sind ausschließlich Kinder, keine Erwachsenen, die in ihrer Kindheit eine Ehescheidung der Eltern erlebt haben.

Es liegt auf der Hand, dass die hier getroffenen Entscheidungen insbesondere die Generalisierbarkeit der Ergebnisse („externe Validität") betreffen. Je breiter die Fragestellung gewählt wird, je weniger Bedingungen an ein bestimmtes Studiendesign gestellt werden, desto größer ist der Anwendungsbereich, den die unterschiedlichen Untersuchungen abdecken (Hall et al. 1994: 19). Umgekehrt sinkt damit aber auch die (statistische) Sicherheit, mit der Aussagen über soziale Tatbestände getroffen werden können, da die Ergebnisse im Allgemeinen heterogener werden (Rustenbach 2003, vgl. auch Abschnitt II.4.b).

2. Literaturrecherche

Bei einer publikationsbasierten Meta-Analyse besteht die Datenerhebung in einer systematischen und transparenten Literaturrecherche. Das grundsätzliche Ziel dieser Suche ist, alle im Regelkatalog aufgeführten und verfügbaren Forschungsergebnisse zu sammeln. In den meisten Fällen wird Repräsentanz nicht durch eine zufallsbasierte Stichprobenziehung von Publikationen gewährleistet, sondern durch deren Vollerhebung (Lipsey und Wilson 2001).

Um an relevante Literatur zu gelangen, lassen sich mindestens drei Suchstrategien unterscheiden: 1. die stichwortbasierte Suche in elektronischen (Fach-)Datenbanken, 2. das „Schneeballsystem", das bei einem (relevanten) Literaturtitel beginnt und systematisch alle darin enthaltenen Referenzen („footnote chasing") auf ihre Eignung prüft sowie 3. eine systematische Befragung von Fachkollegen nach geeigneten Untersuchungen. Im Prinzip bauen die drei aufgezählten Suchstrategien aufeinander auf. So lassen sich Fachdatenbanken oder das World Wide Web nach thematischen Stichworten durchsuchen. In den meisten Fällen wird die Literatur im Volltext vorliegen müssen, um entscheiden zu können, ob sie den Kriterien genügt. Diese Publikationen lassen sich im Hinblick auf bislang unentdeckte Verweise durchsuchen.

Von der Suchstrategie zu unterscheiden sind die möglichen Quellen, die entsprechende Angaben liefern können und die sich den Strategien nur teilweise eindeutig zuordnen lassen. Hier lassen sich a) (qualitative oder quantitative) Forschungsreviews, b) Literaturverweise in anderen Texten, c) Fachdatenbanken (einschließlich des World Wide Web), d) Bibliographien, e) thematisch relevante Zeitschriften, f) Konferenzpro-

gramme, g) ausgewiesene Experten sowie h) Regierungsbehörden (amtliche Gutachten oder Expertisen) unterscheiden (Lipsey und Wilson 2001).

Bei unserer Meta-Analyse der Scheidungsursachen auf europäischer Ebene (Wagner und Weiß 2003b) ergaben sich bei der Literaturrecherche zum Beispiel die folgenden Probleme: 1. Es ist häufig schwierig, die geeigneten Suchbegriffe zu bestimmen, 2. viele Datenbanken enthalten überwiegend englischsprachige Artikel und leisten damit einer Verzerrung zugunsten eines bestimmten Sprach- und Kulturraumes Vorschub, 3. für die Meta-Analyse können Artikel relevant sein, die in einer Sprache geschrieben sind, der die Forscher nicht mächtig sind, 4. Datenbanken sind unvollständig und enthalten nicht immer alle Studien, die ein Autor zu einem Thema verfasst hat.

3. Datenerfassung und -organisation

Neben den statistischen Befunden gibt es weitere Publikations- oder Studieninformationen, die im Rahmen einer Meta-Analyse von Interesse sind. In diesem Abschnitt soll kurz erläutert werden, welche Daten auf welche Art und Weise extrahiert und organisiert werden. Die Datenerfassung im Rahmen einer publikationsbasierten Meta-Analyse lässt sich in mancher Hinsicht und für einen Teil der Merkmale mit dem Vorgehen bei einer Inhaltsanalyse von Printmedien vergleichen (Merten 1995).

a) Empirische Befunde und ihre Vergleichbarkeit

Eine Meta-Analyse integriert und analysiert empirische Befunde. Empirische Befunde im klassisch meta-analytischen Sinn lassen sich als „statistical representation of one empirical relationship involving the variable(s) of interest measured on a single subject sample" (Lipsey und Wilson 2001: 35) begreifen. Dabei interessieren vor allem zwei Aspekte: die Stärke des Zusammenhangs (s.u. Effektstärke) sowie die Reliabilität des Zusammenhangs (Signifikanz) (Rosenthal 1991).

Auf Verfahren, um univariate Statistiken (Mittelwerte, Anteile) zusammenzufassen, gehen beispielsweise Lipsey und Wilson (2001) oder Rosenthal (1994) ein. Lipsey und Wilson (2001) weisen darauf hin, dass sich Meta-Analysen selten auf univariate Statistiken richten, da es in den meisten Meta-Analysen um die Synthese von Zusammenhangsmaßen geht. Darüber hinaus finden sich in der Literatur Techniken, um Befunde aus Regressionen und Faktorenanalysen zu integrieren (Fleiss 1993; Greenland 1987; Greenwald et al. 1994).

Neben der Unterscheidung nach dem Befundtypus ist für die Meta-Analyse von Belang, wie informationshaltig die empirischen Resultate dargestellt werden. 1. Im einfachsten Fall ist ein empirischer Befund lediglich eine Aussage über das Vorhandensein statistischer Signifikanz und die Richtung des Wirkzusammenhangs. Die entsprechenden Informationen lauten dann etwa „signifikant negativ", „nicht signifikant" und „signifikant positiv". Dieser Befundtyp ist die Grundlage für Verfahren des so genannten „vote countings". 2. Informationshaltiger sind Ergebnisdarstellungen, aus denen sich die entsprechenden Irrtumswahrscheinlichkeiten („p-Wert") beziehungsweise die dazu-

gehörigen Teststatistiken ableiten lassen. 3. Präzise meta-analytische Schätzer setzen jedoch die vollständige Angabe der einzelnen Statistiken voraus (Bushman 1994). Der folgende Abschnitt geht detaillierter auf die verschiedenen Befundtypen dieser dritten Gruppe ein.

In seinem Pionierbeitrag für die Meta-Analyse spricht Glass (1976) von Effektstärke („effect size") im Sinne von standardisierten („skalenfreien") Mittelwertsdifferenzen zwischen der Untersuchungs- und Kontrollgruppe. Etwa zehn Jahre später weisen Hedges und Olkin (1985) auf die gestiegene Vielfalt an (standardisierten) Statistiken hin, um Behandlungs- oder Therapieeffekte zu messen.[3] Inzwischen wird in der meta-analytischen Literatur der Begriff Effektstärke für eine ganze Reihe höchst unterschiedlicher statistischer Maßzahlen verwendet. In der Psychologie wird dabei zwischen Kennziffern der r- und der d-Familie unterschieden. In die r-Familie etwa fallen sämtliche Korrelationsmaße. Hierzu gehören beispielsweise die Produkt-Moment-Korrelation r, die punktbiseriale Korrelation r_{pb}, der Φ-Koeffizient für dichotome Merkmale sowie die Rangkorrelation ρ. Die d-Familie dagegen setzt sich aus Unterschiedsmaßen wie *Hedge's g, Glass'* Δ oder *Cohen's d* zusammen (Rosenthal 1994). Es handelt sich hierbei um standardisierte Mittelwertsdifferenzen, die jedoch anhand unterschiedlicher Varianzmaße (zusammengefasste Standardabweichung, Standardabweichung der Kontrollgruppe) standardisiert werden. Medizin und Epidemiologie erweitern diese Vielfalt noch um Risikomaße wie *odds ratio*, *relative risk* oder *risk difference* (Fleiss 1993; 1994; Sutton et al. 2000).

Eine der zentralen Voraussetzungen einer Meta-Analyse ist die Vergleichbarkeit der Befunde in statistischer wie auch inhaltlicher Hinsicht. Um die statistische Vergleichbarkeit zu gewährleisten, finden sich in der Literatur eine Vielzahl von Transformationsregeln, um unterschiedliche Befund- und Effektstärkentypen ineinander umwandeln zu können. Im Wesentlichen sind davon die verschiedenen Effektstärkentypen betroffen. Als die zentrale „Einheit" wird in den meisten Fällen die Produkt-Moment-Korrelation herangezogen (siehe dazu das Beispiel unten). Das heißt, es existieren Umrechnungsregeln, um andere Statistiken direkt oder approximativ in Produkt-Moment-Korrelationen umzurechen (Fricke und Treinies 1985: 121). Empirische Befunde der Art „signifikant/nicht signifikant", Irrtumswahrscheinlichkeiten oder Anteile sind dagegen „direkt" vergleichbar.

Es würde an dieser Stelle zu weit führen, sämtliche Umrechnungsformeln aufzulisten, gute Übersichten finden sich beispielsweise bei Fricke und Treinies (1985) oder Lipsey und Wilson (2001). Ein kurzes, fiktives Beispiel soll dennoch das Vorgehen demonstrieren: Angenommen, es liegen die Ergebnisse von drei Untersuchungen vor, in denen bei Jugendlichen der bivariate Zusammenhang von elterlichem Berufsprestige und der Häufigkeit des Schulschwänzens ermittelt wurde.[4] In einer Arbeit wird auf Grundlage des metrisch gemessenen Berufsprestiges und der Anzahl der Fehltage eine Produkt-Moment-Korrelation von $r_1 = -0{,}08$ ($N = 1630$, $p = 0{,}002$) ermittelt. In einer weiteren Arbeit wurde ein t-Test durchgeführt, wobei hier Prestige in Form eines

3 Um dieser Entwicklung Rechnung zu tragen, schlagen sie alternativ den Begriff „effect magnitude" (Hedges und Olkin 1985: XVI) vor. Dieser Begriff konnte sich allerdings nicht durchsetzen.

4 Zur Anwendung kommt hier ein und derselbe Datensatz (vgl. genauer bei Wagner et al. 2004).

zweistufigen Faktors vorlag. Berichtet wird lediglich, dass die Irrtumswahrscheinlichkeit des Testes unter 0,01 liegt und dass die Fallzahl ebenfalls 1630 beträgt. Schließlich gibt es noch eine Untersuchung, in der das Schulschwänzen anhand eines Schwellenwertes dichotomisiert wurde, gleiches gilt – wie zuvor – für die Prestigevariable. Auf Grundlage einer Vierfeldertafel wird ein χ^2 von 4,93 (df = 1) ermittelt sowie dem Text der Hinweis entnommen, dass zwischen den beiden Merkmalen ein negativer Zusammenhang besteht. Es gilt nun, die Ergebnisse der zweiten und dritten Untersuchung in Produkt-Moment-Korrelations-Äquivalente umzurechnen. Nach Rosenthal (1994) gilt für den Zusammenhang zwischen t-Wert und Produkt-Moment-Korrelation: $r = [t^2/(t^2+df)]^{1/2}$, wobei $df = N_1+N_2–2$. Auf Grundlage des Niveaus der Irrtumswahrscheinlichkeit und der Fallzahl lässt sich ein t-Wert von –2,33 ermitteln. Das geschätzte r_2 beträgt dann $r_2 = [2{,}33^2/(2{,}33^2+1630–2)]^{1/2} = 0{,}06$ beziehungsweise bei Berücksichtigung des Vorzeichens $r_2 = –0{,}06$. Für das r_3 gilt dagegen der folgende Zusammenhang (vgl. Fricke und Treinies 1985): $r = [\chi^2/(\chi^2+N)]1/2$. Für das Beispiel gilt dann: $r_3 = [4{,}93/(4{,}93+1630)]1/2 = 0{,}06$ beziehungsweise $r_3 = –0{,}06$. Damit lauten die drei Korrelationskoeffizienten: $r_1 = –0{,}08$, $r_2 = –0{,}06$ sowie $r_3 = –0{,}06$.

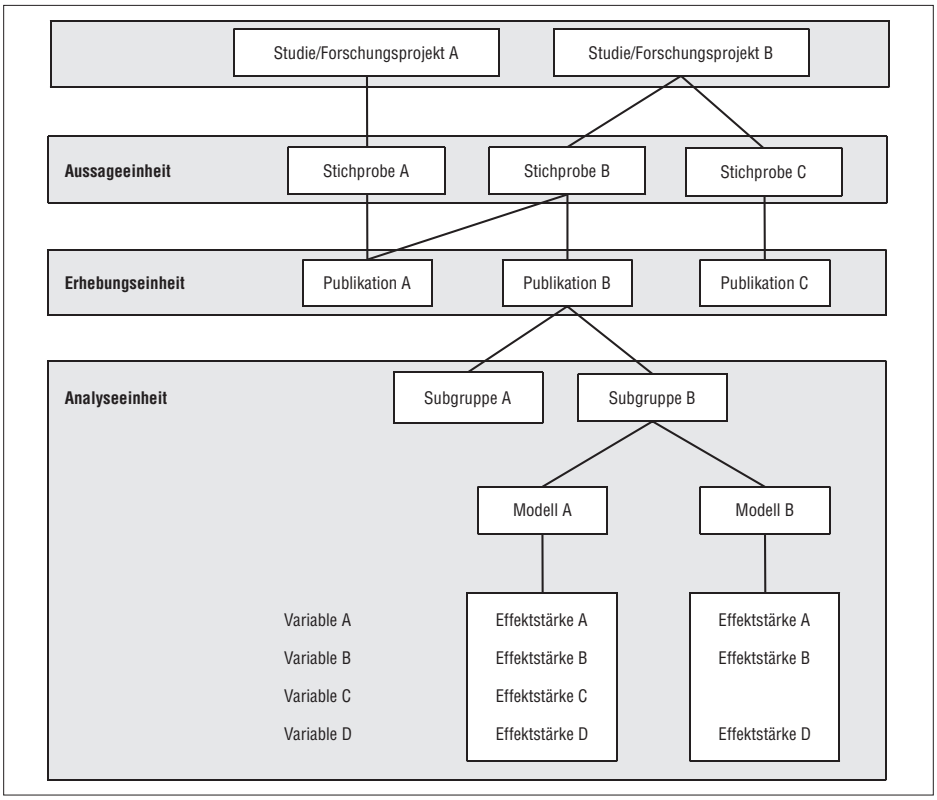
In der einschlägigen Literatur findet sich jedoch der Hinweis, dass für weitere Rechenschritte nicht die originalen Korrelationskoeffizienten, sondern deren Fishers z-Transformation $Z_r = 0.5 \cdot \ln[(1+r)/(1–r)]$ herangezogen werden sollte. Je stärker die Populationskorrelation von 0 abweicht, desto schiefer ist die Verteilung der Stichprobenkoeffizienten (Rosenthal 1994). Für inferenzstatistische Zwecke werden deshalb die Korrelationskoeffizienten via Fishers z-Transformation in eine Normalverteilung überführt (Rustenbach 2003).

b) Datenstruktur und Ebenen der Analyse

Grundsätzlich müssen drei Analyseebenen unterschieden werden: 1. Stichprobe, 2. Publikation sowie 3. numerischer Befund beziehungsweise Effektstärke. Diese drei Konzepte stehen in einer hierarchischen Ordnung zueinander (*Abbildung 1*). Die Stichprobe wird hier als die eigentliche Datenquelle verstanden, die analysiert wird und als Grundlage von Publikationen dient, in denen empirische Befunde mitgeteilt werden. Damit verbunden ist die (notwendige) Unterscheidung von Erhebungs- (Publikation), Analyse- (Effektstärke) und Aussageeinheit (Stichprobe), die sich auf je unterschiedlichen Ebenen wieder finden. In einigen Fällen lässt sich eine weitere, vierte Ebene ausmachen, die noch oberhalb der Stichprobe anzusiedeln ist. Eine Reihe von Stichproben sind wiederum nur Teil eines größeren Forschungsvorhabens und stehen deshalb (teilweise) in konzeptioneller Abhängigkeit zueinander. Prominente Beispiele sind etwa die verschiedenen Befragungswellen des Sozioökonomischen Panels oder die in verschiedenen Ländern durchgeführten Befragungen des „Family and Fertility Surveys“ (vgl. Probleme Abschnitt II.4.b).

Ferner ist zu beachten, dass sich die in den Publikationen berichteten Befunde auf verschiedene Subgruppen beziehen können. So werden etwa in der Ehescheidungsforschung manche Analysen für West- und Ostdeutschland oder für Ehejahrgänge getrennt durchgeführt. Auf der Basis spezifischer Subgruppen werden dann in der Regel

Abbildung 1: Ebenen der Analyse



mehrere Modelle aus dem Repertoire der Ereignisanalyse geschätzt. Jedes dieser Modelle umfasst eine unterschiedliche Kombination von Kovariaten. Die Effektstärken hängen damit nicht nur von der Studie beziehungsweise der Stichprobe ab, sondern auch von der jeweiligen Subgruppe sowie der Modellspezifikation (Wagner und Weiß 2003a).[5]

Für eine Meta-Analyse werden Informationen aus allen Ebenen benötigt. Ein solchermaßen konstruierter Datensatz enthält zum einen die eigentlichen Statistiken (zum Beispiel Produkt-Moment-Korrelationen) sowie zum anderen Merkmale, die die Charakteristika der Publikationen abbilden, wie beispielsweise das Veröffentlichungsjahr, Informationen zur Stichprobe etc. Vereinfacht lässt sich an dieser Stelle festhalten, dass die Merkmale der ersten Gruppe für die Befundsynthese benötigt werden, während die Merkmale der zweiten Gruppe vor allem für die Heterogenitätsanalysen wichtig sind.

5 Die zwischen den verschiedenen Ebenen bestehenden Beziehungen sind nicht notwendigerweise hierarchische Beziehungen. Beispielsweise ist es möglich, dass Publikation A Ergebnisse berichtet, die auf der Basis der beiden Stichproben A und B ermittelt wurden. So beschreiben Wagner und Weiß (2003a) den Fall, dass eine Publikation „doppelt" gezählt wird, da ihre Auswertungen auf zwei unterschiedlichen Stichproben basieren.

4. Datenanalyse

Die verschiedenen meta-analytischen Methoden lassen sich nach Tweedie (2001) in vier Gruppen unterteilen: So gibt es Verfahren, um die

a) Gesamtsignifikanz mehrerer, unabhängiger Befunde zu ermitteln. Liegen lediglich Aussagen über das Vorhandensein von statistischer Signifikanz sowie der Richtung der Effekte vor, dann kommt das so genannte Vote counting zum Einsatz (vgl. Abschnitt II.4.a).
b) Eine weitere Gruppe stellt Methoden zur Verfügung, um eine (mittlere) Gesamtstatistik berechnen zu können.
c) Bevor ein derartiger mittlerer Befund ermittelt werden kann, ist es häufig noch nötig, das Ausmaß der Heterogenität der Statistiken zu bestimmen und zu bewerten.
d) Darüber hinaus gibt es Methoden, um systematische Verzerrungen des Datenmaterials zu entdecken.

a) Inferenzstatistische Verfahren

Vote-counting. Das Vote-counting steht an der Grenze zwischen qualitativem und quantitativem Review. Light und Smith (1971) beschreiben das Verfahren – sie selbst sprechen von „taking a vote" – bereits Anfang der 1970er Jahre. Hier werden die in den Publikationen berichteten Zusammenhänge zwischen Variablen danach sortiert, ob das Ergebnis signifikant positiv, signifikant negativ oder nicht signifikant ist. Von der Modalkategorie wird behauptet, dass sie die beste Schätzung der Richtung des Zusammenhangs von unabhängiger und abhängiger Variable repräsentiere. Das Verfahren weist unter anderem die Schwäche auf, dass sich damit die Stärke von Zusammenhängen zwischen Variablen nicht bestimmen lässt (Glass 1976). Ein weiterer Nachteil des Verfahrens besteht darin, alle Effektstärken unterschiedslos auszuzählen und nicht zu berücksichtigen, dass Effektstärken aus kleineren Stichproben schlechtere Schätzer des Populationseffektes als solche aus großen Stichproben sind. In einer systematischen Analyse des Verfahrens konnten Hedges und Olkin (1985) zeigen, dass Vote-counting-Verfahren mit einem hohen Fehler zweiter Art (β-Fehler) verbunden sind. Demnach wird in den meisten Analysen, die sich des Vote-countings bedienen, der „wahre" Effekt nicht entdeckt (vgl. ausführlich auch Bushman 1994).

Zusammenfassen von Irrtumswahrscheinlichkeiten. In den Agrarwissenschaften entwickelte Ronald A. Fisher 1932 ein Verfahren, um auf Grundlage einzelner Signifikanztests in verschiedenen Untersuchungen, denen aber eine gemeinsame Hypothese zu Grunde liegt, eine Gesamtsignifikanz zu bestimmen. Üblicherweise wird dabei eine der Hypothese H_1 entgegengesetzte, so genannte Nullhypothese H_0 formuliert. In der empirischen Untersuchung wird dann die Teststatistik sowie die dazugehörige Fehlerwahrscheinlichkeit berichtet, sich auf Grundlage der aktuellen Stichprobe fälschlicherweise für die H_1 entschieden zu haben, obgleich „eigentlich" die H_0 Gültigkeit besitzt.

Für die Durchführung dieses integrativen Signifikanztests finden sich in der Literatur verschiedene Vorschläge (für eine Übersicht vgl. Fricke und Treinies 1985). Eine der bekanntesten Varianten dieser Technik wurde nach Fricke und Treinies (1985) von

Stouffer und seinen Mitarbeitern im Jahr 1949 entwickelt. Hier basiert der zusammengefasste Signifikanztest auf der Aufsummierung einzelner p_i-Werte (einseitig!) nach

$$\bar{Z} = \frac{\sum_{1=1}^{k} z(p_i)}{\sqrt{k}}, \qquad (1)$$

wobei $z(p_i)$ dem entsprechenden Abszissenwert einer Standardnormalverteilung entspricht und k der Anzahl der p_i.

Für die oben eingeführte Frage nach dem Zusammenhang zwischen Prestige des Elternhauses und der Häufigkeit des Schulschwänzens lauten die auf den Teststatistiken basierenden Irrtumswahrscheinlichkeiten wie folgt: $p_1 = 0{,}002$; $p_2 < 0{,}010$ und $p_3 = 0{,}026$. Eine sehr konservative Schätzung im zweiten Fall beträgt $p_2 = 0{,}010$. Es gilt nun, die empirischen z-Werte zu ermitteln beziehungsweise im Fall des χ^2-Tests den mit der Irrtumswahrscheinlichkeit korrespondierenden z-Wert. Für zweiseitige Fragestellungen müssen die z-Werte auf der Basis von p/2 geschätzt werden (Hedges und Olkin 1985; Rustenbach 2003). Weiterhin gilt: die Irrtumswahrscheinlichkeiten von Effekten, die einen im Vergleich mit den übrigen Befunden entgegengerichteten Effekt darstellen, müssen zunächst von 1 subtrahiert werden und auf Grundlage dieser Differenz wird anschließend der z-Wert bestimmt. Für das oben eingeführte Beispiel lauten die dazugehörigen z-Werte dann $z_1 = -3{,}09$, $z_2 = -2{,}58$ und $z_3 = -1{,}94$. Die Summe beträgt –7,61, diese geteilt durch die Standardabweichung $k^{0{,}5}$ ergibt ein Gesamt-z von –4,39. Dieser Wert wird schließlich mit dem kritischen Wert für einen einseitigen Test von $z = -1{,}64$ ($\alpha = 0{,}05$) verglichen. Der hier ermittelte Betrag von z liegt weit darunter, so dass die H_0, die keinen Zusammenhang von Prestige und Schwänzhäufigkeit unterstellt, zurückgewiesen werden kann.

Diesen Verfahren ist gemein, dass sie keinerlei Aussagen über die Größe des vorliegenden Effektes gestatten. Fricke und Treinies (1985: 70) kommen zu dem Schluss, dass die „Anwendung integrativer Signifikanztests in Metaanalysen u.E. noch problematischer als die Anwendung in Primäranalysen“ ist.

Auf Rosenthal (1991) geht eine Methode zurück, die es auf Grundlage des zusammengefassten p-Wertes erlaubt, ein so genanntes Fail-safe N zu berechnen. Damit lässt sich bei Vorlage eines signifikanten Gesamteffekts ermitteln, wie viele nicht signifikante Effekte noch zusätzlich vorhanden sein müssten, damit der Gesamteffekt nicht signifikant wird (Fricke und Treinies 1985: 70; für eine soziologische Anwendung vgl. Künzler 1994, vgl. zum File-drawer-Problem auch Abschnitt IV.5). Diese Anzahl X lässt sich nach folgender Gleichung ermitteln:

$$X = \left(\frac{k \cdot \left| \bar{Z} \right|}{1{,}96} \right)^2 - k. \qquad (2)$$

Für das obige Beispiel ergibt sich ein X von 42 zusätzlichen und statistisch nicht bedeutsamen Effekten, die gefunden werden müssten, um von einem nichtsignifikanten Zusammenhang zwischen Berufsprestige und Schulverweigerung ausgehen zu können.

b) Befundsynthese und Heterogenitätstests

Wenngleich in Abschnitt II.3.a eine Reihe von unterschiedlichen Befundtypen aufgelistet wurden, so ist das Vorgehen, einen gemeinsamen Schätzer zu gewinnen, relativ einfach und grundsätzlich auch überall anwendbar. Dabei sind die folgenden drei Bedingungen zu beachten:

a) Die Bildung eines synthetischen Befundes ist nur sinnvoll, wenn mindestens zwei einzelne Statistiken berichtet werden,
b) diese müssen statistisch unabhängig sein (Fricke und Treinies 1985; Lipsey und Wilson 2001) und
c) in die Mittelwertbildung gehen die einzelnen Befunde nach dem Kriterium ihrer Reliabilität gewichtet ein.

Bedingung b) ist besonders bedeutsam, da es – zumindest in der Soziologie – relativ häufig vorkommt, dass sich unterschiedliche Veröffentlichungen auf einer *gemeinsamen Datenbasis* mit identischen oder ähnlichen Fragestellungen befassen und damit *abhängige* Effektstärken resultieren. Wenn beispielsweise mehrere Publikationen, in denen Daten des Familiensurveys verwendet werden, den Einfluss des Heiratsalters auf die Scheidungsrate ermitteln, so sind die Effektstärken des Heiratsalters statistisch abhängig (Wagner und Weiß 2003a). Um diesem Problem zu begegnen, werden in einem ersten Schritt inhaltlich gleiche Effektstärken innerhalb derselben Stichproben gemittelt. In einem zweiten Schritt werden diese Mittelwerte zwischen Studien erneut gemittelt (Beelmann und Bliesener 1994; Bortz und Döring 1995). Die Mittelwertbildung innerhalb von Stichproben ist auch dann angebracht, wenn sich mehrere Veröffentlichungen auf unterschiedliche, aber voneinander abhängige Stichproben einer Studie beziehen. Dies trifft vor allem auf Panelstudien zu.

Um Bedingung c) zu genügen, werden die einzelnen Effektstärken mit der inversen Fehlervarianz der jeweiligen Effektstärke gewichtet. Dazu wird das gewichtete arithmetische Mittel eingesetzt. Dieses Verfahren wird von zahlreichen Autoren vorgeschlagen (Hedges und Olkin 1985; Lipsey und Wilson 2001; Normand 1999; Shadish und Haddock 1994). Effektstärken, die auf der Basis großer Stichproben gewonnen wurden, einen kleineren Standardfehler besitzen und damit eine höhere Reliabilität aufweisen, werden somit entsprechend stark gewichtet.

Für k unabhängige Effektstärken wird die mit der inversen Varianz V_i gewichtete mittlere Effektstärke $\overline{ES}$ wie folgt berechnet:

$$\overline{ES} = \frac{\sum_{i=1}^{k} w_i \times ES_i}{\sum_{i=1}^{k} w_i}, \text{ wobei } w_i = \frac{1}{v_i} = \frac{1}{SE_i^2}. \tag{3}$$

v_i selbst ergibt sich aus dem quadrierten Standardfehler der Effektstärke SE_i. Der zusammengefasste Standardfehler $SE_{\overline{ES}}$ ergibt sich aus:

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}}. \tag{4}$$

Mit Hilfe des so bestimmten Standardfehlers lassen sich auch Signifikanztests durchführen und Konfidenzintervalle ermitteln. Der Quotient aus der Effektstärke sowie dem Standardfehler ist standardnormalverteilt und lässt sich mit dem kritischen z-Wert eines vorher festgelegten Signifikanzniveaus vergleichen.

Bei jeder Meta-Analyse werden sich die einzelnen Effektstärken schon wegen des jeweiligen Stichprobenfehlers unterscheiden. Wenn die Heterogenität der Effektstärken *allein* vom Stichprobenfehler herrührt, dann spricht man von einer *homogenen* Effektstärkenverteilung und die Annahme ist gerechtfertigt, dass allen Effektstärken ein „*wahrer*" Populationseffekt zu Grunde liegt. Dieses Verteilungsmodell wird auch als Modell mit festen Effekten („fixed effects model") bezeichnet. Hingegen wird bei einem Modell mit zufälligen Effekten angenommen, dass es mehrere Populationsparameter gibt, diese selbst eine Verteilung aufweisen und um den Parameter einer „Superpopulation" streuen. In diesem Fall ergibt sich die Gesamtvarianz v_i^* einer Effektstärke aus der durch den Stichprobenfehler bedingten Varianz v_i der einzelnen Studie sowie aus einer Zwischenstudienvarianz σ^2_{ES}, so dass gilt (Shadish und Haddock 1994):

$$v_i^* = \sigma^2_{ES} + v_i. \tag{5}$$

In der tatsächlichen Anwendung unterscheiden sich die beiden Ansätze also in der Berechnung der Gesamtvarianz, so dass sich je nach Verteilungsmodell unterschiedliche Gewichtungen ergeben. Beim Modell mit zufälligen Effekten wird bei der Mittelung der einzelnen Befunde mit v_i^* gewichtet, wobei die Vergrößerung der Varianz einen deutlichen Verlust an statistischer Signifikanz zur Folge hat.

Die geschätzte Zwischenstudienvarianz $\hat{\sigma}^2_{ES}$ berechnet sich nun nach (zur Berechnung von Q siehe unten):

$$\hat{\sigma}^2_{ES} = \frac{Q-(k-1)}{c} \quad \text{mit} \quad c = \sum_{i=1}^{k} w_i - \left[\frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right]. \tag{6}$$

Heterogenitätstests wie etwa Cochrans χ^2-Test (Cochran 1954) geben Aufschluss darüber, ob ein Verteilungsmodell mit zufälligen oder mit festen Effekten angemessen ist. Dieser Test basiert auf der so genannten Q-Statistik (Normand 1999; Olkin 1996). Die Q-Statistik ist mit k–1 Freiheitsgraden χ^2 verteilt, wobei k die Anzahl der einzelnen Befunde ist:

$$Q = \sum_{i=1}^{k} w_i (ES_i - \overline{ES})^2 \tag{7}$$

Wenn Q den kritischen Wert der χ^2-Verteilung mit k–1 Freiheitsgraden übersteigt, muss die Nullhypothese, die eine homogene Verteilung annimmt, verworfen werden.

Mit anderen Worten: die einzelnen Statistiken unterscheiden sich signifikant voneinander. Demnach wäre in diesem Fall die Verteilung der Befunde heterogen.[6] Eine Konsequenz ist, wie bereits oben beschrieben, die Änderung der Gewichte zu v_i^* und damit auch der Betrag des als gewichteten Mittels berechneten Gesamtbefundes. Es stehen nun verschiedene Techniken der Heterogenitätsanalyse zur Verfügung, um Heterogenitätsquellen zu identifizieren und die Stichprobe dergestalt zu gruppieren, dass die Streuung innerhalb der Gruppen homogen ist. Je nach Skalenniveau der potenziellen Heterogenitätsquellen lassen sich varianz- und regressionsanalytische Verfahren unterscheiden.[7]

Ein Beispiel für einen varianzanalytischen Ansatz findet sich bei Amato und Keith (1991b). Sie gehen in ihrer Meta-Analyse dem Zusammenhang von Scheidung der Eltern und dem Wohlbefinden („well-being") der Kinder nach. Dazu vergleichen sie das psychische Wohlbefinden von Kindern aus stabilen Familien mit solchen aus Scheidungsfamilien. Für die Messung des Wohlbefindens existieren verschiedene Skalen und berichtet wurden dann etwa die Mittelwertdifferenzen (d, g) zwischen beiden Gruppen oder t-Werte. Die Autoren unterscheiden acht Bereiche des „Wohlbefindens", denen die empirischen Ergebnisse zugeordnet werden können: (1) Academic achievement, (2) Conduct, (3) Psychological adjustment, (4) Self-concept, (5) Social adjustment, (6) Mother-child relations, (7) Father-child relations und (8) Others. Insgesamt gehen 284 Effektstärken in die Untersuchung ein und mehr als zwei Drittel der Befunde sind negative Effektstärken. Scheidungskinder erreichen geringere Werte auf den verschiedenen Skalen des Wohlbefindens als solche aus stabilen Familien. Die entsprechenden mittleren Effekte für die einzelnen Bereiche des Wohlbefindens enthält *Tabelle 1*.

Die mit H_T bezeichneten Werte entsprechen dem oben eingeführten Q, und signifikante Werte indizieren eine heterogene Effektstärkenverteilung. „This statistic serves as a test for study by effect size interactions; that is, it tests to see if the effects of divorce on children are stronger in some studies than in other. When significant, it is advisable to assess study characteristics to determine if they account for variation in effect sizes" (Amato und Keith 1991b: 30).

Eine graphische Exploration des Materials erlauben so genannte Forestplots. Dazu wird jeder Effektstärkenschätzer in Form eines Quadrats sowie das dazugehörige Konfidenzintervall gegen die Abszisse geplottet. Die Fläche des einzelnen Quadrats ist dabei umgekehrt proportional zum Standardfehler. Größere Quadrate indizieren bessere Schätzer (vgl. *Abbildung 2*).

Für verschiedene Merkmale wie etwa Stichprobendesign, Stichprobengröße, Datenquelle, Stichprobenzusammensetzung rechnen die beiden Autoren nach Subgruppen getrennte Analysen. Interessant ist eine Subgruppenanalyse, die unterscheidet, ob die Ef-

6 Einige Autoren unterstellen diesem Heterogenitätstest eine geringe statistische Testpower (hoher β-Fehler). Demnach wird die Nullhypothese, die eine homogene Verteilung unterstellt, beibehalten, obgleich für die Population von einer heterogenen Verteilung ausgegangen werden muss (Hardy und Thompson 1998; Higgins und Thompson 2002). Fleiss (1986) schlägt daher vor, erst bei einer Irrtumswahrscheinlichkeit von mehr als 10 Prozent von einer homogenen Verteilung auszugehen.

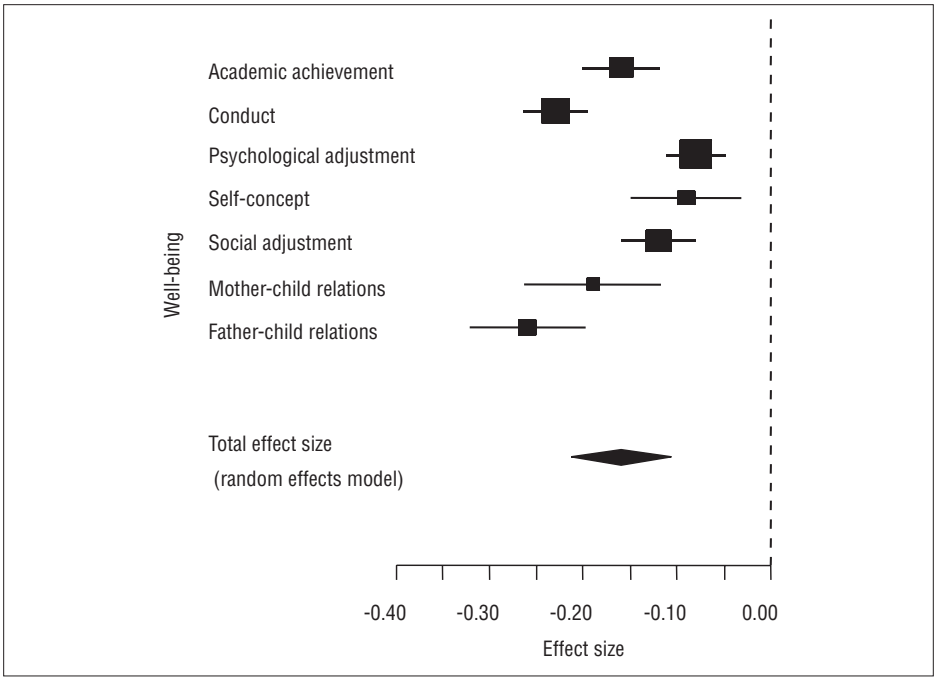
7 Rustenbach (2003) wendet sich speziell dem Problem der „Ausreißer" zu.

Tabelle 1: Vergleich von Kindern aus Scheidungsfamilien und stabilen Familien für verschiedene Dimensionen des Wohlbefindens: Mittlere Korrelationskoeffizienten, Fail-safe N und Heterogenitätstest

Domain of well-being	N	Mean effect size	Fail-safe N	H_T
School achievement	39	–0,16*	855	85,6*
Conduct	56	–0,23*	3474	221,5*
Psychological adjustment	50	–0,08*	464	98,6*
Self-concept	34	–0,09*	111	87,2*
Social adjustment	39	–0,12*	506	219,4*
Mother-child relations	22	–0,19*	212	86,8*
Father-child relations	18	–0,26*	456	73,0*
Other	26	0,06	–	51,0*

Quelle: Amato und Keith (1991b: 30).

Abbildung 2: Forestplot von Korrelationskoeffizienten für den Vergleich von Kindern aus Scheidungsfamilien und stabilen Familien nach verschiedenen Bereichen des Wohlbefindens (Quelle: Amato und Keith 1991b: 30)



fektstärken unter der Kontrolle weiterer Variablen (Nettoeffekte) ermittelt wurden oder ob es sich um Bruttoeffekte handelt (*Tabelle 2*).

In *Tabelle 2* beschreibt $H_{between}$ das Ausmaß der Heterogenität, das den Studienmerkmalen (hier „No controls“ und „Controls“) zugeschrieben werden kann. Getestet wird damit, ob sich die subgruppenspezifischen mittleren Effektstärken signifikant voneinander unterscheiden. H_{within} erfasst dagegen die innerhalb der Gruppen verbleiben-

Tabelle 2: Mittlere Korrelationskoeffizienten nach dem Publikationsmerkmal „Statistische Drittvariablenkontrolle“

Domain of well-being	N	Mean effect size	H_{within}	$H_{between}$
Academic achievement				12,2***
No controls	22	–0,25***	41,9**	
Controls	17	–0,10**	31,5*	
Conduct				2,5
No controls	39	–0,25***	170,5***	
Controls	17	–0,20***	48,5***	
Psychological adjustment				4,1*
No controls	42	–0,07***	85,0***	
Controls	8	–0,16***	9,5	
Self-concept				2,6
No controls	25	–0,11***	72,4***	
Controls	9	–0,04	12,2	
Social adjustment				4,5*
No controls	27	–0,10***	181,6***	
Controls	12	–0,23***	33,3***	
Mother-child relations				28,0***
No controls	18	–0,23***	48,7***	
Controls	4	0,18	10,1*	
Father-child relations				1,2
No controls	14	–0,26***	62,9***	
Controls	4	–0,38***	8,9*	

Quelle: Amato und Keith (1991b: 32); * $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.

de Heterogenität (Amato und Keith 1991b). Für „Academic achievement“ lässt sich beispielsweise erkennen, dass die unter der Kontrolle von Drittvariablen berechnete Effektstärke deutlich schwächer ausfällt.

III. Probleme von Meta-Analysen

1. „Apples and Oranges“ und „Garbage in and Garbage out“

Eine der häufigsten Kritiken an der Meta-Analyse lautet, dass hier Studien verglichen werden, obwohl sie unvergleichbar sind. Die unterstellte mangelnde Vergleichbarkeit kann sich dabei auf ganz unterschiedliche Sachverhalte beziehen: die Definition der unabhängigen und abhängigen Variablen, die angewandten Methoden, spezielle Eigenschaften der Stichproben (Olkin 1996). Eine Meta-Analyse durchzuführen ähnelt dann einem Vorgehen, bei dem „apples and oranges“ hergenommen und dann vermessen werden, etwa im Hinblick auf deren Gewicht und Größe. Rosenthal (1994) hält dem entgegen, dass bei einem Review immer Studien verglichen werden, die sich nicht völlig ähneln. Viel triftiger ist aber sein Argument, dass die Generalisierbarkeit von Befunden in diesem Fall auch gering wäre. Gerade Aussagen über „Früchte“ setzen eine Mischung aus „apples and oranges“ voraus. Es kommt hinzu, dass detaillierte Meta-Ana-

lysen methodologische Unterschiede zwischen den Studien berücksichtigen können, indem sie diese als Moderator-Variablen in ihr Studiendesign integrieren und quantitativ bestimmen, ob diese Variablen einen signifikanten Beitrag zur Aufklärung unterschiedlicher Untersuchungsergebnisse leisten.

Unter „garbage in, garbage out" wird verstanden, dass in Meta-Analysen Studien unterschiedlicher Qualität eingehen. Wenn aber „schlechte" Studien in die Analysen einbezogen werden, so könne das Ergebnis einer Meta-Analyse ebenfalls nicht überzeugen. Dieser Kritik kann durch eine Gewichtung der Studien nach Qualitätsmerkmalen begegnet werden. Solche Merkmale könnten sich auf die Art der Stichprobe oder deren Größe beziehen, ebenso können andere Qualitätsmerkmale in einem Ratingverfahren durch Experten herangezogen werden. Bei der Zusammenfassung der Befunde einzelner Studien werden diese Gewichte dann eingesetzt, wobei „schlechte" Studien zu einem geringeren Anteil berücksichtigt werden als „gute" Studien. Darüber hinaus ist es möglich, schlechte Studien aus der Meta-Analyse auszuschließen: „Dropping studies is merely the special case of zero weighting" (Rosenthal 1991: 130). Bevor dieser Schritt jedoch in Erwägung gezogen wird, muss untersucht werden, inwieweit es dadurch zu Verzerrungen des Datenmaterials kommt.

2. Unvollständiges Datenmaterial

Nach Pigott (1994) können Meta-Analysen in zweierlei Hinsicht von unvollständigem Datenmaterial betroffen sein: 1. Es fehlen komplette Studien/Publikationen (vgl. hierzu genauer Abschnitt III.3); 2. Meta-Analysen werden häufig durch mangelhaft publizierte Auswertungen (keine Standardfehler, ungenaue Variablendarstellung etc.) erschwert (Sutton et al. 2000).

Von den beispielsweise in Wagner und Weiß (2003a) untersuchten 42 Publikationen enthielten nur 9 (21,4 Prozent) numerische Angaben zum Standardfehler oder zum t-Wert. In den restlichen Publikationen wurde mit den bekannten „Sternchen" oder anderen Symbolen das Signifikanzniveau (nur sehr grob) angegeben. Um die Gewichte berechnen zu können, müssen die vorhandenen Informationen soweit als möglich genutzt und die Standardfehler geschätzt werden. Eine genaue Darstellung des angewandten Schätzverfahrens findet sich in Wagner und Weiß (2003a).

Ein Teil dieser Probleme wäre lösbar, wenn sich die soziologische Forschungs- und Publikationspraxis verbessern würden. Besonders nachteilig ist dabei, dass selten die bivariaten Effekte dargestellt und die Standardfehler der Effektstärken nicht angegeben werden.

3. Verzerrungen durch selektives Publizieren

Ein häufig mit Meta-Analysen in Zusammenhang gebrachtes Problem betrifft bestimmte Formen der Selektivität in der Publikationspraxis. Es ist bekannt, dass einerseits von Seiten der Autoren die Neigung besteht, nicht signifikante Resultate eher nicht zu publizieren (sondern in der Schublade verschwinden zu lassen: file-drawer-

Problematik) und andererseits diese Neigung von Seiten der Herausgeber und Lektoren in der Weise „gefördert" wird, dass bevorzugt hypothesenbestätigende und signifikante Studienbefunde herausgegeben werden (Publication bias) (Diekmann 1997: 599; Drinkmann 1990). Hinter einer solchen Verhaltensweise verbirgt sich, wie Begg (1994: 400) schreibt, die Vorstellung, dass „[s]tatistical significance, if it is achieved, may be used as a ‚proof' of a theory".[8]

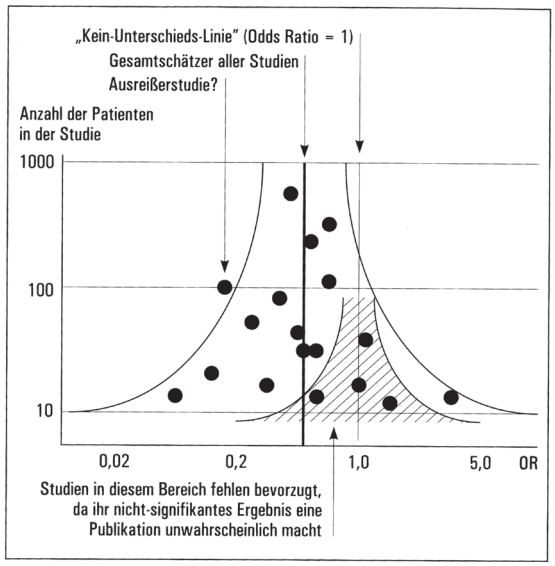
Diekmann (1997: 599) illustriert dies durch ein extremes Beispiel, indem er annimmt, 100 Forscher würden unabhängig voneinander eine Hypothese testen. Besitzt H_0 in der Population Gültigkeit, so werden im Durchschnitt etwa fünf Forscher signifikante und 95 Forscher nicht signifikante Ergebnisse erhalten. Würden im Folgenden nur die signifikanten Ergebnisse publiziert, so würde in der Fachliteratur fälschlicherweise der Eindruck einer einhelligen Zurückweisung der Nullhypothese entstehen. Im Weiteren führt Diekmann Ergebnisse einer älteren Untersuchung an, in der die Anteile signifikanter Ergebnisse in soziologischen und psychologischen Fachzeitschriften ermittelt wurden. Diese Angaben schwanken zwischen 60 Prozent und 100 Prozent und weisen auf eine einseitige Publikationspraxis hin.

Die Meta-Analyse stellt nun eine Reihe von Techniken zur Verfügung, mit denen sich systematische Verzerrungen durch die Publikationspraxis identifizieren lassen (für eine Übersicht und Beispiele vgl. Begg 1994: 401ff.). Ein gebräuchliches, grafisches Verfahren zur Identifikation eines Publication Bias stellt der so genannte Funnelplot dar. In einem Koordinatensystem werden zunächst die Effektgrößen der einzelnen Studien in Abhängigkeit ihrer Studiengröße (oder Standardfehler) abgetragen. Die zu Grunde liegende Idee dabei ist, dass die Effektmaße mit zunehmender Studiengröße beziehungsweise abnehmendem Standardfehler weniger stark streuen. Im Idealfall liegt keine Verzerrung vor und der Plot nimmt die Form eines Trichters (engl. „funnel") an. Liegt ein Publication Bias vor, dann enthält das Diagramm ein „Loch", meistens im Bereich kleinerer Fallzahlen und Effektstärken und damit abnehmender statistischer Signifikanz.

Abbildung 3 veranschaulicht schematisch das Prinzip eines Funnelplots. Die einzelnen Effektstärken – hier Odds Ratios – streuen im Idealfall so um den Gesamtschätzer, dass die oben genannte Trichterform entsteht. Ein Publication Bias entsteht dann, wenn, wie hier durch die schraffierte Fläche illustriert, vor allem kleine, nicht signifikante Effektstärken „in der Schublade" verschwinden und damit keinen Eingang in die Meta-Analyse finden.

8 Erste Untersuchungen, die sich mit dem Phänomen des Publication Bias befassen, wurden 1959 von Sterling durchgeführt und förderten zu Tage, dass von 294 in psychologischen Fachzeitschriften veröffentlichten Artikeln 286 (das heißt 97 Prozent) ein signifikantes Ergebnis berichteten. Auch neuere Untersuchungen stützen den Befund, dass Studien mit signifikanten Ergebnissen eine größere Veröffentlichungswahrscheinlichkeit haben (Begg 1994: 400).

Abbildung 3: Schematischer Aufbau eines Funnelplot (Sauerland und Neugebauer 2001)



4. Bivariate und multivariate Effektstärken

Gerade in der Soziologie kommt es häufiger vor, dass die einzelnen Koeffizienten, die zu mittleren Effektstärken aggregiert werden sollen, unterschiedlich spezifizierten Modellen entstammen. Koeffizienten aus einem bivariaten und einem multivariaten Modell unterscheiden sich aber im Hinblick auf ihren Betrag und den Standardfehler. Liegen etwa Regressionsmodelle mit jeweils unterschiedlichen unabhängigen Variablen vor, dann könnte jeder Regressionskoeffizient einer bestimmten Variablen einen unterschiedlichen Populationsparameter schätzen (Becker und Schram 1994: 362; Lipsey und Wilson 2001: 67). In der meta-analytischen Literatur wird noch diskutiert, ob derartige Partialkoeffizienten umstandslos in die Befundsynthese aufgenommen werden können: „Multivariate relationships present special challenges to meta-analysis" (Lipsey und Wilson 2001: 67, siehe auch die Diskussion zwischen Brüderl 2004 und Wagner und Weiß 2004a).

Mit diesem Sachverhalt kann man auf zwei Weisen umgehen (dazu ausführlich Wagner und Weiß 2004a). Erstens beziehen viele Autoren ohne besondere Vorbehalte Koeffizienten aus multivariaten Modellen in die Meta-Analyse ein. Man kann es sogar als einen Vorteil ansehen, wenn bi- und multivariate Effekte vorliegen, weil dies die Aussagekraft und Generalisierbarkeit der Befunde erhöht. Amato und Gilbreth (1999) transformierten in ihrer Meta-Analyse zum Zusammenhang von Vaterabwesenheit und dem Wohlbefinden von Kindern standardisierte und nichtstandardisierte Regressionskoeffizienten in Korrelationskoeffizienten und partielle Korrelationskoeffizienten, um Effektgrößen zu berechnen (vgl. auch Abschnitt II.4.b). Karney und Bradbury (1995),

die Befunde von Längsschnittstudien zur Ehequalität und -stabilität analysieren, argumentieren, dass sich bestimmte Assoziationen zwischen zwei Variablen A und B im Längsschnitt nur dann sicher identifizieren lassen, wenn die querschnittliche Korrelation von A und B kontrolliert wird. Hier ist es also sogar wünschenswert, wenn Partialkorrelationen in die Meta-Analyse eingehen. Ein solches Vorgehen lässt sich dadurch rechtfertigen, dass es auch sonst in der Meta-Analyse üblich ist, Effektgrößen zu aggregieren, die sich auf unterschiedliche Subgruppen beziehen (Kohorten, Regionen, Jahre), zwischen denen Effektgrößen signifikant variieren können. Zweitens lässt sich aber auch unter Anwendung von Heterogenitätsanalysen direkt überprüfen, inwiefern sich bivariate und multivariate Effekte unterscheiden. So kommen Amato und Keith (1991b) und Amato (2001) in ihren Arbeiten zu dem Ergebnis, dass methodologisch differenziert angelegte Studien weniger gravierende Folgen einer Ehescheidung für das Wohlbefinden der Kinder feststellen als andere Studien. Die Scheidungseffekte verringern sich, wenn Drittvariablen kontrolliert werden: „These results suggest that studies that do not use control variables tend to overestimate the effects of parental divorce" (Amato und Keith 1991b).

IV. Anwendungspotenzial der Meta-Analyse in der Soziologie

Meta-Analysen werden in den Natur- und den Sozialwissenschaften angewandt, sie dienen dem Erkenntnisfortschritt in jeder Wissenschaft, die sich quantitativer Verfahren bedient. Damit ist auch die Soziologie eine Wissenschaft, die von der Meta-Analyse profitiert und weiter profitieren kann.

Die Soziologie benötigt Meta-Analysen, und dieser Bedarf wird zukünftig weiter anwachsen. Hierfür gibt es zwei allgemeine Gründe. Erstens nimmt der Bestand an empirischen Studien zu, die sich einem bestimmten sozialen Tatbestand widmen. Es wird damit immer schwieriger, den Forschungsstand auf einem Gebiet sicher zu bestimmen. Der Forschungsstand lässt sich im Hinblick auf drei Elemente spezifizieren: ein theoretisches, ein empirisches und ein methodologisches Element. Zur Theorieprüfung und damit auch zur Theorieentwicklung trägt die Meta-Analyse dann bei, wenn sie explizit als ein hypothesentestendes Verfahren eingesetzt wird. Der Stand des empirischen Wissens wird bestimmt, indem zunächst einmal der Bestand an geeigneten Studien gesichtet und evaluiert wird. Die systematische Integration empirischer Befunde liefert genaueres und zuverlässigeres Wissen als jede einzelne Primärstudie für sich genommen. Zur Kritik und Weiterentwicklung methodologischer Aspekte der Forschung verhilft die Meta-Analyse, da sie methodische Defizite der empirischen Forschung aufdeckt, Methodeneffekte identifiziert und ein Instrument zur Qualitätskontrolle soziologischer Forschung darstellt.

Zweitens dient die Meta-Analyse nicht nur dazu, den Forschungsstand festzustellen, sondern ihn auch signifikant zu erweitern und neue erklärende Faktoren zu identifizieren. Die Soziologie kann hier in besonderer Weise von Heterogenitätsanalysen profitieren, also von der Beschreibung und Erklärung der Unterschiedlichkeit empirischer Befunde. Diese Unterschiedlichkeit lässt sich beispielsweise an spezifischen Publikations- und Studienmerkmalen festmachen, die im weitesten Sinn den sozialen, regionalen

oder historischen Kontext der jeweiligen Studie kennzeichnen. Wenn sich empirische Befunde deutlich danach unterscheiden, ob es sich um eine Stichprobe von Männern oder Frauen oder älteren oder jüngeren Geburtsjahrgängen handelt, dann liefert die Meta-Analyse originäre Kenntnisse über die Kontextabhängigkeit empirischer Befunde, die nur in der Zusammenschau verschiedener Primärstudien gewonnen werden können. Ganz analog kann man im Rahmen komparativer Fragestellungen untersuchen, ob die Heterogenität empirischer Befunde auf den Länderkontext zurückzuführen ist (De Leeuw und Hox 2003). Und im historischen Vergleich ist es bedeutsam zu wissen, ob etwa der Zusammenhang zwischen zwei Variablen über die Zeit hinweg zu- oder abgenommen hat. Schließlich gibt es erste Ansätze, die Meta-Analyse im Kontext von Mehrebenenproblemen anzuwenden (Goldstein 1995; Turner et al. 2000).

Um die Meta-Analyse in der Soziologie optimal anwenden zu können, müssen einige Bedingungen erfüllt sein, die vor allem die Forschungs- und Publikationspraxis betreffen. Nach unseren eigenen Erfahrungen mit der Durchführung von Meta-Analysen deutscher Scheidungsstudien weist die Publikationspraxis eine Reihe von Defiziten auf. So genügt es nicht, die Signifikanzniveaus von Effektstärken nur mit „Sternchen" zu kennzeichnen. Vielmehr muss der Standardfehler der Effektstärken angeben werden, will man nicht die Durchführung von Meta-Analysen und damit eine systematische Beurteilung des empirischen Forschungsstands erschweren. Ein weiteres Defizit ist auch die alleinige Berechnung multivariater Modelle. Es ist immer notwendig, auch die bivariaten Assoziationen zu berichten. Ein Vergleich empirischer Befunde zwischen Studien steht auf einem unsicheren Fundament, wenn Effektstärken von Variablen zueinander in Beziehung gesetzt werden sollen, die unterschiedlich spezifizierten Modellen entstammen. Derartige Schwächen der empirischen Forschungspraxis sind dem Ziel abträglich, den Wissensstand durch eine Kumulation von Forschungsergebnissen zu vergrößern.

Literatur

Altman, Douglas G., 2000: Statistics in Medical Journals: Some Recent Trends. Statistics in Medicine 19: 3275–3289.

Amato, Paul R., 2001: Children of Divorce in the 1990s: An Update of the Amato and Keith (1991) Meta-Analysis. Journal of Family Psychology 15: 355–370.

Amato, Paul R., und *Joan G. Gilbreth,* 1999: Nonresident Fathers and Children's Well-Being: A Meta-Analysis. Journal of Marriage and the Family 61: 557–573.

Amato, Paul R., und *Bruce Keith,* 1991a: Parental Divorce and Adult Well-being: A Meta-Analysis. Journal of Marriage and the Family 53: 43–58.

Amato, Paul R., und *Bruce Keith,* 1991b: Parental Divorce and the Well-Being of Children: A Meta-Analysis. Psychological Bulletin 110: 26–46.

Andrews, Gavin, Barry Guitar und *Pauline Howie,* 1980: Meta-Analysis of the Effects of Stuttering Treatment. Journal of Speech and Hearing Disorder 45: 287–307.

Bausell, Barker R., Yu-Fang Li, Meei-Ling Gau und *Karen L. Soeken,* 1995: The Growth of Meta-Analytic Literature from 1980 to 1993. Evaluation and the Health Professions 18: 238–251.

Becker, Betsy Jane, und *Larry V. Hedges,* 1984: Meta-Analysis of Cognitive Gender Differences: A Comment on an Analysis by Rosenthal and Rubin. Journal of Educational Psychology 76: 583–587.

Becker, Betsy Jane, und *Christine M. Schram,* 1994: Examining Exploratory Models through Research Syntheses. S. 357–381 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russel Sage Foundation.

Beelmann, A., und *T. Bliesener,* 1994: Aktuelle Probleme und Strategien der Metaanalyse. Psychologische Rundschau 45: 211 – 233.

Begg, Colin B., 1994: Publication Bias. S. 399–409 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Blettner, Maria, Willi Sauerbrei, Brigitte Schlehofer, Thomas Scheuchenpflug und *Christine Friedenreich,* 1997: Vergleich von traditionellen Reviews, Metaanalysen und gepoolten Analysen zur Bewertung von Risikofaktoren. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 28: 148–166.

Bortz, Jürgen, und *Nicola Döring,* 1995: Forschungsmethoden und Evaluation für Sozialwissenschaftler, 2. Auflage. Berlin/Heidelberg: Springer-Verlag.

Brüderl, Josef, 2004: Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsforschung oder „statistischer Fruchtsalat"? Zeitschrift für Soziologie 33: 84–86.

Bushman, Brad, 1994: Vote-counting Procedures in Meta-Analysis. S. 193–213 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Cochran, William G., 1954: The Combination of Estimates from Different Experiments. Biometrics 10: 101–129.

Cohn, Lawrence D., 1991: Sex Differences in the Course of Personality Development: A Meta-Analysis. Psychological Bulletin 109: 252–266.

Cooper, Harris M., 1982: Scientific Guidelines for Conducting Integrative Research Reviews. Review of Educational Research 52: 291–302.

Cooper, Harris M., und *Larry V. Hedges* (Hg.), 1994a: The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Cooper, Harris M., und *Larry V. Hedges,* 1994b: Research Synthesis as a Scientific Enterprise. S. 3–14 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Crain, Robert-L., und *Rita-E. Mahard,* 1983: The Effect of Research Methodology on Desegregation-Achievement Studies: A Meta-Analysis. American Journal of Sociology 88: 839–854.

De Leeuw, Edith, und *Joop J. Hox,* 2003: The Use of Meta-Analysis in Crossnational Studies. S. 329–399 in: *Janet A. Harkness, Fons van de Vijver* und *Peter Ph. Mohler* (Hg.), Cross-Cultural Survey Methods. Hoboken: John Wiley & Sons.

Diekmann, Andreas, 1997: Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 3. Auflage. Hamburg: Rowohlt.

Drinkmann, Arno, 1990: Methodenkritische Untersuchungen zur Metaanalyse. Weinheim: Deutscher Studienverlag.

Druckman, Daniel, 1994: Determinants of Compromising Behavior in Negotiation. A Meta-Analysis. Journal of Conflict Resolution 38: 507–556.

Eagly, Alice H., und *Wendy Wood,* 1991: Explaining Sex Differences in Social Behavior: A Meta-analytic Perspective. Personality and Social Psychology Bulletin 17: 306–315.

Fleiss, Joseph L., 1986: Analysis of Data from Multiclinic Trials. Controlled Clinical Trials 7: 267–275.

Fleiss, Joseph L., 1993: The Statistical Basis of Meta-analysis. Statistical Methods in Medical Research: 121–145.

Fleiss, Joseph L., 1994: Measures of Effect Size for Categorical Data. S. 245–260 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Fricke, Reiner, und *Gerhard Treinies,* 1985: Einführung in die Metaanalyse. Bern: Huber.

Friedrichs, Jürgen, 1973: Methoden der empirischen Sozialforschung. Reinbek: Rowohlt.

Ganong, Lawrence H., Marilyn Coleman und *Dennis Mapes,* 1990: A Meta-analytic Review of Family Structure Stereotypes. Journal of Marriage and the Family 52: 287–297.

Gartrell, C. David, und *John-W. Gartrell,* 1985: Social Status and Agricultural Innovation: A Meta-Analysis. Rural Sociology 50: 38–50.

Glass, Gene V., 1976: Primary, Secondary and Meta-Analysis of Research. Educational Researcher 5: 3–8.

Glass, Gene V., und *Mary Lee Smith,* 1979: Meta-Analysis of Research on Class Size and Achievement. Educational Evaluation and Policy Analysis 1: 2–16.

Goldstein, Harvey, 1995: Multilevel Statistical Models, 2. Auflage. London/New York: Oxford University Press.

Greenland, Sander, 1987: Quantitative Methods in the Review of Epidemiologic Literature. Epidemiologic Reviews 9: 1–30.

Greenwald, Rob, Larry V. Hedges und *Richard Laine,* 1994: When Reinventing the Wheel is Not Necessary: A Case Study in the Use of Meta-Analysis in Education Finance. Journal of Education Finance 20: 1–20.

Hall, Judith H., Linda Tickle-Degnen, Robert Rosentha und *Frederick Mosteller,* 1994: Hypotheses and Problems in Research Synthesis. S. 17–28 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Hardy, Rebecca J., und *Simon G. Thompson,* 1998: Detecting and Describing Heterogeneity in Meta-Analysis. Statistics in Medicine 17: 841–856.

Hedges, Larry V., 1987: How Hard is Hard Science, How Soft is Soft Science? American Psychologist 42: 443–455.

Hedges, Larry V., 1994: Statistical Considerations. S. 29–38 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Hedges, Larry V., und *Ingram Olkin,* 1985: Statistical Methods for Meta-Analysis. Orlando: Academic Press.

Heitmann, Kai U., 2001: Literaturrecherche im Internet. S. 41–47 in: *Karl W. Lauterbach* und *Matthias Schrappe* (Hg.), Gesundheitsökonomie, Qualitätsmanagement und Evidence-based Medicine. Eine systematische Einführung. Stuttgart: Schattauer.

Higgins, Julian P. T., und *Simon G. Thompson,* 2002: Quantifying Heterogeneity in a Meta-analysis. Statistics in Medicine 21: 1539–1558.

Hunt, Merton M., 1997: How Science Takes a Stock. New York: Russell Sage Foundation.

Hunter, John, und *Frank L. Schmidt,* 1990: Methods of Meta-Analysis. Newbury Park: Sage Publications.

Hyde, Janet Shibley, 1981: How Large Are Cognitive Gender Differences? A Meta-Analysis Using w^2 and d. American Psychologist 36: 892–900.

Karney, Benjamin R., und *Thomas N. Bradbury,* 1995: The Longitudinal Course of Marital Quality and Stability: A Review of Theory, Method and Research. Psychological Bulletin 118: 3–34.

Künzler, Jan, 1994: Familiale Arbeitsteilung: Die Beteiligung von Männern an der Hausarbeit. Bielefeld: Kleine Verlag.

Light, R. J., und *P. V. Smith,* 1971: Accumulating Evidence: Procedures for Resolving Contradictions among Different Research Studies. Harvard Educational Review 41: 429–471.

Light, Richard J., und *David B. Pillemer,* 1984: Summing Up. The Science of Reviewing Research. Cambridge, MA/London: Harvard University Press.

Lipsey, Mark W., und *David B. Wilson,* 2001: Practical Meta-Analysis. Thousand Oaks, CA/London/New Delhi: Sage Publications.

Markus, Evelyn, Alfred Lange und *Thomas F. Pettigrew,* 1990: Effectiveness of Family Therapy: A Meta-Analysis. Journal of Family Therapy 12: 205–221.

Merten, Klaus, 1995: Inhaltsanalyse. Einführung in Theorie, Methode und Praxis, 2. Auflage. Opladen: Westdeutscher Verlag.

Normand, Sharon-Lise T., 1999: Tutorial in Biostatistics. Meta-Analysis: Formulating, Evaluating, Combining, and Reporting. Statistics in Medicine 18: 321–359.

Olkin, Ingram, 1996: Meta-Analysis: Current Issues in Research Synthesis. Statistics in Medicine 15: 1253–1257.

Pigott, Therese D., 1994: Methods for Handling Missing Data in Research Synthesis. S. 163–175 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Raftery, Adrian E., 2001: Statistics in Sociology, 1950–2000: A Selective Review. Sociological Methodology 31: 1–45.

Raudenbush, Stephan W., 1984: Magnitude of Teacher Expectancy Effects on Pupil IQ as a Function of the Credibility of Expectancy Induction: A Synthesis of Findings From 18 Experiments. Journal of Educational Psychology 76: 85–97.

Rosenthal, Robert, 1991: Meta-Analytic Procedures for Social Research, 2. Auflage. Newbury Park: Sage Publications.

Rosenthal, Robert, 1994: Parametric Measures of Effect Size. S. 231–244 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Rustenbach, Stephan Jeff, 2003: Metaanalyse. Eine anwendungsorientierte Einführung. Bern: Verlag Hans Huber.

Sackett, D. L., W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes und *W. S. Richardson,* 1997: Was ist Evidenz-basierte Medizin und was nicht? Münchner medizinische Wochenschrift 139: 644–645.

Sally, David, 1992: Conversation and Cooperation in Social Dilemmas. A Meta-Analysis of Experiments from 1958 to 1992. Rationality and Society 7: 58–92.

Sauerland, Stefan, und *Daniel Galandi,* 2001: Cochrane Collaboration und Cochrane Library. S. 84–88 in: *Karl W. Lauterbach* und *Matthias Schrappe* (Hg.), Gesundheitsökonomie, Qualitätsmanagement und Evidence-based Medicine. Eine systematische Einführung. Stuttgart: Schattauer.

Sauerland, Stefan, und *Edmund Neugebauer,* 2001: Qualität von Meta-Analysen. S. 89–95 in: *Karl W. Lauterbach* und *Matthias Schrappe* (Hg.), Gesundheitsökonomie, Qualitätsmanagement und Evidence-based Medicine. Eine systematische Einführung. Stuttgart: Schattauer.

Schnell, Rainer, Paul B. Hill und *Elke Esser,* 1995: Methoden der empirischen Sozialforschung. München: Oldenbourg Verlag.

Schulze, Ralf, Heinz Holling und *Dankmar Böhning* (Hg.), 2003: Meta-Analysis. New Developments and Applications in Medical and Social Science. Göttingen: Hogrefe und Huber.

Shadish, William R., und *Keith C. Haddock,* 1994: Combining Estimates of Effect Size. S. 261–281 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Smith, Mary Lee, und *Gene V. Glass,* 1977: Meta-Analysis of Psychotherapy Outcome Studies. American Psychologist 32: 752–760.

Sutton, Alex J., Keith R. Abrams, Michael Jonas, Trevor A. Sheldon und *Fujian Song,* 2000: Methods for Meta-analysis in Medical Research. Chichester/New York: J. Wiley.

Turner, Rebecca M., Rumana Z. Omar, Min Yang, Harvey Goldstein und *Simon G. Thompson,* 2000: A Multilevel Model Framework for Meta-analysis of Clinical Trials with Binary Outcomes. Statistics in Medicine 19: 3417–3432.

Tweedie, Richard L., 2001: Meta-analysis: Overview. S. 9717–9724 in: *Neil J. Smelser* und *Paul B. Baltes* (Hg.), International Encyclopedia of the Social and Behavioral Sciences. Amsterdam/New York: Elsevier.

Twenge, Jean M., W. Keith Campbell und Craig A. Foster, 2003: Parenthood and Marital Satisfaction: A Meta-Analytic Review. Journal of Marriage and the Family 65: 574–583.

Vemer, Elizabeth, Marilyn Coleman, Lawrence H. Ganong und *Harris M. Cooper,* 1989: Marital Satisfaction in Remarriage: A Meta-Analysis. Journal of Marriage and the Family 51: 713–725.

Wagner, Michael, Imke Dunkake und *Bernd Weiß,* 2004: Schulverweigerung. Empirische Analysen zum abweichenden Verhalten von Schülern. Kölner Zeitschrift für Soziologie und Sozialpsychologie 56: 457–489.

Wagner, Michael, und *Bernd Weiß,* 2001: Meta-Analyse in der Scheidungsforschung. Abschlussbericht für die Deutsche Forschungsgemeinschaft. Köln (unveröffentliches Manuskript).

Wagner, Michael, und *Bernd Weiß,* 2003a: Bilanz der deutschen Scheidungsforschung. Versuch einer Meta-Analyse. Zeitschrift für Soziologie 32: 29–49.

Wagner, Michael, und *Bernd Weiß,* 2003b: Divorce Risks in Europe: An Overview. The Second Conference of the European Research Network on Divorce, 13. bis 14. November 2003 in Tilburg, Universität zu Köln (unveröffentliches Manuskript).

Wagner, Michael, und *Bernd Weiß,* 2004a: Die Bilanz der Scheidungsforschung bleibt eine Bilanz. Eine Antwort auf Josef Brüderl. Zeitschrift für Soziologie 33: 87–89.

Wagner, Michael, und *Bernd Weiß,* 2004b: Scheidungsrisiken in Deutschland: Eine Meta-Analyse von empirischen Befunden der Jahre 1987 bis 2003. S. 381–409 in: *Robert Kecskes, Michael Wagner* und *Christof Wolf* (Hg.), Angewandte Soziologie. Wiesbaden: Verlag für Sozialwissenschaften.

Wells, Edward L., und *Joseph H. Rankin,* 1991: Families and Delinquency: A Meta-Analysis of the Impact of Broken Homes. Social Problems 38: 71–93.

White, Howard D., 1994: Scientific Communication and Literature Retrival. S. 41–55 in: *Harris M. Cooper* und *Larry V. Hedges* (Hg.), The Handbook of Research Synthesis. New York: Russell Sage Foundation.

Wolf, Fredric, 1986: Meta-Analysis. Beverly Hills: Sage Publications.

Wortman, Paul M., und *Fred B. Bryant,* 1985: School Desintegration and Black Achievement. An Integrative Review. Sociological Methods and Research 13: 289–324.

VII. Modellbildung und Simulation

DYNAMISCHE SYSTEMMODELLE

Klaus G. Troitzsch

Zusammenfassung: Nach einem kurzen Überblick über die Geschichte der sozialwissenschaftlichen formalen Modellierung und über die verschiedenen Arten von Ansätzen der Computersimulation, die in den letzten 40 Jahren verfolgt worden sind, charakterisiert der Beitrag zwei grundlegend unterschiedliche Typen formaler Modellierung: Einerseits ist Simulation nicht mehr als die numerische Lösung mathematischer Gleichungssysteme, von denen die interessantesten meist keine analytische Lösung haben und über die man ohne numerische Simulation kaum etwas aussagen könnte. Hierher gehört vor allem der System-Dynamics-Ansatz, der vor 30 Jahren durch die Weltmodelle des Club of Rome bekannt geworden ist. Andererseits kann man von Computersimulation im eigentlichen Sinne erst sprechen, wenn die Möglichkeiten moderner, z.B. objektorientierter Softwaresysteme in der Weise ausgeschöpft werden, dass alle möglichen realweltlichen Entitäten auf programmiersprachliche Objekte abgebildet werden, die dann im Simulationsmodell auf ähnliche Weise miteinander in gegenseitige Abhängigkeiten treten, wie dies die jeweils entsprechenden realweltlichen Entitäten auch tun. Hierzu gehören Mehrebenenmodelle, die klassische Mikrosimulation, zellulare Automaten und vor allem die Multiagentenmodelle.

I. Einleitung

1. Systeme und Systemmodelle

Alle Sozialwissenschaften beschäftigen sich mit komplexen Systemen, in denen große Zahlen von Komponenten einander beeinflussen, und dies in höchst vielfältiger Form. Manche dieser Einflüsse mögen stochastischer Natur sein, jedenfalls können sie leicht stochastisch modelliert werden; manche Einflüsse, vor allem solche auf der Makroebene, mögen auch deterministisch sein; wieder andere sind von regelhaftem Verhalten und regelgeleitetem Handeln gekennzeichnet. Sozialwissenschaftliche Theoriebildung hat allen diesen Aspekten Rechnung getragen, und das gilt im Wesentlichen auch für die sozialwissenschaftlichen Ansätze formaler Modellierung und der Computersimulation.

Man kann kaum einen Beitrag, in dessen Titel das Wort „System“ vorkommt, schreiben, ohne wenigstens den Versuch einer Definition des Begriffs „System“ zu unternehmen. In verschiedenen Wissenschaftszweigen wird der Begriff „System“ nämlich

auf höchst verschiedene Weise verwendet, aber es kann nicht Aufgabe dieses Beitrages sein, einen Überblick über die verschiedenen Begriffsverwendungen zu geben. Hier soll – soweit irgend möglich – der Systembegriff von Mario Bunge (1979) zugrunde gelegt werden. Um ein realweltliches System zu beschreiben, hat man nach Bunge zunächst drei Vorfragen zu klären:
– Welche Dinge gehören zum System,
– welche Dinge gehören zu seiner Umwelt und
– wie stehen diese Dinge in Beziehung zueinander?
Dazu muss man festlegen, auf welche Art von Dingen man eigentlich abstellen will – für den Physiker werden das Partikel sein, für die Chemikerin Moleküle, für den Biologen vielleicht Zellen, für die Sozialwissenschaftlerin Menschen (die somit ebenfalls unter die sehr allgemeine Kategorie „Ding" fallen). Etwas formaler benutzt Bunge zur Definition seines Systembegriffs zunächst eine Grundmenge von Dingen der realen Welt eines oder mehrerer Typen. Eine Teilmenge dieser Grundmenge ist die Zusammensetzung (composition) des Systems. Das Komplement dieser Zusammensetzung (d.h. die Menge aller Dinge, die nicht zur Zusammensetzung des Systems gehören) ist die Umwelt (environment) des Systems. In welcher Beziehung schließlich all diese Dinge (im System und in der Umwelt) stehen, wird in der Struktur (structure) des Systems beschrieben. Dies ist eine Menge von Relationen, die auf den Elementen der Grundmenge definiert sind, und unter diesen kann man zwei Arten von Relationen unterscheiden: solche, die mit einer Wirkung eines Elements der Grundmenge auf ein anderes verbunden sind (bonding relations) von solchen, bei denen solche Wirkungen nicht auftreten. Eine solche bonding relation wäre z.B. die Relation „liebt" („hasst", „stört", „belehrt" ...), während „ist größer als" oder „steht neben" einfache (non-bonding) Relationen sind. Bunge spricht von einem System nur dann, wenn die Struktur bonding relations enthält, vermöge deren die Elemente (der Zusammensetzung und der Umwelt) eines Systems Wirkungen in dem Sinne aufeinander ausüben, dass ihre Zukunft von den Einflüssen anderer Elemente abhängt.

Dynamische Systemmodelle in den Sozialwissenschaften, wie sie dieser Beitrag in seinem Titel zu behandeln verspricht, sind also solche, in denen Menschen, aber auch Gruppen, Verbände, Organisationen in einer Weise modelliert werden, dass sie sich über die Zeit gegenseitig beeinflussen.

2. Überblick

Nach einem kurzen Überblick über die Geschichte der sozialwissenschaftlichen formalen Modellierung und der Computersimulation und über die verschiedenen Arten von Ansätzen, die in den letzten 40 Jahren verfolgt worden sind, charakterisiert der Beitrag[1] zwei grundlegend unterschiedliche Typen formaler Modellierung: Einerseits ist Simulation nicht mehr als die numerische Lösung mathematischer Gleichungssysteme, von denen die interessantesten meist keine analytische Lösung haben und über die man ohne numerische Simulation kaum etwas aussagen könnte. Andererseits kann man

1 Einige Abschnitte dieses Artikels wurden übernommen aus Troitzsch (2003).

von Computersimulation im eigentlichen Sinne erst sprechen, wenn die Möglichkeiten moderner, z.B. objektorientierter Softwaresysteme in der Weise ausgeschöpft werden, dass alle möglichen realweltlichen Entitäten auf programmiersprachliche Objekte abgebildet werden, die dann im Simulationsmodell auf ähnliche Weise miteinander in gegenseitige Abhängigkeiten treten, wie dies die jeweils entsprechenden realweltlichen Entitäten auch tun. Diesen zweiten Typ von Simulation untersucht der vorliegende Beitrag nicht bis in alle Details, hierzu sei auf den folgenden Beitrag von Andreas Flache und Michael Macy verwiesen.

3. Dynamische Modelle sozialer Prozesse

Zu den ersten formalisierten dynamischen Modellen in den Sozialwissenschaften dürften demographische Modelle wie etwa die von Thomas Robert Malthus (1798) zählen, in denen vom geometrischen Wachstum der Bevölkerung und vom arithmetischen Wachstum der Nahrungsmittelproduktion die Rede ist, deren Kombination notwendig zu Armut und Hungersnöten führe. Diese Tradition setzt sich fort über Richardsons (1948) Rüstungsmodelle, Fucks' (1965) „Formeln zur Macht" und Forresters System Dynamics (1968, 1971) zu Meadows' „Grenzen des Wachstums" (1972, 1974, 1992). In diesen Modellen werden typischer Weise einige wenige Eigenschaften eines hochkomplexen „Systems", dessen Zusammensetzung (im Sinne Bunges) gar nicht betrachtet wird, über eine kleine Zahl von Differential- oder Differenzengleichungen und eine gelegentlich unüberschaubare Zahl von Hilfsgleichungen miteinander in Beziehung gesetzt (vgl. den nächsten Abschnitt).

Erst in den 1950er wurden Differentialgleichungen auch dazu eingesetzt, Beziehungen in kleineren Systemen zu modellieren – vgl. etwa Simons (Simon 1952) Modellierung der Gruppentheorie von Homans (Homans 1950). In all diesen bisher skizzierten Fällen kann man im Grunde noch nicht von „Systemmodellen" in Mario Bunges Definition von „System" sprechen; vielmehr ist der Gegenstand der Modellierung auch in System Dynamics (siehe den Abschnitt II.1) ein im Wesentlichen nicht in Elemente gegliedertes Ganzes, bei dem infolgedessen die Interaktionen zwischen den Elementen („Struktur" des Systems in Bunges Sinne) keine Rolle spielen können; stattdessen werden in diesen Modellen die gegenseitigen Abhängigkeiten von Zustandsvariablen des Gesamt„systems" untersucht.

Parallel zu diesen Ansätzen kamen stochastische Modellierungen sozialer Prozesse auf, die sich lange auf die reine mathematische Analyse solcher Prozesse beschränken konnten, weil für eine große Zahl einfacherer Prozesse geschlossene Lösungen möglich sind, wenn auch gelegentlich unter Annahmen, die in problematisch vereinfachender Weise vorgenommen werden (müssen). Mindestens zwei Richtungen lassen sich hier unterscheiden, die Modellierung von Prozessen, die durch eine Abfolge diskreter Ereignisse gekennzeichnet sind – vgl. z.B. Doreian und Hummon (1976), Diekmann (1980), Bartholomew (1982), Tuma und Hannan (1984), Diekmann und Mitter (1984) –, und die Modellierung von Prozessen in großen, als homogen angenommenen Populationen (Synergetik, Sociophysics) – vgl. die frühen Beiträge von Haken (1978, 1982, 1996) oder Weidlich (Weidlich und Haag 1983; Helbing und Weidlich

1995) –, wobei in die letztere Variante vielfach Elemente der Spieltheorie eingegangen sind, vor allem, wenn die Modelle Dilemma-Situationen mit einer größeren Zahl von Spielern untersuchen. Hier kann man denn auch in Mario Bunges Sinne von Systemmodellen sprechen.

4. Simulationsansätze in den Sozialwissenschaften

Computersimulation ist in den Sozialwissenschaften fast ebenso alt wie in allen anderen Wissenschaften auch: Kaum waren die ersten Computer verfügbar, wurden sie auch schon von Sozialwissenschaftlern eingesetzt. Dies gilt sowohl für Ansätze, bei denen es lediglich darauf ankam, Differential- oder Differenzengleichungssysteme numerisch zu berechnen, weil eine geschlossene Lösung nicht möglich oder mit ähnlich großem Rechenaufwand verbunden war oder die graphische Darstellung eines Prozesses ohnehin die Berechnung einer großen Zahl von Zwischenschritten erforderlich machte, als auch für Ansätze, die von vornherein nicht darauf angelegt waren, mit den klassischen Methoden der mathematischen Analyse zu arbeiten. Zu den ersteren zählen neben den genannten deterministischen Gleichungssystemen stochastische Prozesse, wie man sie u.a. in Warteschlangensystemen, aber auch Geburts- und Todesprozessen findet, bei denen befriedigende analytische Lösungen immerhin für einfache Fälle gegeben sind. Zu den letzteren zählen stochastische Prozesse, die den Geburts- und Todesprozessen zwar strukturell ähnlich sind, bei denen aber geschlossene analytische Lösungen nicht möglich sind und bei denen schon eine Aussage über qualitative Aspekte nicht ohne numerische Rechnung möglich ist – ein Beispiel findet sich in dem Abschnitt über die Entwicklung von Verwandtschaftsnetzwerken weiter unten.

Für den deutschen Sprachraum hat schon in den 1970er Jahren Steffen Harbordt die bis dahin vorhandene, fast ausschließlich amerikanische Literatur zur „Computersimulation in den Sozialwissenschaften“ aufgearbeitet (Harbordt 1974a, b; siehe auch Hauff und Latzelsberger 1968). Renate Mayntz (1967) hatte schon vorher in einem Sammelband mit Übersetzungen zahlreicher amerikanischer Beiträge zur mathematischen oder anderweitig formalisierten Soziologie diesen Ansatz im deutschen Sprachraum bekannt gemacht. Ziegler (1972) hatte mit seinem „Theorie und Modell“ den „Beitrag der Formalisierung zur soziologischen Theorienbildung“ beschrieben, aber auch kritisch beurteilt. Auch Rapoport (1980) hat jedenfalls die mathematischen Methoden in den Sozialwissenschaften einem deutschsprachigen Publikum zugänglich gemacht; weitere Versuche einer Formalisierung sozialwissenschaftlicher Theorien mit den Methoden vor allem der Mathematik (und weniger noch der Informatik) sind Norbert Müllers Lehrbuch der „Logik und Mathematik für Sozialwissenschaftler“ (Kliemann und Müller 1973, 1976; vor allem Kapitel 4 des zweiten Bandes) und viel später seine „Civilization Dynamics“ (Müller 1989, 1991).

Eine heute gängige Einteilung der verschiedenen Ansätze unterscheidet (vgl. Gilbert und Troitzsch 2005) die Tradition von System Dynamics (vgl. Abschnitt II.1), die Warteschlangen- oder ereignisorientierte Modellierung (II.2), die klassische Mikrosimulation (III.1), die Mehrebenensimulation (III.2.a-c, e und f), die zellularen Automaten (III.2.d und e) und die Multi-Agenten-Simulation (III.3), der ein eigener Beitrag in diesem Band gewidmet ist.

II. Computersimulation zur numerischen Lösung mathematischer Modelle

1. Der System Dynamics-Ansatz

a) Quantitative Vorhersage: Weltmodelle. System Dynamics ist wohl unter den sozialwissenschaftlichen Simulationsansätzen derjenige, der noch am ehesten einem breiten Publikum auch außerhalb der Wissenschaft bekannt geworden ist. Er wurde, angeregt durch den Club of Rome, Ende der 1960er und Anfang der 1970er Jahre von verschiedenen Forschergruppen erstmals eingesetzt, um die Entwicklung der Weltbevölkerung und ihrer Ressourcen bis ans Ende des 21. Jahrhunderts vorherzusagen (Forrester 1971; Meadows et al. 1972, 1974, 1992). Dazu benutzten sie sehr umfangreiche Gleichungssysteme, aus denen jeweils aus dem Zustand zu einem bestimmten Zeitpunkt die innerhalb einer kurzen Zeitspanne (von typischerweise einem Monat oder einem zehntel Jahr) zu erwartenden Zustandsänderungen berechnet wurden, woraus sich ein neuer Zustand zu einem etwas späteren Zeitpunkt ergab, aus dem für die nächste Zeitspanne berechnet werden konnte, wie sich der Zustand verändern würde. Diese Art von Modell hat zwei zentrale Charakteristika:

Es beschreibt den sich ändernden Zustand eines einzigen Objekts – hier der ganzen Welt – ohne irgendeine Unterteilung in Bestandteile. Die Systeme von System Dynamics sind in Bunges Sinne also gar keine Systeme; ihr Zustand wird mit einer (meist sogar ziemlich großen) Zahl von variablen Attributen dieses Objekts beschrieben – von der Kopfzahl der Weltbevölkerung bis zum Ausmaß der globalen Umweltverschmutzung.

Es unterstellt eine große Zahl von Invarianten, Zusammenhängen also, die sich im Laufe der Modellzeit nicht verändern – die Abhängigkeit der Geburtenrate von der Kopfzahl der Weltbevölkerung, vom Ausmaß der Umweltverschmutzung, vom Lebensstandard (und einer Reihe weiterer solcher variabler Attribute der Welt als ganzer) ist zu jeder Zeit dieselbe.

In Forresters WORLD2 (vgl. *Abbildung 1*) wird etwa die Welt mit gerade einmal fünf Zustandsvariablen beschrieben: der Kopfzahl der Weltbevölkerung, dem Ausmaß der Umweltverschmutzung, dem eingesetzten Kapital (nur gerade aufgeschlüsselt nach Landwirtschaft einerseits und dem Rest der Weltwirtschaft andererseits) und dem Umfang der (noch verfügbaren) natürlichen Ressourcen, wobei diese fünf Zustandsvariablen durch eine große Zahl von Zustandsänderungsfunktionen und Hilfsgrößen miteinander in Beziehung stehen.

Es ist schnell ersichtlich, dass die Ungegliedertheit der modellierten Welt die reale Welt schlecht wiedergibt. Es macht wohl nicht viel Sinn, die mittlere weltweite Geburtenrate in Beziehung zum mittleren weltweiten Lebensstandard oder zur mittleren weltweiten Umweltverschmutzung zu setzen (nicht davon zu reden, dass die Umweltverschmutzung eine Reihe von durchaus verschiedenartigen Aspekten hat oder dass Geburtenraten im Zusammenhang mit der Altersstruktur der Bevölkerung gesehen werden müssen).

Wenn trotzdem vor allem in den 1970er Jahren diese Weltmodelle einen nicht unerheblichen Einfluss auf die politische Diskussion gehabt haben (und in Teilbereichen noch immer haben), so hat das zwei Arten von Gründen: System Dynamics-Modelle

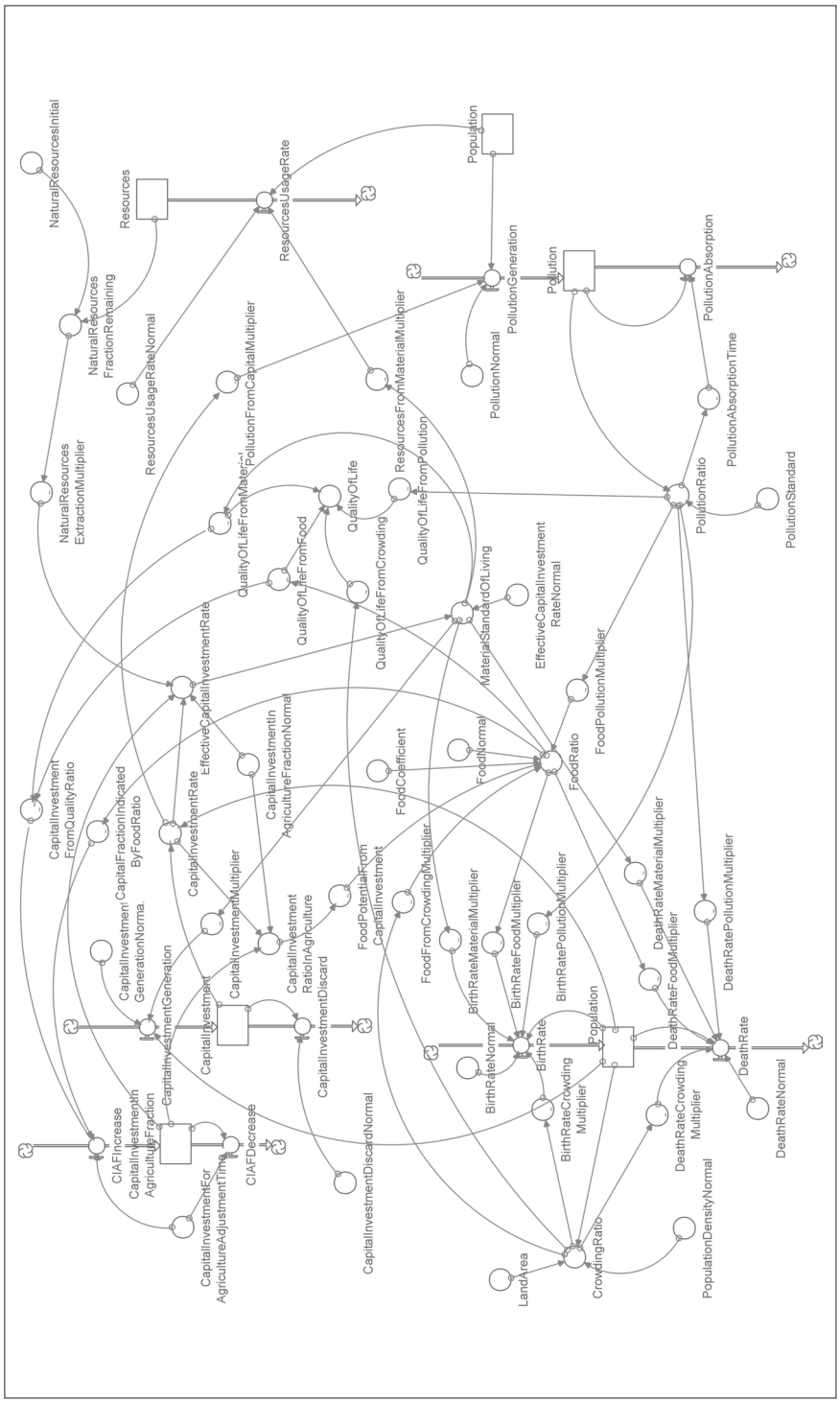


Abbildung 1: System Dynamics-Diagramm von Forresters Weltmodell WORLD2

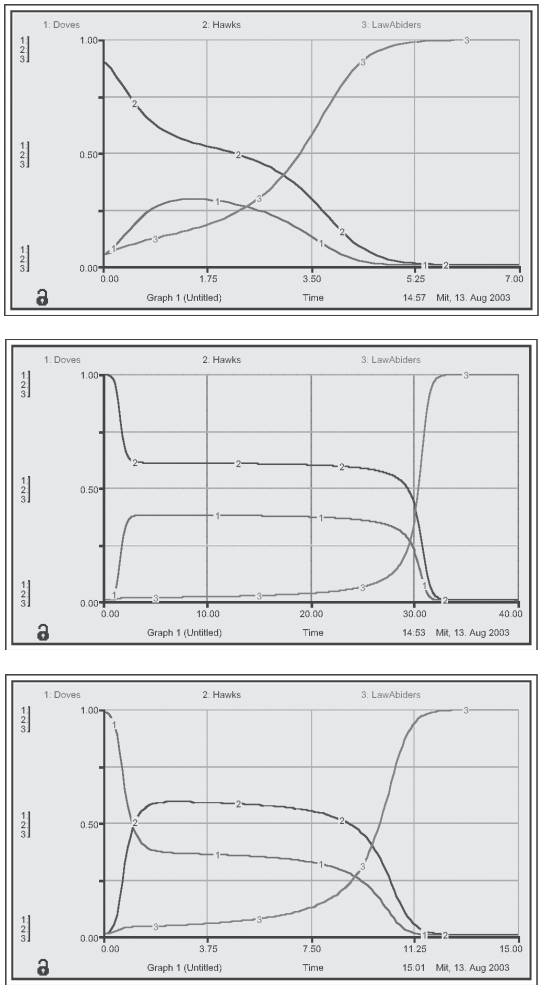
stellen einstweilen die einzige Möglichkeit dar, weitreichende Prognosen zu machen; zugleich basieren sie nicht auf einer einfachen Trendfortschreibung, sondern machen eine in vielen, wenn nicht in allen Einzelheiten nachvollziehbar begründete Annahme über die Zusammenhänge zwischen den einzelnen Variablen. Auch wenn die Weltmodelle von Forrester und Meadows sich jeweils über mehrere Druckseiten Programmcode erstrecken, sind sie doch immer noch so übersichtlich, dass sie von Menschen nachvollzogen und modifiziert werden können. Sie haben also das Argument für sich, dass sie das Ergebnis einer im Prinzip offenen Diskussion sind und somit das gesamte Wissen ihrer Zeit über ihren Gegenstand repräsentieren. Dies lässt die oben genannten, eher grundsätzlichen Einwände in den Hintergrund treten.

b) Qualitative Vorhersage. System Dynamics-Modelle haben ihren Ursprung im Grunde in Differenzengleichungs- und Differentialgleichungssystemen, mathematischen Modellen also, die den Zusammenhang zwischen Zustand und Zustandsänderung beschreiben und seit dem 17. Jahrhundert, zunächst in der Physik, dann in immer mehr Wissenschaften eingesetzt wurden. Malthus' Überlegungen zur Entwicklung von Bevölkerung und Nahrungsvorrat gehen ebenso auf Differentialgleichungssysteme zurück wie das Modell der Rüstungswettläufe, das Richardson (1948) kurz nach Ende des Zweiten Weltkriegs beschrieben hat. Andere Anwendungen haben diese Modelle bei vielen Soziologen der zwei oder drei Jahrzehnte nach dem Zweiten Weltkrieg gefunden (Simon 1957; Ziegler 1972; Martinez Coll 1986), wobei es ihnen meist nicht um die konkrete Vorhersage künftiger Zustände komplexer Weltausschnitte ging, sondern vielmehr um die Analyse, welche Arten von Verhalten ein Sozialsystem unter variierten Bedingungen wohl würde zeigen können.

Eins der erwähnten Beispiele möge hier etwas ausführlicher dargestellt werden: Die Modellierung des Übergangs vom Hobbesschen Naturzustand in einen Zustand, in dem Gewalt weitgehend überflüssig geworden ist. Martinez Coll hat Hobbes' Naturzustand als einen solchen beschrieben, in dem (fast) alle Akteure die Strategie des „Falken" anwenden, d.h. Besitzer von Ressourcen attackieren, um in den Besitz dieser Ressourcen zu gelangen, und dabei gegebenenfalls Verletzungen in Kauf nehmen. Dieser Strategie lässt sich eine Nutzenberechnung zuordnen, die die wahrscheinlichen Kosten möglicher Verletzungen mit dem wahrscheinlichen Nutzen des Besitzes der erworbenen Ressource miteinander verrechnet. Alternative Strategien sind die der „Taube" – „Tauben" greifen nie an und verteidigen sich auch nicht; angegriffen, geben sie auf, was sie besitzen, und neue Ressourcen erwerben sie nur, indem sie abwarten, bis ein anderer Akteur seine Ressource zufällig aufgibt – und die des „Gesetzestreuen" – „Gesetzestreue" greifen, wie die „Tauben" nie an, verteidigen sich aber, wenn angegriffen, wie die „Falken".

Die drei Strategien verbreiten sich in einer Population umso mehr oder weniger, je profitabler oder weniger profitabel es für ein (freilich nicht individuell modelliertes) Individuum ist, von einer Strategie zu einer anderen zu wechseln. Mathematisch modelliert, entsteht aus Martinez Colls Annahmen ein kubisches Differentialgleichungssystem, welches unter anderem auch mit den Werkzeugen von System Dynamics numerisch gelöst und graphisch veranschaulicht werden kann. *Abbildung 2* zeigt, dass unter allen Umständen bei vernünftigen zahlenmäßigen Besetzungen der Kosten für Verletzung und Abwarten und des Wertes der im Kampf oder mit Geduld auszutauschenden

Abbildung 2: Drei Verläufe der Entwicklung der Strategien „Falke“, „Taube“ und „Gesetzestreuer“ in Martinez Colls Modell des Übergangs aus Hobbes' Naturzustand



Ressourcen am Ende praktisch alle Akteure die Strategie der „Gesetzestreuen“ gewählt haben werden. In dieser Modellierung ist also die explizite Wahl eines Souveräns und die explizite Unterwerfung aller Menschen unter diesen Souverän überflüssig (oder im Sinne von Adam Smiths „unsichtbarer Hand“ bereits implizit modelliert). Es ist übrigens interessant festzustellen (und keinesfalls selbstverständlich), dass bei einer anderen Modellierungsstrategie, bei der das „System“ Gesamtpopulation ganz in Bunges Sinn tatsächlich auch als aus einzelnen Individuen zusammengesetzt modelliert wird – siehe weiter unten – qualitativ gleichwertige Simulationsergebnisse erzielt werden.

2. Warteschlangenmodelle

In weiten Bereichen der Ingenieurwissenschaften und der Betriebswirtschaftslehre wird Simulation mit Warteschlangensimulation gleichgesetzt (Kheir 1988); es gibt viele Simulationswerkzeuge, die in erster Linie für diese Art von Simulation entworfen worden sind. Die zentralen Elemente einer Warteschlangen- oder ereignisorientierten Simulation sind Bedienstationen, zugehörige Warteschlangen und Kunden, die das simulierte System irgendwann, d.h. mit realistisch verteilten Ankunftszeiten, betreten, auf Bedienung warten und schließlich – nach Bedienung – wieder aus dem System entfernt werden, wobei sie in der Regel mehr als eine Bedienstation einschließlich der zugehörigen Warteschlangen passiert haben. Dieser Ansatz ist geeignet, eine Reihe praktischer Probleme im Workflow Management zu bearbeiten, bei denen es darum geht, die Anzahl und Leistungsfähigkeit von Bedienstationen so zu optimieren, dass Wartezeiten möglichst gering gehalten oder ganz vermieden werden.
Die wesentlichen Elemente eines Warteschlangenmodells sind
– als dynamische Elemente: die Kunden;
– als statische Elemente: Bedienstationen und Warteschlangen;
– weitere Elemente: Quellen und Senken, die neue Kunden generieren oder sie nach Bedienung wieder aus dem System entfernen;
– eine Ereignisliste.

Die Ereignisliste ist das zentrale Element einer Warteschlangensimulation. In ihr werden alle voraussehbaren Ereignisse (Kunde betritt System, Kunde verlässt Bedienstation usw.) in ihrer zeitlichen Reihenfolge eingetragen und abgearbeitet. Der Systemzustand, d.h. die Zusammenfassung aller Zustände der einzelnen Elemente des gesamten modellierten Systems, ändert sich nur zum Zeitpunkt eines Ereignisses – zwischen Ereignissen passiert nichts. Während die meisten anderen Simulationsansätze sozusagen eine Simulationsuhr haben, die zu jedem Zeitschritt Zustandsänderungen auslöst, ist die ereignisorientierte Simulation insofern sparsamer. Dafür muss bereits am Beginn einer Kette von Ereignissen bekannt sein, wann das jeweils nächste Ereignis eintreten wird – mit dem Eintritt des Ereignisses „Kunde kommt an“ wird bereits das nächste Ereignis vom gleichen Typ („nächster Kunde kommt an“) in die Ereignisliste eingetragen; ebenso wird mit dem Ereignis „Bedienstation ruft nächsten Kunden aus der Warteschlange“ bereits das Ereignis „dieser Kunde ist zu Ende bedient“ für einen Zeitpunkt in der Zukunft eingetragen. Der Abstand zwischen zwei Kundenankünften wird ebenso wie die Bedienzeit als Zufallsgröße angenommen, die bestimmten – möglichst empirisch zu rechtfertigenden – Verteilungen folgen, wobei Bedienzeiten durchaus von der Art des Kunden abhängig sein können.

Es leuchtet ein, dass mit der Technik der Warteschlangensimulation Realitätsausschnitte modelliert und simuliert werden können, in denen Warte- und Bedienprozesse vorkommen, etwa
– ein Flughafen mit einer Anzahl von verschiedenen Informations- und Abfertigungsschaltern sowie Flugsteigen, mit Passagieren unterschiedlicher Art (First Class oder Economy, mit oder ohne Gepäck) (vgl. Gilbert und Troitzsch 2005: 87–97) oder
– eine Behörde, die sich damit beschäftigt, eingehende Anträge zu bearbeiten, sie zunächst auf Vollständigkeit zu prüfen, gegebenenfalls zur Wiedervorlage bei Vollstän-

digkeit zurückzustellen und schließlich zu bearbeiten, wobei Posteingangsbearbeitung, Vollständigkeitsprüfung, Zusammentragung und Bescheidung an unterschiedlichen Bedienstationen von unterschiedlichen Bedienern durchgeführt werden (vgl. Gilbert und Troitzsch 1999: 88–89)

und dass eine Simulation Aufschlüsse über Wartezeiten, Engpässe und Optimierungsmöglichkeiten geben kann. Gleichwohl handelt es sich bei dem Gegenstandsbereich der ereignisorientierten Simulation eigentlich nicht um den Kernbereich der Sozialwissenschaften – der Zwang, die Entitäten der Realität auf einige wenige Arten von Modellobjekten (Kunden, Bedienstationen, Warteschlangen) abbilden zu müssen, lässt wenig Raum für eine gegenstandsadäquate Modellierung. Gleichwohl lassen sich (siehe weiter unten) einige Techniken der ereignisorientierten Simulation zweckmäßig und effizienzsteigernd in andere sozialwissenschaftlich relevante Simulationsansätze exportieren, etwa in die mikroanalytische Simulation (Antcliff 1993; Galler 1997) oder in die agentenbasierten Ansätze (Troitzsch et al. 1998; Troitzsch 2003).

III. Computersimulation im eigentlichen Sinne

1. Die klassische mikroanalytische Simulation

Einen ähnlichen Einfluss auf Politik und Öffentlichkeit wie System Dynamics hat ein Simulationsansatz gehabt, der ebenfalls schon ziemlich früh entwickelt worden ist (vgl. Orcutt et al. 1986; Hauser et al. 1994a/b; Harding 1996). Die mikroanalytische Simulation nahm die Herausforderung an, vor der System Dynamics scheitern musste. Um es an einem Beispiel zu sagen: Wenn die Bevölkerungszahl des nächsten Jahres von altersspezifischen Geburten- und Todesraten und der Altersstruktur einer Bevölkerung abhängt, dann wird der Versuch, die Aufgabe der Vorhersage der Bevölkerungsentwicklung mit einem Differenzen- oder Differentialgleichungsmodell zu lösen, sehr schnell unübersichtlich und die Integration weiterer demographischer Aspekte – etwa die Beteiligung an Bildung und Erwerbstätigkeit – sehr schnell sehr schwierig. Hier bot sich an, im Rechner eine Population mit allen ihren einzelnen Mitgliedern abzubilden. Der Zustand eines einzelnen Mitglieds ist nun vergleichsweise übersichtlich, während der Zustand der gesamten Population jederzeit durch Aggregation (sozusagen eine Volkszählung im Simulationsmodell) errechnet werden kann. Der Gesamtzustand der Population wirkt sich in dieser Art von Modell typischerweise nicht unmittelbar auf die modellierten Individuen aus; Interaktionen zwischen den modellierten Individuen werden jedoch abgebildet (insofern erfüllen die Weltausschnitte, die unter diesem Simulationsansatz betrachtet werden, Bunges Definition von Systemen).

Wichtige Einsatzbereiche der mikroanalytischen Simulation sind neben der demographischen Prognose Vorausberechnungen der Effekte von Steuer- und Transfergesetzen über kürzere oder längere Zeiträume. Insofern Steuer- und Transfergesetze meist Schwellenwerte definieren (z.B. Steuerfreibeträge, unterschiedliche Proportionalzonen, Elternfreibeträge beim Bundesausbildungsförderungsgesetz, Beitragsbemessungsgrenzen in den verschiedenen Zweigen der Sozialversicherung), lässt sich aus der Änderung des Steuerfußes oder eines Beitragssatzes weder unmittelbar ausrechnen, wie viel mehr oder

weniger eine bestimmte Personengruppe im Mittel oder die Bevölkerung insgesamt an Steuern oder Sozialversicherungsbeiträgen zahlen muss, noch lässt sich so das Mehr- oder Minderaufkommen an Steuern oder Beiträgen abschätzen. Hierzu muss vielmehr für alle einzelnen Betroffenen getrennt das neue Recht angewandt werden, hernach muss aus den Einzeleffekten der Gesamteffekt aggregiert werden. In der sogenannten statischen Mikrosimulation geschieht genau das für ein oder wenige Jahre im voraus, freilich nicht auf der Basis der Daten der Gesamtbevölkerung, sondern auf der Basis einer (sehr) großen Stichprobe. Angesichts der kurzen Lebensdauer von Änderungsgesetzen im Steuer- oder Sozialversicherungsrecht ist es kaum notwendig, die Änderungen in der Zusammensetzung der Bevölkerung nach Alter, Einkommen, Erwerbstätigkeit oder Bildungsbeteiligung in das Modell einzubauen (statische Mikrosimulation). Simulationen dieser Art werden zur Politikberatung in allen entwickelten Staaten praktisch ständig eingesetzt (für einen Überblick siehe z.B. Harding 1996).

Anders sieht die Situation aus, wenn es darauf ankommt, Aussagen über eine eher ferne Zukunft zu machen, etwa darüber, ob und wie die Altersrenten für diejenigen, die heute am Anfang ihres Erwerbsleben stehen, dann noch gezahlt werden können, wenn sie sie beanspruchen können. Hier genügt es nicht, von den Individuen (oder Haushalten) einer Stichprobe unter Verwendung festgelegter Algorithmen, die den neuen oder geänderten Steuer- oder Transfergesetzen entsprechen, hochzurechnen, welche Effekte diese Gesetze haben werden, und dabei eventuell die Stichprobe umzugewichten, um einer etwa veränderten Altersstruktur Rechnung zu tragen. Für eine Abschätzung von langfristigen Auswirkungen von Änderungen in der Altersversorgung müssen die Biographien der Individuen einer Stichprobe fortgeschrieben werden. Hierzu bedient man sich der sogenannten dynamischen Mikrosimulation, deren Vorgehensweise etwa folgendermaßen charakterisiert werden kann.

In einer Datenbank sei der aktuelle Zustand einer großen Zahl von Haushalten und der zu ihnen gehörigen Individuen zum Zeitpunkt der Erhebung einer Stichprobe – etwa des Deutschen Sozioökonomischen Panels – gespeichert. Für jeden Zeitschritt werden für jeden Haushalt bzw. jedes Individuum folgende Teilberechnungen ausgeführt:

- Das Lebensalter jedes Individuums wird fortgeschrieben.
- Unter Benutzung von Angaben oder Annahmen über die aktuelle Sterbewahrscheinlichkeit in Abhängigkeit von Geschlecht, Alter, oder anderen bekannten Eigenschaften jedes einzelnen Individuums wird festgestellt, welche modellierten Individuen diesen Zeitschritt nicht überleben werden.
- Unter Benutzung von Angaben oder Annahmen über aktuelle alters-, bildungs-, einkommens-, haushaltsgrößenabhängige Reproduktionswahrscheinlichkeiten wird festgestellt, in welchen modellierten Haushalten ein Kind geboren wird.
- Unter Benutzung von Angaben oder Annahmen über die Wahrscheinlichkeit einer Ehescheidung oder einer sonstigen Auflösung einer langfristigen Partnerschaft in Abhängigkeit von der Art der Partnerschaft, des Alters, Berufs, Bildungsstands usw. der Beteiligten wird festgelegt, welche modellierten Haushalte geteilt werden.
- Unter Benutzung von Angaben oder Annahmen über Wahrscheinlichkeiten der Eheschließung oder der Bildung sonstiger langfristiger Partnerschaften wird festgelegt, welche modellierten Haushalte zusammengefasst werden.

In ähnlicher Weise kann für jedes Individuum bestimmt werden, wie sich seine Bildungsbeteiligung, seine Erwerbstätigkeit, sein Einkommen aus verschiedenen Einkommensquellen und je nach Bedarf seine weiteren Attribute verändern. Mit dem Sozioökonomischen Panel (und ähnlichen Datenquellen in anderen Ländern) steht der Forschung heute eine Datenquelle zur Verfügung, die den Ausgangspunkt für vielgestaltige Simulationen bilden kann, umso mehr als dieses seit fast zwei Jahrzehnten fortgeführte Panel zugleich Schätzungen von Übergangswahrscheinlichkeiten für praktisch sämtliche erhobenen Zustandsvariablen ermöglicht.

Soweit nur Geburts- und Todesprozesse, nicht aber zugleich auch Haushaltsauflösungs- und Vereinigungsprozesse beschrieben werden sollen, ließe sich eine solche Fortschreibung noch stochastisch oder deterministisch (mit praktisch gleichen Ergebnissen) in einem großen Gleichungssystem oder in einer Matrix von Übergangswahrscheinlichkeiten darstellen und geschlossen lösen, so dass man es insoweit noch mit einer Computersimulation der ersten Art (numerische Lösung eines mathematisch formulierten Modells) zu tun hat. Da aber bereits die Hereinnahme der Abhängigkeit der Reproduktionswahrscheinlichkeit vom Bildungsstand der Mutter oder vom Gesamteinkommen des Haushalts die mathematische Formulierung des Modells reichlich schwierig macht, wird man schon für diese Simulation von vornherein eine Repräsentation in einer geeigneten Programmiersprache vorsehen.

Am Ende eines Zeitschritts enthält die Datenbank ein mehr oder weniger realistisches Abbild der Stichprobe einen Zeitschritt später, wobei natürlich die individuellen Attribute der Personen oder Haushalte in der simulierten Stichprobe nicht mehr mit den am Ende des realen Zeitschritts zu beobachtenden Eigenschaften der realen Individuen und Haushalte übereinstimmen, aber die Eigenschaften des Aggregats dürften im Modell und in der Realität ziemlich ähnlich sein – wie ähnlich, hängt lediglich von der Güte und Verlässlichkeit der Angaben oder Annahmen über die verschiedenen Übergangswahrscheinlichkeiten ab und nicht davon, welche Individuen und Haushalte sich wie verändert haben.

Software für diese Art von Simulation war über viele Jahre auf Großrechner beschränkt, musste in einer Allzweck-Programmiersprache geschrieben werden und konnte so nur im Rahmen großer Forschungsprojekte wie etwa des Sonderforschungsbereichs 3 (Hauser et al. 1994a, b) oder in den interessierten Ministerien (Liedtke und Quinke 1997) genutzt werden (wobei im letzteren Falle Modell und Ergebnisse teilweise auch bei Konferenzpräsentationen vertraulich behandelt werden mussten). Erst seit wenigen Jahren (Heike et al. 1996a, b; Merz 1996a, b; Heike und Sauerbier 1997; Sauerbier 2002) steht mit UMDBS ein Simulationswerkzeug zur Verfügung, welches es recht einfach macht, Daten aus dem Sozioökonomischen Panel zu importieren und ein Mikrosimulationsmodell zu schreiben, ablaufen zu lassen und die Ergebnisse auszuwerten.

Bei bestimmten Fragestellungen ist es nicht notwendig, Zeitschritt für Zeitschritt (d.h. etwa Monat für Monat oder Jahr für Jahr) vorzugehen; stattdessen ist es manchmal möglich, ereignisorientiert zu modellieren, d.h. etwa bei Eintreten des Ereignisses „Geburt“ gleich das Ereignis „Beginn Schulbesuch“ für einen geeigneten Zeitpunkt ca. sechs Jahre später einzutragen oder anstelle der jährlichen Anwendung der altersabhän-

gigen Sterbewahrscheinlichkeit schon zum Zeitpunkt der Geburt den Todeszeitpunkt aus einer geeigneten Zufallsvariablen zu ziehen (vgl. Antcliff 1993; Galler 1997).

Da Annahmen über künftige Übergangswahrscheinlichkeiten (etwa Reproduktionswahrscheinlichkeiten) im Allgemeinen nur unter Vorbehalten gemacht und nur für kurze Zeiträume als konstant angenommen werden können, bleiben Vorhersagen über sehr lange Zeiträume selbstverständlich spekulativ. Gerade bei demographischen Vorhersagen ist es daher üblich, die Simulation mit unterschiedlichen Annahmen über die künftige Entwicklung von Übergangswahrscheinlichkeiten durchzuführen, um auf diese Weise zu sehen, welche Auswirkungen unterschiedliche Annahmen auf die Prognosen haben. Bei langfristigen Prognosen kommt es vielfach auch auf quantitative Präzision gar nicht an, vielmehr kann es dort genügen, Aufschlüsse etwa in der folgenden Art zu bekommen (vgl. Hauser et al. 1994a: 130–133; Galler 1990): Während 1986, d.h. zum Zeitpunkt der Erhebung der Stichprobe von den über 60-Jährigen, 60,2 Prozent mit einem Partner zusammenlebten, 82,7 Prozent Kinder und nur 9,2 Prozent weder Partner noch Kinder hatten, ist für das Jahr 2050 und die dann über 60-Jährigen zu erwarten, dass von ihnen nur noch 43,2 Prozent einen Partner und 68,6 Prozent Kinder haben werden, dass aber 22,3 Prozent weder Partner noch Kinder haben werden, wenn man die Übergangswahrscheinlichkeiten (Scheidung, Eheschließung, Geburt, Tod) der Jahre 1982/1983 zugrunde legt. Mit zusätzlichen Annahmen über die Entwicklung dieser Wahrscheinlichkeiten (z.B. weiter steigende Scheidungswahrscheinlichkeiten, sinkende Wahrscheinlichkeiten für Eheschließung, Geburten und frühen Tod – oder umgekehrt, oder in noch anderen Kombinationen) kommt man zu Abschätzungen, wie sich unter unterschiedlichen, aber heute für realistisch gehaltenen Annahmen Verwandtschaftsnetzwerke entwickeln werden. Diese Abschätzungen lassen sich dann zum Beispiel bei der Planung von Gesetzen zur Pflegeversicherung verwenden – sie mögen etwa plausibel machen, dass eine Pflegepflichtversicherung grundsätzlich erforderlich ist; für die Berechnung des Beitrags-/Leistungsverhältnisses nach vielen Jahren wird eine so langfristige Prognose aber nicht sehr hilfreich sein.

Die Entwicklung des Verwandtschaftsnetzwerks, die hier im Zentrum des Interesses steht, macht es erforderlich, nicht nur Eigenschaften einzelner Individuen oder Haushalte, sondern außerdem auch Beziehungen zwischen Individuen untereinander (z.B. ... ist Vater von ...) und zwischen Individuen und Haushalten (z.B. ... gehört zu ...) zu modellieren. Dies sprengt die Möglichkeiten zwar nicht der mathematischen Formalisierung (der von Sneed 1979; Stegmüller 1979, Balzer et al. 1987 verfolgte Ansatz der strukturalistischen Wissenschaftstheorie beschreibt eine sehr mathematische Formalisierung gerade auch solcher Theorien), aber doch den klassischer mathematischer Analyse – hier kommt in Ostroms (1988) Sinne die Computersimulation zu ihrem eigenen Recht.

2. Mehrebenenmodelle und zellulare Automaten

In den klassischen mikroanalytischen Simulationsmodellen ließen sich bereits mindestens zwei Modellierungsebenen unterscheiden: die Ebene der Individuen und die Ebene der Gesamtbevölkerung; in komplexeren Modellen kommt als dritte Ebene die der

Haushalte hinzu (oder bei Modellen, in denen außer Haushalten auch Unternehmen modelliert sind, die der Unternehmen). Zwischen diesen Ebenen finden allerdings keine (oder nur in sehr beschränktem Umfang) Wechselwirkungen statt: In den im vorigen Abschnitt genannten Beispielen wird gerade nicht unterstellt, dass die Veränderungen von Steuersätzen und/oder -freibeträgen zu Veränderungen im Konsum- oder Kapitalexportverhalten der Steuerpflichtigen führen oder dass die Veränderungen in der Struktur von Verwandtschaftsnetzwerken Auswirkungen auf das generative Verhalten der Individuen haben. Die modellierten Individuen reagieren also nicht auf das Geschehen, das sich in ihrer simulierten Umwelt abspielt, sondern lediglich auf die Eingriffe des Experimentators. Mit dieser Beschränkung werden die mikroanalytischen Simulationsmodelle zentralen Fragestellungen der Sozialwissenschaften nicht gerecht.

An dieser Stelle kommen verschiedene Arten von Mehrebenenmodellen ins Spiel, bei denen Komponenten mindestens zweier verschiedener Ebenen miteinander in Wechselwirkung stehen – im einfachsten Fall beeinflussen die modellierten Individuen durch ihre individuellen Zustandsänderungen den Gesamtzustand der Population, auf den sie anschließend wieder mit veränderten Zustandsänderungen reagieren. Die Untersuchung von Wechselwirkungen dieser Art ist eigentlich die Domäne der Synergetik (Haken 1978, 1982; zur Rezeption der Synergetik in den Sozialwissenschaften siehe Haken 1996). Stehen bei den Modellen der Synergetik Wechselwirkungen zwischen den Ebenen im Vordergrund, so sind es bei den zellularen Automaten vor allem die Wechselwirkungen zwischen benachbarten Elementen der gleichen Ebene, die auf der Aggregatebene zur Bildung neuer Strukturen führen.

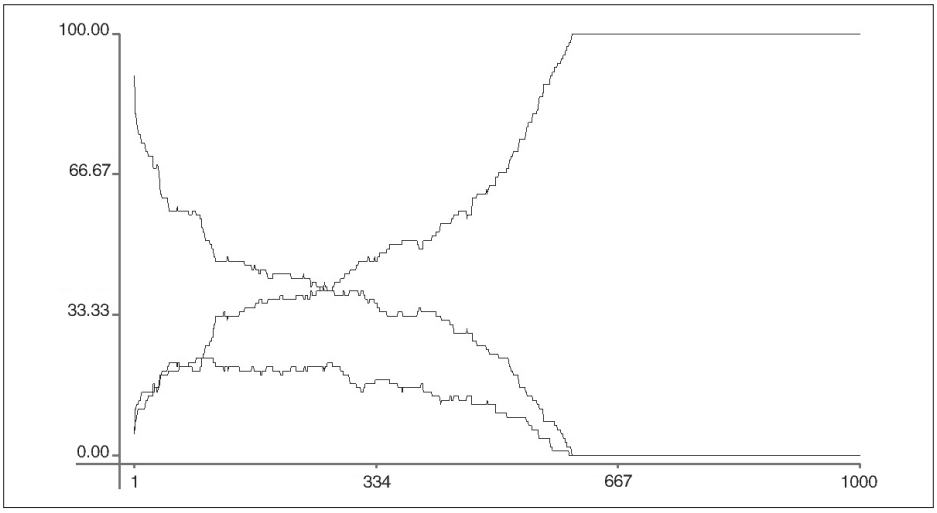
a) Die Überwindung des Hobbesschen Naturzustandes im Mehrebenenmodell. Weiter oben war der Ansatz von Martinez Coll zur Modellierung der Überwindung des Hobbesschen Naturzustandes referiert worden. Seine Grundgedanken lassen sich auch in einem Mehrebenenmodell wiedergeben. Im System Dynamics-Modell (vgl. oben S. 509ff.) waren die Veränderungen der Größen der drei nach angewandten Strategien unterschiedenen Subpopulationen berechnet worden. Demgegenüber stellt ein Mehrebenenmodell jedes Individuum einzeln dar und macht die Entscheidung eines Individuums abhängig von der Verteilung der Strategien unter allen Individuen, wobei selbstverständlich jedes einzelne Individuum mit der Änderung seiner Strategie auch die Verteilung der Strategien ändert.

In der Mehrebenenvariante des Modells (Gilbert und Troitzsch 2005: 123–128) ist es die Wahrscheinlichkeit eines Strategiewechsel, welche von der Verteilung der Strategien abhängt: Je profitabler eine alternative Strategie im Vergleich zur bisher angewandten ist, umso wahrscheinlicher ist dieser Wechsel (dabei gibt es sogar minimale Wahrscheinlichkeiten dafür, dass ein Wechsel zu einer weniger profitablen Strategie erfolgt, denn die Wechselwahrscheinlichkeit wird (wie vielfach in solchen Modellen) als

$$p(A \rightarrow B) = \nu \exp(U_B - U_A)$$

berechnet, worin U_B und U_A die erwarteten Profite aus den Strategien A und B sind (ν ist ein Faktor, der einerseits verhindert, dass die berechnete Wahrscheinlichkeit größer als 1 wird, andererseits aber auch die Geschwindigkeit bestimmt, mit der sich Veränderungen im Modell vollziehen).

Abbildung 3: Die Entwicklung der Verteilung der Strategien „Falke“, „Taube“ und „Gesetzestreu“ im Mehrebenenmodell



Die Entwicklung der Verteilung der Strategien sieht in diesem Modell (*Abbildung 3*) ganz ähnlich aus wie im Makromodell (vgl. *Abbildung 2*); allerdings sind die Kurven hier nicht glatt, weil die Zustandsvariablen der Gesamtpopulation jetzt (richtigerweise!) ganzzahlig sein müssen, und von irgendeinem Zeitpunkt an gibt es keine Individuen mehr, die die Strategien „Falke“ und „Taube“ benutzen, und damit verschwindet auch die Wahrscheinlichkeit fast vollständig (wegen der oben aufgeführten Formel bleibt sie aber positiv!), dass „Gesetzestreue“ sich jemals wieder für eine der beiden anderen Strategien entscheiden.

Es sei an dieser Stelle darauf hingewiesen, dass die Ergebnisse von Simulationen, die zwei unterschiedliche Simulationsansätze benutzen, keineswegs ähnlich sein müssen; vielmehr sind durchaus Fälle bekannt, in denen auch schon die qualitativen Vorhersagen eines Modells davon abhängen, welcher Modellierungsansatz benutzt wurde (Troitzsch 1989).

b) Ein Mehrebenenmodell des Wählerverhaltens. Dem Muster der Synergetik-Modelle folgt ein Simulationsmodell, das zu erklären sucht, wie es in Wählerschaften zu Polarisierungen kommt – eine Frage, der bereits Downs nachgegangen ist, wie folgende beide Zitate und Graphiken belegen: „Falls man die Wähler nicht irgendwie zur Mitte der Skala bringen und ihre Polarisierung überwinden kann... Der für Revolutionen typische politische Zyklus lässt sich als eine Reihe von Bewegungen von Menschen entlang der politischen Skala betrachten. Vor dem Umsturz beginnt sich die zunächst zentralisierte Verteilung in zwei Extreme zu polarisieren ...“ (Downs 1968: 116–117).

Die Graphiken bei Downs zeigen eine schematische Darstellung der Verteilung von Wählern über eine Links-rechts-Skala (genauer müsste man sagen: Verteilung eines Merkmals wie der Links-rechts-Selbsteinschätzung in der Wählerschaft), für die es ein empirisches Korrelat gibt. Während die Selbsteinstufung der Befragten auf einer Links-

Abbildung 4: Unimodale Verteilung von Wählern (Downs 1968: 114, Abbildung 2)

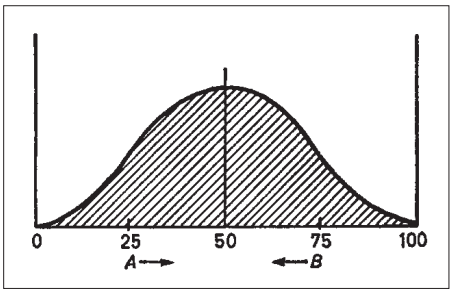
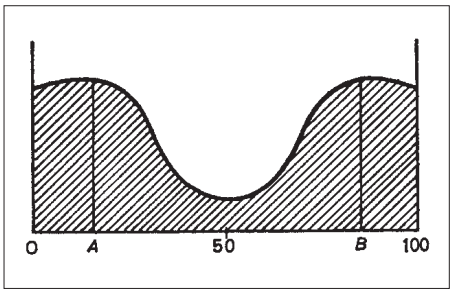


Abbildung 5: Bimodale Verteilung von Wählern (Downs 1968: 115, Abbildung 3)



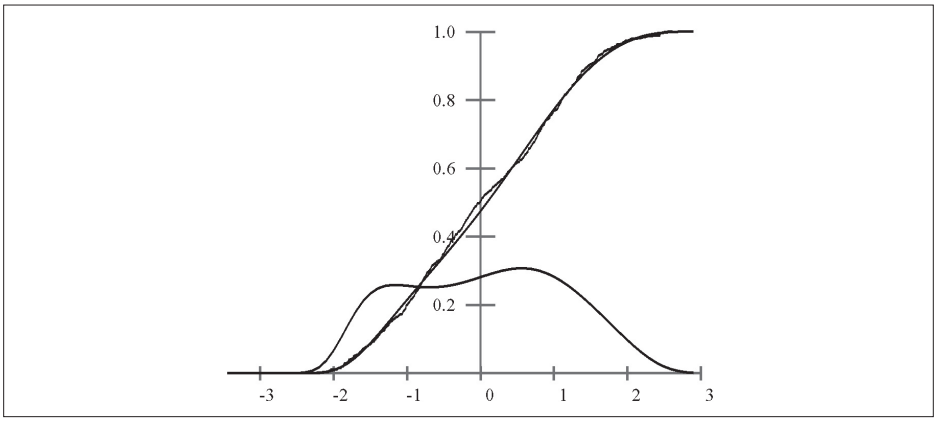
rechts-Skala von (z.B.) 0 bis 10 im Allgemeinen eine eingipflige Verteilung ergibt, die der Normalverteilung ähnlich sieht, erbrachte vor allem in den 1970er und 1980er Jahren die erste Hauptkomponente, die man aus den Skalometerfragen der Politbarometer-Umfragen extrahieren konnte, jedenfalls in Wahlkampfzeiten sehr häufig bimodale Verteilungen.

Auch die zweidimensionale Verteilung der ersten beiden Hauptkomponenten war in Wahlkampfzeiten meist bi- oder gar trimodal verteilt, was auf eine Polarisierung der Wählerschaft hindeutet. Diese Polarisierung kann man sich – mit Downs – so erklären, dass die Parteien auf die Wähler sozusagen Anziehungskräfte ausüben, so dass diese sich „entlang der politischen Skala“ bewegen (für ein Modell, das die Wechselwirkungen zwischen Wählern und Parteien in den Blick nimmt, siehe Troitzsch 1987a, für ein weiteres, das Wechselwirkungen zwischen Parteimitgliedern und Parteifunktionären betrachtet, siehe Troitzsch 1987b). Es lässt sich jedoch zeigen, dass sogar schon bei noch einfacheren Annahmen Polarisierungs- und Depolarisierungseffekte zu beobachten sind. Unterstellt man, dass die Individuen auf irgendeine Weise merken, ob ihre gegenwärtige Einstellung (zu wichtigen politischen Fragen, zu Parteien oder zu Kandidaten) eher eine Minderheiten- oder eher eine Mehrheitenmeinung ist, und unterstellt man ferner, dass diese Individuen ein Interesse daran haben, ihre (jedenfalls in Umfragen zum Ausdruck gebrachte) Einstellung so zu modifizieren, dass sie sie mit einer größeren Anzahl von Menschen gemeinsam haben als vorher, so lässt sich das – ein wenig unscharf, aber vielleicht dafür anschaulicher – so modellieren, dass sich die Individuen auf Downs' „politischer Skala“ in Richtung auf die Häufigkeitsberge bewegen – bei unimodaler Verteilung also zur Mitte, bei bimodaler Verteilung also aus der Mitte und vom Rand weg.

Die Frage ist nun, ob es zu erwarten ist (und wenn ja, unter welchen Bedingungen), dass aus einer unimodalen, nahezu perfekt normalen Anfangsverteilung durch den beschriebenen Mechanismus Polarisierungseffekte erwachsen. Mit verbaler Argumentation und mit mathematischer Analyse lässt sich diese Frage nicht beantworten, sie lässt sich aber sehr wohl in Computerexperimenten untersuchen (vgl. Troitzsch 1990).

Ausgangspunkt der nachstehend beschriebenen Simulation war die Feststellung, dass sich in der dritten Welle der Wahlstudie 1976 die Befragten längs der Links-

Abbildung 6: Häufigkeitsverteilung der Befragten der dritten Welle der Wahlstudie 1976 (Erläuterungen siehe Text)



rechts-Skala deutlich bimodal verteilten. In *Abbildung 6* sind die geschätzte Wahrscheinlichkeitsdichtefunktion (deutlich zweigipflig) und die empirische und die geschätzte Verteilungsfunktion aufgetragen (die beiden letzteren sind kaum zu unterscheiden, was für die Güte der Schätzung spricht).[2]

Ohne dass es hier der Ort ist, ein vollständiges Computersimulationsmodell vorzustellen, lässt es sich abgekürzt folgendermaßen beschreiben:

– Am Beginn eines Simulationslaufs wird eine große Zahl von modellierten Individuen mit zwei voneinander unabhängig normalverteilten Attributen initialisiert, so dass in der Population also die oben postulierte unimodale, nahezu perfekte Normalverteilung zweier Variabler herrscht.
– In jedem Schritt werden nun zunächst aus den Attributen aller einzelnen Individuen die Parameter der simulierten Verteilung berechnet. Wie in jeder Stichprobe weicht die simulierte Verteilung von der Normalverteilung ab.

Nota bene werden nicht die Parameter einer Normalverteilung (deren Wahrscheinlichkeitsdichtefunktion die Exponentialfunktion eines Polynoms bis zum zweiten Grade ist) berechnet, sondern die einer etwas vielfältigeren Verteilung, deren Wahrscheinlichkeitsdichtefunktion die Exponentialfunktion eines Polynoms bis zum vierten (oder sechsten) Grade ist.

Anschließend wird für jedes Individuum berechnet, welche Änderung seiner Attribute es am schnellsten an eine Position höherer Wahrscheinlichkeitsdichte bringt. Diese Änderung der Attribute wird vollzogen, zusätzlich wird dem noch eine zufällige Änderung überlagert. Am Ende eines jeden Schrittes haben alle Individuen ihre Positionen im Einstellungsraum verändert, und im Allgemeinen wird sich auch die Verteilung der Population über dem Einstellungsraum verändert haben.

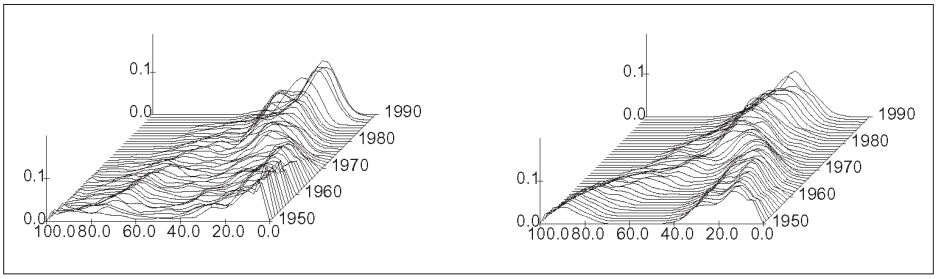
2 Für die Schätzung der Wahrscheinlichkeitsdichtefunktion wurden die Parameter der Exponentialfunktion eines Polynoms bis zum Grade 6 benutzt (die Wahrscheinlichkeitsdichtefunktion der Normalverteilung ist die Exponentialfunktion eines quadratischen Polynoms). Für Einzelheiten sei auf Troitzsch (1987a und 1987b) verwiesen.

Wiederholt man in jedem Schritt die Berechnung der aktuellen Dichtefunktion für die gesamte Population und dann die Veränderung der individuellen Positionen im Einstellungsraum, so wird man – je nach dem, wie stark sich die systematische Positionsveränderung durch den Gradienten der Dichtefunktion gegen die zufällig überlagerte Positionsveränderung durchsetzen kann – eine mehr oder weniger deutliche Polarisierung oder bei einem Überwiegen der zufälligen Komponente eben gar keine nennenswerte Veränderung feststellen.

c) Die Überwindung der Geschlechtertrennung in Gymnasien. In diesem Beispiel wird der Versuch unternommen, die Entwicklung des Frauenanteils in den Lehrerkollegien an sämtlichen Gymnasien des Landes Rheinland-Pfalz in einem möglichst einfachen Simulationsmodell nachzuvollziehen (Kraul und Wirrer 1993; Kraul et al. 1995). Dazu standen jährliche Angaben über die Zahlen der Lehrerinnen und Lehrer an den Gymnasien und Gesamtschulen des Landes Rheinland-Pfalz vom Schuljahr 1950/51 bis 1989/90 zur Verfügung. Zur anschaulicheren Darstellung der Häufigkeitsverteilungen der einzelnen Jahre (*Abbildung 7*) wurden alle Schulen für jedes Schuljahr nach dem Geschlechteranteil bei den Lehrern in eine von 40 Klassen (0 bis 2,5 Prozent Lehrerinnen, ..., 97,5 bis 100 Prozent Lehrerinnen) eingeteilt. Dabei wurde eine Glättung über sieben Klassen hinweg vorgenommen, um die Häufigkeitsverteilung über die Klassen besser sichtbar zu machen. Zur Ermittlung der Häufigkeitsverteilung wurden die Schulen mit den jeweiligen Lehrerzahlen gewichtet.

Die Entwicklung der Häufigkeitsverteilungen ist im linken Teil von *Abbildung 7* dargestellt.

Abbildung 7: Entwicklung der Häufigkeitsverteilung des Lehrerinnenanteils 1950–1990, alle Gymnasien in Rheinland-Pfalz (links: empirische Daten, rechts: Simulation)



1950 gab es zwei etwa gleich große und wichtige Gruppen von Schulen. In der einen – dies ist die Gruppe der als Mädchenschulen gegründeten und mittlerweile koedukativen Gymnasien – betrug der Lehrerinnenanteil typischerweise etwa 80 Prozent, in der anderen – dies ist die Gruppe der als Knabenschulen gegründeten und mittlerweile koedukativen Gymnasien – typischerweise etwa 10 Prozent. Während in der ersten Gruppe der Lehrerinnenanteil bis zum Schuljahr 1973/74 auf etwa 55 Prozent sank, stieg er in der anderen – vor allem zwischen 1967 und 1970 – auf etwa 25 Prozent. Ab dem Schuljahr 1975/76 lässt sich die Gruppeneinteilung nicht mehr darstellen: die Häufigkeitsverteilung ist jetzt eingipflig geworden, Mitte der 1980er Jahre mit einem

Maximum (Modus) bei rund 30 Prozent. In den letzten beiden untersuchten Jahren ist der Modus weiter auf unter 20 Prozent abgesunken.

Es stellte sich nun die Frage, ob sich die empirischen Ergebnisse mit einfachen Annahmen rekonstruieren lassen. Dass der Frauenanteil in den Lehrerkollegien bei etwa einem Drittel stagniert, obwohl seit Jahren etwa gleich viele Frauen und Männer in den Schuldienst eintreten, könnte an der im Allgemeinen geringeren Verweildauer der Frauen liegen – dies war also eine zentrale Annahme.

Simuliert wurden 150 Schulen mit jeweils 20 bis 40 Lehrern, 30 Prozent „Mädchenschulen" mit zu Anfang – d.h. im Jahre 1950 – ca. 88 Prozent Lehrerinnen, 30 Prozent „gemischte Schulen" mit ca. 22 Prozent Lehrerinnen, 40 Prozent „Knabenschulen" mit ca. 8 Prozent Lehrerinnen. Dies entspricht ungefähr der empirischen Ausgangslage. Der Simulation (und damit *Abbildung 7*, rechts) liegen folgende drei Annahmen zugrunde:

– Ausscheidende Lehrer werden je etwa zur Hälfte[3] durch männliche Lehrer und durch Lehrerinnen ersetzt, die bei Eintritt in den Schuldienst 25 bis 30 Jahre alt sind. (Das Wachstum der Lehrerzahlen wird nicht berücksichtigt.)
– Die Verweildauer im Schuldienst ist normalverteilt, bei den Männern mit Mittelwert 30 Jahre, bei den Frauen mit Mittelwert 15 Jahre und bei beiden mit Standardabweichung 5 Jahre; 65-Jährige scheiden stets aus.
– Die Zuweisung von Lehrern und von Lehrerinnen zu Schulen erfolgt nach dem an einer Schule herrschenden Zahlenverhältnis ζ, und zwar nach der Formel

$$P(W \mid \zeta) = \nu(t) \exp(\kappa).$$

Dabei ist $P(W \mid \zeta)$ die Wahrscheinlichkeit, dass an einer Schule mit dem Geschlechterverhältnis ζ eine Frau als Ersatz für einen ausscheidenden (männlichen oder weiblichen) Lehrer eingestellt wird. Der Parameter $\nu(t)$ stellt dabei sicher, dass $P(W)$ über alle Schulen in jedem einzelnen Jahr t gerade 0.5 ist (es ist also zu erwarten, dass auch im Simulationsmodell jedes Jahr gleich viele Männer und Frauen neu in den Schuldienst eintreten).

Für $\kappa > 0$ bedeutet die Formel, dass die Wahrscheinlichkeit, dass ein ausscheidender (männlicher oder weiblicher) Lehrer durch eine Frau ersetzt wird, umso höher ist, je höher der Frauenanteil ($0 \leq \zeta \leq 1$) im jeweiligen Lehrerkollegium bereits ist.

In dem in *Abbildung 7* rechts dargestellten Simulationslauf betrug $\kappa = 0.5$. Damit ist an Schulen mit rein männlichem Kollegium die Wahrscheinlichkeit, dass ein pensionierter Lehrer durch eine Frau ersetzt wird, gerade $\nu(t)$, an Schulen mit rein weiblichem Kollegium ist sie $\nu(t) \exp(0.5)$, also ungefähr $1.65\ \nu(t)$ und damit um knapp zwei Drittel größer.

Für $\kappa >> 1$ gibt es nach einiger Zeit nur noch rein weibliche und rein männliche Lehrerkollegien. Je kleiner κ ist, desto schneller vollzieht sich die Mischung der Lehrerkollegien. Für $\kappa = 0$ ist natürlich die Wahrscheinlichkeit, dass eine Frau eingestellt wird, unabhängig vom bisherigen Geschlechterverhältnis an einer Schule.

---

3 Dies entspricht nicht ganz den empirischen Gegebenheiten, denn bei den in den Vorbereitungsdienst Eingestellten stieg in der Realität der Frauenanteil von 31 Prozent im Jahre 1965/66 auf über 60 Prozent seit dem Schuljahr 1988/89. Ältere Zahlen lagen nicht vor.

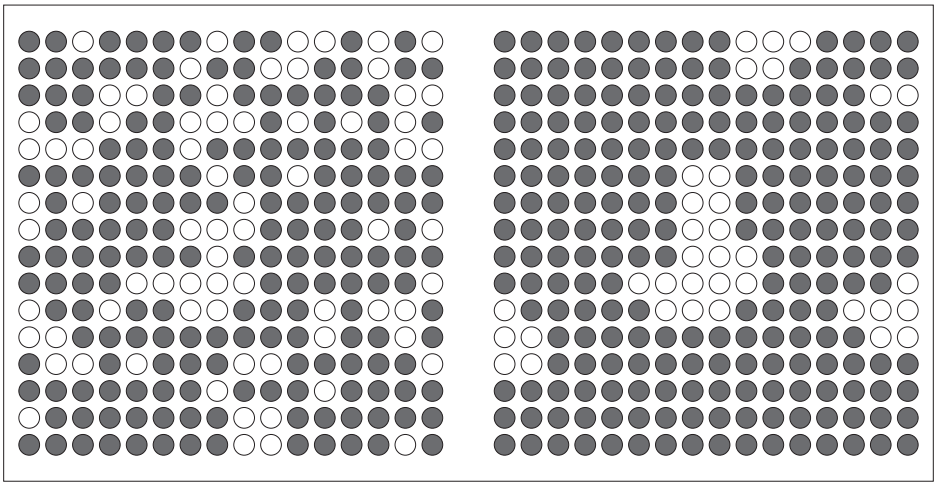
Die Ähnlichkeit zwischen den beiden Graphiken in *Abbildung 7* ist verblüffend. Zusammenfassend lässt sich damit feststellen, dass sich die Entwicklung der Anteile der Lehrerinnen in den Kollegien von Gymnasien in Rheinland-Pfalz mit einem verblüffend einfachen Modell fast quantitativ nachzeichnen lässt: Mit der offenbar früh einsetzenden Einebnung der Barriere zwischen reinen Mädchengymnasien und reinen Knabengymnasien[4] konnten Lehrerinnen und Lehrer in die bisherigen Domänen des jeweils anderen Geschlechts eindringen; der Prozess der Homogenisierung der Kollegien dauerte wie in der Realität etwa 20 Jahre – dann war der Lehrerinnenanteil in allen Schulen im Wesentlichen gleich. Obwohl mindestens seit diesem Zeitpunkt – in der Simulation sogar von Anfang an – etwa gleich viele Männer und Frauen in den Schuldienst an rheinland-pfälzischen Gymnasien eintraten, liegt der Lehrerinnenanteil nach wie vor nur bei etwa 25 bis 30 Prozent, was nach den Ergebnissen der Simulationsstudie allein auf die durchschnittliche längere Verweildauer der männlichen Lehrer zurückgeführt werden kann.

d) Segregation in ursprünglich homogenen Populationen. Zu den frühen Versuchen sozialwissenschaftlicher Computersimulation zählen auch die sogenannten zellularen Automaten (Hegselmann 1996a, b, c), mathematische oder programmiersprachliche Gebilde, die aus einer großen Zahl gleichartiger und lokal interagierender Komponenten bestehen, die man sich am einfachsten als die Zellen eines quadratischen Gitters vorstellen kann, in dem jede Zelle mit ihren vier (von-Neumann-Nachbarschaft) oder acht Nachbarn (Moore-Nachbarschaft) kommuniziert. Der bekannteste zellulare Automat (außerhalb der Sozialwissenschaften) ist zweifellos das „Game of Life" (Gardener 1970), in dem jede Zelle einen von zwei Zuständen – tot oder lebendig – annehmen kann. Die Regel für die Zustandsänderung in jedem Schritt ist einfach: Ist eine Zelle lebendig und hat sie zwei oder drei lebendige Nachbarn, so bleibt sie lebendig, sonst stirbt sie; eine tote Zelle bleibt tot, wenn sie nicht genau zwei lebendige Nachbarn hat. Das „Game of Life" zeigt interessante raum-zeitliche Strukturen; so gibt es etwa Muster, die sich nach einer bestimmten Zahl von Schritten an der gleichen Stelle wiederholen, oder solche, die sich nach einer bestimmten Zahl von Schritten (und Gestaltveränderungen) an einer anderen Stelle wiederholen.

Nach dem gleichen Prinzip der lokalen Wechselwirkungen kann man auch zellulare Automaten bauen, die menschliches Migrationsverhalten abbilden. Anders als in den drei vorhergehenden Abschnitten finden hier die wesentlichen Interaktionen direkt zwischen den Elementen des Systems – also ganz im Sinne der eingangs erwähnten Definition von Mario Bunge – statt und nicht mehr vermittelt über einen Aggregatzustand (auf den die individuellen Systemelemente einwirken und der auf sie zurückwirkt). Ein Klassiker ist der Aufsatz von Schelling (1971) über ethnische Segregation in Wohngebieten, in dem unterstellt wird, dass es zwei Bevölkerungsgruppen gibt, deren

4 Gerade dieser Gesichtspunkt der Einebnung der Barriere zwischen reinen Mädchengymnasien und reinen Knabengymnasien weist allerdings unübersehbar darauf hin, dass der historische Prozess doch ein wenig komplexer war als in dem stark abstrahierenden Simulationsmodell. Diese Barriere musste im Wesentlichen wegfallen, weil es sich Land und Kommunen nicht leisten konnten, außerhalb der Großstädte parallele Gymnasien für Mädchen und für Jungen anzubieten – aber dieser Gesichtspunkt kommt in der Simulation gar nicht vor.

Abbildung 8: Schellings Segregationsmodell; wenn 5 von 8 Nachbarzellen von „andersfarbigen" Nachbarn besetzt sind, wird umgezogen (links: zufällige Anfangsverteilung; rechts der Endzustand: für niemand gibt es mehr einen besseren Platz)



Mitglieder gegenüber den Mitgliedern der jeweils anderen Gruppe eine bestimmte Toleranzschwelle aufweisen, die sie veranlasst, ein anderes Wohnquartier zu suchen, wenn in ihrem bisherigen der Anteil der anderen Gruppe diese Toleranzschwelle übersteigt – jeder solcher Umzug erhöht natürlich den Anteil der anderen Gruppe, was weitere Mitglieder der ersten Gruppe zum Umzug in eine Gegend mit geringerem Anteil der anderen Gruppe veranlasst, mit dem Ergebnis, dass es schließlich nur noch mehr oder weniger einheitlich von der einen oder der anderen Gruppe bewohnte Gebiete gibt. Aus den Experimenten lässt sich ablesen, dass schon ein eher niedriger Schwellenwert für eine weitgehende Segregation ausreicht (bei dem die Leute bereit sind, in einer mehrheitlich andersartigen Umgebung zu leben, vgl. *Abbildung 8*).

e) Ein Modell zur Untersuchung der Entstehung von Zweisprachigkeit. An Hand eines weiteren Modells soll abschließend noch einmal auf die Frage eingegangen werden, in welchem Maße die gewählte Modellierungsstrategie die Ergebnisse von Simulationsmodellen beeinflussen kann. In diesem Modell wird untersucht, ob und gegebenenfalls wie sich Zweisprachigkeit in einer kleinen isolierten Population (z.B. von Auswanderern aus zwei Sprachgemeinschaften in ein noch unbewohntes Territorium) durchsetzt (Troitzsch 2000: 162–168; 2003). In einem Mehrebenenmodell, welches unterstellt, dass in dem modellierten Realsystem alle Individuen ständig miteinander interagieren, ohne dass zwischen ihnen stabile und langfristige Beziehungen entstehen, wird der Erwerb der Zweitsprache abhängig sein davon, wie viele Sprecher der einen oder der anderen Sprache es gibt und wie viele Zweisprachige vorhanden sind: Befindet sich ein Individuum in einer Population, in der fast alle seine Sprache sprechen, besteht kein Bedürfnis, die andere Sprache zu lernen, während die Angehörigen der sprachlichen Minderheit die zweite Sprache erlernen werden. Modelliert man dies zunächst unter

Außerachtlassung von Geburts- und Todesprozessen (der Zeitbedarf für das Erlernen einer zweiten Sprache ist gemessen an der Lebenszeit von Menschen relativ gering), so wird man zunächst ohne komplizierte Formalisierungen feststellen können, dass für jede positive Wahrscheinlichkeit, die andere Sprache zu lernen, über kurz oder lang alle Mitglieder einer solchen Population zweisprachig sein werden: Zweisprachigkeit ist (wenn überhaupt nur zwei Sprachen in Frage stehen) ein „absorbierender Zustand", er wird nie wieder in Richtung Einsprachigkeit verlassen werden. Hier liegt übrigens der Fall vor, bei dem schon im Symbolsystem der natürlichen Sprache (Ostrom 1988) eine eindeutige Schlussfolgerung möglich ist – was natürlich der Tatsache geschuldet ist, dass das bisher beschriebene Modell von extremer Vereinfachung gekennzeichnet ist.

Nun beobachtet man freilich eine bevölkerungsweite Zweisprachigkeit nur außerordentlich selten. Vielmehr sind zwei Phänomene wesentlich häufiger: Entweder sind die Sprecher der Minderheitssprache (z.B. des Sorbischen oder des Rätoromanischen) alle oder fast alle zweisprachig, während die Sprecher der Mehrheitssprache (des Deutschen) weit überwiegend einsprachig bleiben, oder es bildet sich eine mehr oder weniger klare Sprachgrenze heraus (nicht nur entlang von Staatsgrenzen, sondern auch innerhalb eines Staates wie etwa der Schweizer Kantone Bern, Freiburg oder Wallis). Offenbar ist also die bisher beschriebene Modellierung noch zu stark vereinfachend, um die Realität einigermaßen angemessen abzubilden.

Schon die Einführung von Geburts- und Todesprozessen sorgt für eine größere Realitätsnähe; die Einführung des Todes in das Modell macht verständlich, dass für kleine positive Wahrscheinlichkeiten für den Erwerb der anderen Sprache die Lebenszeit von Menschen nicht ausreicht, als dass sich Zweisprachigkeit überall durchsetzen könnte; und die Tatsache, dass auch Kinder zweisprachiger oder verschiedensprachiger Eltern nicht notwendigerweise mit Sicherheit zweisprachig werden, tut ein übriges, um die Durchsetzung von Zweisprachigkeit zu verzögern. Sobald man auch nur feste Wahrscheinlichkeiten einführt, denen gemäß Kinder die eine oder die andere oder gegebenenfalls beide Sprachen ihrer Eltern lernen (wobei es neun Kombinationen von mütterlicher und väterlicher Sprachkombination gibt), und gleichzeitig einen späteren Zweitspracherwerb ausschließt, stellt man fest, dass Zweisprachigkeit überhaupt nur eine Chance hat, wenn wenigstens die Kinder zwei- oder gemischtsprachiger Eltern mit einer Wahrscheinlichkeit größer als 0.5 beide Sprachen lernen (vgl. Troitzsch 2004: 40).

Nun ist der Ausschluss eines späteren Zweitsprachenerwerbs sicher nicht realitätsadäquat. Aber auch die Mischung – die Zweitsprache kann sowohl im Kindesalter als auch später erworben werden – kann noch vollständig im Symbolsystem der Mathematik untersucht werden. Hier ist – weitgehend unabhängig von der mathematischen Formulierung der verschiedenen Wahrscheinlichkeiten, eine zweite Sprache als Kind von Eltern mit der einen oder anderen Kombination von sprachlichen Kompetenzen oder als Erwachsener in einer jeweils unterschiedlich zusammengesetzten Gesamtbevölkerung zu lernen – die qualitative Prognose, dass Zweisprachigkeit bei mindestens der Hälfte der Bevölkerung die wahrscheinlichste Folge sein wird.

Diese qualitative Prognose beruht immer noch auf stark vereinfachenden Annahmen: Sie geht nämlich immer noch davon aus, dass alle Individuen einander ständig begegnen. Damit können sie (im Modell) einerseits also ihre individuelle Entschei-

dung, die zweite Sprache zu erlernen, von einer realistischen Erfahrung über die Sprecheranteile in der Gesamtbevölkerung abhängig machen. Andererseits nimmt das mathematische Modell implizit an, dass Mütter und Väter völlig zufällig zueinander finden, während in der Realität vieles dafür spricht, dass Sprachkompetenz und Partnerwahl miteinander in Beziehung stehen (dies auch noch mathematisch zu modellieren, dürfte zwar möglich sein, aber keine geschlossene Lösung mehr erlauben). In realen Bevölkerungen begegnet aber nicht jede jedem, und nicht alle Mitglieder einer gemischtsprachigen Bevölkerung sind der Notwendigkeit ausgesetzt, die zweite Sprache auch noch zu lernen, weil sie z.B. in einem geographisch abgegrenzten Teilgebiet leben, in dem ihre eigene Sprache eine überwältigende Mehrheit hat (siehe das oben beschriebene Modell der Segregation – in dem wiederum die Fähigkeit von Menschen, ihre Gruppenzugehörigkeit zu ändern, nicht repräsentiert ist).

Erst ein Simulationsmodell, welches sowohl ein Mehrebenenmodell ist als auch die Topographie des Lebensraums der zu modellierenden Bevölkerung und zugleich die Interaktionen zwischen den Individuen in Betracht nimmt, ist in der Lage, das Phänomen realitätsnäher abzubilden – und damit auch die empirisch kaum messbaren Wahrscheinlichkeiten des Zweitspracherwerbs unter den unterschiedlichen Randbedingungen abschätzbar zu machen. In einem solchen Modell werden (anders als im mathematischen Modell, das im Grunde als Variable in einem stochastischen Differentialgleichungssystem nur die Anteile oder Zahlen der verschiedenen Sprachgruppen kennt und damit dem eingangs erwähnten Ansatz der Synergetik zuzurechnen ist) alle einzelnen Individuen durch Softwareagenten repräsentiert,

- die durch Attribute wie Alter, Geschlecht, Sprachkompetenz und Wohnsitz gekennzeichnet sind,
- die älter werden und die mit einem mehr oder weniger dauerhaften Partner Kinder zeugen,
- die die zweite Sprache als Kinder, auf der Partnersuche oder in ihrer jeweiligen Lebensumgebung lernen oder nicht,
- die ihren Wohnsitz verlegen, um einen gleichsprachigen Partner zu finden oder mit Partner und Kindern in eine Gegend zu ziehen, in der ihre eigene Sprache Mehrheitssprache ist.

Mit der Einführung des Attributs „Wohnsitz" wird aus dem synergetischen Modell zugleich auch ein Modell nach der Art der zellularen Automaten, mit der Modellierung von Entscheidungsprozessen (Partner suchen, heiraten, Nachkommen zeugen, umziehen) fällt es zugleich in die Kategorie zumindest einfacher Multiagentenmodelle. Ein solches Modell macht unter der Annahme hoher Wahrscheinlichkeiten des kindlichen Zweitsprachenerwerbs ähnliche qualitative Prognosen wie das mathematische Modell; bei geringeren Wahrscheinlichkeiten kindlichen Zweitsprachenerwerbs zeigt es aber die empirisch eigentlich erwartete geringe Tendenz zur bevölkerungsweiten Zweisprachigkeit.

f) Interaktionen in kleinen Gruppen. Ebenfalls zu den frühen Versuchen der Computersimulation zu Problemen, für die es aussichtslos schien, ein geeignetes mathematisches Modell zu konstruieren, zählt ein Modell von Kirk und Coleman (1967), das untersucht, ob sich in einer Drei-Personen-Gruppe langfristig eine feste Interaktionsstruktur herausbildet – sei es, dass sich ein stabiles Paar herausbildet, das das dritte Mitglied

isoliert, sei es, dass sich eine Hierarchie in der Weise herausbildet, dass ein Mitglied mit beiden anderen interagiert, diese aber nicht mehr miteinander. Wir haben es hier – wie bei den zellularen Automaten – mit lokalen Interaktionen zu tun, die auf der Ebene der Gruppe zur Strukturbildung führen. Das Modell von Kirk und Coleman hat folgenden Aufbau: Die drei Mitglieder haben jeweils eine bestimmte Neigung, mit einem der beiden anderen zu interagieren. Diese Neigung wird durch jede realisierte Interaktion verstärkt. Ob eine Interaktion zustandekommt, entscheidet zu jedem Zeitpunkt ein zufällig vom Simulationssystem ausgewähltes Individuum (das jeweils „dominante"), das sodann mit demjenigen Partner interagiert, zu dem es die höchste Neigung hat. Der dritte Partner geht (in dieser Runde) leer aus. Folgende Variationen sind schon von Kirk und Coleman untersucht worden (vgl. auch Troitzsch 1996: 220–222):

– Die Wahrscheinlichkeit, die Initiative zu ergreifen (dominant zu sein), ist für alle Mitglieder gleich und konstant (kappa = 0), oder sie wächst (kappa > 0) oder sie schrumpft (kappa < 0) mit der bisherigen Anzahl der Interaktionen.
– Das Ausmaß, in dem eine Interaktion die Neigung verstärkt, mit dem gleichen Partner wieder zu interagieren, ist unabhängig davon, ob der Partner der Wunschpartner war (RA = RB); oder die Interaktion mit dem nicht gewünschten Partner erhöht die Neigung, mit ihm noch einmal zu interagieren, stärker (RB > RA) – oder weniger stark (RB < RA) – als das bei einer Interaktion mit dem Wunschpartner der Fall ist.

Je nachdem, in welchem Teil des Parameterraums das Modell spielt, ergeben sich alle möglichen Szenarien – von einer gleichmäßigen Beteiligung aller Mitglieder an den Interaktionen (kappa = 0 und RA = RB) bis zur Paarbildung unter Ausschluss des Dritten (kappa >> 0 und RA = RB) und der Hierarchiebildung, bei dem zwei Mitglieder nicht miteinander, sondern nur mit dem Dritten interagieren (kappa < 0 und RA >> RB).

Mit nur geringer Mühe lässt sich die Simulation leicht auf größere Gruppen erweitern, hier können dann alle denkbaren Formen von Cliquenbildung beobachtet werden (Troitzsch 1996: 223–224).

3. Multiagentenmodelle

Alle bisher erwähnten Ansätze der sozialwissenschaftlichen Computersimulation im engeren Sinne haben gemeinsam, dass menschliche Individuen mit relativ einfachen stochastischen oder deterministischen Automaten modelliert werden, die nur über wenige verschiedene Zustände verfügen. Immerhin entstehen aber selbst in solch einfachen Modellen Verhaltensmuster oder Gruppenstrukturen, wie sie in realen menschlichen Gesellschaften unter den modellierten Bedingungen ebenfalls vorkommen. Die Ergebnisse des zuletzt berichteten Modells der Paarbildung in einer Drei-Personen-Gruppe oder der Cliquenbildung in größeren Gruppen erinnern an Vorgänge, die z.B. unter Kindern, die einander neu kennen lernen, durchaus vorkommen können. Die Parameter des Modells (bei Kirk und Coleman: kappa, RA und RB) sind aber in realen Modellen schwerlich messbar, und selbst darüber hinaus muss man erhebliche Zweifel anmelden, dass ein solches Modell das Zustandekommen von Binnenstrukturen inner-

halb einer solchen Kleingruppe befriedigend erklärt. Denn reale menschliche Individuen sind in der Lage, sich ein Bild von der Gruppe, der sie angehören, zu machen, dieses Wissen strategisch zu nutzen und eine Gruppenstruktur aufrechtzuerhalten, auch wenn sich die äußeren Parameter ändern. Genau dies können die bisher betrachteten stochastischen oder deterministischen Automaten gerade nicht. Im Falle des Modells von Kirk und Coleman „kennt" jeder der drei Automaten gerade einmal seine aktuelle Neigung, mit einem der beiden anderen Automaten zu interagieren, und er „weiß", wie häufig er schon an Interaktionen beteiligt war und wie er nach erfolgter Interaktion seine aktuelle Neigung, mit den beiden anderen zu interagieren, fortschreiben muss, vermag sich aber kein Bild der entstehenden Gruppenstruktur zu machen.

Zu den frühesten Versuchen, in den Sozialwissenschaften Computersimulation einzusetzen, gehören Abelsons und Bernsteins Modell von Referendumskampagnen auf Gemeindeebene (Abelson und Bernstein 1963; vgl. auch Abelson und Caroll 1965; Abelson 1968). Im Sinne einer modernen Klassifikation von Simulationsmodellen dürfte es sich hier um ein Mehrebenenmodell oder um einen frühen Vorläufer von Multiagentenmodellen gehandelt haben. 50 Individuen sind einerseits Informationen ausgesetzt, die verschiedene Kommunikationskanäle über das Referendumsthema verbreiten, andererseits tauschen sie solche Informationen auch untereinander aus. In welchem Maße sie welche Informationen tatsächlich aufnehmen und in welchem Maße sie gegebenenfalls zu Veränderungen in ihren Einstellungen zum Referendumsthema führen, hängt von ihren simulierten Kommunikationsgewohnheiten, aber vor allem auch von den (aktuellen) eigenen Einstellungen ab. Insgesamt haben Abelson und Bernstein 51 Regeln definiert, von denen sich 22 auf Kommunikationskanäle und Informationsquellen und 27 auf den Informationsaustausch der simulierten Individuen untereinander beziehen; die zwei verbleibenden bestimmen das Abstimmungsverhalten am Ende der Referendumskampagne; die Regeln für allgemein zugängliche Kanäle und Quellen und die für die zwischenmenschliche Kommunikation entsprechen einander weitgehend; eine dieser Regeln – A3 bzw. B2 – lautet zum Beispiel: Je extremer die eigene Einstellung ist, desto geringer ist die Empfänglichkeit für neue Informationen.

Auch dieses frühe Modell konnte die modellierten Individuen noch keineswegs mit einem angemessenen Verhaltensrepertoire ausstatten; immerhin zeigt es aber schon ein relativ breites Spektrum an Kommunikationsmöglichkeiten zwischen den modellierten Individuen, das in der klassischen Mikrosimulation nicht angestrebt wurde und auch in Mehrebenenmodellen und zellularen Automaten nicht erreicht wird.

Es ist von einigem Interesse zu untersuchen, wie sich eine künstliche „Gesellschaft" von Automaten verhält, die über ein breiteres Verhaltensrepertoire verfügen. Agentenbasierte soziale Simulation, die diese Untersuchungen ermöglicht, ist unter dem Einfluss der Künstlichen Intelligenz (KI) und insbesondere der Verteilten Künstlichen Intelligenz (VKI) entstanden (Wooldridge und Jennings 1995).

Multiagentensimulation ist seit einigen Jahren dabei, das beherrschende Paradigma der sozialwissenschaftlichen Simulation zu werden (vgl. Gilbert und Doran 1994; Gilbert und Conte 1995; Epstein und Axtell 1996; Conte et al. 1997 und das seit Anfang 1998 erscheinende elektronische Journal of Artificial Societies and Social Simulation (JASSS, http://www..soc.surrey.ac.uk/JASSS/JASSS.html). Es spricht vieles dafür, dass dieser Ansatz alle bisherigen in sich aufnehmen und miteinander integrieren wird (vgl.

Brassel et al. 1997; Mentges 1999 sowie den folgenden Beitrag von Andreas Flache und Michael Macy).

IV. Ausblick

Nicht zuletzt an den Beispielen sollte klar geworden sein, dass Simulation mindestens zwei sehr verschiedenartige Absichten verfolgen kann. Die eine lässt sich als explanatorische Absicht fassen (dies schließt auch Simulation in der Lehre ein), die andere als prognostische. Explanatorische Simulation wird in aller Regel den zeitlichen Vorrang vor der prognostischen haben – Vorhersage ist nicht gut möglich, solange man den Prozess, der vorhergesagt werden soll, noch nicht verstanden hat.

Explanatorische Modelle im Sinne z.B. von Casti (1996: 14) dienen eher dazu, vergangene Beobachtungen zu erklären als künftige vorherzusagen. Casti benutzt, um zu erklären, was ein explanatorisches Modell ist, ein Beispiel, das viel Ähnlichkeit mit vielen konzeptuellen Modellen hat, die in den Sozialwissenschaften in den letzten zwanzig Jahren verwendet worden sind. Es handelt sich um ein Modell des Prozesses der Planetenentstehung in frühen Sonnensystemen, das es uns erlaubt vorherzusagen, was für Planetensysteme man unter verschiedenartigen physikalischen Umgebungsbedingungen wohl im Weltraum entdecken könnte. Ganz ähnliche Sätze könnte man formulieren – „was für Sozialsysteme man unter verschiedenartigen physischen, psychischen und sozialen Umgebungsbedingungen wohl in menschlichen Gesellschaften entdecken könnte ...“ –, wenn man z.B. von den zellularen Automaten spricht, die Hegselmann, Latané und Nowak entwickelt haben, oder von den künstlichen Welten, die Castelfranchi, Conte, Doran, Drogoul und andere geschaffen und beschrieben haben (siehe zahlreiche Beiträge in Gilbert und Doran 1994; Gilbert und Conte 1995; Hegselmann et al. 1996; Troitzsch et al. 1996; Conte et al. 1997) und sogar auch von einem so frühen Modell wie dem Papierkorb-Modell der Organisation (Cohen et al. 1972). Sie alle berühren das Problem, welche Muster aus den Handlungen und Interaktionen individueller Agenten entstehen, die bei ihren Handlungen und Interaktionen gar keine Pläne im Hinblick auf irgendwelche Muster und Strukturen in den Gruppen, Organisationen oder Gesellschaften haben, zu denen sie gehören. Und das ist ja nach Hayek (1942: 288) eins der zentralen Probleme der theoretischen Sozialwissenschaften.

Einige der Beispiele haben jedoch auch gezeigt, dass die Verknüpfung von Simulationsmodell und empirischer Realität nicht immer im Vordergrund der Überlegung der verschiedenen Autoren gestanden hat. Die klassischen Modelle – Weltmodelle und mikroanalytische Simulationsmodelle – waren und sind immer auf die Vorhersage künftiger Zustände ausgerichtet und werden daher immer auf ihre prädiktive oder replikative (Zeigler 1976: 5) Validität geprüft; für Warteschlangenmodelle, auch wenn ihr Anwendungsbereich eher am Rande der Sozialwissenschaften liegt, gilt das gleiche, denn sie dienen im Allgemeinen zur Optimierung von Abläufen oder der Untersuchung der Konsequenzen organisatorischer Maßnahmen, und man wird sich überzeugen, ob sich die im Modell vorgenommene Optimierung auch in der Realität erreichen lässt oder ob die vorhergesagten Konsequenzen tatsächlich eingetreten sind. Das Beispiel der Simulation der Lehrerinnenanteile an rheinland-pfälzischen Gymnasien hat eine hohe re-

plikative, aber eine bescheidene strukturelle (Zeigler 1976: 5) Validität erkennen lassen: Zwar wird die Entwicklung quantitativ recht gut nachgebildet, wir wissen aber, dass das Kultusministerium gerade keine von Wahrscheinlichkeitsfunktionen diktierten Zufallsauswahlen unter Bewerberinnen und Bewerbern um eine Stelle an einem Gymnasium vorgenommen hat, wie das Simulationsmodell es unterstellt. Und was die Validierung der übrigen Beispielsimulationen angeht, so steht beim Segregationsmodell oder beim Bilingualismus-Modell keinesfalls Vorhersage oder Replikation historischer Prozesse in realen Städten oder auf realen zuvor unbewohnten Inseln im Vordergrund, sondern hier wird ganz im Gegenteil versucht, Modelle zu entwerfen, die „nicht nur das beobachtete Systemverhalten reproduzieren, sondern wahrhaftig die Art und Weise reflektieren, in der das jeweilige reale System operiert um dieses Verhalten hervorzubringen“ (Zeigler 1976: 5) – auch wenn hier in Wirklichkeit nicht „das beobachtete Systemverhalten“ reproduziert wird, sondern nur der „stylised fact“ (Kaldor 1961: 178).

Literatur

Abelson, Robert P., 1968: Simulation of Social Behavior. S. 274–356 in: *Gardner Lindzey* und *Elliot Aronson* (Hg.), Handbook of Social Psychology 2. Reading, MA: Addison-Wesley.

Abelson, Robert P., und *A. Bernstein*, 1963: A Computer Simulation of Community Referendum Controversies. Public Opinion Quarterly 27(1): 93–122.

Abelson, Robert P., und *J. Douglas Carroll*, 1965: Computer Simulation of Individual Belief Systems. American Behavioral Scientist 8: 24–30.

Antcliff, Susan, 1993: An Introduction to DYNAMOD: A Dynamic Microsimulation Model. Technical Report 1. National Centre for Social and Economic Modelling (NATSEM). Canberra: University of Canberra.

Balzer, Wolfgang, C. Ulises Moulines und *Joseph D. Sneed*, 1987: An Architectonic for Science. The Structuralist Program. Dordrecht: Reidel.

Bartholomew, David J., 1982: Stochastic Models for Social Processes, 3. Auflage. Ann Arbor: UMI Books on Demand.

Brassel, Kai H., Michael Möhring, Elke Schumacher und *Klaus G. Troitzsch*, 1997: Can Agents Cover all the World? S. 55–72 in: *Rosaria Conte, Rainer Hegselmann* und *Pietro Terna* (Hg.), Simulating Social Phenomena. Bd. 456 of Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag.

Bunge, Mario, 1979: Ontology II: A World of Systems. Treatise on Basic Philosophy 4. Dordrecht: Reidel.

Casti, John L., 1996: Would-Be Worlds. How Simulation Is Changing the Frontiers of Science. New York: Wiley.

Cohen, Michael D., James G. March und *Johan P. Olsen*, 1972: A Garbage Can Model of Organizational Choice. Administrative Sciences Quarterly 17: 1–25.

Conte, Rosaria, Rainer Hegselmann und *Pietro Terna*, 1997: Simulating Social Phenomena. Berlin: Springer-Verlag.

Diekmann, Andreas, 1980: Dynamische Modelle sozialer Prozesse. München: Oldenbourg.

Diekmann, Andreas, und *Peter Mitter*, 1984: Stochastic Modelling of *Social Processes.* Orlando: Academic Press.

Doreian, Patrick, und *Norman P. Hummon*, 1976: Modeling Social Processes. New York: Elsevier.

Downs, Anthony, 1968: Ökonomische Theorie der Demokratie. Tübingen: J.C.B. Mohr (Paul Siebeck).

Epstein, Joshua M., und *Robert Axtell*, 1996: Growing Artificial Societies – Social Science from the Bottom up. Cambridge, MA: MIT Press.

Forrester, Jay W., 1968: Principles of Systems. Cambridge, MA: MIT Press.
Forrester, Jay W., 1971: World Dynamics. Cambridge, MA: MIT Press.
Fucks, Wilhelm, 1965: Formeln zur Macht. Prognosen über Völker, Wirtschaft, Potentiale. Stuttgart: Deutsche Verlagsanstalt.
Galler, Heinz P., 1990: Verwandtschaftsnetzwerke im demographischen Modell – Ergebnisse einer Modellrechnung. Acta Demographica: 1.
Galler, Heinz P., 1997: Discrete-time and Continuous-time Approaches to Dynamic Microsimulation Reconsidered. Technical Report 13. National Centre for Social and Economic Modelling (NATSEM). Canberra: University of Canberra.
Gardener, Martin, 1970: The Game of Life. Scientific American 223(4): April.
Gilbert, Nigel, und *Rosaria Conte*, 1995: Artificial Societies: The Computer Simulation of Social Life. London: UCL Press.
Gilbert, Nigel, und *Jim E. Doran*, 1994: Simulating Societies: The Computer Simulation of Social Phenomena. London: UCL Press.
Gilbert, Nigel, und *Klaus G. Troitzsch*, 2005: Simulation for the Social Scientist. 2. Aufl. Maidenhead/New York: Open University Press.
Haken, Hermann, 1978: Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology. Springer Series in Synergetics. Vol. 1. 2nd enlarged edition. Berlin: Springer-Verlag.
Haken, Hermann, 1982: Synergetik. Eine Einführung. Berlin: Springer-Verlag.
Haken, Hermann, 1996: Synergetik und Sozialwissenschaften. Ethik und Sozialwissenschaften. Streitforum für Erwägungskultur 7(4): 587–594.
Harbordt, Steffen, 1974: Computersimulation in den Sozialwissenschaften. 1: Einführung und Anleitung. Reinbek: Rowohlt.
Harbordt, Steffen, 1974: Computersimulation in den Sozialwissenschaften. 2: Beurteilung und Modellbeispiele. Reinbek: Rowohlt.
Harding, Ann (Hg.), 1996: Microsimulation and Public Policy. Volume 232 of Contributions to Economic Analysis. Amsterdam: Elsevier North Holland.
Hauff, Volker, und *Friedrich Latzelsberger*, 1968: Simulationsmodelle in der soziologischen Forschung. S. 127–147 in: *Rul Gunzenhäuser* (Hg.), Nichtnumerische Informationsverarbeitung. Beitäge zur Behandlung nicht-numerischer Probleme mit Hilfe von Digitalrechenanlagen. Wien: Springer-Verlag.
Hauser, Richard, Uwe Hochmuth und *Johannes Schwarze*, 1994a: Mikroökonomische Grundlagen der Gesellschaftspolitik. Band 1: Ausgewählte Probleme und Lösungsansätze. Ergebnisse aus dem gleichnamigen Sonderforschungsbereich an den Universitäten Frankfurt und Mannheim. Berlin: Akademie-Verlag.
Hauser, Richard, Notburga Ott und *Gert Wagner*, 1994b: Mikroökonomische Grundlagen der Gesellschaftspolitik. Band 2: Erhebungsverfahren, Analysemethoden und Mikrosimulation. Ergebnisse aus dem gleichnamigen Sonderforschungsbereich an den Universitäten Frankfurt und Mannheim. Berlin: Akademie-Verlag.
Hayek, Friedrich A., 1942–1944: Scientism and the Study of Society. Economica 9, 10, 11: 267–291, 34–63, 27–39.
Hegselmann, Rainer, 1996a: Cellular Automata in the Social Sciences. Perspectives, Restrictions and Artefacts. S. 209–230 in: *Rainer Hegselmann, Ulrich Mueller* und *Klaus G. Troitzsch* (Hg.), Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Theory and Decision Library, Series A: Philosophy and Methodology of the Social Sciences. Dordrecht: Kluwer.
Hegselmann, Rainer, 1996b: Solidarität unter Ungleichen. S. 105–128 in: *Rainer Hegselmann* und *Heinz-Otto Peitgen* (Hg.), Modelle sozialer Dynamiken. Ordnung, Chaos und Komplexität. Wien: Hölder-Pichler-Tempsky.
Hegselmann, Rainer, 1996c: Understanding Social Dynamics: The Cellular Automata Approach. S. 282–306 in: *Klaus G. Troitzsch, Ulrich Mueller, Nigel Gilbert* und *Jim E. Doran* (Hg.), Social Science Microsimulation. Berlin: Springer-Verlag.

Hegselmann, Rainer, Ulrich Mueller und *Klaus G. Troitzsch* (Hg.), 1996: Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Theory and Decision Library, Series A: Philosophy and Methodology of the Social Sciences. Dordrecht: Kluwer.

Helbing, Dirk, und *Wolfgang Weidlich*, 1995: Quantitative Soziodynamik. Gegenstand, Methodik, Ergebnisse und Perspektiven. Kölner Zeitschrift für Soziologie und Sozialpsychologie 47: 114–140.

Heike, Hans-Dieter, C. Fleck, Harald Ritz und *Thomas Sauerbier*, 1996a: Object-Oriented Approaches in Micro and Macro Simulation. S. 417–427 in: *Frank Faulbaum* und *Wolfgang Bandilla* (Hg.), SoftStat'95. Advances in Statistical Software 5. Stuttgart: Lucius & Lucius.

Heike, Hans-Dieter, *Kai Beckmann, Achim Kaufmann, Harald Ritz* und *Thomas Sauerbier*, 1996b: A Comparison of a 4GL and an Object-Oriented Approach in Micro and Macro Simulation. S. 3–32 in: *Klaus G. Troitzsch, Ulrich Mueller, Nigel Gilbert* und *Jim E. Doran* (Hg.), Social Science Microsimulation. Berlin: Springer-Verlag.

Heike, Hans-Dieter, und *Thomas Sauerbier*, 1997: MISTRAL – A New Object-based Micro Simulation Language. S. 403–410 in: *Frank Faulbaum* und *Wolfgang Bandilla* (Hg.), SoftStat'97. Advances in Statistical Software 6. Stuttgart: Lucius & Lucius.

Homans, George C., 1950: The Human Group. New York: Harpers.

Kaldor, Nicholas, 1961: Capital Accumulation and Economic Growth. S. 177–222 in: *F.A. Lutz* und *D.C. Hagur* (Hg.), The Theory of Capital. New York: St. Martin's Press.

Kheir, Naim A., 1988: Systems Modeling and Computer Simulation. Vol. 1 of Electrical Engineering and Electronics. New York: Marcel Dekker.

Kirk, Jerome, und *James S. Coleman*, 1967: Formalisierung und Simulation von Interaktionen in einer Drei-Personen-Gruppe. S. 169–190 in: *Renate Mayntz* (Hg.), Formalisierte Modelle in der Soziologie. Soziologische Texte Band 39. Neuwied: Luchterhand.

Kliemann, Wolfgang, und *Norbert Müller*, 1973: Logik und Mathematik für Sozialwissenschaftler 1. Grundlagen formalisierter Modelle in den Sozialwissenschaften. München: Fink.

Kliemann, Wolfgang, und *Norbert Müller*, 1976: Logik und Mathematik für Sozialwissenschaftler 2. Bestandteile und Methoden formalisierter Modelle in den Sozialwissenschaften. München: Fink.

Kraul, Margret, Klaus G. Troitzsch und *Rita Wirrer*, 1995: Lehrerinnen und Lehrer an Gymnasien: Empirische Ergebnisse aus Rheinland-Pfalz und Resultate einer Simulationsstudie. S. 334–340 in: *Heinz Sahner* und *Stefan Schwendtner* (Hg.), 27. Kongress der Deutschen Gesellschaft für Soziologie, Halle an der Saale 1995. Kongressband II: Berichte aus den Sektionen und Arbeitsgruppen. Opladen: Westdeutscher Verlag.

Kraul, Margret, und *Rita Wirrer*, 1993: Einführung der Koedukation – Bildungspolitische oder programmatische Entscheidung? Die Deutsche Schule 85: 84–97.

Liedtke, B.H., und *Hermann Quinke*, 1997: AsA – Analysesystem Alterssicherung. S. 419–426 in: *Frank Faulbaum* und *Wolfgang Bandilla* (Hg.), SoftStat'97. Advances in Statistical Software 6. Stuttgart: Lucius & Lucius.

Malthus, Thomas R., 1798: An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers. London.

Mayntz, Renate (Hg.), 1967: Formalisierte Modelle in der Soziologi*e*. Soziologische Texte 39. Luchterhand: Neuwied.

Martinez Coll, Juan Carlos, 1986: A Bioeconomic Model of Hobbes' ‚State of Nature'. Social Science Information 25(2): 493–505.

Meadows, Dennis, Donella Meadows, Erich Jahn und Peter Milling, 1972: Die Grenzen des Wachstums. Bericht des Club of Rome zur Lage der Menschheit. Stuttgart: Deutsche Verlagsanstalt.

Meadows, Dennis L., William W. Behrens III, Donella H. Meadows, *Roger F. Naill, Jørgen Randers* und *Erich K.O. Zahn*, 1974: The Dynamics of Growth in a Finite World. Cambridge, MA: MIT Press.

Meadows, Donella H., Dennis L. Meadows und *Jørgen Randers*, 1992: Die neuen Grenzen des Wachstums. Die Lage der Menschheit: Bericht und Zukunftschancen. Stuttgart: Deutsche Verlagsanstalt.

Mentges, Elke, 1999: Concepts for an Agent-based Framework for Interdisciplinary Social Science Simulation. Journal of Artificial Societies and Social Simulation 2(2). http://www.soc.surrey.ac.uk/JASSS/2/2/4.html.

Merz, Joachim, 1996a: MICSIM – A PC Microsimulation Model for Research and Teaching: A Brief View on its Concept and Program Developments. S. 433–442 in: *Frank Faulbaum* und *Wolfgang Bandilla* (Hg.), SoftStat'95. Advances in Statistical Software 5. Stuttgart: Lucius & Lucius.

Merz, Joachim, 1996b: MICSIM: Concept, Developments, and Applications of a PC Microsimulation Model for Research and Teaching. S. 33–65 in: *Klaus G. Troitzsch, Ulrich Mueller, Nigel Gilbert* und *Jim E. Doran* (Hg.), Social Science Microsimulation. Berlin: Springer-Verlag.

Müller, Norbert, 1989*:* Civilization Dynamics I – Fundamentals of a Model-Oriented Description. Aldershot: Avebury.

Müller, Norbert, 1991: Civilization Dynamics II – Nine Simulation Models. Aldershot: Avebury.

Orcutt, Guy H., Joachim Merz und *Hermann Quinke* (Hg.), 1986: Microanalytic Simulation Models to Support Social and Financial Policy. Information Research and Resource Reports, Band 7. Amsterdam: North-Holland.

Ostrom, Thomas, 1988: Computer Simulation: The Third Symbol System. Journal of Experimental Social Psychology 24: 381–392.

Rapoport, Anatol, 1980: Mathematische Modelle in den Sozialwissenschaften. Würzburg/Wien: Physica-Verlag.

Richardson, Lewis F., 1948: War Moods. Psychometrika 13: 147–174, 197–232.

Sauerbier, Thomas, 2002: UMDBS – A New Tool for Dynamic Microsimulation. Journal of Artificial Societies and Social Simulation 5: no. 2 <http://jasss.soc.surrey.ac.uk/5/2/5.html>

Schelling, Thomas C., 1971: Dynamic Models of Segregation. Journal of Mathematical Sociology 1: 143–186.

Simon, Herbert A., 1952: A Formal Theory of Interaction in Social Groups. American Sociological Review 17: 202–221.

Simon, Herbert A., 1957: Models of Man, Social and Rational. Mathematical Essays on Rational Human Behavior in a Social Setting. New York: Wiley.

Sneed, Joseph D., 1979: The Logical Structure of Mathematical Physics. Synthese Library 35. 2nd edition. Dordrecht: Reidel.

Stegmüller, Wolfgang, 1979: The Structuralist View of Theories. A Possible Analogue of the Bourbaki Programme in Physical Science. Berlin: Springer-Verlag.

Troitzsch, Klaus G., 1987a: Bürgerperzeptionen und Legitimierung. Anwendung eines formalen Modells des Legitimations-/Legitimierungsprozesses auf Wählereinstellungen und Wählerverhalten im Kontext der Bundestagswahl 1980. Frankfurt a.M.: Lang.

Troitzsch, Klaus G., 1987b: Interactions Between Party Members and Local Party Activists: A Formal Model and Its Empirical Application. Journal of Mathematical Sociology 12: 415–438.

Troitzsch, Klaus G., 1989: Chaotisches Verhalten in einem Sozialsystem. Gegenüberstellung eines Makro- und eines Mikromodells. S. 173–191 in: *Ali B. Çambel, Bruno Fritsch* und *Jürgen W. Keller (*Hg.), Dissipative Strukturen in Integrierten Systemen. Baden-Baden: Nomos.

Troitzsch, Klaus G., 1990: Self-organisation in Social Systems. S. 353–377 in: *Johannes Gladitz* und *Klaus G. Troitzsch* (Hg.), Computer Aided Sociological Research. Proceedings of the Workshop „Computer Aided Sociological Research" (CASOR'89), Holzhau/DDR, October 2nd–6th, 1989. Berlin: Akademie Verlag.

Troitzsch, Klaus G., 1996: Individuelle Einstellungen und kollektives Verhalten. S. 200–228 in: *Günter Küppers* (Hg.), Chaos und Ordnung. Formen der Selbstorganisation in Natur und Gesellschaft. Reclams Universalbibliothek, Band 9434. Stuttgart: Reclam.

Troitzsch, Klaus G. 2000: Dynamik interagierender Bevölkerungen. S. 135–183 in: *Ulrich Mueller, Bernhard Nauck* und *Andreas Diekmann* (Hg.), Handbuch der Demographie 1: Modelle und Methoden. Berlin: Springer.

Troitzsch, Klaus G., 2003: Simulation in den Sozialwissenschaften. S. 353–363 in: *Barbara Orth, Thomas Schwietring* und *Johannes Weiß* (Hg.), Soziologische Forschung: Stand und Perspektiven. Opladen: Leske + Budrich.

Troitzsch, Klaus G., 2004: A Multi-agent Model of Bilingualism in a Small Population. S. 38–43 in: *Helder Coelho, Bernard Espinasse* und *Martina-M. Seidel* (Hg.), 5th Workshop on Agent-Based Simulation, Erlangen. San Diego: SCS Publishing House.

Troitzsch, Klaus G., Ulrich Mueller, Nigel Gilbert und *Jim E. Doran,* 1996: Social Science Microsimulation. Berlin: Springer-Verlag.

Troitzsch, Klaus G., Michael Möhring, Ulrich Meyer, Friedebert Kröger, Günter Schiepek und *A. Hendrischke*, 1998: Verhaltenskoordination im Katastrophenfall – Ein Simulationsmodell. S. 443–450 in: *Max Engeli* und *Veronika Hrdlicka* (Hg.), ASIM'98 Zürich. 12. Symposium Simulationstechnik. ASIM, FA 4.5 der Gesellschaft für Informatik. Zürich: vdf Hochschulverlag an der ETH Zürich.

Tuma, Nancy B., und *Michael T. Hannan*, 1984: Social Dynamics. Models and Methods. Orlando: Academic Press.

Weidlich, Wolfgang, und *Günter Haag*, 1983: Concepts and Models of a Quantitative Sociology. The Dynamics of Interacting Populations. Berlin: Springer.

Wooldridge, Michael, und *Nick R. Jennings*, 1995: Intelligent Agents: Theory and Practice. Knowledge Engineering Review 10(2): 115–152.

Zeigler, Bernhard P., 1976: Theory of Modelling and Simulation. New York: Wiley.

Ziegler, Rolf, 1972: Theorie und Modell. Der Beitrag der Formalisierung zur soziologischen Theoriebildung. München: Oldenbourg.

„BOTTOM-UP“ MODELLE SOZIALER DYNAMIKEN

Agentenbasierte Computermodellierung und methodologischer Individualismus*

Andreas Flache und Michael W. Macy

Zusammenfassung: Agentenbasierte Computermodellierung ist ein noch junger Ansatz, der theoretisch zu modellieren versucht, wie komplexe soziale Makrodynamiken durch Interaktionen autonomer, aber interdependenter individueller Akteure („Agenten“) „bottom-up“ entstehen. Unser Beitrag beschreibt die zentralen Prinzipien agentenbasierter Computermodellierung und setzt sie dann zum Paradigma des methodologischen Individualismus in Beziehung. Wir argumentieren, dass die agentenbasierte Computermodellierung methodologischen Individualisten ein Werkzeug bietet, das die Striktheit formaler Modellbildung mit einer größeren Freiheit in der Wahl der Verhaltensannahmen verbinden kann als es der alternative Ansatz der formalen Spieltheorie erlaubt. Wir illustrieren diese These an einem Vergleich agentenbasierter Modelle sozialer Ordnung mit spieltheoretischen Analysen. Dabei werden insbesondere Modelle besprochen, die den Einfluss von stabilen Beziehungen, Netzwerkstrukturen und Verhaltensmechanismen auf spontan entstehende soziale Ordnung untersuchen. Wir schließen unseren Beitrag mit einer Diskussion methodologischer Prinzipien für eine fruchtbare Anwendung agentenbasierter Computermodellierung in der Sozialtheorie ab.

I. Einleitung

Was verbindet Vogelschwärme, Modewellen, Krawalle und ethnisch segregierte Wohnviertel miteinander? Die Antwort lautet: Selbstorganisation. Die tanzartigen Bewegungen eines Vogelschwarmes entstehen, ohne dass ein Leitvogel sie choreographiert. Krawalle entwickeln sich auch ohne Anführer. Ethnisch homogene Ghettos entstehen, auch ohne dass Hypothekbanken und Makler die Entscheidungen der Wohnungssuchenden entsprechend beeinflussen.

Soziologen versuchen traditionell, soziale Strukturen als Folge eines Systems von Institutionen und Normen zu begreifen, die „top-down“ das Verhalten von Individuen steuern. Im Gegensatz dazu geht ein neuerer Ansatz zur Modellierung sozialer Systeme davon aus, dass soziale Strukturen größtenteils „bottom-up“ entstehen, eher mit improvisiertem Jazz vergleichbar als mit einem klassischen Symphoniekonzert. Soziale Akteure spielen nicht einfach die Rollen, die ihnen von Eliten zugedacht werden. Wir schrei-

* Die Forschungsarbeit von Andreas Flache wurde durch die Königliche Niederländische Akademie der Wissenschaften ermöglicht. Die Autoren danken Robb Willer, Damon Centola und einem anonymen Gutachter für wichtige Anregungen zu dieser Arbeit und Andreas Diekmann für hilfreiche Kommentare und seine Geduld bei der Begleitung der Manuskripterstellung.

ben unsere Rollen erst im Laufe der Aufführung. Wie aber ist soziale Ordnung ohne zentrale Koordination möglich?

Seit kurzem verwenden Sozialtheoretiker agentenbasierte Computermodellierung (kurz: ABCM), um zu neuen Antworten zu gelangen. Agentenbasierte Computermodellierung wurde durch die Entwicklung so genannter Multiagentensysteme (MAS) in der Künstlichen Intelligenz (KI) inspiriert. Multiagentensysteme lösen komplexe Informationsverarbeitungsprobleme durch das Zusammenwirken autonomer Softwareagenten. „Agenten" können eigene Berechnungen ausführen und haben einen internen Zustand und ein eigenes lokales Wissen, wechseln aber Informationen mit anderen Agenten aus und reagieren auf deren Input. In der KI wird gefordert, dass autonome Softwareagenten vier Hauptmerkmale aufweisen: Autonomie, Reaktivität, Proaktivität und Sozialfähigkeit (Troitzsch 2000; vgl. Wooldridge und Jennings 1995; Schmidt 2001). Autonome Agenten kontrollieren ihre eigenen Zielsetzungen, ihr Verhalten und ihren internen Zustand. Reaktive Agenten nehmen ihre Umgebung wahr und reagieren auf sie. Proaktivität bezeichnet die Fähigkeit, aktiv auf Aspekte der Umgebung einzuwirken, so dass individuelle Ziele besser erreicht werden. Sozialfähigkeit meint schließlich das Vermögen, auf die Einflüsse anderer Agenten zu reagieren und diese wiederum gezielt zu beeinflussen.

In der agentenbasierten Computermodellierung wurden Multiagentensysteme bald auch als Modell sozialer Systeme aufgefasst. Das Hauptinteresse soziologischer Anwender richtet sich dabei meist auf die Entstehung von Makrophänomenen aus einfachen individuellen Verhaltensregeln. Durch die Betonung der Einfachheit von Verhaltensregeln wurden in sozialwissenschaftlichen Anwendungen der ABCM die Agenten zwar generell als autonom und adaptiv modelliert, aber nicht unbedingt als proaktiv und sozialfähig (vgl. Macy und Willer 2002). Damit gehören selbst frühe Modelle sozialer Selbstorganisation zur ABCM, wie etwa Schellings (1971) bekanntes Modell sozialer Segregation. Neuere Anwendungen der ABCM richten sich zum Beispiel auf die Entstehung und Dynamik kultureller Differenzierung (z.B. Mark 2003; Axelrod 1997), soziale Netzwerkstrukturen (z.B. Klüver und Stoica 2003; Stokman und Zeggelink 1996), Normkonformität und kollektives Handeln (Mosler und Brucks 2003; Mosler und Tobias 2000; Lepperhof 2000; Macy 1990), Gerechtigkeitsnormen bei Verhandlungen und Verteilungsproblemen (Vieth 2003; Vetschera 2003) oder Kooperation in dyadischen Tauschbeziehungen (Macy und Flache 2002).

Die Vorgehensweise bei agentenbasierter Computermodellierung soll an Sakodas frühem Computermodell der Segregation sozialer Gruppen illustriert werden (Sakoda 1971). Sakoda untersucht in seiner Arbeit, wie soziale Segregation, etwa zwischen ethnischen Gruppierungen, mit den unterschiedlichen Einstellungsmustern individueller Gruppenmitglieder zur eigenen und zur anderen Gruppe zusammenhängt.[1] Zwei zentrale Bausteine eines agentenbasierten Computermodells sind die formale Beschreibung der Handlungsmöglichkeiten individueller Akteure und die Spezifikation des zu untersuchenden Makrophänomens als Funktion individueller Handlungen. In Sakodas Modell „leben" Angehörige zweier Gruppen in einer schachbrettartigen, so genannten zel-

1 Das unter Soziologen viel bekanntere Segregationsmodell von Schelling (1971) ist dem Modell von Sakoda sehr ähnlich, unterscheidet aber nicht zwischen verschiedenen möglichen Einstellungsmustern.

lularen Welt (vgl. Hegselmann und Flache 1998), deren Felder (Zellen) nicht vollständig besetzt sind. Individuen können zu leeren Zellen innerhalb eines kleinen Migrationsfensters der Größe 3×3 wandern. Falls dort keine leeren Zellen vorhanden sind, kann innerhalb einer 5×5 Nachbarschaft gewandert werden. Segregation auf der Makroebene ist umso stärker, je näher Individuen sich bei eigenen Gruppenmitgliedern befinden und je weiter weg sie sich von Mitgliedern der anderen Gruppe positionieren.

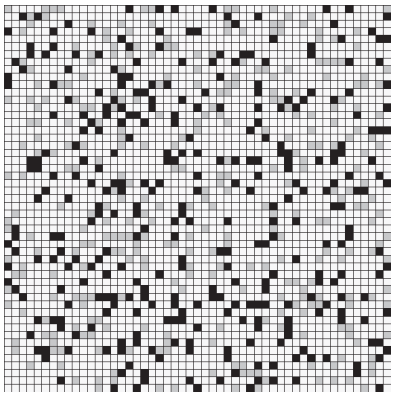
Das Kernstück der ABCM ist die Spezifikation der individuellen Entscheidungsregeln. Diese Regeln legen fest, wie Agenten auf ihre Umgebung reagieren. In Sakodas Modell haben alle Akteure positive, neutrale bzw. negative Einstellungen („Valenzen": +1, 0, –1) gegenüber den Mitgliedern der eigenen bzw. der anderen Gruppe. Sakodas Individuen wählen ihre Standorte nach der Regel, dass dort die Summe der gewichteten Valenzen, aufsummiert über die ganze Welt, maximal ist.[2] Das Gewicht jeder Valenz nimmt mit der Entfernung ab. Agenten verfolgen also das Ziel, sich im sozialen Raum möglichst nahe bei anderen Agenten zu befinden, die sie positiv bewerten und möglichst weit entfernt von jenen zu sein, die negativ bewertet werden. In Sakodas Modell können Agenten ihre Standorte sequentiell, in zufälliger Reihenfolge, verändern. Hierbei reagieren sie auf die jeweils aktuelle Situation in ihrer Umgebung.

Sakodas Model illustriert, dass ABCM kontraintuitive Makrokonsequenzen einfacher individueller Verhaltensregeln aufzeigen kann. Sakoda vergleicht in seiner Analyse zwei unterschiedliche Einstellungsmuster, die er „segregation" und „mutual suspicion" nennt. Gemeinsam ist beiden Einstellungsmustern die jeweils *negative* Einstellung gegenüber Angehörigen der jeweils *anderen* Gruppe. Unter dem *segregation* genannten Einstellungsmuster stehen beide Gruppen ihren *eigenen* Angehörigen *positiv*, unter dem *mutual suspicion* genannten Muster hingegen *neutral* gegenüber.

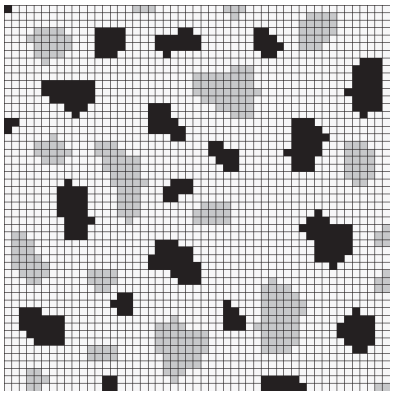
ABCM ist vor allem dann fruchtbar, wenn Forscher zuvor explizit gemachte Hypothesen über das zu erwartende Modellverhalten mit Computersimulationen oder analytischen Methoden überprüfen. Die von Sakoda gewählten Bezeichnungen für die Einstellungsmuster deuten seine Erwartungen an. Es erscheint zunächst plausibel, dass das Einstellungsmuster *segregation* zu mehr Segregation auf der Makroebene führt als *mutual suspicion*. Beide Einstellungsmuster sehen für das Verhalten der Individuen die Tendenz vor, sich von den Angehörigen der jeweils anderen Kategorie zu entfernen. Bei *segregation* versuchen die Agenten aber darüber hinaus, den vor der anderen Kategorie fliehenden Mitgliedern der eigenen Gruppe nahe zu kommen. Dies könnte zu der Vermutung führen, dass sich die Gruppen im Endergebnis bei *segregation* räumlich stärker voneinander trennen werden. Sakodas ursprüngliche Simulation der Modelldynamik (Sakoda 1971: 126ff.) in einer Welt mit 8x8 Feldern und 2 Gruppen mit je 6 Agenten hat diese Vermutung aber nicht bestätigt. Sakoda programmierte sein Modell in der Sprache FORTRAN und ließ das Programm ausgehend von zufällig gewählten Startsituationen die Veränderung der Verteilung der Agenten berechnen, die sich aus den von ihm definierten Entscheidungsregeln ergab. Sakoda konnte im Makroergebnis keinen Unterschied in den stabilen Endsituationen erkennen, die sich aus der Modelldynamik ergaben. In beiden Fällen konvergierte das Modell gegen die maximale soziale

2 Agenten wählen zufällig zwischen Standorten mit gleicher maximaler Summe der gewichteten Valenzen.

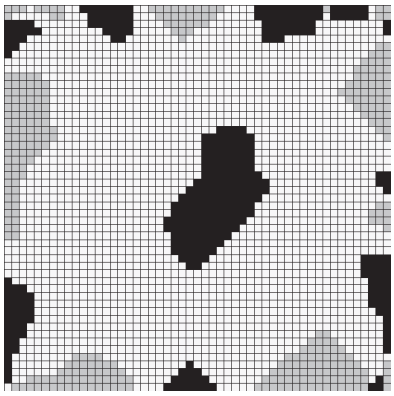
Abbildung 1: Replikation von Sakodas Modell. Effekte unterschiedlicher individueller Einstellungsmuster auf Gruppenbildung (N = 51 × 51 Felder und 2 Gruppen mit je 325 Agenten)



Startsituation: zufällig erzeugte „Ursuppe“



Stabiler Endzustand bei Einstellungsmuster „Segregation“



Stabiler Endzustand bei Einstellungsmuster „Suspicion“

Segregation, die Gruppen versammelten sich in jeweils einander gegenüberliegenden Ecken der schachbrettartigen Welt.

Flache und Hegselmann (2000) zeigten, dass eine Replikation von Sakodas Modell in einer größeren Welt mit größeren Gruppen auf ein noch überraschenderes Ergebnis führt. Abbildung 1 zeigt das Resultat einer *Ceteris-paribus*-Replikation der Dynamik Sakodas in einer Welt mit 51 × 51 Feldern und 2 Gruppen mit je 100 Agenten.

Abbildung 1 macht deutlich, dass Neutralität gegenüber der eigenen Gruppe sogar zu stärkerer Segregation führen kann als eine positive Einstellung zur eigenen Gruppe. Diese Replikation der Simulation Sakodas ist nur eines von vielen Beispielen dafür, wie ABCM kontraintuitive Makroimplikationen einfacher Mikromechanismen aufzeigt. Das Demonstrieren eines überraschenden Modellverhaltens mit einer Computersimula-

tion ist allerdings erst fruchtbar, wenn das Resultat auch unabhängig vom Computerprogramm verstehbar ist. Im Falle Sakodas ist eine einfache Erklärung möglich. Bei *segregation* und *mutual suspicion* fliehen alle Agenten vor den Angehörigen der jeweils anderen Gruppe. Begegnen sich bei *segregation* Angehörige der gleichen Gruppe auf dieser Flucht vor den jeweils anderen, dann gewinnt der Begegnungsort eben dadurch bereits an Attraktivität. Genau das ist aber bei *mutual suspicion* nicht der Fall. Während es unter dem Einstellungsmuster *segregation* dann zu einer Art „Ausflockung" der gesamten Population kommt, entstehen unter *mutual suspicion* auch sehr viel größere Cluster. Legt man allerdings – wie Sakoda – eine kleine Welt mit kleinen Gruppengrößen zu Grunde, dann kommt dieser Effekt noch nicht zum Tragen.

Können Soziologen etwas aus Modellen selbstorganisierten Verhaltens lernen, die zunächst für die Erforschung von Computernetzwerken, Vogelschwärmen oder chemischen Oszillationen entwickelt worden sind? Unserer Meinung nach ist dies aus mehreren Gründen der Fall. Erstens zeigen agentenbasierte Computermodelle, wie einfache Regeln lokaler Interaktionen komplexe Populationsdynamiken generieren können, die mit herkömmlichen Methoden nur schwerlich (oder gar nicht) modellierbar sind. Zweitens machen die Modelle deutlich, dass Durkheimsche „soziale Fakten" auf der Populationsebene entstehen können, ohne dass sie schon durch Eigenschaften der Individuen a priori festliegen. Drittens bieten sich agentenbasierte Computermodelle als virtuelles Laboratorium an, in dem experimentell erforscht werden kann, welche Mikromechanismen hochkomplexe soziale Phänomene erklären können.

Schon lange vor der Entwicklung agentenbasierter Computermodellierung wurden die Kernideen der zu Grunde liegenden Modellierungsstrategie durch den methodologischen Individualismus (vgl. Udehn 2002) in die Soziologie eingeführt. Aus Sicht des methodologischen Individualismus sind soziale Phänomene Makrokonsequenzen des zielgerichteten Handelns autonomer, aber interdependenter Individuen. Diese Individuen sind in ihren Handlungsmöglichkeiten wiederum durch die Makrobedingungen eingeschränkt, die sie kollektiv erzeugt haben (vgl. Coleman 1990). Autonome, reaktive, proaktive und sozialfähige „Agenten", die aufgrund einfacher Entscheidungsheuristiken handeln, sind offensichtlich mit dem Individualmodell des methodologischen Individualismus vereinbar (vgl. Lindenberg 2001). Im methodologischen Individualismus wurde aber bis vor kurzem zur Modellierung kollektiver Phänomene vor allem das aus der neoklassischen Wirtschaftstheorie übernommene Verhaltensmodell des rationalen, egoistischen Nutzenmaximierers verwendet (Coleman 1990: 14). Der Grund dafür ist, dass die Theorie der „Rationalen Wahl" es in Verbindung mit Annahmen über die Homogenität individueller Ziele erlaubt, die Modellierung kollektiven Verhaltens in mathematisch lösbare Gleichgewichtsanalysen umzusetzen (vgl. Voss und Abraham 2000). Allgemeine Aussagen über Eigenschaften und Bedingungen von Gleichgewichtslösungen sind damit erreichbar, ohne dass explizit die zu Grunde liegende Dynamik individueller Entscheidungen berechnet werden muss. In der formalen Spieltheorie wird das Gleichgewichtsprinzip dann so ausgearbeitet, dass auch strategisches rationales Verhalten modelliert werden kann, bei dem rationale Akteure das rationale Verhalten anderer antizipieren und optimal darauf reagieren (z.B. Raub und Buskens, in diesem Band).

Spieltheoretische Modelle sind agentenbasierte Modelle, aber sie sind nicht unbedingt Computermodelle. Was kann agentenbasierte Computermodellierung dann zum

methodologischen Individualismus beitragen? Die Antwort ist, dass agentenbasierte Computermodelle weit weniger restriktive mathematische Annahmen benötigen, weil sie auf die Anforderung der analytischen Lösbarkeit verzichten. Agentenbasierte Computermodellierung ist dennoch eine formale Methode, mit der Makroimplikationen individueller Verhaltensannahmen auf systematische Weise strikt deduziert werden können, indem Computerprogramme Schritt für Schritt die Dynamiken errechnen, die sich aus der Interaktion von Individuen nach den vom Modellierer festgelegten Verhaltensregeln ergeben.

Das zunehmende Interesse von Soziologen an agentenbasierter Computermodellierung hat zwei Gründe. Zum einen teilen viele Soziologen die individualistische Perspektive, dass soziale Phänomene eine – möglicherweise unbeabsichtigte – Konsequenz zielgerichteten individuellen Handelns sind. Zum anderen greift aber die Befürchtung um sich, dass der Analyseapparat der neoklassischen Gleichgewichtsanalyse theoretische Restriktionen erzwingt, die oftmals inhaltlich problematisch oder zumindest fragwürdig sind. Erkenntnisse der experimentellen Spieltheorie (Camerer 2003; Fehr und Gächter 2000) und neuere Theorien „sozialer Rationalität" (Esser 2001; Lindenberg 2001) hinterfragen das neoklassische Handlungsmodell, indem sie eine Reihe empirischer Anomalien aufzeigen. Insgesamt scheint soziales Entscheidungsverhalten nach diesen Arbeiten eher auf einfachen Entscheidungsheuristiken zu beruhen. Diese Heuristiken ändern sich situationsspezifisch (z.B. Todd und Gigerenzer 2003; Vanberg 2002) und machen es insbesondere von sozialen Verhaltenssignalen abhängig, wie eine bestimmte Situation perzipiert oder „geframed" wird (Esser 2001; Lindenberg 2001). Modelle solcher individuellen Verhaltensheuristiken mögen plausibler sein als Nutzenmaximierung, sie machen es aber schwierig oder vielleicht sogar unmöglich, analytisch Gleichgewichtslösungen für Modelldynamiken zu ermitteln.

Bis vor kurzem neigten Vertreter der Theorie der rationalen Wahl dazu, unrealistische Verhaltensannahmen nicht als Problem für die Analyse von Makroregelmäßigkeiten zu sehen (Hechter 1988: 31–33; Coleman 1987: 184). Es wurde argumentiert, dass individuelle Abweichungen von rationalem Verhalten unsystematische Zufallsereignisse seien, und dass adaptive Verhaltensheuristiken, wie zum Beispiel Lernen oder Imitation, im Endergebnis auch dazu führten, dass sich Akteure verhielten „als ob" sie perfekt rationale Entscheidungen treffen. Dieser Optimismus wird heute zunehmend in Frage gestellt. Neuere theoretische Arbeiten zeigen, dass globale Strukturen, die aus lokalen Interaktionen entstehen, sehr stark von Details der Modellierung individuellen Verhaltens abhängen können. Schon Voss (1990) wies darauf hin, dass die Berücksichtigung begrenzter Rationalität die Ergebnisse einer Rationalen-Wahl-Modellierung inhaltlich entscheidend verändern kann. Wittek und Flache (2003) haben kürzlich eine Reihe von Arbeiten in der Organisationsforschung diskutiert, die aufzeigen, dass Vorhersagen über Kooperationsverhalten und Effekte der Organisationsstruktur empfindlich davon beeinflusst werden, welche der Kernannahmen der orthodoxen Theorie der rationalen Wahl in welcher Weise modifiziert werden.

Einigen neueren Ansätzen ist es gelungen, Verhaltensheuristiken mit mathematischen Gleichgewichtsanalysen zu vereinbaren. Dies gilt insbesondere für die evolutionäre Spieltheorie (z.B. Bendor und Swistak 2001), spieltheoretische Modelle von Lernverhalten (z.B. Fudenberg und Levine 1998) und Anwendungen physikalischer Modelle

auf soziale Dynamiken (z.B. Helbing und Nagel 2003; Schweitzer 2003). Aber auch in diesen Ansätzen müssen stark vereinfachende Annahmen gemacht werden, um Gleichgewichtsanalysen zu ermöglichen. So werden etwa soziale Netzwerkstrukturen typischerweise als vollständige oder perfekt randomisierte Graphen abgebildet (Akteure interagieren mit zufällig gewählten Partnern), um die Modelle handhabbar zu halten.

Agentenbasierte Computermodelle bieten einen Zwischenweg. Sie ermöglichen es, kognitiv unrealistische Annahmen der Rationalen-Wahl-Theorie abzuschwächen, erlauben aber dennoch zu modellieren, wie soziale Strukturen aus den lokalen Interaktionen zahlreicher, in komplexe Netzwerkstrukturen eingebetteter Individuen entstehen. Außerdem hat die zunehmende Verfügbarkeit leistungsfähiger Rechner und die Verbreitung von Programmierkenntnissen und -werkzeugen Computermodellierungen für viele Soziologen zugänglicher gemacht als die mathematisch-analytischen Modelle der traditionellen Rationale-Wahl-Theorie. Einige unserer publizierten Lernmodelle (z.B. Macy und Flache 2002) haben wir inklusive graphischer Ausgabe mit dem elementaren Spreadsheetprogramm Excel nachprogrammieren können.[3] Weitaus leistungsfähigere, speziell für ABCM entworfene Werkzeuge sind verfügbar, von denen wir einige in Abschnitt 3 kurz vorstellen werden.

Die Leistungsfähigkeit moderner Computer und Software bringt paradoxerweise auch eine Gefahr mit sich. Praktisch unbegrenzte Rechenkapazitäten verführen Forscher dazu, hochkomplexe, „realistische" Modelle zu entwickeln. Es ist vergleichsweise einfach, ein Programm zu schreiben, das eine lange Liste plausibler Verhaltensannahmen enthält, wie etwa Youngers Computermodell einer Jäger- und Sammler-Gemeinschaft (Younger 2004; vgl. auch Lepperhoff 2000). Diese Modelle enthalten Details wie den Zeitpunkt, zu dem Agenten einschlafen und welche Ereignisse sie zum Aufwachen veranlassen. Es ist aber wesentlich schwieriger, das Verhalten solcher Modelle zu analysieren und die ihm zu Grunde liegenden Mechanismen und ihre Relevanz für theoretische Fragestellungen zu verstehen. Agentenbasierte Computermodellierung könnte im Endeffekt als Instrument deduktiver Theoriebildung nutzlos werden, wenn Modellbauer nicht ernsthaft nach soliden Intuitionen und Erklärungen für das Verhalten ihrer Modelle suchen. Nur so kann ausgeschlossen werden, dass Modellergebnisse nicht Artefakte oder gar Folgen von Programmierfehlern sind. Wir formulieren daher zum Abschluss unseres Beitrages eine Reihe methodologischer Empfehlungen für den Einsatz agentenbasierter Computermodellierung als Instrument theoretischer Forschung.

Es ist uns unmöglich, alle Beiträge zur agentenbasierten Computermodellierung in der Soziologie in einem Übersichtsartikel hinreichend zu würdigen. Wir beschränken uns auf eine Selektion von Arbeiten, die sich auf das fundamentale soziologische Problem der Entstehung sozialer Ordnung und der ihr zu Grunde liegenden Mikromechanismen richten. Zur Ergänzung unseres Beitrages empfehlen wir kürzlich erschienene Übersichtsartikel zur agentenbasierten Computermodellierung und Computersimulation, insbesondere die Arbeiten von Gotts et al (2003), Sawyer (2003), Macy und Willer (2002), Troitzsch (in diesem Band) und Moretti (2002).[4]

3 Ein Beispielprogramm ist verfügbar unter: http://www.ppsw.rug.nl/~flache/

4 Neue Beiträge zur ABCM erscheinen regelmäßig. Einige einschlägige Zeitschriften sind: Journal

Im zweiten Abschnitt unseres Beitrages diskutieren wir agentenbasierte Computermodelle sozialer Ordnung im Vergleich zu spieltheoretischen Arbeiten. Abschnitt III richtet sich auf methodologische Fragen und Modellierungswerkzeuge. Der vierte Abschnitt schließt den Artikel mit einer Diskussion der agentenbasierten Computermodellierung und ihrer inhaltlichen Beiträge zur Soziologie ab.

II. Die spontane Entstehung sozialer Ordnung

Das Problem der sozialen Ordnung war eine der ersten Fragestellungen, die mit agentenbasierter Computermodellierung untersucht wurde. Im methodologischen Individualismus entspricht das Problem der Frage, wie und unter welchen Bedingungen Kooperation unter rationalen Egoisten in einer sozialen-Dilemma Situation möglich ist. Ein soziales Dilemma ist eine Situation, in der allseitige Kooperation (Beitragen zum Gemeinwohl) kollektiv wünschenswert ist (Pareto-effizient), aber dennoch daran scheitern kann, dass individuelle Akteure Ausbeutung fürchten oder versucht sind, die Kooperationsbereitschaft anderer auszunutzen (Defektion). Probleme wie mangelndes Vertrauen in Geschäftsbeziehungen, Trittbrettfahrerverhalten in Arbeitsgruppen, das Scheitern kollektiver Interessenvertretung oder der Verfall sozialer Solidarität in modernisierenden Gesellschaften sind als Beispiele sozialer Dilemma-Situationen beschrieben worden. Aus Sicht des methodologischen Individualismus ist vor allem die Frage bedeutend, wie die Entstehung von Kooperation und sozialen Normen „bottom-up", als Ergebnis selbstorganisierender individueller Interaktionen, erklärt werden kann (vgl. z.B. Hechter 1988; Voss 1985).

Die Spieltheorie bietet Erklärungen erfolgreicher Kooperation an, die weder inhärenten Altruismus noch zentrale soziale Kontrolle voraussetzen (für eine Übersicht vgl. Raub und Buskens in diesem Band; Voss 2001). Spieltheoretische Modellierungen betonen zwei zentrale Bedingungen der Kooperation, stabile Beziehungen (wiederholte Interaktion derselben Akteure) und Netzwerkeinbettung. Gemäß spieltheoretischer Modelle ist Beziehungsstabilität wichtig, weil sie Reziprozität in Form bedingter Kooperation ermöglicht. Wenn egoistische Akteure auf die langfristigen Vorteile dauerhafter Kooperation hinreichend viel Wert legen, dann kann es in einer stabilen Beziehung mit bedingt kooperativen Partnern für sie rational sein, die zukünftige Kooperation nicht durch Ausbeutungsversuche zu gefährden. Spieltheoretische Analysen von Netzwerkeinbettung greifen ebenfalls auf bedingte Kooperation zurück, erweitern sie aber um Reputationseffekte (Buskens 2002; Raub und Weesie 1990). Hier entspricht bedingte Kooperation der Strategie, mit einem Partner dann zu kooperieren, wenn man noch nicht über Dritte Hinweise auf früheres unkooperatives Verhalten dieses Partners in anderen Beziehungen erhalten hat. Wenn Kommunikationsnetzwerke hinreichend dicht sind, kann die Erwartung, dass andere Spieler solch einer Strategie folgen, einen rationalen Akteur davon abhalten, sich selbst eine schlechte Reputation zu erwerben. Erweiterungen des Arguments der bedingten Kooperation haben gezeigt, dass die Erklärung auch

---

of Artificial Societies and Social Simulation (JASSS, http://jasss.soc.surrey.ac.uk), Computational and Mathematical Organization Theory (CMOT), Journal of Mathematical Sociology (JMS).

auf kollektive Handlungsprobleme mit zahlreichen Teilnehmern erweiterbar ist (z.B. Raub 1988; Taylor 1987) und auf Spiele mit imperfekter Information angewendet werden kann, bei denen Individuen nur unvollkommen über das Verhalten anderer (z.B. Bendor und Mookherjee 1987) oder deren Präferenzen (z.B. Buskens 2003) informiert sind. Im Allgemeinen zeigt sich aber in diesen Studien auch, dass zunehmende Gruppengröße und Informationsunsicherheit bedingte Kooperation erschweren.

Spieltheoretische Modellierungen haben gezeigt, dass perfekt oder zumindest weitgehend rationale, egoistische Akteure der sozialen Falle mangelnder Kooperation entrinnen können. Eines der zentralen Themen in der agentenbasierten Computermodellierung ist die Überprüfung der Robustheit spieltheoretischer Resultate, wenn stark eingeschränkte kognitive Kapazitäten der Akteure angenommen werden. Wie kann soziale Ordnung als Ergebnis der Interaktionen autonomer Agenten erklärt werden, wenn Agenten Entscheidungen aufgrund einfacher, in ihren Konsequenzen möglicherweise irrationaler Heuristiken treffen? Diese Frage hat Entwickler agentenbasierter Computermodelle dazu angeregt, die strukturellen Bedingungen zu erforschen, unter denen adaptive Agenten zu kooperativen Lösungen sozialer Dilemmata gelangen können. Wir betrachten im Folgenden Arbeiten aus der ABCM zur Stabilität von Beziehungen, Effekten von Netzwerkeinbettung und Effekten unterschiedlicher Verhaltensannahmen.

1. Stabilität von Beziehungen

Axelrods *Evolution der Kooperation* (Axelrod 1984) ist die klassische Studie der Entstehung sozialer Ordnung ohne explizite individuelle Rationalität. Axelrod testete theoretisch, ob stabile Beziehungen Kooperation auf Basis von Reziprozität auch dann ermöglichen, wenn individuelle Agenten „fest verdrahteten" Verhaltensstrategien folgen, die nicht notwendigerweise auch strategisch rational sind. Die Strategien wurden von Experten speziell dafür entworfen, ein Turnier gegen unbekannte andere Strategien zu gewinnen. Die Spieler kannten in jeder Runde des Spiels nur die bisherigen Handlungen ihres Gegenübers.

Axelrod modellierte die Selektion von Strategien durch Konkurrenzdruck auf dem Niveau der Gesamtpopulation, ein Prinzip, das er aus der Evolutionsbiologie importierte. Strategien, deren durchschnittliche Auszahlung über all ihre Spiele hinweg höher ist als die der Gesamtpopulation, überleben gemäß des Selektionsprinzips mit größerer Wahrscheinlichkeit und vermehren sich schneller. Axelrod verglich dies mit Imitation erfolgreicher Rollenmodelle in menschlichen Sozialsystemen.

Der Sieger des Turniers war eine der einfachsten Strategien, die teilnahmen, das bedingt kooperative „Tit-for-Tat". Tit-for-Tat, entwickelt von dem Sozialpsychologen und Mathematiker Anatol Rapoport, kooperiert im ersten Zug des Spiels und kopiert danach den vorhergehenden Zug des Gegenspielers. Axelrods Computerexperimente haben gezeigt, dass die spieltheoretische Lösung bedingter Kooperation in stabilen Beziehungen auch dann robust ist, wenn Agenten nur minimale kognitive Ressourcen in Entscheidungsfindung investieren. Die Resultate widersprachen spieltheoretischen Ergebnissen allerdings insofern, dass bedingte Kooperation nicht etwa ein Spiel unbestimmter Länge voraussetzte, sondern selbst bei hinreichender Spieldauer schon evolu-

tionär erfolgreich war. Strikt rationale Spieler würden statt dessen in einem endlichen Spiel immer defektieren, da Kooperation in der letzten Runde keine zukünftigen Vorteile mehr erbringt, und dementsprechend auch in der vorletzten und damit allen weiteren Runden nicht auf zukünftige Kooperativität der Gegenpartei gerechnet werden kann („backward-induction"). Laborexperimente mit menschlichen Versuchspersonen in Gefangenendilemma-Spielen endlicher und bekannter Dauer sprechen eher gegen diese Vorhersage und für das Resultat Axelrods (Fehr und Gächter 2000).

Eine lange Reihe von Nachfolgearbeiten (vgl. Gotts et al. 2003 für eine Übersicht) haben weiter zum theoretischen Nachweis beigetragen, dass spontane Entstehung sozialer Ordnung in stabilen Beziehungen auch ohne die restriktiven Annahmen der klassischen Spieltheorie erklärbar ist. Axelrods inhaltliche Argumente wurden dabei auch weit über die spieltheoretische Literatur hinaus kritisch rezipiert und diskutiert (Etzioni 2001).

Die *Evolution der Kooperation* traf auch auf Kritik von Spieltheoretikern (Binmore 1998), die zwei ernstzunehmende Schwächen der ursprünglichen Arbeit aufgedeckt haben, die Abhängigkeit der Ergebnisse von arbiträren Annahmen über die Zusammenstellung der Strategienpopulation und die restriktive Voraussetzung perfekter Information. Wir diskutieren im Folgenden agentenbasierte Computermodellierungen, die beide Schwächen untersucht haben.

a) Die Strategienpopulation. Axelrod erklärte den bemerkenswerten Erfolg von Tit-for-Tat mit drei Eigenschaften der Strategie, ihrer Freundlichkeit (keine Defektion ohne vorhergehende Provokation), Provozierbarkeit (jede Defektion des Gegenspielers wird bestraft) und Vergebungsbereitschaft (nach Kooperation des Gegenspielers wird auch selbst wieder kooperiert). Kritiker (z.B. Binmore 1998) haben jedoch darauf hingewiesen, dass viele andere Strategien ebenfalls diese Eigenschaften aufweisen und dass die Ergebnisse jeder einzelnen Strategie stark von der spezifischen Population von Strategien abhängen.

In einer Nachfolgearbeit versuchte Axelrod (1997: 14–29) mit einem so genannten genetischen Algorithmus (GA) nachzuweisen, dass Tit-for-Tat sich auch in einer nahezu unbegrenzten Strategienpopulation mit zufällig gewählter Ausgangssituation durchsetzt. Der GA erweitert die Strategienmenge, indem er durch einen Zufallsmechanismus vollständig neue und nicht zuvor festgelegte Strategien erzeugt. Genetische Algorithmen beschreiben die individuelle Strategie mit Sequenzen von Symbolen, die – analog zu den Genen eines Chromosoms – bestimmte Eigenschaften der Strategie festlegen, z.B. ob die Strategie freundlich ist, oder nach wie vielen Defektionen des Partners selbst zur Defektion übergegangen wird. Die Instruktionen eines Genes generieren das Verhalten im Spiel, welches seinerseits den relativen Erfolg eines Agenten in der Population und damit die Reproduktionschancen seines Genes bestimmt. „Paarung" erfolgreicher Sequenzen kann dann durch zufällige Rekombination von Eigenschaften zu neuen und möglicherweise noch besseren Strategien führen. Wenn dies der Fall ist, verdrängt die neue Strategie langfristig ihre „Eltern" aus der Population. Genetische Algorithmen verwenden darüber hinaus oft Zufallsmutationen der kopierten Gene, um damit die Heterogenität der Population zu erhalten.

In Zusammenarbeit mit John Holland fand Axelrod eine Reihe von Tit-for-Tat ähnlichen langfristig erfolgreichen Strategien. Diese Strategien entsprachen Tit-for-Tat

in den Eigenschaften der Provozierbarkeit und Vergebungsbereitschaft, verhielten sich aber zugleich in Interaktionen mit naiven Gegenspielern deutlich ausbeuterischer als Tit-for-Tat. Dieses Resultat wurde im Wesentlichen in einer Reihe weiterer Studien mit der Methode so genannter selbstprogrammierender Moore-Maschinen repliziert (Probst 1996; Linster 1992). Bei einer Moore Maschine hängt das Verhalten eines Agenten (Kooperation oder Defektion) von seinem jeweiligen internen Zustand ab. Die eigentliche Strategie wird durch Regeln bestimmt, die festlegen, wie das wahrgenommene Verhalten des Gegenspielers den internen Zustand verändert. Probst (1996) verwendete einen Evolutionsmechanismus mit Zufallsmutationen, der die Entstehung von Moore-Maschinen mit bis zu 16 internen Zuständen zuließ. Freundliche, aber einfache Strategien wie Tit-for-Tat waren in diesen Computerexperimenten in der Regel nur anfänglich erfolgreich. Im Verlauf der Evolutionsdynamik entwickelte sich aber ein evolutionäres Wettrüsten, in dem Strategien zunehmend komplexer wurden und aggressivere Varianten Tit-for-Tat allmählich aus der Population verdrängten. Gleich Tit-for-Tat waren diese Strategien in der Lage, mit Partnern zu kooperieren, die nicht leicht ausgebeutet werden konnten. Anders als Tit-for-Tat nutzen diese Strategien aber „naive" Partner aus.

b) Imperfekte Information, Unsicherheit und Toleranz. Axelrod ging davon aus, dass Spieler vollständige und perfekte Information über das Verhalten ihrer Partner in der Vergangenheit haben. Spieltheoretische Analysen zeigen, dass Unsicherheit, zum Beispiel durch fehlende oder fehlerhafte Information, die Effizienz von Reziprozitätsstrategien stark beeinträchtigen kann (z.B. Bendor und Mookherjee 1987). Bendor und Mookherjee haben analytisch die Effekte exogener Störungen untersucht, durch die Spieler unabsichtlich defektieren, ohne dass ihre Interaktionspartner wissen, ob die Defektion intendiert war oder nicht. Die Autoren zeigen, dass unter diesen Bedingungen bedingte Kooperation zwar noch rational sein kann, aber mit größerer Unsicherheit auch „toleranter" und zunehmend ineffizient werden muss, weil „versehentliche" wechselseitige Bestrafungen trotz Toleranz nicht vermeidbar sind.

Diese spieltheoretischen Erkenntnisse haben agentenbasierte Computermodellierungen veranlasst, die untersuchen, ob imperfekte Information ähnlich schwerwiegende Folgen hat, wenn Agenten ihre Strategien nicht rational wählen, sondern Strategien durch evolutionäre Selektionsmechanismen auf der Populationsebene selektiert werden (z.B. Nowak und Sigmund 1993; Kollock 1993). Im Großen und Ganzen bestätigen diese Studien, dass Unsicherheit die Durchsetzungschancen toleranter, bedingt kooperativer Strategien verbessert. Tolerante Strategien schlagen erst nach mehreren Defektionen der Gegenpartei zurück und auch dann nur begrenzt. Damit können sie besser als intolerantere, leichter provozierbare Strategien langfristig profitable Kooperationsbeziehungen erhalten.

Weitere agentenbasierte Computermodellierungen haben zu neuen Einsichten geführt. So wurde aufgezeigt, dass adaptive (backward-looking) Strategien – wie etwa „Win-Stay, Lose-Shift" (WSLS) – besonders effektiv mit Unsicherheit umgehen können (Macy 1996; Nowak und Sigmund 1993). WSLS verändert sein Verhalten (Kooperation oder Defektion) nur dann, wenn das Ergebnis der vorigen Spielrunde als verlustreich angesehen wird und bleibt ansonsten bei seinem Verhalten. Ausbeutung durch den Gegenspieler oder beidseitige Defektion sind typischerweise verlustreiche Er-

gebnisse im Gefangenendilemma. Entsprechend werden Ausbeutung des Gegenspielers oder beidseitige Kooperation als erfolgreich bewertet. Im Gegensatz zu Tit-for-Tat, welches versucht, dem Partner eine Lehre zu erteilen, lernt WSLS vom Verhalten des Partners. In wiederholten Spielen mit Unsicherheit ist es eine besondere Stärke adaptiver Strategien, dass sie schnell wieder zu beidseitiger Kooperation zurückfinden können, wenn die Beziehung mit einem freundlichen Gegenspieler durch einen externen Schock gestört wurde. Unzufriedenheit mit den sozialen Kosten wechselseitiger Defektion bringt adaptive Spieler nämlich dazu, ihrem Gegenüber ein Friedensangebot zu machen, das beiden Parteien erlaubt, aus dem Teufelskreis auszubrechen.

Unsicherheit kann nicht nur Kooperation in bestehenden Beziehungen negativ beeinflussen, sondern auch Partnerselektion bei der Bildung von Beziehungen erschweren. Die Möglichkeit des Abbruchs und der Neuformung von Beziehungen wurde ursprünglich in der theoretischen Untersuchung sozialer Ordnung nicht berücksichtigt, weil sich das Hauptaugenmerk auf Beziehungsstabilität richtete. Eine Reihe von agentenbasierten Computermodellierungen (Vanberg und Congleton 1992; Schüssler 1989) hat aufgezeigt, dass Kooperation auch ohne stabile Beziehungen möglich ist. In diesen Modellierungen evolutionärer Selektionsprozesse entsteht soziale Ordnung, weil unkooperative Agenten von Beziehungen ausgeschlossen werden, gemäß der Strategie „sei kooperativ, aber verlasse deinen Partner sobald er defektiert". Wenn hinreichend viele Mitglieder der Population einer solchen Regel folgen, dann bilden auf die Dauer freundliche Strategien stabile Beziehungen miteinander aus und zwingen aggressive Agenten in unvorteilhafte Beziehungen mit ihresgleichen. In einer Nachfolgestudie fanden Schüssler und Sandten (2000) – ähnlich zu Probst (1996) und Axelrod (1997) –, dass sich bei komplexeren Agentenarchitekturen in der Evolutionsdynamik auch eher aggressive Varianten dieser bedingt kooperativen Strategien durchsetzen. Insbesondere werden dann unfreundliche Strategien erfolgreich. Diese Strategien kooperieren einerseits mit Partnern, die sich nicht ausbeuten lassen, gehen aber andererseits grundsätzlich zu Defektion über, wenn es leicht möglich erscheint, naive Opfer zu finden. Das ist insbesondere dann der Fall, wenn durch Populationszuwachs die Anzahl möglicher neuer Partner schnell zunimmt.

2. Netzwerkeinbettung und die Entstehung sozialer Ordnung

Spieltheoretische Analysen der kooperationsstabilisierenden Effekte von Beziehungsstabilität und Netzwerkeinbettung durch Reputationseffekte (Buskens 2002; Raub und Weesie 1990) verwenden gleichermaßen unplausible Annahmen perfekter Information und Rationalität. Modelle von Reputationsmechanismen unterstellen insbesondere, dass Akteure die gesamte Struktur des sie umgebenden sozialen Netzwerkes kennen und in der Lage sind, exakt vorherzusehen, wie sich Information über ihr Verhalten selbst in entfernten Regionen des Netzwerkes verbreitet. Agentenbasierte Computermodellierungen legen jedoch nahe, dass diese Annahmen für die Wirksamkeit von Reputationsmechanismen nicht unbedingt notwendig sind. Insbesondere konnte Takahashi (2000) zeigen, dass Strategien der bedingten Kooperation auf Basis von Reputation auch ohne die Annahme rationaler Strategiewahl, allein durch evolutionäre Mechanismen der

Strategiemutation und -selektion auf der Populationsebene entstehen und sich stabilisieren können. Eine Reihe ähnlicher agentenbasierter Computermodellierungen bestätigt im Wesentlichen Takahashis Resultat (Younger 2004; Saam und Harrer 1999; Castelfranchi et al. 1998). Diese Arbeiten haben allerdings Effekte unterschiedlicher Netzwerkstrukturen noch nicht systematisch erforscht, so dass hier ein Vergleich agentenbasierter Modellierungen mit spieltheoretischen Resultaten noch aussteht.

Ein weiteres Problem spieltheoretischer Modellierungen von Netzwerkeffekten ist die ihnen zu Grunde liegende Annahme stabiler Netzwerkstrukturen. Diese Annahme ist schwerlich mit neueren Modellen dynamischer Netzwerke vereinbar, die explizit davon ausgehen, dass rationale Akteure Beziehungen gezielt anbahnen und wieder beenden, um damit ihre Position im Netzwerk zu optimieren (vgl. Jackson 2004). Aber auch diesen Modellierungen ist es bislang nicht gelungen, Dynamiken der gleichzeitigen Verhaltens- und Netzwerkveränderung analytisch zu integrieren. Einige agentenbasierte Computermodellierungen haben genau diese Integration unternommen. Hegselmann (1996; vgl. Flache und Hegselmann 2000, 1999) hat Kooperationsprobleme in dynamischen Netzwerken untersucht, in denen Agenten als Interaktionspartner unterschiedlich attraktiv sind und Interaktionspartner wechseln können. Ein Beispiel hierfür sind Technologiekooperationen zwischen Firmen. Podolny und Page (1998) haben empirisch aufgezeigt, dass Firmen um Kooperationsbeziehungen mit attraktiven Partnern konkurrieren, die die meisten technologischen Ressourcen und den höchsten Status in ihrer Industrie haben. Ausgehend von der Standardannahme strategischer Rationalität leitet Hegselmann aus seinem Modell ab, dass Akteure in wiederholten Tauschinteraktionen bedingt kooperativ sind, aber nur wenn sie eine hinreichende Dauer der Beziehung erwarten und den Partner hinreichend attraktiv finden. Hegselmann modelliert die Partnersuche mit Hilfe eines zellularen Automaten, der gleichen Technik, die auch Sakoda anwendete. Agenten können in Hegelmanns Modell nur mit ihren unmittelbaren Nachbarn in der Gitterstruktur Kooperationsbeziehungen unterhalten. Wechsel aller oder eines Teils der Interaktionspartner geschieht durch Veränderung des Standortes im zellularen Gitter. Hegselmann modelliert die Entscheidung der Standortwahl im sich ständig verändernden Netzwerk mit einer myopischen, nur beschränkt rationalen Entscheidungsheuristik. Ein Agent zieht in die Nachbarschaft der aus seiner Sicht attraktivsten erreichbaren Partner, die gerade noch bereit sind, mit dem fokalen Agenten zu kooperieren. Die Computerexperimente ergeben, dass diese Dynamik unter einem breiten Spektrum von Randbedingungen zur Entstehung dichter Kooperationsnetzwerke führt, in denen vor allem Agenten ähnlicher Attraktivität miteinander kooperieren. Dabei bildet sich durch die wechselseitige Selektion von Interaktionspartnern ein Segregationsmuster. Die attraktivsten Agenten streben zueinander und formen den Kern des Netzwerkes. Die am wenigsten attraktiven Populationsmitglieder werden von fast allen anderen gemieden und an den Rand des Netzwerkes gedrängt. Dort können sie nur noch Beziehungen mit ähnlich unattraktiven Partnern eingehen. Dies Resultat spiegelt teilweise Strukturen von Tauschnetzwerken wieder, die Komter (1996) in einer empirischen Studie gefunden hat.

Smith und Stevens (1999) haben in einer ähnlichen Arbeit die Dynamik von Netzwerken sozialer Unterstützungsbeziehungen modelliert. In ihrem Modell suchen Agenten Partner, die ihnen bei der Bewältigung von Angstzuständen helfen. Agenten ent-

scheiden in einem Prozess wechselseitig selektiver Partnerwahl, mit wem sie Beziehungen eingehen. Die Autoren stellen fest, dass in ihren Computerexperimenten im Endergebnis homophile Cluster entstehen, in denen Agenten Partner haben, die ihnen in ihrer Fähigkeit zur Stressbewältigung ähneln.

3. Emergente soziale Ordnung und Verhaltensannahmen

Im vorangegangenen Abschnitt haben wir Arbeiten besprochen, die die Effekte struktureller Randbedingungen wie Beziehungsstabilität und Netzwerkeinbettung auf die Entstehung und Aufrechterhaltung sozialer Ordnung modellieren. Wir wenden uns jetzt Studien zu, die die Konsequenzen der zu Grunde liegenden Verhaltensannahmen untersuchen. Wir betrachten insbesondere Arbeiten, die untersuchen, welche Konsequenzen der Lernmechanismus hat, der das adaptive Verhalten der Agenten steuert.

Die evolutionäre Spieltheorie hat gezeigt, dass spontan entstehende soziale Ordnung unter Egoisten auch ohne rationale Agenten möglich ist. Kritiker bezweifeln allerdings, dass evolutionäre Selektion ein geeignetes Modell sozialer Lernmechanismen ist. Chattoe (1998) argumentiert zum Beispiel, dass kulturelle Evolution nicht mit genetischer Evolution vergleichbar ist. Welcher Mechanismus merzt in kultureller Evolution weniger erfolgreiche Strategien aus einer Population aus und sorgt für die Verbreitung ihrer erfolgreichen Konkurrenten? „Imitation der Besten" scheint ein besseres Modell zu sein als physische Reproduktion und Aussterben der individuellen Träger von Verhaltensstrategien. Imitationsmechanismen können aber nur für die Weiterverbreitung sichtbarer Verhaltensmerkmale sorgen und nicht für die Replikation der ihnen zu Grunde liegenden, nicht beobachtbaren Verhaltenstrategie. Biologische Metaphern verschleiern die Bedeutung dieses Unterschiedes.

Die Zweifel an der Eignung biologischer Evolutionsmechanismen als Modell sozialen Lernens haben das Interesse an Modellen verstärkt, die die Selektion von Verhaltenstrategien von der Populationsebene auf die kognitive Ebene einzelner Individuen verlagern. Als einfaches Ausgangsmodell haben neuere agentenbasierte Modellierungen dazu das bereits von Rapoport und Chammah (1965) angewendete Reinforcement-Lernen aufgegriffen. Reinforcement-Lerner entscheiden probabilistisch zwischen Handlungsoptionen. Handlungen, die in der Vergangenheit gemessen an einem Erwartungsniveau erfolgreich waren, werden mit größerer Wahrscheinlichkeit wiederholt und zuvor erfolglose Strategien eher vermieden. Je erfolgreicher eine Strategie war, umso wahrscheinlicher wird es also, dass ein Agent sie wieder benutzt. Lernmodelle siedeln damit die Selektion von Strategien auf der Ebene des Individuums an, während diese bei Modellen evolutionärer Selektion auf der Populationsebene stattfindet.

Macy (1996) hat mit Hilfe von Computermodellierung systematisch die Effekte evolutionärer Selektion und individuellen Lernens auf Kooperationsverhalten verglichen. Macy repräsentierte evolutionäre Selektion mit einem genetischen Algorithmus (GA) und verwendete ein neuronales Netzwerk (NN) zur Modellierung individuellen Lernens. Ein neuronales Netzwerk ist in ähnlicher Weise wie ein genetischer Algorithmus eine selbstoptimierende Einheit. Anders als beim GA wird aber beim NN nicht eine erfolgreiche Strategie aus verschiedenen Strategiekandidaten selektiert. Stattdessen

wird in wiederholten Entscheidungssequenzen die interne Struktur eines Entscheidungsnetzwerkes durch Verstärkung und Abschwächung seiner Netzwerkverbindungen schrittweise verändert. Dabei lernt das Netzwerk durch Reinforcement-Lernen, zu einem gegeben Inputvektor (zum Beispiel dem Ergebnis der vorigen Runde eines Gefangenendilemmaspieles) eine optimale Reaktion zu erzeugen (die Strategie in der nächsten Runde). Macy fand mit Hilfe des GA und in Übereinstimmung mit früheren Arbeiten von Nowak und May (1993), dass unter evolutionärer Selektion langfristig Strategien erfolgreich sind, die dem Reinforcement-Lernen entsprechen. Ein Beispiel ist das weiter oben besprochene Win-Stay, Lose-Shift. Interessanterweise entwickelten sich diese Strategien aber nicht, wenn Agenten in Form neuronaler Netzwerke ihre Verhaltensstrategie selbst erlernten. Das Problem war, dass der Erfolg dieser „zurückschauenden" Strategien sich erst dann einstellt, wenn sie mit ihresgleichen interagieren. Miteinander interagierende, aber autonom lernende Agenten könnten die Strategie also nur durch deren zufällige gleichzeitige Verwendung entdecken. Dies ist jedoch insbesondere bei komplexeren Lernmechanismen sehr unwahrscheinlich.

Agentenbasierte Modelle mit einfachen Reinforcement-Lernregeln werden in der Soziologie und den Wirtschaftswissenschaften zunehmend eingesetzt, um die Bedingungen für Kooperation in sozialen Dilemmata theoretisch zu erforschen (Macy und Flache 2002; Flache und Macy 2002; Erev und Roth 1998; Roth und Erev 1995). Macy (1990) modellierte in Nachfolge von Rapoport und Chammah (1965) das Verhalten interagierender Spieler in einem wiederholten Gefangenendilemma mit einem stochastischen Lernalgorithmus. Macy konnte zeigen, dass die Lerndynamik Agenten in ein stabiles Gleichgewicht mit wechselseitiger Kooperation führen kann. Ein Beispiel ist die Dynamik in einem Gefangendilemma mit zwei Spielern, die beide dann und nur dann mit ihrer Handlung (Kooperation oder Defektion) zufrieden sind, wenn der jeweils andere sich kooperativ verhalten hat. Wenn beide Spieler anfänglich defektieren, erhöht sich für beide durch Unzufriedenheit mit dem Ergebnis ihre Neigung, bei der nächsten Interaktion zu kooperieren. Wenn dann beidseitige Kooperation die Folge ist, erhöht Zufriedenheit mit dem Ergebnis die Kooperationsbereitschaft weiter. Computerexperimente haben gezeigt, dass dieser Mechanismus der „stochastischen Kollusion" (Macy 1990) unter einem breiten Spektrum von Randbedingungen langfristig zu stabiler Kooperation führt. Macy konnte insbesondere zeigen, dass stochastische Kollusion auch kollektive Handlungsprobleme bei sozialen Dilemmata mit zahlreichen Spielern lösen kann. Weitere Studien haben eine Reihe von Bedingungen ergeben, die die Lösung begünstigen. Insbesondere höhere individuelle Lernschnelligkeit, geringere Gruppengrößen und „small-world" Netzwerkstrukturen, die die Anzahl der Interaktionspartner pro Akteur begrenzen, erleichtern die spontane Koordination kooperativer Spieler, die als Ausgangspunkt stochastischer Kollusion notwendig ist.

Macy und Flache (2002) haben in einer Nachfolgestudie gezeigt, dass die „zurückschauende" Lösung des Kooperationsproblems auf eine größere Klasse sozialer Dilemmata generalisierbar ist. Flache und Macy (2002) haben weiterhin untersucht, inwieweit die inhaltlichen Resultate von der spezifischen Form des Lernalgorithmus abhängen. Die Autoren haben dazu zwei verschiedene Implementationen von Reinforcement-Lernen theoretisch integriert, das Bush-Mosteller-Modell stochastischen Lernens und das Modell von Roth und Erev (1995), welches insbesondere in den Wirtschafts-

wissenschaften und der Spieltheorie als Modell stark begrenzter Rationalität diskutiert wurde (vgl. Erev und Roth 1998). Die agentenbasierten Computermodelle von Flache und Macy haben gezeigt, dass beide Varianten Kooperation durch stochastische Kollusion erzeugen. Sie weisen aber auch nach, dass der von Roth und Erev vorgeschlagene Mechanismus weit empfindlicher auf Veränderungen des Aspirationsniveaus der Agenten reagiert.

Insgesamt zeigt sich, dass agentenbasierte Computermodelle ein hilfreiches theoretisches Werkzeug sind, mit dem die Sensitivität makrosozialer Vorhersagen gegenüber Veränderungen der ihnen zu Grunde liegenden Verhaltensannahmen überprüft werden kann. Dies hilft insbesondere dabei festzustellen, welche Hypothesen als robust gelten dürfen. Ein bekanntes Beispiel ist Olsons (1965) Vorhersage, dass in kollektiven Handlungen mit zunehmender Gruppengröße weniger Kooperation stattfindet. Dieser Effekt kann für ein breites Spektrum von Randbedingungen spieltheoretisch aus strikt rationalem Verhalten abgeleitet werden (z.B. Raub 1988), folgt aber auch aus Computermodellierungen von Reinforcement-Lernen (Macy 1990) und evolutionärer Selektion (Diekmann und Manhart 1989).

III. Methodologische Prinzipien und Werkzeuge

Agentenbasierte Computermodellierungen suchen nach mikrosozialen kausalen Mechanismen, die beobachtbaren Makroregularitäten zu Grunde liegen könnten. Ein Vorteil des Ansatzes ist, dass Modellspezifikationen nicht der restriktiven Anforderung mathematischer Analysierbarkeit der Modelldynamik unterliegen. Dies ist zugleich eine mögliche Schwäche, weil Forscher dazu verleitet werden können, komplexe Modelle zu erzeugen, deren Verhalten nicht mehr unabhängig von ihrer spezifischen Computerimplementation verstehbar ist. Modellbildung ist sinnlos, wenn die Erklärung der Dynamik eines Modells kaum weniger schwierig ist als die Erklärung des modellierten realweltlichen Phänomens. Wir formulieren im Folgenden eine Reihe erfahrungsbasierter Prinzipien zur Vermeidung dieses Problems. Im Anschluss daran werden wir kurz auf Softwarewerkzeuge eingehen.

1. Methodologische Prinzipien

a) Deutliche Spezifikation des Problems und der Theorie. Wir stimmen Colemans (1990) Empfehlung zu, individualistische Modellbildung – und damit auch agentenbasierte Computermodellierung – mit dem Ziel zu betreiben, beobachtbare Beziehungen zwischen deutlich umschriebenen Makrophänomenen zu erklären oder vorherzusagen. Dies sichert eine klare Definition der Fragestellung und schafft eine natürliche Verbindung zwischen virtuellen Computermodellen und empirischen Anwendungen. Dabei bleibt das Hauptziel herauszufinden, wie sich Vorhersagen durch unterschiedliche Kombinationen von Annahmen verändern. Wir schließen uns damit auch Sawyers Standpunkt an (Conte et al. 2001), dass nicht formalisierte, inhaltliche soziologische Theorien für formale Modellbauer ein nützlicher Ausgangspunkt sind. Diese Theorien

liefern oftmals empirische Beobachtungen und inhaltliche Ideen über die ihnen zu Grunde liegenden sozialen Mechanismen. Je deutlicher diese Ideen formuliert sind, desto nützlicher sind sie als Richtschnur bei der Entwicklung agentenbasierter Computermodelle.

b) Mit einfachen Modellen beginnen. Manche Anhänger der agentenbasierten Modellierung würden die meisten von uns besprochenen Modelle für übermäßig einfache Verhaltensannahmen kritisieren. Diese Autoren optieren für realistischere Modelle menschlicher Kognitionsprozesse (Conte et al. 2001). Insbesondere werden in der Literatur Modelle vorgeschlagen, die von situational veränderlichen Kognitionsmechanismen ausgehen, wie etwa einem Wechsel zwischen Reinforcement-Lernen, Imitation, bewusstem Nachdenken oder sozialem Vergleich (Jager et al. 2000; vgl. Schmidt 2001). Die Forderung, Verhaltensmodelle komplexer zu gestalten, ist nach unserer Ansicht fehl am Platze, wenn dadurch die Dynamik der Interaktion vieler komplexer Agenten nicht mehr verstehbar ist. Der Wert experimenteller Methoden in der Computermodellierung geht weitgehend verloren, wenn nur noch mit Hilfe fortgeschrittener statistischer Verfahren erkennbar ist, welche Mikromechanismen welchen Einfluss auf das Makroresultat haben. Die Analyse extrem einfacher, unrealistischer Modelle hat schon oftmals zu neuen theoretischen Einsichten geführt, die viel weiter anwendbar sind als die stark stilisierten Modelle, die die Einsichten hervorgebracht haben. Wir empfehlen, mit einfachen Verhaltensmodellen anzufangen und Komplikationen erst dann Schritt für Schritt hinzuzufügen, wenn die Dynamiken der einfacheren Modelle voll verstanden sind.

c) Systematisch experimentieren statt nur explorieren. Agentenbasierte Computermodellierung ist ein Werkzeug für theoretische Experimente. Die Exploration unbegrenzter theoretischer Möglichkeiten kann zwar zu wichtigen Einsichten führen, aber Modellierer sollten der Versuchung widerstehen, sich auf ziellose Abenteuerreisen in künstliche Welten zu begeben. Sorgfältige, systematische, mit statistischen Tests wie etwa Varianz- oder Regressionsanalyse abgesicherte Erkundung der Effekte verschiedener Parameterkombinationen mag weniger faszinierend sein, ist aber bessere Wissenschaft. Voraussetzungen dafür sind, dass Parametermanipulationen durch sorgfältige Studien des einschlägigen theoretischen und empirischen Wissens motiviert sind und den Simulationsexperimenten klar formulierte Hypothesen zu Grunde liegen.

d) Robustheit testen. Auch wenn Simulationsstudien Experimentaldesigns statt post-hoc statistische Analysen verwenden sollten, bleiben statistische Analysen notwendig. Die Robustheit agentenbasierter Computersimulationen muss insbesondere bei stochastischen Modellen durch Replikationen getestet werden. Wenn möglich sollten bei Replikationen diejenigen Modellparameter variiert werden, die theoretisch beliebig oder nebensächlich sind. Modellierer müssen dabei experimentelle Manipulationen, bei denen bestimmte Effekte von Parameterveränderungen erwartet werden, von Robustheitsanalysen unterscheiden, deren Resultate im Wesentlichen stabil sein sollten (vgl. Chattoe et al. 2000; Saam 1999).

e) Unabhängige Replikation. Empirische Tests der Implikationen agentenbasierter Computermodelle sind nur dann sinnvoll, wenn Modellierer darauf vertrauen können, dass sie auch die beabsichtigten theoretischen Annahmen implementiert haben. Edmonds und Hales (2003) weisen auf die Schwierigkeit der Validierung von Computer-

modellen hin, die auf verschiedene Weisen implementiert werden können. Die Autoren empfehlen, Modellresultate mit wenigstens zwei unabhängigen Programmierungen zu replizieren. Sie begründen dies mit dem instruktiven Beispiel ihrer Reimplementation eines Modells der ähnlichkeitsbasierten Partnerselektion von Riolo, Cohen und Axelrod (2001). Edmonds und Hales zeigen, dass ihre Reimplementation zunächst kleine Unterschiede zu den ursprünglich publizierten Resultaten generiert. Die Autoren konnten diese Unterschiede dann damit beseitigen, dass sie einen spezifischen Replikationsmechanismus verwendeten, der der Beschreibung des Originalmodells nicht widersprach. Es stellte sich aber heraus, dass in der neuen Modellversion die Bedingungen, unter denen ähnlichkeitsbasierte Partnerselektion zu Kooperation führte, weitaus restriktiver waren als Riolo und seine Koautoren angegeben hatten.

f) Modelle angleichen und vergleichen. Modellreplikation ist hilfreich um sicherzustellen, dass eine Implementation das theoretisch intendierte Modell repräsentiert. Axtell und andere (Axtell et al. 1996) schlagen vor, darüber hinaus verschiedene Modelle des gleichen Phänomens einander anzugleichen (model alignment), so dass ein genauer Vergleich diejenigen Modellannahmen aufdecken kann, die Unterschiede in den Resultaten verursachen. Im Idealfall kann dies dazu führen, dass verschiedene Designs als Spezialfälle eines allgemeineren Modells theoretisch integriert werden. Ein Beispiel ist Flache und Macys (2002) analytische Integration zweier unterschiedlicher Modelle von Reinforcement-Lernen. Vollständige Integration ist vermutlich bei komplexeren Multiagenten-Modellen nicht immer möglich. Wir meinen aber, dass Forscher versuchen sollten, verschiedene konzeptuelle Modelle des gleichen sozialen Phänomens so weit wie möglich einander anzugleichen, um damit die wichtigen Unterschiede zwischen Modellen von den unwichtigen zu trennen. Wenn dies gelingt, wissen Modellbauer mit größerer Sicherheit, welche Details vernachlässigbar sind.

g) Externe Validität testen. Computerexperimente testen die interne Validität oder Konsistenz einer Theorie. Ohne interne Validität ist es sinnlos, externe Validität zu testen. Wenn Forscher hinreichend überzeugt sind, eine konsistente Theorie zu haben, ist es an der Zeit, über empirische Tests in Laborexperimenten oder mit Felddaten nachzudenken. Die Arbeit von Mosler und Brucks (2001) ist ein gutes Beispiel für eine agentenbasierte Computermodellierung, die auf eine Feldanwendung abzielt. Die Autoren schlagen ein Modell sozialer Dynamik bei der Nutzung begrenzter natürlicher Ressourcen vor. Mosler und Brucks gehen über frühere Arbeiten (z.B. Macy 1990) hinaus, indem sie in ihrer Modellierung individueller Nutzungsentscheidungen Kosten-Nutzen-Abwägungen mit normativen Handlungsbewertungen auf der Basis sozialen Vergleichs theoretisch integrieren. Die Autoren setzen dann die Parameter ihres Modells systematisch zu Bedingungen in Beziehung, die in früheren Laborexperimenten mit kollektiven Handlungsdynamiken untersucht wurden. Mosler und Brucks argumentieren, dass sie so zahlreiche bekannte empirische Resultate mit einem einheitlichen Modell erklären können. In einer weiteren Arbeit setzen Mosler und Tobias (2000) ein ähnliches Modell gezielt zur Analyse der Effektivität verschiedener praktisch anwendbarer Mobilisierungsstrategien für kollektive Handlungen ein.

2. Werkzeuge

Die Implementation fast aller Modelle, die wir in diesem Beitrag besprochen haben, ist mit einer objektorientierten Allzweckprogrammiersprache wie JAVA, Delphi oder C++ und guten Basiskenntnissen dieser Sprachen möglich. Das Konzept eines Objektes in einer objektorientierten Programmiersprache entspricht in natürlicher Weise dem Konzept eines Agenten. Der Modellierer gibt einem Objekt bestimmte Eigenschaften, die symbolisch repräsentierbar sind. Beim Modell Sakodas muss ein Agent etwa einen Standort im zellularen Gitter in Form von numerischen Koordinaten haben und eine Gruppenzugehörigkeit, die zum Beispiel durch einen festen nominalen Wert definiert wird. Darüber hinaus haben Objekte Methoden, d.h. berechenbare Funktionen, nach denen Eigenschaften sich verändern. Methoden berechnen Veränderung in Abhängigkeit von den eigenen, aktuellen Eigenschaften eines Agenten und denen anderer Objekte (Agenten). Dabei kann ein Teil der Eigenschaften eines Objektes „privat" sein, d.h. für Methoden anderer Objekte nicht sichtbar. Bei einer Implementation des Modells von Sakoda müssen Gruppenzugehörigkeit und Standort eines Agenten von außen sichtbar sein. Agenten sollten darüber hinaus über eine Methode verfügen, die den Attraktivitätswert potentieller Standorte aus Sicht des Agenten berechnet. Das Ergebnis dieser Berechnung braucht aber anderen Agenten nicht bekannt zu sein.

Der Nachteil allgemeiner Programmiersprachen ist, dass Funktionen, die bei ABCM immer wieder gebraucht werden, nicht als Standardmodule verfügbar sind. Um ABCM zu unterstützen, wurden und werden eine Reihe von Softwarewerkzeugen entwickelt und leicht zugänglich gemacht, die etwa graphische und statistische Module enthalten, aber vor allem auch fertig verwendbare Schablonen für bestimmte Agententypen zur Verfügung stellen oder für das Erzeugen von Netzwerkstrukturen, genetischen Algorithmen oder neuronalen Netzwerken in Agentenpopulationen. Viele Werkzeuge sind eine Softwarebibliothek, die eine allgemeine Programmiersprache erweitert. Tobias und Hofman (2004) besprechen vergleichend eine Reihe von JAVA-basierten Bibliotheken, geben in ihrem Artikel aber auch eine Übersicht über eine Vielzahl weiterer ABCM-Werkzeuge und die dazugehörigen Webseiten. Wir wollen an dieser Stelle keine bestimmten Empfehlungen aussprechen, da die Nützlichkeit einer Software stark von Erfahrung und Kenntnissen des Anwenders sowie dem Modellierungsziel abhängt. Ein typisches Beispiel für ein vielseitiges und mächtiges Werkzeug ist das am Santa Fe-Institut entwickelte SWARM,[5] das aber relativ viel Programmiererfahrung erfordert. Ein recht einfach erlernbares, aber eingeschränkteres Werkzeug ist NetLogo.[6] MIMOSE[7] ist schließlich eine an der Universität Koblenz entwickelte Programmiersprache, die zwar nur in einer Unix-Umgebung verfügbar ist, sich dafür aber sehr gut zur Darstellung der konzeptionellen Struktur eines Modells eignet (siehe z.B. Saam 1999).

5 http://www.swarm.org
6 http://ccl.northwestern.edu/netlogo/
7 http://www.uni-koblenz.de/~moeh/projekte/mimose.html

IV. Ausblick

Wir haben in unserer Betrachtung agentenbasierter Computermodellierung spieltheoretische Modelle sozialer Ordnung mit Computermodellen verglichen. Durch diesen Schwerpunkt konnten andere, gleichermaßen wichtige und interessante Anwendungen der agentenbasierten Computermodellierung nicht behandelt werden. Wir haben insbesondere Modellierungsansätze vernachlässigt, die explizit auf Politikberatung und empirische Anwendung zielen. Dazu gehören neben anderen auch Arbeiten, die Verkehrsdynamiken in großem Maßstab vorherzusagen versuchen (Helbing und Nagel 2003), oder Analysen der Nutzung und des Schutzes natürlicher Ressourcen in einem Flussbett (Doran 2001). Solche Modelle ähneln Mikrosimulationen, die ab etwa 1970 angewendet wurden, weil sie große Mengen an empirischen Daten in die Modellierung der Populationsdynamik einfüttern. Anders als traditionelle Mikrosimulationen beschreiben agentenbasierte Modelle aber individuelle Akteure explizit als autonome Agenten, die ihren individuellen Zielsetzungen nachstreben und aktiv auf Veränderungen in ihrer Handlungssituation reagieren, die aus dem Handeln anderer Agenten entstehen. Wir sind davon überzeugt, dass solche Arbeiten langfristig von großem Nutzen für Politikberatung sein können. Wir teilen aber Dorans (2001) Besorgnis, dass datenbasierte Modelle komplexer empirischer Anwendungen beim heutigen Stand der Modellentwicklung nur unter großem Aufwand, wenn überhaupt, auf Konsistenz getestet und validiert werden können.

Agentenbasierte Computermodellierung ist ein leistungsfähiger neuer Ansatz individualistischer Theorieentwicklung. Er verbindet formale Modellierung kollektiven Verhaltens mit komplexeren Verhaltensannahmen als sie die orthodoxe Spieltheorie berücksichtigen kann. Agentenbasierte Computermodelle sind damit ein ideales Werkzeug, um Hypothesen über makrosoziale Dynamiken aus Mikromodellen abzuleiten, die Prinzipien sozialer Rationalität und „schneller und fruchtbarer" Entscheidungsheuristiken implementieren. In systematischen Computerexperimenten können theoretische Aussagen über Effekte struktureller Randbedingungen, zum Beispiel Netzwerkstrukturen, auf soziale Dynamiken abgeleitet werden. Wenn die Anwendung agentenbasierter Computermodellierung durch disziplinierende methodische Prinzipien gesteuert wird, dann kann sie unserer Meinung nach erheblich zum Verständnis der „bottom up" Entstehung sozialer Strukturen aus individuellem Verhalten beitragen.

Literatur

Axelrod, Robert, 1997: The Complexity of Cooperation. Princeton, NJ: Princeton University Press.

Axelrod, Robert, 1984: The Evolution of Cooperation. New York: Basic Books.

Axelrod, Robert, Joshua Epstein und *Michael Cohen*, 1996: Aligning Simulation Models: A Case Study and Results. Computational and Mathematical Organization Theory 1: 123–141.

Bendor, Jonathan, und *Dilip Mookherjee,* 1987: Institutional Structure and the Logic of Ongoing Collective Action. American Political Science Review 81: 129–154.

Bendor, Jonathan, und *Piotr Swistak,* 2001: The Evolution of Norms. American Journal of Sociology 106: 1493–1545.

Binmore, Ken, 1998: Review of ‚R. Axelrod, The Complexity of Cooperation: Agent-based Models of Competition and Collaboration; Princeton: UP 1997'. Journal of Artificial Societies and Social Simulation. http://www.soc.surrey.ac.uk./JASSS/1/1/review1.html

Buskens, Vincent, 2002: Social Networks and Trust. Dordrecht: Kluwer Academic Publishers.

Buskens, Vincent, 2003: Trust in Triads: Effects of Exit, Control, and Learning. Games and Economic Behavior 42: 235–252.

Camerer, Colin F., 2003: Behavioral Game Theory. Experiments in Strategic Interaction. New York: Russel Sage Foundation.

Castelfranchi, Cristiano, Rosaria Conte und *Marco Paolucci*, 1998: Normative Reputation and the Costs of Compliance. Journal of Artificial Societies and Social Simulation 1. http://jasss.soc.surrey.ac.uk/1/3/3.html

Chattoe, Edmund, 1998: Just How (Un)realistic are Evolutionary Algorithms as Representations of Social Processes? Journal of Artificial Societies and Social Simulation. 1. http://jasss.soc.surrey.ac.uk/1/3/2.html

Chattoe, Edmund, Nicole J. Saam und *Michael Möhring*, 2000: Sensitivity Analysis in the Social Sciences: Problems and Prospects. S. 243–273 in: *Ramzi Suleiman, Klaus G. Troitzsch* und *Nigel Gilbert* (Hg.), Tools and Techniques for Social Science Simulation. Heidelberg: Physica.

Coleman, James S., 1987: Psychological Structure and Social Structure in Economic Models. S. 181–185 in: *Robin M. Hogarth* und *Melvin W. Reder* (Hg.), Rational Choice. The Contrast between Economics and Psychology. Chicago: University of Chicago Press.

Coleman, James S., 1990: Foundations of Social Theory. Cambridge, MA: Harvard University Press.

Conte, Rosaria, Bruce Edmonds, Scott Moss und *R. Keith Sawyer,* 2001: Sociology and Social Theory in Agent Based Simulation: A Symposium. Computational and Mathematical Organization Theory 7: 183–205.

Diekmann, Andreas, und *Klaus Manhart*, 1989: Kooperative Strategien im Gefangenendilemma. Analyse & Kritik 11: 134–153.

Doran, Jim, 2001: Can Agent-based Modelling REALLY be Useful? S. 57–108 in: *Nicole J.Saam* und *Bernd Schmidt* (Hg.), Cooperative Agents. Applications in the Social Sciences. Dordrecht. Kluwer Academic Publishers.

Edmonds, Bruce, und *David Hales*, 2003: Replication, Replication and Replication: Some Hard Lessons from Model Alignment. Journal of Artificial Societies and Social Simulation 6. http://jasss.soc.surrey.ac.uk/6/4/11.html

Erev, Ido, und *Alvin E. Roth*, 1998: Predicting how People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. American Economic Review 88: 848–879.

Esser, Hartmut, 2001: Soziologie. Spezielle Grundlagen. Band 6: Sinn und Kultur. Frankfurt a.M.: Campus.

Etzioni, Amitai, 2001: On Social and Moral Revival. Journal of Political Philosophy 9: 356–371.

Fehr, Ernst, und *Simon Gächter*, 2000: Cooperation and Punishment in Public Goods Experiments. American Economic Review 90: 980–994.

Flache, Andreas, und *Rainer Hegselmann*, 1999: Rationality vs. Learning in the Evolution of Solidarity Networks: A Theoretical Comparison. Computational and Mathematical Organization Theory 5: 97–127.

Flache, Andreas, und *Rainer Hegselmann,* 2000: Abschlussbericht zum DFG Projekt „Dynamik sozialer Dilemma-Situationen". Universität Bayreuth, Institut für Philosophie. http://pe.uni-bayreuth.de/downloads/dfg/index.html

Flache, Andreas, und *Michael W. Macy*, 2002: Stochastic Collusion and the Power Law of Learning: A General Reinforcement Learning Model of Cooperation. Journal of Conflict Resolution 46: 629–653.

Fudenberg, Drew, und *David K. Levine*, 1998: The Theory of Learning in Games. Boston: MIT Press.

Gotts, Nick M., J. Gareth Polhill und *Alistair N.R Law*, 2003: Agent-Based-Simulation in the Study of Social Dilemmas. Artificial Intelligence Review 19: 3–92.

Hechter, Michael, 1988: Principles of Group Solidarity. Berkeley: University of California Press.

Hegselmann, Rainer, 1996: Solidarität unter Ungleichen. S. 105–128 in: *Rainer Hegselmann* und *Hans O. Peitgen* (Hg.), Modelle sozialer Dynamiken – Ordnung, Chaos und Komplexität. Wien: Hölder-Pichler-Tempsky.

Hegselmann, Rainer, und *Andreas Flache*, 1998: Understanding Complex Social Dynamics: A Plea For Cellular Automata Based Modelling. Journal of Artificial Societies and Social Simulation 1. http://www.soc.surrey.ac.uk/JASSS/1/3/1.html

Helbing, Dirk, und *Kai Nagel*, 2003: Verkehrsdynamik und urbane Systeme. Physik Journal 2: 35–41.

Homans, George C., 1974: Social Behavior. Its Elementary Forms. New York: Harcourt Brace Jovanovich.

Jackson, Matthew O., 2004: A Survey of Models of Network Formation: Stability and Efficiency. In: *Gabrielle Demange* und *Myrna Wooders* (Hg.), Group Formation in Economics: Networks, Clubs and Coalitions. Cambridge: Cambridge University Press.

Jager, Wander, Marco A. Janssen, Bert J.M. de Vries, Jodi De Greef und *Charles A.J. Vlek.* 2000: Behavior in Commons Dilemmas: Homo-economicus and Homo-psychologicus in an Ecological-economic Model. Ecological Economics 35: 357–379.

Jaffe, Klaus, 2002: An Economic Analysis of Altruism: Who Benefits from Altruistic Acts? Journal of Artificial Societies and Social Simulation 5. http://jasss.soc.surrey.ac.uk/5/3/3.html

Kollock, Peter, 1993: „An Eye for an Eye Leaves Everyone Blind". American Sociological Review 58: 768–786.

Komter, Aafke, 1996: Reciprocity as a Principle of Exclusion. Sociology 30: 299–316.

Klüver, Jürgen, und *Christina Stoica*, 2003: Simulations of Group Dynamics with Different Models. Journal of Artificial Societies and Social Simulation 6. http://jasss.soc.surrey.ac.uk/6/4/8.html

Lepperhoff, Niels, 2000: Dreamscape: Simulation der Entstehung von Normen im Naturzustand mittels eines computerbasierten Modells des Rational-Choice-Ansatzes. Zeitschrift für Soziologie 29: 463–484.

Lindenberg, Siegwart M., 2001: Social Rationality versus Rational Egoism. S. 635–668 in: *Jonathan Turner* (Hg.), Handbook of Sociological Theory. New York: Kluwer.

Linster, Bruce G., 1992: Evolutionary Stability in the Repeated Prisoners' Dilemma Played by Two-state Moore Machines. Southern Economic Journal 58: 880–903.

Macy, Michael W., 1990: Learning Theory and the Logic of Critical Mass. American Sociological Review 55: 809–826.

Macy, Michael W., 1996: Natural Selection and Social Learning in Prisoner's Dilemma: Co-adaptation with Genetical Algorithms and Artificial Neural Networks. Sociological Methods and Research 25: 103–137.

Macy, Michael W., und *Andreas Flache*, 2002: Learning Dynamics in Social Dilemmas. Proceedings of the National Academy of Sciences USA 99: 7229–7236.

Macy, Michael W., und *Robb Willer*, 2002: From Factors to Actors: Computational Sociology and Agent-based Modelling. Annual Review of Sociology 28: 143–66.

Mark, Noah, 2003: Culture and Competition: Homophily and Distancing Explanations for Cultural Niches. American Sociological Review 68: 319–345.

Moretti, Sabrina, 2002: Computer Simulation in Sociology: What Contribution? Social Science Computer Review 20: 43–57.

Mosler, Hans Joachim, und *Wernher M. Brucks*, 2003: Integrating Commons Dilemma Findings in a General Dynamic Model of Cooperative Behavior in a Resource Crisis. European Journal of Social Psychology 33: 119–133.

Mosler, Hans Joachim, und *Robert Tobias*, 2000: Die Organisation kollektiver Aktionen durch Beeinflussung der individuellen Teilnahmeentscheidung. Eine Simulationsstudie. Kölner Zeitschrift für Soziologie und Sozialpsychologie 52: 264–290.

Nowak, Martin A., und *Robert May,* 1993: The Spatial Dilemmas of Evolution. International Journal of Bifurcation and Chaos 3: 35–78.

Nowak, Martin A., und *Karl Sigmund*, 1993: A Strategy of Win-Stay, Lose-Shift that Outperforms Tit-for-Tat in the Prisoner's Dilemma Game. Nature 364: 56–58.

Olson, Mancur, 1965: The Logic of Collective Action. Cambridge, Mass.: Harvard University Press.

Orcutt, Guy H., Joachim Merz und *Hermann Quinke* (Hg.), 1986: Microanalytic Simulation Models to Support Social and Financial Policy. Information Research and Resource Reports 7. Amsterdam: North-Holland.

Podolny, Joel M., und *Karen L. Page*, 1998: Network Forms of Organizations. Annual Review of Sociology 24: 57–76.

Probst Daniel A., 1996: On Evolution and Learning in Games. Universität Mannheim, Dissertation.

Rapoport, Anatol, und *Albert M. Chammah*, 1965: Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: Michigan University Press.

Raub, Werner, 1988: Problematic Social Situations and the Large Number Dilemma. Journal of Mathematical Sociology 13: 311–357.

Raub, Werner, und *Jeroen Weesie*, 1990: Reputation and Efficiency in Social Interaction: an Example of Network Effects. American Journal of Sociology 96: 626–654.

Raub, Werner, und *Vincent Buskens*, 2004. Spietheoretische Modellierungen und empirische Anwendungen in der Soziologie. In: *Andreas Diekmann* (Hg.), Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Wiesbaden: VS-Verlag für Sozialwissenschaften.

Riolo, Rick L., Michael Cohen, und *Robert Axelrod*, 2001: Evolution of Cooperation without Reciprocity. Nature 411: 441–443.

Roth, Alvin E., und *Ido Erev*, 1995: Learning in Extensive-form Games: Experimental Data and Simple Dynamic Models in Intermediate Term. Games and Economic Behavior, Special Issue: Nobel Symposium 8: 164–212.

Saam, Nicole J., 1999: Simulating the Micro-Macro Link: New Approaches to an Old Problem and an Application to Military Coups. Sociological Methodology 29: 43–79.

Saam, Nicole J., und *Andreas Harrer*, 1999: Simulating Norms, Social Inequality, and Functional Change in Artificial Societies. Journal of Artificial Societies and Social Simulation 2. http://jasss.soc.surrey.ac.uk/2/1/2.html

Saam, Nicole J., und *Bernd Schmidt* (Hg.), 2001: Cooperative Agents. Applications in the Social Sciences. Dordrecht. Kluwer Academic Publishers.

Sakoda, James M., 1971: The Checkerboard Model of Social Interaction: Journal of Mathematical Sociology 1: 119–132.

Sawyer, R. Keith, 2003: Artificial Societies: Multiagent Systems and the Micro-Macro Link in Sociological Theory. Sociological Methods and Research 31: 325–363.

Schelling, Thomas C., 1971: Dynamic Models of Segregation. Journal of Mathematical Sociology 1: 143–186.

Schmidt, Bernd, 2001: What are Agents and What are they for? S. 5–20 in: *Nicole J. Saam* und *Bernd Schmidt* (Hg.), Cooperative Agents. Applications in the Social Sciences. Dordrecht. Kluwer Academic Publishers.

Schüssler, Rudolf, 1989: Exit Threats and Cooperation under Anonymity. The Journal of Conflict Resolution 33: 728–749.

Schüssler, Rudolf, und *Ulrich Sandten*, 2000: Exit, Anonymity and the Chances of Egoistical Cooperation. Analyse & Kritik 22: 114–129.

Smith, Thomas S., und *Gregory T. Stevens*, 1999: The Architecture of Small Networks: Strong Interaction and Dynamic Organization in Small Social Systems. American Sociological Review 64: 403–420.

Schweitzer, Frank, 2003: Meinungsbildung, Kommunikation und Kooperation aus physikalischer Perspektive. Physik Journal 2: 57–62.

Stokman, Frans N., und *Evelien P.H. Zeggelink*, 1996: Self-organizing Friendship Networks. S. 385–418 in: *Wim B.G. Liebrand* und *David M. Messick* (Hg.), Frontiers in Social Dilemmas Research. Berlin: Springer.

Taylor, Michael, 1987: The Possibility of Cooperation. Cambridge: Cambridge University Press.

Takahashi, Noboyuki, 2000: The Emergence of Generalized Exchange. American Journal of Sociology 105: 1105–1134.

Tobias, Robert, und *Carole Hofmann*, 2004: Evaluation of Free Java-libraries for Social-scientific Agent Based Simulation. Journal of Artificial Societies and Social Simulation 7. http://jasss.soc.surrey.ac.uk/7/1/6.html

Todd, Peter M., und *Gerd Gigerenzer*, 2003: Bounding Rationality to the World. Journal of Economic Psychology 24: 143–165.

Troitzsch, Klaus G., 2000: Computersimulationen in den Sozialwissenschaften. S. 181–203 in: *Dietmar Herz* und *Andreas Blätte* (Hg.), Simulation und Planspiel in den Sozialwissenschaften. Eine Bestandsaufnahme der internationalen Diskussion. Münster: Lit Verlag.

Troitzsch, Klaus G., 2004: Dynamische Systemmodelle. In: *Andreas Diekmann* (Hg.), Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Wiesbaden: VS-Verlag für Sozialwissenschaften.

Udehn, Lars, 2002: The Changing Faces of Methodological Individualism. Annual Review of Sociology 28: 479–507.

Vanberg, Viktor J., 2002: Rational Choice vs. Program-Based Behavior. Alternative Theoretical Approaches and their Relevance of the Study of Institutions. Rationality and Society 14: 7–54.

Vanberg, Viktor J., und *Roger D. Congleton*, 1992: Rationality, Morality and Exit. American Political Science Review 86: 418–431.

Vetschera, Rudolf, 2003: Experimentation and Learning in Repeated Cooperation. Computational & Mathematical Organization Theory 9: 37–60.

Vieth, Manuela, 2003: Die Evolution von Fairnessnormen im Ultimatumspiel. Zeitschrift für Soziologie 32: 346–367.

Voss, Thomas, 1985: Rationale Akteure und Soziale Institutionen. München: Oldenbourg.

Voss, Thomas, 1990: Eine Individualistische Theorie der Evolution von Regeln und einige Anwendungsmöglichkeiten in der Organisationsforschung. München: Sozialwissenschaftliche Fakultät der Universität München (Habilitationsschrift).

Voss, Thomas, 2001: Game Theoretical Perspectives on the Emergence of Social Norms. S. 105–136 in: *Michael Hechter* und *Karl Dieter Opp* (Hg.), Social Norms. New York: Russel Sage Foundation.

Voss, Thomas, und *Martin Abraham*, 2000: Rational Choice Theory in Sociology: A Survey. S. 50–83 in: *Stella R. Quah* und *Arnaud Sales* (Hg.), The International Handbook of Sociology. London: Sage.

Wittek, Rafael P. M., und *Andreas Flache*, 2003: Rational Choice und Organisationstheorie. S. 55–87 in: *Jutta Allmendinger* und *Thomas Hinz* (Hg.), Organisationssoziologie. Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderheft 42.

Wooldridge Michael, und *Nicholas R. Jennings*, 1995: Intelligent Agents: Theory and Practice. Knowledge Engineering Review 10: 115–152.

Younger, Stephen, 2004: Reciprocity, Normative Reputation, and the Development of Mutual Obligation in Gift-Giving Societies. Journal of Artificial Societies and Social Simulation 7. http://jasss.soc.surrey.ac.uk/7/1/5.html

SPIELTHEORETISCHE MODELLIERUNGEN UND EMPIRISCHE ANWENDUNGEN IN DER SOZIOLOGIE*

Werner Raub und Vincent Buskens

Zusammenfassung: Die Spieltheorie ist ein Instrument zur Beschreibung und Analyse von sozialen Situationen, in denen Akteure interdependent, also voneinander wechselseitig abhängig sind. In unserem Beitrag präsentieren wir nach einer Skizze zentraler Konzepte und Annahmen der Spieltheorie einige Beispiele für Anwendungen spieltheoretischer Modellierungen in der Soziologie. Diese Anwendungen betreffen ein zentrales Thema der soziologischen Theorie und spezieller Soziologien wie der Organisations-, Wirtschafts- und Rechtssoziologie, nämlich Effekte sozialer Bedingungen auf Vertrauen und Kooperation und die Organisation des wirtschaftlichen Handelns und des ökonomischen Tauschs. Wir zeigen, wie spieltheoretische Modellierungen zu empirisch prüfbaren Hypothesen führen, und präsentieren Überprüfungen aus unterschiedlich angelegten Studien, nämlich einer Feldstudie, einer Vignettenstudie und einem Experiment. Wir wollen also mit Hilfe von Beispielen zeigen, dass spieltheoretische Modellierungen einen Beitrag leisten können zur Integration von systematischer Theoriebildung in Gestalt des Rational Choice-Ansatzes und empirischer Forschung in der Soziologie.

I. Einleitung

„Soziologie ... soll heißen: eine Wissenschaft, welche soziales Handeln deutend verstehen und dadurch in seinem Ablauf und seinen Wirkungen ursächlich erklären will ... ‚Soziales' Handeln ... soll ein solches Handeln heißen, welches seinem von dem oder den Handelnden gemeinten Sinn nach *auf das Verhalten anderer bezogen wird und daran in seinem Ablauf orientiert ist*" (Weber 1976: 1; Hervorhebungen so nicht im Origi-

* Wir danken Jeroen Weesie, Chris Snijders und den anderen Mitgliedern unserer Utrechter Arbeitsgruppe „Cooperation in Social and Economic Relations" (früher „The Management of Matches") für nützliche Kommentare und Diskussionen zu zahlreichen Themen dieses Beitrags. Auch einem anonymen Gutachter der KZfSS verdanken wir einige nützliche Hinweise. An den empirischen Studien, die wir verwenden, haben Ronald Batenburg, Boris Blumberg, Gideon Keren, Gerrit Rooks, Chris Snijders, Frits Tazelaar und Jeroen Weesie mitgewirkt. Die vorliegende Arbeit entstand im Rahmen eines durch die Niederländische Organisation für wissenschaftliche Forschung (NWO) unterstützten Forschungsprogramms „The Management of Matches" (NWO PGS 50-370 und S96-168). Finanzielle Unterstützung für die Feldstudie zu IT-Transaktionen niederländischer Betriebe leistete die Forschungsstiftung (NEVI Research Stichting, NRS) der Niederländischen Vereinigung für Einkaufsmanagement (Nederlandse Vereniging voor Inkoopmanagement, NEVI). Der Beitrag von Buskens ist Teil eines durch die Königlich Niederländische Akademie der Wissenschaften (KNAW) geförderten Projekts „Third-Party Effects in Cooperation Problems". Einige Passagen der Abschnitte IV. und V. überlappen mit den entsprechenden Teilen in Buskens und Raub (2004).

nal). Webers berühmte Definition der Soziologie hebt ab auf Interdependenzen zwischen Akteuren und darauf, dass die Akteure diese Interdependenzen „in Rechnung stellen". Die Spieltheorie ist der Zweig der Theorie rationalen Handelns, der Situationen mit Interdependenzen zwischen Akteuren modelliert. Die Theorie liefert uns „Bausteine" (Konzepte, Annahmen und Theoreme), mit denen wir spezifizieren können, wie sich rationale Akteure in solchen Situationen verhalten, also Akteure, die sich so verhalten, als ob sie versuchen, in Entscheidungssituationen unter Restriktionen ihre Präferenzen möglichst gut zu realisieren (vgl. Diekmann und Voss 2004 für eine knappe neuere Darstellung des Rational Choice-Ansatzes) und als ob sie dabei Interdependenzen und das rationale Verhalten der anderen Akteure berücksichtigen. Die Spieltheorie ist damit ein maßgeschneidertes Instrument für eine Soziologie in Webers Sinn. Sie ist auch deshalb ein für die Soziologie interessantes Instrument, weil einerseits Interdependenzen ein paradigmatisches Beispiel für *soziale* Bedingungen individuellen Verhaltens sind und weil – wie wir noch sehen werden – *soziale* Konsequenzen interdependenten Handelns, nämlich kollektive Effekte wie z.B. die Pareto-Optimalität oder Suboptimalität der „Lösung" eines Spiels zu den zentralen Explananda der Theorie zählen. In Essers (1993: 94–97) griffiger Beschreibung der drei Schritte bei der Erklärung kollektiver Phänomene entsprechend Colemans (1991: Kap. 1) bekanntem Schema: Die Spieltheorie bietet uns Instrumente für die Analyse der „Logik der Situation" (wie beeinflussen soziale Bedingungen die Präferenzen und Informationen der Akteure?), der „Logik der Selektion" (welche Handlungen führen die Akteure aus?) und der „Logik der Aggregation" (welche kollektiven Effekte resultieren aus diesen Handlungen?).

In diesem Beitrag wollen wir an einigen Beispielen zeigen, dass spieltheoretische Modellierungen durchaus dazu beitragen können, den oft (vgl. z.B. Green und Shapiro 1999) beklagten Graben zwischen dem Rational Choice-Ansatz und einer empirisch orientierten Soziologie wenn nicht zu schließen, dann doch jedenfalls schrittweise und systematisch zu verkleinern, ganz im Sinn von Goldthorpes (2000: Kap. 5) Plädoyer für eine Allianz zwischen dem Rational Choice-Ansatz (RAT) und der quantitativen Analyse sozialwissenschaftlicher Daten (QAD).[1] Wir möchten also mit Beispielen belegen, dass die Spieltheorie mit einem gewissen Recht zu den „Methoden der Sozialforschung" gezählt werden kann. Wir werden dazu in einem ersten Schritt zentrale Konzepte und Annahmen der Spieltheorie skizzieren. Danach präsentieren wir Anwendungen aus eigenen Arbeiten in einem Feld, das sowohl aus der Perspektive der soziologischen Theorie als auch aus der Perspektive spezieller Soziologien wie der Organisations-, Wirtschafts- und Rechtssoziologie interessant ist, nämlich Effekte sozialer Bedingungen auf Vertrauen und Kooperation und die Organisation des wirtschaftlichen Handelns und des ökonomischen Tauschs. Wir behandeln sowohl die Formulierung empirisch prüfbarer Hypothesen, die mit Hilfe spieltheoretischer Modelle generiert werden können, als auch die empirische Überprüfung dieser Hypothesen. Wir wählen unsere Beispiele so, dass empirische Studien mit unterschiedlichen Designs die Revue passieren, nämlich eine Feldstudie mit Surveydaten, eine Vignettenstudie und ein Ex-

1 Goldthorpes Plädoyer ist natürlich nicht unumstritten (vgl. diverse Beiträge in Blossfeld und Prein 1998), aber eine ausführliche Diskussion würde in diesem Beitrag zu weit führen.

periment. Wir komplettieren unseren Beitrag mit einem Anhang, in dem wir einige weiterführende Literaturhinweise zur Spieltheorie und ihren Anwendungen in der Soziologie und in anderen sozialwissenschaftlichen Disziplinen zusammenstellen.

II. Einige Bausteine der Spieltheorie

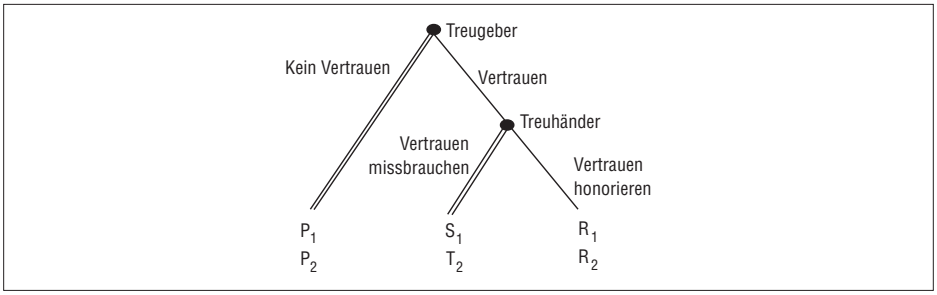
Gegenstand der Spieltheorie sind Situationen strategischer Interdependenz zwischen zwei oder mehr Akteuren. Damit ist gemeint, dass die Entscheidungen eines Akteurs Folgen haben für den oder die „Mitspieler" und umgekehrt. Wir beginnen mit einer knappen Skizze einiger Konzepte und Annahmen der Spieltheorie. Dabei gehen wir informell und exemplarisch vor. Uns ist daran gelegen, intuitives Verständnis zu fördern. Es geht uns nicht um den exakten Aufbau von Terminologie oder um den Beweis von Theoremen (vgl. ein beliebiges modernes Lehrbuch der Spieltheorie wie z.B. Rasmusen 1994 für detaillierte Diskussionen aller im Folgenden behandelten spieltheoretischen Konzepte und Annahmen, einschließlich verschiedener mehr technischer Annahmen, die wir in diesem Beitrag der Kürze halber implizit lassen).[2]

Als erstes Beispiel verwenden wir das *elementare Vertrauensspiel* (vgl. Camerer und Weigelt 1988; Dasgupta 1988; Kreps 1990a; Snijders 1996: Kap. 1–4; Buskens 2002: Kap. 1–3 für ausführliche Diskussionen). Das Spiel (vgl. *Abbildung 1*) modelliert zentrale Vertrauensprobleme im Sinn Colemans (1991: Kap. 5). An dem Spiel sind zwei Akteure beteiligt, der Treugeber und der Treuhänder. Das Spiel beginnt mit einem Zug des Treugebers. Dieser muss entscheiden, ob er dem Treuhänder vertraut oder nicht. Wird Vertrauen verweigert, dann endet die Interaktion, der Treugeber erhält eine „Auszahlung" P_1, der Treuhänder hingegen P_2. Wird Vertrauen gegeben – in Colemans Terminologie: der Treugeber überantwortet dem Treuhänder bestimmte wertvolle Ressourcen –, dann entscheidet danach der Treuhänder, ob er Vertrauen honoriert oder durch opportunistisches Verhalten missbraucht. Wird Vertrauen honoriert, dann erhalten Treugeber und Treuhänder Auszahlungen $R_i > P_i$, $i = 1,2$. Wird Vertrauen missbraucht, dann ist die Auszahlung für den Treugeber $S_1 < P_1$ und für den Treuhänder $T_2 > R_2$.

Abbildung 1 präsentiert die extensive Form des Vertrauensspiels. Die *extensive Form* enthält die zentralen „Bausteine" für spieltheoretische Analysen. Zu diesen Bausteinen gehört der *Spielbaum*. An jedem *Knoten* des Baums wird angegeben, welcher *Akteur (Spieler)* am Zug ist. Am Rande bemerken wir, dass in Anwendungen der Theorie die Akteure durchaus nicht immer natürliche Personen sein müssen. Vielfach wird vereinfachend und im Gegensatz zu einer radikalen Interpretation des methodologischen Individualismus angenommen, dass auch „korporative Akteure" (im Sinn von Coleman

2 Wir beschränken uns auf Grundzüge der nichtkooperativen Spieltheorie. In einem nichtkooperativen Spiel sind bindende Vereinbarungen zwischen Akteuren und bindende einseitige Verpflichtungen eines Akteurs ausgeschlossen, es sei denn, Vereinbarungen und Verpflichtungen werden explizit als Züge in der extensiven Form des Spiels modelliert. Man beachte – wir werden darauf zurückkommen –, dass damit nicht ausgeschlossen wird, dass sich die Akteure in einem nichtkooperativen Spiel kooperativ verhalten. Wir werden vielmehr sehen, dass Modelle für nichtkooperative Spiele gerade dazu verwendet werden können, Bedingungen für kooperatives Verhalten zu analysieren.

Abbildung 1: Das Vertrauensspiel ($S_1 < P_1 < R_1$, $P_2 < R_2 < T_2$)



1992) wie Firmen oder Staaten als Akteure behandelt werden können. Ein *Zug* eines Akteurs ist eine Handlung, die dieser Akteur an einem gegebenen Knoten ausführt. Am Ende des Baums finden wir die *Auszahlungen* der Akteure. Wir interpretieren diese Auszahlungen als kardinale Nutzenniveaus. Wir bemerken, dass die Spieltheorie *selbst* keine Annahmen darüber enthält, wie die Nutzenfunktionen der Akteure beschaffen sind. Lindenbergs (1981: 26) Bemerkung, dass die Nutzentheorie ohne Brückenannahmen über Nutzenargumente ein „leerer Sack" sei, gilt selbstverständlich auch für die Spieltheorie. Die Spieltheorie *selbst* sagt z.B. nichts darüber, ob der Nutzen eines Akteurs ausschließlich abhängt von materiellen (gegebenenfalls monetären) Belohnungen, wie die Risikopräferenzen des Akteurs beschaffen sind oder ob der Nutzen des Akteurs (auch) abhängt von den Resultaten der Interaktion für den anderen Akteur, von Fairnesserwägungen, „sozialen Orientierungen" o.Ä. Wir sehen nun auch deutlicher, in welchem Sinn Treugeber und Treuhänder im Vertrauensspiel interdependent sind. Ob der Treugeber nach gegebenem Vertrauen das Nutzenniveau R_1 oder aber S_1 realisiert, hängt ab vom Zug des Treuhänders. Ob umgekehrt der Treuhänder das Nutzenniveau T_2 bzw. R_2 oder aber das Nutzenniveau P_2 realisiert, ist abhängig vom Zug des Treugebers.

Ein zentraler Begriff der Spieltheorie ist der der Strategie eines Akteurs. Eine *Strategie* ist eine Verhaltensvorschrift für jeden Knoten, an dem der Akteur eine Entscheidung treffen muss (diese Charakterisierung des Strategiebegriffs enthält noch eine Ungenauigkeit, die wir etwas weiter unten beseitigen werden). Im elementaren Vertrauensspiel verfügt der Treugeber also über die beiden Strategien „Vertrauen geben" und „Vertrauen verweigern", während „Vertrauen honorieren" und „Vertrauen missbrauchen" die Strategien des Treuhänders sind. Man beachte, dass jeder der beiden Akteure im elementaren Vertrauensspiel jeweils nur an einem einzigen Knoten einen Zug macht. Dadurch kann der Eindruck entstehen, dass eine Strategie nichts anderes ist als ein Zug. Wir werden noch sehen, dass das im Allgemeinen natürlich nicht so ist: Strategien sind sehr viel „komplexere" Entitäten, nämlich Pläne für Verhalten, die in dem Sinn vollständig sind, dass sie für alle möglichen Umstände, die sich im Lauf eines Spiels ergeben können, das Verhalten des betreffenden Akteurs spezifizieren.

Bemerkung: In Einführungen in die Spieltheorie für Soziologen und auch in „soziologienahen" Anwendungen findet man häufig an Stelle der extensiven Form eines Spiels eine andere Darstellung, nämlich die *Normalform.* Die Normalform spezifiziert die Akteure, die Strategien jedes Akteurs

Abbildung 2: Normalform des Vertrauensspiels ($S_1 < P_1 < R_1$, $P_2 < R_2 < T_2$)

		Treuhänder	
		Vertrauen honorieren	Vertrauen missbrauchen
Treugeber	Vertrauen	R_1, R_2	S_1, T_2
	kein Vertrauen	P_1, P_2	P_1, P_2

und die Auszahlungsfunktion jedes Akteurs, also den erwarteten Nutzen eines Akteurs in Abhängigkeit von den gewählten Strategien aller Akteure. Bei Spielen mit zwei Akteuren lässt sich die Normalform übersichtlich in einer *Auszahlungsmatrix* zusammenfassen (vgl. *Abbildung 2* für den Fall des Vertrauensspiels). Wir bemerken, dass die Normalform bestimmte Details des Spiels „verbirgt", die in spieltheoretischen Analysen wichtig werden können, insbesondere die Reihenfolge, in der die Akteure ziehen.

Bislang haben wir Terminologie skizziert. Eine Theorie ist aber bekanntlich kein System von Begriffen, sondern ein System von Aussagen. Wenden wir uns nun also Annahmen der Spieltheorie zu. In der Spieltheorie wird *rationales Verhalten* der Akteure unterstellt: Jeder Akteur wählt eine Strategie, die seinen erwarteten Nutzen maximiert, gegeben die Opportunitäten bzw. Restriktionen, wie sie in der extensiven Form repräsentiert werden. In Situationen mit strategischer Interdependenz ist aber weniger deutlich als in Situationen ohne wechselseitige Abhängigkeiten, wie die Rationalitätsannahme zu spezifizieren ist. Die Konsequenzen der Strategiewahl von Akteur A für A hängen ja auch ab von den Strategien der Akteure B, C ... und umgekehrt. Man kann dies auch so ausdrücken, dass die Wahrscheinlichkeiten, mit denen bestimmte Konsequenzen auftreten, mit denen also A einen bestimmten Endpunkt des Baums erreicht, anders als im Fall von Entscheidungen unter Risiko („Spiele gegen die Natur") nicht nur von exogenen Ereignissen abhängen, sondern auch von den Entscheidungen anderer Akteure. A muss also antizipieren auf die Entscheidungen von B, C ... und umgekehrt. Es entsteht das Problem, dass rationale Akteure Erwartungen bilden müssen über das Verhalten anderer Akteure, deren Verhalten selbst abhängt von ihren Erwartungen ...

Man sieht, dass die zentrale Aufgabe der Spieltheorie, nämlich die Spezifikation der *Lösung* eines Spiels, also die Spezifikation der Kombination von Strategien, die rationale Akteure wählen werden, nicht ganz trivial ist. Ein wesentlicher und eleganter Beitrag zur Bearbeitung dieser Aufgabe ist der Gleichgewichtsbegriff von Nash (1951). Ein *Nash-Gleichgewicht* ist eine Kombination von Strategien derart, dass jeder Akteur mit seiner Gleichgewichtsstrategie seinen erwarteten Nutzen maximiert, gegeben die Gleichgewichtsstrategien aller anderen Akteure. Im Gleichgewicht ist die Strategie eines jeden Akteurs *beste Antwort* (nutzenmaximierend) gegen die Strategien aller anderen Akteure, und mithin hat keiner der Akteure einen Anreiz, einseitig von seiner Gleichgewichtsstrategie abzuweichen. Im elementaren Vertrauensspiel existiert offensichtlich genau ein solches Gleichgewicht (in *Abbildung 1* markieren wir die Züge, die die Akteure im Gleichgewicht ausführen, mit doppelten Linien, und in *Abbildung 2* markieren wir das Gleichgewicht durch fette Umrandung der entsprechenden Zelle der Auszahlungsmatrix): Der Treuhänder würde gegebenes Vertrauen missbrauchen und der Treugeber verweigert (sozusagen in Antizipation des Verhaltens des Treuhänders) Ver-

trauen. Man kann sich leicht klar machen, dass aus naheliegenden, wenn auch „starken" Annahmen folgt, dass die Lösung eines Spiels ein Nash-Gleichgewicht sein muss. Wenn (1) eine eindeutige Lösung existiert, wenn (2) alle Akteure sich so verhalten, als ob sie diese Lösung antizipieren und wenn (3) alle Akteure rational sind, dann folgt deduktiv (Beweis: indirekt), dass die Lösung ein Nash-Gleichgewicht sein muss.

Nashs Leistung (die ihm 1994 den Nobelpreis für Ökonomie einbrachte) liegt aber nicht nur darin, dass er den Gleichgewichtsbegriff entwickelt hat (für Begriffsbildung gibt es keine Nobelpreise, auch nicht in „soziologienahen" Disziplinen). Nash konnte beweisen, dass die *Existenz* eines Gleichgewichts in endlichen Spielen (also Spielen mit endlich vielen Akteuren, wobei jeder Akteur über endlich viele Strategien verfügt) gesichert ist.[3] Viele Spiele haben jedoch mehr als ein Gleichgewicht. Dass eine Strategienkombination ein Gleichgewicht ist, ist also lediglich eine notwendige, aber noch keine hinreichende Bedingung dafür, dass diese Strategienkombination auch die Lösung des Spiels ist. Dieses *„Gleichgewichtsauswahlproblem"* wird als ein, wenn nicht *das* fundamentale Problem der Spieltheorie angesehen (vgl. Kreps 1990b: Kap. 5 für eine gut zugängliche Diskussion und für Literaturhinweise).

Zur Abrundung unserer Skizze und auch als theoretischen Hintergrund für unsere empirischen Anwendungen betrachten wir noch zwei Varianten unseres elementaren Vertrauensspiels. Eine bislang implizite Annahme in unserer Charakterisierung des elementaren Vertrauensspiels war die der vollständigen Information der Akteure über alle Aspekte der Situation. Tatsächlich sind Akteure in sozialen Interaktionen aber häufig nur mehr oder weniger unvollständig informiert. Insbesondere sind sie häufig unvollständig informiert über die Handlungsmöglichkeiten und die Anreize des Partners. Der Treugeber kann etwa unsicher sein über den Nutzen des Missbrauchs von Vertrauen für den Treuhänder und es für möglich halten, dass es für den Treuhänder attraktiver ist, gegebenes Vertrauen zu honorieren, als gegebenes Vertrauen zu missbrauchen. Man denke etwa an die Möglichkeit, dass der Treuhänder Normen und Werte internalisiert haben könnte, die den Missbrauch von Vertrauen mit ausreichend schweren „inneren Sanktionen" bestrafen. Unter unvollständiger Information kann der Treugeber typischerweise nicht mehr sicher vorhersagen, ob ein rationaler Treuhänder gegebenes Vertrauen honorieren oder missbrauchen wird. Mit spieltheoretischen Mitteln können wir dies in einem *Vertrauensspiel mit unvollständiger Information* („incomplete information" im technischen Sinn von Rasmusen 1994: Kap. 2) modellieren (vgl. die extensive Form in *Abbildung 3*). Das Spiel beginnt mit einem Zufallszug der „Natur", bei dem der Anreiz des Treuhänders für den Missbrauch von Vertrauen bestimmt wird (man sagt auch, dass die Natur den *Typ* des Treuhänders bestimmt). Mit einer Wahrscheinlichkeit π ist die Auszahlung beim Missbrauch von Vertrauen $T_2^* < R_2$ (der Treuhänder hat Normen

3 Man beachte, dass wir in diesem Beitrag mit „Strategien" stets „reine Strategien" im technischen Sinn der Spieltheorie meinen. Ein Akteur wählt eine „gemischte Strategie", wenn er mit Hilfe eines Zufallsmechanismus eine Wahl trifft zwischen reinen Strategien, wenn er also etwa als Treugeber im elementaren Vertrauensspiel mit einer Wahrscheinlichkeit p Vertrauen gibt (die reine Strategie „Vertrauen geben" spielt) und mit der komplementären Wahrscheinlichkeit $1 - p$ Vertrauen verweigert (die reine Strategie „Vertrauen verweigern" spielt). Der Satz von Nash besagt, dass in endlichen Spielen jedenfalls die Existenz eines Gleichgewichts in gemischten Strategien gesichert ist.

und Werte der Vertrauenswürdigkeit internalisiert) und mit der komplementären Wahrscheinlichkeit $1 - \pi$ ist diese Auszahlung $T_2 > R_2$ (mit der Wahrscheinlichkeit $1 - \pi$ hat der Treuhänder die entsprechenden Normen und Werte nicht internalisiert). Der Treuhänder kennt seine eigenen Handlungsanreize, der Treugeber kennt lediglich die Wahrscheinlichkeit π, kann den Ausgang des Zufallszugs aber selbst nicht direkt beobachten. In *Abbildung 3* wird dies dadurch angedeutet, dass die beiden Knoten für den Zug des Treugebers nach dem Anfangszug der Natur Elemente derselben Informationsmenge (in der deutschsprachigen Literatur ist auch die Bezeichnung „Informationsbezirk" verbreitet) sind, die wir durch das Oval angeben, in dem die beiden Knoten liegen: Der Treugeber weiß nicht, ob er sich am rechten oder am linken Knoten befindet. Wir sehen jetzt auch die notwendige Verfeinerung des Strategiebegriffs, auf die wir bereits vorsorglich hingewiesen haben: Eine Strategie eines Akteurs ist eine Verhaltensvorschrift für jede Informationsmenge, bei der der Akteur einen Zug macht. Unser Beispiel zeigt nicht nur, wie man Informationsunvollkommenheiten modellieren kann. Es zeigt auch, wie man externe „Zufallselemente" in ein spieltheoretisches Modell „einbaut". Man erkennt leicht das Gleichgewicht im Vertrauensspiel mit unvollständiger Information. Ein Treuhänder mit internalisierten Normen und Werten der Vertrauenswürdigkeit honoriert gegebenes Vertrauen, während der andere Typ Treuhänder gegebenes Vertrauen missbraucht. Für den Treugeber ist die Verweigerung von Vertrauen stets mit der Auszahlung P_1 verbunden, während die erwartete Auszahlung beim Geben von Vertrauen $\pi R_1 + (1 - \pi)S_1$ beträgt. Die eindeutige Gleichgewichtsstrategie des Treugebers ist also Verweigerung von Vertrauen, wenn $\pi < (P_1 - S_1)/(R_1 - S_1)$. Umgekehrt ist Geben von Vertrauen eindeutige Gleichgewichtsstrategie, wenn

$$(1) \qquad \pi > (P_1 - S_1)/(R_1 - S_1).$$

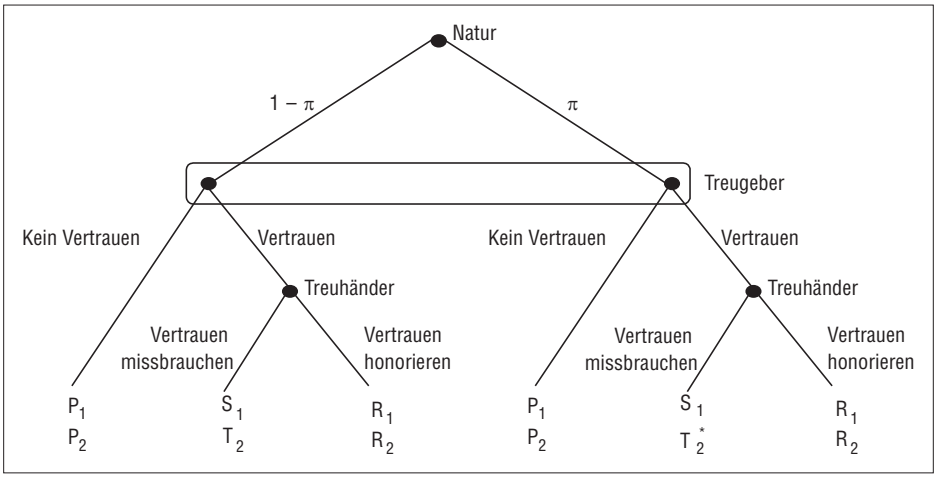
Im Grenzfall $\pi = (P_1 - S_1)/(R_1 - S_1)$ hat das Spiel multiple Gleichgewichte (beide reinen und alle gemischten Strategien des Treugebers sind dann Gleichgewichtsstrategien).[4] Unser Beispiel für eine strategische Situation mit unvollständiger Information ist bewusst einfach gewählt. Harsanyis (1967/68) zentraler Beitrag zur Spieltheorie (der ihn zum Teilhaber des Nobelpreises für Leistungen auf dem Gebiet der Spieltheorie gemacht hat) liegt darin, dass er allgemein gezeigt hat, wie man Spiele mit unvollständiger Information analysieren kann.[5]

Als letzte Variante betrachten wir das *wiederholte Vertrauensspiel*. Treugeber und Treuhänder spielen das Vertrauensspiel „immer wieder". Genauer nehmen wir an, dass nach jeder Runde *t,* in der das elementare Vertrauensspiel gespielt wurde, mit Wahrscheinlichkeit *w* eine weitere Runde gespielt wird. Wir nehmen an, dass beide Akteure nach jeder Runde über das Verhalten des Partners in der betreffenden Runde informiert sind, etwa weil der Akteur das Verhalten des Partners in der betreffenden Runde

4 Wir bemerken am Rande, dass Colemans (1991: Kap. 5) bekannte Bedingung für das Geben von Vertrauen unmittelbar aus der skizzierten Analyse folgt.

5 Dabei sollte man nicht übersehen, dass gerade bei der Analyse von Spielen mit unvollständiger Information häufig besonders weitgehende Rationalitätsannahmen verwendet werden (vgl. Rasmusen 1994), über deren Adäquatheit für empirische Anwendungen man sicherlich geteilter Meinung sein kann.

Abbildung 3: Vertrauensspiel mit unvollständiger Information
($S_1 < P_1 < R_1$, $P_2 < R_2 < T_2$, $T_2^* < R_2$)



direkt beobachten kann. Im wiederholten Vertrauensspiel ist eine Strategie eines Akteurs eine Regel, die das Verhalten des Akteurs in Runde t als Funktion der bisherigen Geschichte des Spiels spezifiziert (man sieht, dass eine Strategie nun in der Tat zu unterscheiden ist von einem Zug in einer Runde des Spiels). Wir verwenden die Standardannahme, dass die Auszahlung eines Akteurs im wiederholten Vertrauensspiel die verdiskontierte Summe seiner Auszahlungen in den einzelnen Runden ist, wobei wir die Fortsetzungswahrscheinlichkeit w als Diskontparameter verwenden. Ein Treugeber, der im wiederholten Vertrauensspiel ständig Vertrauen gibt, das auch ständig honoriert wird, erhält also die erwartete Auszahlung $R_1 + wR_1 + \ldots + w^{t-1}R_1 + \ldots = R_1/(1 - w)$.

Im wiederholten Vertrauensspiel kann der Treugeber das Verhalten des Treuhänders dadurch beeinflussen, dass er Vertrauen in einem bestimmten Sinn *bedingt* gibt. Genauer gesagt kann der Treugeber eine Strategie spielen, die so beschaffen ist, dass er vertrauenswürdiges Verhalten des Treuhänders belohnt durch die erneute Vergabe von Vertrauen in zukünftigen Runden (positive Sanktionen), während er den Missbrauch von Vertrauen bestraft durch Verweigerung zukünftigen Vertrauens (negative Sanktionen). Der Treuhänder muss dann im fokalen Vertrauensspiel in Runde t den kurzfristigen Anreiz $T_2 - R_2$ für den Missbrauch von Vertrauen abwägen gegen die langfristigen Anreize, Vertrauen zu honorieren und dadurch auch in zukünftigen Runden Auszahlungen $R_2 > P_2$ erhalten zu können. Man kann zeigen, dass Reziprozität im skizzierten Sinn eine Basis rationalen Vertrauens in dem Sinn sein kann, dass das wiederholte Vertrauensspiel ein Gleichgewicht derart hat, dass im Gleichgewicht in jeder Runde Vertrauen gegeben und honoriert wird. Dazu betrachte man diejenige bedingte Strategie des Treugebers, die für den Treuhänder mit der größten Belohnung für vertrauenswürdiges Verhalten und der schwersten Strafe für den Missbrauch von Vertrauen verbunden ist. Das ist die Strategie, die in der ersten Runde Vertrauen gibt und auch in allen folgenden Runden, vorausgesetzt, dass in allen vorherigen Runden stets Vertrauen gegeben und honoriert wurde. Wird jedoch in irgendeiner Runde kein Vertrauen gege-

ben und honoriert, dann verweigert der Treugeber das Vertrauen in allen folgenden Runden. Elementare Überlegungen zeigen, dass ständiges Honorieren von Vertrauen (und ständiger Missbrauch, sobald es eine Abweichung gegeben hat vom Pfad „Vertrauen geben und honorieren") beste Antwort des Treuhänders gegen diese bedingte Strategie des Treugebers ist, wenn

$$(2) \qquad w \geq (T_2 - R_2)/(T_2 - P_2),$$

wenn also der „Schatten der Zukunft" (Axelrod 1987), der sich in der Fortsetzungswahrscheinlichkeit w des Spiels niederschlägt, groß genug ist relativ zum kurzfristigen Anreiz $T_2 - R_2$ für den Missbrauch von Vertrauen. Daraus folgt, dass das wiederholte Vertrauensspiel genau dann Gleichgewichte hat, so dass im Gleichgewicht in jeder Runde Vertrauen gegeben und honoriert wird, wenn die Bedingung (2) erfüllt ist.

Bemerkung: Am Beispiel des wiederholten Vertrauensspiels kann man sich auch die Tragweite des Gleichgewichtsauswahlproblems klar machen. So prüft man leicht nach, dass das wiederholte Vertrauensspiel stets auch ein Gleichgewicht derart hat, dass der Treugeber niemals Vertrauen gibt, während der Treuhänder gegebenes Vertrauen immer missbrauchen würde. Aus dem „Folk Theorem" der Spieltheorie folgt, dass das wiederholte Vertrauensspiel für genügend großes w noch zahlreiche weitere Gleichgewichte hat. Man muss also zusätzliche Bedingungen formulieren, denen ein Nash-Gleichgewicht genügen muss, um als Lösung des Spiels in Betracht zu kommen. Zu diesem Zweck wurden verschiedene *Verfeinerungen* des Nash-Gleichgewichtsbegriffs entwickelt. Die zweifellos wichtigste dieser Verfeinerungen (die 1994 ebenfalls mit dem Nobelpreis ausgezeichnet wurde) ist Seltens (1965) Begriff des *teilspielperfekten Gleichgewichts*. Man betrachte dazu einen Teil eines Spielbaums, der selbst wieder ein Spielbaum im technischen Sinn ist und somit ein Teilspiel des Spiels repräsentiert. Im elementaren Vertrauensspiel (*Abbildung 1*) ist der Baum, der mit dem Zug des Treuhänders beginnt, ein solcher Teil des Spielbaums für das elementare Vertrauensspiel. Das Spiel, in dem lediglich der Treuhänder einen Zug ausführt, nämlich zwischen „Vertrauen honorieren" und „Vertrauen missbrauchen" wählt, wobei nach diesem Zug Treugeber und Treuhänder Auszahlungen entsprechend *Abbildung 1* erhalten, ist also ein Teilspiel des elementaren Vertrauensspiels. Im wiederholten Vertrauensspiel stelle man sich z.B. das Teilspiel vor, das in Runde 2 mit dem Zug des Treugebers beginnt, wenn der Treugeber in Runde 1 Vertrauen gegeben hat, das durch den Treuhänder missbraucht wurde (man mache sich klar, dass das wiederholte Vertrauensspiel ersichtlich eine „sehr" große Zahl von Teilspielen hat: jeder Knoten ist Beginn eines neuen Teilspiels). Ein teilspielperfektes Gleichgewicht ist dann eine Strategienkombination, die nicht nur ein Gleichgewicht für das gesamte Spiel ist, sondern auch für jedes Teilspiel ein Gleichgewicht „induziert". Nicht alle Nash-Gleichgewichte sind auch teilspielperfekt. Die Existenz von teilspielperfekten Gleichgewichten ist für endliche Spiele gesichert. Teilspielperfekte Gleichgewichte haben die attraktive Eigenschaft – hier liegt die intuitive „Rechtfertigung" für Seltens Konzept – dass die Gleichgewichtsstrategien glaubwürdige Versprechungen und vor allem glaubwürdige Drohungen enthalten. Im wiederholten Vertrauensspiel kann man leicht zeigen, dass das Gleichgewicht, in dem der Treugeber bedingt Vertrauen gibt und der Treuhänder Vertrauen honoriert, in der Tat teilspielperfekt ist. Das heißt, dass die implizite Drohung des Treugebers, Missbrauch von Vertrauen dadurch zu bestrafen, dass in weiteren Runden kein Vertrauen mehr gegeben wird, in dem Sinn glaubwürdig ist, dass diese Strategie auch Teil eines Gleichgewichts ist für das oben angedeutete Teilspiel, das in Runde 2 beginnt, nachdem der Treuhänder in Runde 1 Vertrauen missbraucht hat. Entsprechend seiner bedingten Strategie muss der Treugeber dann seine Drohung implementieren und Vertrauen verweigern. Das ist jedoch wiederum Gleichgewichtsverhalten in dem nun beginnenden Teilspiel und in diesem Sinn ist die Drohung glaubwürdig.[6] Es gibt Fälle, in denen

6 Das sog. „second-order free-rider problem" (Coleman 1991: Kap. 11) bezeichnet das Problem,

ein Spiel mehrere „einfache" Nash-Gleichgewichte, aber nur ein teilspielperfektes Gleichgewicht hat. In solchen Fällen kann man das teilspielperfekte Gleichgewicht als Lösung betrachten. Im allgemeinen kann ein Spiel aber auch mehrere und gegebenenfalls zahlreiche teilspielperfekte Gleichgewichte haben. Im wiederholten Vertrauensspiel ist z.B. auch das Gleichgewicht teilspielperfekt, niemals Vertrauen zu geben und gegebenes Vertrauen immer zu missbrauchen. Um bedingtes Geben und Honorieren von Vertrauen als Lösung des wiederholten Vertrauensspiels auszuzeichnen, bedarf es also weiterer Argumente. Ein typisches Argument ist in diesem Zusammenhang z.B. das der *Auszahlungsdominanz*. Ein Gleichgewicht wird auszahlungsdominiert, wenn es ein anderes Gleichgewicht gibt, das für mindestens einen Akteur mit einer höheren und für keinen Akteur mit einer niedrigeren Auszahlung verbunden ist. Im wiederholten Vertrauensspiel wird das Gleichgewicht, in dem der Treugeber bedingt Vertrauen gibt und der Treuhänder Vertrauen honoriert, ersichtlich nicht auszahlungsdominiert durch ein anderes Gleichgewicht. Das Gleichgewicht, in dem Vertrauen permanent verweigert wird, erfüllt das Kriterium der Auszahlungsdominanz hingegen nicht.

III. Von spieltheoretischen Modellen zu empirisch prüfbaren Hypothesen: Einbettungseffekte

Wie kann man nun spieltheoretische Modelle verwenden, um empirisch prüfbare soziologische Hypothesen abzuleiten? Im Kern geht es darum, dass wir die extensive Form eines Spiels spezifizieren, und zwar auf Grund empirischer Annahmen über zentrale Merkmale der sozialen Situation und der handelnden Akteure. Man mache sich dazu klar, dass die extensive Form einerseits soziale Bedingungen, also Opportunitäten und Restriktionen, für die Akteure modelliert (also die Annahmen, die in Colemans 1991 Schema für die Mikrofundierung soziologischer Makrohypothesen „links oben" stehen). Die extensive Form spezifiziert auch die „unabhängigen Variablen" der Theorie rationalen Handelns auf der Mikro-Ebene („links unten" in Colemans Schema), nämlich die Präferenzen der Akteure (die durch die Auszahlungen am Ende des Spielbaums repräsentiert werden) und die Information, über die ein Akteur verfügt, wenn er am Zug ist. Weiterhin fasst die extensive Form des Spiels die „Logik der Situation" zusammen (der linke vertikale Pfeil in Colemans Schema), weil sie zeigt, wie das Nutzenniveau eines Akteurs und seine Informationen abhängen von sozialen Bedingungen (den Merkmalen der Interdependenz zwischen den Akteuren). So kann man etwa die Auszahlungsfunktion ohne weiteres als eine Brückenannahme (im Sinn Lindenbergs 1981) interpretieren, die den Nutzen eines Akteurs als Funktion sozialer Bedingungen spezifiziert.

Auf Grund empirischer Annahmen spezifizieren wir die extensive Form des Spiels. Die spieltheoretische Analyse selbst besteht dann im Beweis von Theoremen über Gleichgewichte des Spiels und deren Eigenschaften. Die spieltheoretische Analyse des wiederholten Vertrauensspiels liefert uns z.B. ein Theorem, das besagt, dass die Bedingung (2) notwendig und hinreichend dafür ist, dass im wiederholten Vertrauensspiel ein teilspielperfektes Gleichgewicht derart existiert, dass im Gleichgewicht Vertrauen gegeben und honoriert wird. Dieser Analyseschritt betrifft die „Logik der Selektion"

Anreize sicherzustellen für die Implementation von Sanktionssystemen. Man mache sich klar, dass die Teilspielperfektheit eines Gleichgewichts die Lösung dieses Problems impliziert.

und wird in Colemans Schema durch den unteren horizontalen Pfeil repräsentiert. Aus den Theoremen über Gleichgewichte und zusätzlichen spieltheoretischen Annahmen, mit denen man in der Menge der Gleichgewichte die Lösung des Spiels auszeichnet, leitet man dann Implikationen darüber ab (typischerweise dadurch, dass man auf die eine oder andere Weise komparative Statik betreibt), wie sich soziale Bedingungen bzw. Veränderungen in sozialen Bedingungen auswirken auf die Existenz und die Eigenschaften von Gleichgewichten und damit auf das Verhalten rationaler Akteure. Dies sind Hypothesen über die Auswirkungen sozialer Bedingungen auf individuelles Verhalten, wie wir sie in diesem Beitrag ausarbeiten und überprüfen. In einem letzten Schritt kann man – wiederum mit Hilfe der extensiven Form des Spiels – Aussagen ableiten über kollektive Effekte. Aus der extensiven Form des Vertrauensspiels folgt z.B., dass das Geben und Honorieren von Vertrauen Pareto-optimal ist. Der kollektive Effekt „Pareto-Optimalität" ist in diesem speziellen Fall eine leicht ersichtliche kollektive Folge individueller Handlungen.[7] In anderen strategischen Situationen kann die „Logik der Aggregation" (bei Lindenberg 1977 sind dies die Transformationsregeln und in Colemans Schema geht es um den rechten vertikalen Pfeil) natürlich sehr viel komplexer sein.[8]

Man sieht, dass die Generierung von soziologischen Hypothesen aus spieltheoretischen Modellen der Konstruktion einer „Theorie mittlerer Reichweite" im Sinn Mertons ähnelt (auf diesen Umstand weisen auch Diekmann und Voss 2004: 20 hin) bzw. der Analyse von sozialen Mechanismen im Sinn von Hedström und Swedberg (1998). Ein offensichtlicher Vorteil der systematischen Verwendung spieltheoretischer Modelle bei der Entwicklung von Theorien mittlerer Reichweite für verschiedene Felder des sozialen Lebens ist, dass diese Theorien miteinander verbunden sind, weil sie einen gemeinsamen Kern von Rationalitätsannahmen enthalten.

In unserem Beitrag wenden wir die skizzierte Methode für die Generierung von Hypothesen aus spieltheoretischen Modellen an. Bei unseren Anwendungen ist der Umstand hilfreich, dass die Vertrauensspiele, die wir in unserer Skizze von Bausteinen der Spieltheorie untersucht haben, als einfache Modelle für die Analyse von Grundproblemen der soziologischen Theorie und für Anwendungen in z.B. der Organisations-, Wirtschafts- und Rechtssoziologie dienlich sind.

Um das einzusehen, mache man sich einerseits klar, dass das elementare Vertrauensspiel als ein einfaches Modell für das Hobbessche Ordnungsproblem im Fall von zwei Akteuren betrachtet werden kann (vgl. Voss 1985). Ein elementares Vertrauensspiel, in dem die Akteure keine bindenden Vereinbarungen abschließen können, welche durch einen externen „Erzwingungsstab" überwacht werden, ist ein Beispiel für den Hobbesschen Naturzustand. Rationale Akteure werden im elementaren Vertrauensspiel kein

7 Wir heben hier ab auf Pareto-Optimalität für die beiden im Vertrauensspiel involvierten Akteure. Man beachte, dass Vertrauen zwischen Treugeber und Treuhänder durchaus ungünstige Folgen haben kann für dritte Parteien. Vertrauen zwischen den Mitgliedern eines Kartells ist z.B. schädlich für andere Marktteilnehmer.

8 Wir verzichten an dieser Stelle auf eine wissenschaftstheoretische Diskussion der Frage, ob spieltheoretische Analysen im Prinzip als deduktiv-nomologische Erklärungen aufzufassen sind oder aber anders interpretiert werden sollten, etwa im Sinn van Friedmans (1953) instrumentalistischer Position (vgl. die knappe Diskussion bei Green und Shapiro 1999: 43–45 für einige weiterführende Hinweise).

Vertrauen geben bzw. würden gegebenes Vertrauen missbrauchen. Dieser Ausgang des elementaren Vertrauensspiels – die Akteure erhalten die Auszahlung P_i – ist aber Pareto-suboptimal, denn honoriertes Vertrauen ist wegen $R_i > P_i$ für beide Akteure ein besserer Ausgang. Rationales Verhalten im elementaren Vertrauensspiel hat damit ähnliche Folgen wie im Hobbesschen Naturzustand, in dem es zum Krieg aller gegen alle kommt. In anderer Terminologie: das elementare Vertrauensspiel ist ein soziales Dilemma, in dem individuelle Rationalität (Gleichgewichtsverhalten) und kollektive Rationalität (Pareto-Optimalität) auseinanderfallen (Rapoport 1974).

Andererseits ist auch der mögliche Beitrag spieltheoretischer Modelle zu Fragen der Organisations-, Rechts- und Wirtschaftssoziologie leicht ersichtlich. Dazu bedenke man, dass Vertrauensprobleme typische Probleme beim sozialen und ökonomischen Tausch repräsentieren. Wir verleihen z.B. Bücher an Kollegen und an Studenten im Vertrauen auf rechtzeitige und unbeschädigte Rückgabe. Wenn wir uns alle weigern, Bücher zu verleihen, dann verzichten wir auf die Vorteile im Sinn gewonnener Zeit und gesparten Geldes, die sich aus dem Verleih von Büchern ergeben können: Vertrauen ist ein „Schmiermittel" für die Effizienz von Interaktionen (Arrow 1974). Der Verleih von Büchern an Kollegen und Studenten ist ein Beispiel des sozialen Tauschs (Blau 1964). Vertrauensprobleme beim ökonomischen Tausch ergeben sich u.a. bei zwischenbetrieblichen Beziehungen. Man denke an einen Abnehmer, der seinem Lieferanten vertraut, dass dieser rechtzeitig und in adäquater Qualität liefert, auch wenn es keinen schriftlichen Vertrag gibt, der alle Details der Transaktion für alle denkbaren Umstände vollständig spezifiziert (Macaulay 1963). Die Aushandlung und Niederschrift eines solchen vollständigen und expliziten Vertrags wäre kostspielig und würde für beide Parteien den möglichen Transaktionsgewinn reduzieren. Vertrauen kann also auch hier effizienzfördernd wirken, aber das Problem für den Abnehmer bleibt, dass der Lieferant Gelegenheiten und Anreize hat – oder jedenfalls haben könnte –, die Lieferung zu verzögern bzw. ein Produkt minderer Qualität zum Preis eines Produkts guter Qualität zu liefern.

Im Zusammenhang mit dem Hobbesschen Ordnungsproblem geht es um Hypothesen über soziale Bedingungen, die zur „Lösung" des Problems beitragen. Ähnlich geht es in der Organisations-, Wirtschafts- und Rechtssoziologie um soziale Bedingungen, die Vertrauen beim sozialen und wirtschaftlichen Tausch fördern. Derartige soziale Bedingungen beziehen sich typischerweise auf die „Einbettung" (Granovetter 1985) von Vertrauensproblemen. Man meint damit grob gesprochen (vgl. Raub und Weesie 2000a für eine genauere Diskussion), dass ein „fokales" Vertrauensproblem zwischen zwei Akteuren eingebettet ist in eine Beziehung mit vorangegangenen und zukünftigen Interaktionen zwischen den beiden Akteuren und gegebenenfalls auch in Beziehungen der beiden Akteure mit dritten Parteien. Häufig hebt man auch ab auf die Einbettung eines fokalen Vertrauensproblems in einen institutionellen Kontext (vgl. Nohria und Eccles 1992 und Swedberg 1993 für empirische Studien).

Mit spieltheoretischen Mitteln lässt sich zunächst präzisieren, was mit „Einbettung" gemeint ist und man kann verschiedene Formen sozialer Einbettung systematisch unterscheiden. Der Leser kann sich vielleicht schon vorstellen, dass es vielversprechend sein kann, zu diesem Zweck ein fokales elementares Vertrauensspiel als „Teil" (damit kann ein Teilspiel im technischen Sinn gemeint sein, aber das ist keineswegs zwingend)

eines komplexeren Spielbaums zu betrachten. Vor allem aber kann man spieltheoretische Instrumente verwenden, um Bedingungen zu spezifizieren, die in einem „eingebetteten" Vertrauensproblem das Geben und Honorieren von Vertrauen fördern. Man wird also die Frage untersuchen, ob und wann für den komplexeren Spielbaum ein Gleichgewicht existiert, gegebenenfalls auch ein solches, das mit guten Gründen als Lösung des Spiels betrachtet werden kann, wobei im fokalen Vertrauensproblem Vertrauen zustande kommt und honoriert wird. Aussagen über solche Bedingungen sind Hypothesen über Einbettungseffekte auf Vertrauen.[9] Wir präsentieren Studien zu drei verschiedenen Typen sozialer Einbettung, die wir als dyadische Einbettung, Netzwerkeinbettung und institutionelle Einbettung andeuten.

IV. Eine Feldstudie mit Surveydaten: Lern- und Kontrolleffekte der dyadischen Einbettung beim ökonomischen Tausch

Aus der spieltheoretischen Analyse des elementaren Vertrauensspiels folgt, dass ein rationaler Treugeber in einem „isolierten" Vertrauensspiel niemals vertraut, da der Treuhänder in einem solchen Spiel Vertrauen immer missbrauchen würde. „Isoliert" bedeutet hier einerseits, dass Treugeber und Treuhänder ihr Verhalten im Vertrauensspiel nicht abhängig machen können von früheren Interaktionen, sei es früheren Interaktionen zwischen Treugeber und Treuhänder, sei es früheren Interaktionen, bei denen (auch) dritte Parteien beteiligt waren. Andererseits handelt es sich um ein isoliertes Vertrauensspiel in dem Sinn, dass es auch keine zukünftigen Interaktionen gibt, bei denen das Verhalten von Treugeber, Treuhänder oder dritten Parteien abhängig gemacht werden kann vom Ablauf des fokalen Vertrauensspiels (vgl. Weber 1976: 11–14 für den Grundgedanken, Bedingungen des Handelns in einer fokalen sozialen Situation in Merkmalen vergangener und möglicher zukünftiger Interaktionen zu suchen). Erste Hypothesen über Bedingungen des Vertrauens können wir aus spieltheoretischen Modellen gewinnen, wenn wir das Vertrauensspiel mit unvollständiger Information verwenden. Dazu bemerken wir, dass die Größe $(P_1 - S_1)/(R_1 - S_1)$ in (1) ein Maß ist für das Risiko des Treugebers, da dieser Ausdruck das Verhältnis zwischen dem möglichen Verlust $P_1 - S_1$ durch missbrauchtes Vertrauen und dem möglichen Gewinn $R_1 - S_1$ aus honoriertem Vertrauen wiedergibt. Je größer dieses Risiko, desto restriktiver ist die Bedingung $\pi > (P_1 - S_1)/(R_1 - S_1)$ und mithin erhalten wir die Hypothese, dass Vertrauen mit zunehmendem Risiko abnehmen wird (H_1).[10]

Rationales Vertrauen hängt nicht nur ab vom Risiko des Treugebers, sondern auch von der Wahrscheinlichkeit π, dass der Treuhänder gegebenes Vertrauen honoriert. Diese Wahrscheinlichkeit wird u.a. abhängen von Informationen des Treugebers über früheres Verhalten des Treuhänders bei Vertrauensproblemen: Aus früherem Verhalten des Treuhänders kann der Treugeber lernen. Es ist ein charakteristisches Merkmal vieler

9 Man sieht nun (vgl. Fußnote 2), dass es in der nichtkooperativen Spieltheorie auch um die Frage nach Bedingungen für kooperatives Verhalten (hier im Sinn von „Geben und Honorieren von Vertrauen") geht.

10 Snijders (1996) bietet umfangreiche experimentelle Evidenz aus isolierten („one-shot") Vertrauensspielen für diese Hypothese.

soziologischer Analysen von Vertrauensproblemen (vgl. z.B. Blau 1964; Granovetter 1985: 400; Coleman 1991: Kap. 5; Burt und Knez 1995), dass sie sich insbesondere auf die Analyse solcher *Lerneffekte* richten. Lernmodelle sagen dabei generell voraus (vgl. Buskens und Raub 2002 für eine ausführlichere Diskussion und Literaturhinweise), dass positive Information über früheres Verhalten des Treuhänders in dem Sinn, dass der Treuhänder sich in diesen früheren Interaktionen als vertrauenswürdig erwiesen hat, dazu führen werden, dass der Treugeber seine Einschätzung der Wahrscheinlichkeit vertrauenswürdigen Verhaltens des Treuhänders nach oben anpassen wird. Wir können also als eine erste Hypothese über Lerneffekte formulieren, dass Vertrauen zunimmt, wenn der Treugeber selbst in der Vergangenheit positive Erfahrungen bei Vertrauensproblemen mit dem Treuhänder gesammelt hat (H_2). Es handelt sich hierbei um eine Hypothese über Effekte der *dyadischen Einbettung* eines fokalen Vertrauensproblems in eine Beziehung zwischen Treugeber und Treuhänder, die auch frühere Interaktionen zwischen den Akteuren enthält.

Unsere Analyse des wiederholten Vertrauensspiels hat gezeigt, dass die dyadische Einbettung eines fokalen Vertrauensproblems in eine Beziehung zwischen Treugeber und Treuhänder auch dadurch Vertrauen fördern kann, dass der Treuhänder im Hinblick auf zukünftige Interaktionen jedenfalls teilweise abhängig ist vom Verhalten des Treugebers, so dass der Treuhänder die langfristigen Folgen opportunistischen Verhaltens im fokalen Vertrauensproblem in Rechnung stellen muss. Der Treugeber kann den Treuhänder durch die bedingte Vergabe von Vertrauen kontrollieren. Dieser *Kontrolleffekt* wird in der Literatur auch als Reziprozität (Gouldner 1960; Blau 1964; Voss 1985) oder bedingte Kooperation (Taylor 1987) angedeutet. Die Bedingung (2) für ein Gleichgewicht derart, dass Vertrauen gegeben und honoriert wird, ist weniger restriktiv, wenn der Parameter w zunimmt. Eine typische Hypothese über Kontrolleffekte auf der Ebene der dyadischen Beziehung, die daraus folgt, ist die, dass Vertrauen zunimmt, wenn zukünftige Interaktionen zwischen Treugeber und Treuhänder wahrscheinlicher werden (H_3). Eine weitere Hypothese, die wir aus der Bedingung (2) gewinnen können, ist die, dass Vertrauen abnimmt, wenn die kurzfristigen Anreize des Treuhänders für den Missbrauch von Vertrauen zunehmen (H_4), wobei $T_2 - R_2$ ein naheliegendes Maß für diese Anreize ist. Man beachte dazu, dass (2) restriktiver wird, wenn $T_2 - R_2$ zunimmt.[11]

Bemerkung: Unsere Behandlung von Vertrauensproblemen zeigt natürlich auch, dass spieltheoretische Modellierungen typischerweise vereinfachende Annahmen einschließen. Wir abstrahieren z.B. von der Möglichkeit, dass der Treugeber dem Treuhänder *mehr oder weniger* vertrauen und dass der Treuhänder Vertrauen *mehr oder weniger* honorieren kann. Wir haben bislang ausschließlich das Problem betrachtet, dass der Treuhänder Möglichkeiten und Anreize hat, Vertrauen zu missbrauchen und sich opportunistisch zu verhalten. Wir haben also den Fall vernachlässigt, dass Vertrauen in die *Fähigkeiten* des Treuhänders problematisch sein kann. Im wiederholten Vertrauensspiel nehmen wir an, dass *in jeder Runde dasselbe* Spiel gespielt wird, aber das muss natürlich nicht so sein. Im Zuge von „Modellbau mit der Methode der abnehmenden Abstraktion" (Lindenberg 1992) kann man derartige vereinfachende Annahmen schrittweise ersetzen durch realistischere Annahmen, die dann natürlich zu komplexeren Modellen führen. So ist etwa das „investment game"

11 Man sieht, dass wir hier implizit Auszahlungsdominanz als Argument bei der Gleichgewichtsauswahl anwenden.

(Berg et al. 1995) eine Variante des Vertrauensspiels, in der Treugeber und Treuhänder nicht nur binäre Entscheidungen treffen können. Raub (2004) untersucht ein Vertrauensspiel, in dem auch die Fähigkeiten des Treuhänders problematisch sind. In diesem Beitrag betrachten wir derartige Komplikationen nicht im Detail, machen aber gelegentlich Gebrauch von Resultaten bei der Analyse solcher komplexerer Modelle. Etwas ausführlicher sei im Hinblick auf unsere empirische Anwendung jedoch der Fall betrachtet, dass die Akteure nicht in jeder Runde dasselbe Spiel spielen. In unserer Skizze von Lerneffekten haben wir der Einfachheit halber ausgeblendet, dass vergangene Interaktionen nicht nur Lerneffekte haben können. Frühere Interaktionen zwischen Treugeber und Treuhänder können nicht nur dazu führen, dass der Treugeber etwas lernt. Man nehme etwa an, dass Treugeber und Treuhänder einem Vertrauensproblem in einer früheren Interaktion dadurch entgegengetreten sind, dass sie einen Vertrag konzipiert haben, der ihre Rechte und Pflichten näher spezifiziert und dadurch das Vertrauensproblem reduziert. Wenn die beiden Parteien abermals einem Vertrauensproblem ausgesetzt sind, können sie diesen Vertrag – gegebenenfalls in angepasster Form – möglicherweise erneut verwenden. Die vertragliche Planung der früheren Interaktion beeinflusst dann die Anreizstruktur beim fokalen Vertrauensproblem, z.B. in dem Sinn, dass durch die Verfügbarkeit des früher konzipierten Vertrags eine vertragliche „Regelung" des fokalen Vertrauensproblems mit geringeren Kosten verbunden ist. Wenn wir davon ausgehen, dass die Akteure bei der früheren Interaktion voraussehen konnten, dass sie später – jedenfalls möglicherweise – erneut mit einem Vertrauensproblem konfrontiert werden würden, bedeutet dies umgekehrt natürlich auch, dass sie bei der rationalen vertraglichen Planung der früheren Interaktion die Folgen in Rechnung stellen, die ein Vertrag für zukünftige Vertrauensprobleme haben wird (vgl. Raub und Snijders 2001 für explizite spieltheoretische Modellierungen).

Wir wenden uns nun einem ersten Beispiel empirischer Überprüfungen von spieltheoretisch begründeten Hypothesen über Bedingungen des Vertrauens zu. Dazu verwenden wir eine Untersuchung über den Einkauf von IT-Produkten (Hard- und Software) durch niederländische Klein- und Mittelbetriebe (5–200 Angestellte; Batenburg und Raub 1995; Rooks et al. 1998; vgl. Batenburg et al. 2003; Rooks 2002: Kap. 3 für Beschreibungen der Studie). Im Rahmen dieser Studie wurden Schlüsselinformanten der abnehmenden Betriebe – typischerweise für die Anschaffung verantwortliche IT-Manager – mittels eines strukturierten Fragebogens zur Anschaffung eines IT-Produkts befragt. Wir betrachten also zunächst Goldthorpes (2000: 94) „Standardfall" von QAD, nämlich die Analyse von Daten, die durch Surveyforschung gewonnen wurden. Die Studie wurde 1995 bei 788 Abnehmern solcher Produkte durchgeführt. Etwa 25 Prozent der Befragten war bereit, zusätzlich einen Fragebogen über die Anschaffung eines zweiten IT-Produkts durch den Abnehmer – sei es bei demselben oder bei einem anderen Lieferanten – auszufüllen. Wir verfügen dadurch über Daten zu 788 + 183 = 971 Transaktionen. In 1998 wurden die Abnehmer erneut kontaktiert. In dieser Runde wurden Daten über 281 weitere Transaktionen erhoben. Insgesamt entstand so ein Datensatz mit 971 + 281 = 1252 Transaktionen. Im Vergleich zu anderen Organisationsbefragungen war die Teilnahmebereitschaft hoch. Non-Response-Analysen zeigen, dass der Datensatz hinsichtlich zentraler Merkmale der Firmen wie Größe, Industriezweig oder Firmensitz, aber auch hinsichtlich der allgemeinen Zufriedenheit mit IT-Lieferanten nicht verzerrt ist.

Eine IT-Transaktion konzeptualisieren wir als eine Variante eines Vertrauensspiels mit dem Abnehmer in der Rolle des Treugebers und dem Lieferanten in der Rolle des Treuhänders. Wir konzentrieren uns auf Bemühungen des Abnehmers, opportunistisches Verhalten des Lieferanten – man denke an verspätete Lieferung oder Lieferung

inferiorer Qualität – durch (Investitionen in das) ex ante-Management der Transaktion zu begrenzen. Als Indikatoren für ex ante-Management verwenden wir die Anzahl der Arbeitstage von Mitarbeitern des Arbeitnehmers, die in die Aushandlung eines Vertrags investiert wurden, die Anzahl der Abteilungen, die bei den Vertragsverhandlungen beteiligt waren, die Hinzuziehung von Juristen, den Gebrauch eines Standardvertrags bzw. eines spezifisch angepassten Vertrags und schließlich die Anzahl der Klauseln rechtlicher, finanzieller und technischer Art, die im Vertrag enthalten sind bzw. Gegenstand der Vertragsverhandlungen waren (vgl. Batenburg et al. 2003 für Details aller in diesem Beitrag verwendeten Operationalisierungen und deskriptive Statistiken). Derartiges ex ante-Management der Transaktion reduziert einerseits die Möglichkeiten und Anreize des Lieferanten für opportunistisches Verhaltens und andererseits den Schaden des Abnehmers, falls der Lieferant sich opportunistisch verhält (etwa dadurch, dass ex ante-Management zu vertraglichen Vereinbarungen führt, die eine Entschädigung des Abnehmers bei Lieferungsverzögerungen oder Qualitätsmängeln zur Folge haben). Zugleich ist ex ante-Management für den Abnehmer mit Kosten verbunden und daher interpretieren wir die Investitionen in das ex ante-Management als Maß für fehlendes Vertrauen des Abnehmers in den Lieferanten: Je weniger der Abnehmer dem Lieferanten vertraut, desto mehr wird er in das ex ante-Management der Transaktion investieren. Investitionen des Abnehmers in das ex ante-Management der Transaktion verwenden wir in unserer Analyse als abhängige Variable.

Wir betrachten zunächst Merkmale der Transaktion, von denen angenommen werden kann, dass sie das ex ante-Management des Abnehmers beeinflussen. In unserer empirischen Analyse verwenden wir Variablen, die an zentrale Argumente der Transaktionskostentheorie (z.B. Williamson 1985) anschließen. Diese Variablen beeinflussen das Problempotenzial der Transaktion, nämlich einerseits die Risiken des Abnehmers und andererseits die Anreize des Lieferanten für opportunistisches Verhalten. Entsprechend unseren Hypothesen H_1 und H_4 erwarten wir, dass das ex ante-Management des Abnehmers zunimmt, wenn mit zunehmendem Problempotenzial die Risiken des Abnehmers oder die Anreize für opportunistisches Verhaltens des Lieferanten größer werden. Eine erste Determinante des Problempotenzials ist das finanzielle Volumen der Transaktion. Eine weitere Variable ist die Unsicherheit des Abnehmers: das Problempotenzial nimmt zu, wenn es für den Abnehmer schwieriger ist, die Qualität des gelieferten Produkts zu beurteilen. Für die Messung von Unsicherheit verfügen wir über Indikatoren für die „objektive" Komplexität des Produkts (Art der gelieferten Komponenten und Dienste) und über verschiedene subjektive Einschätzungen des Befragten. Das Problempotenzial der Transaktion hängt ebenfalls ab von den Ersatzkosten, d.h. den Kosten, die für den Abnehmer entstehen, wenn er das Produkt ersetzen muss. Wir verfügen über Angaben zu den Kosten bei der Anschaffung eines alternativen Produkts, bei (erneuter) Schulung von Mitarbeitern im Umgang mit einem alternativen Produkt, durch erneute Dateneingabe und durch Stillstand der eigenen Produktion bei Wechsel des Produkts. Unsere beiden letzten Bestimmungsgrößen des Problempotenzials sind die Wichtigkeit einer langen Lebensdauer des Produkts für den Abnehmer und schließlich die Wichtigkeit des Produkts für die Profitabilität des Abnehmers. Man beachte, dass Ersatzkosten, die Wichtigkeit einer langen Lebensdauer des Produkts und die Wichtigkeit des Produkts für die Profitabilität des Abnehmers abhängen von einer zen-

tralen Größe in der Transaktionskostentheorie, den mit der Transaktion verbundenen spezifischen Investitionen, also solchen Investitionen, die wertlos sind für alternative Verwendungen. Wir berücksichtigen mithin die von Williamson und den Transaktionskostentheoretikern betonten Effekte spezifischer Investitionen auf das Problempotenzial.

Lerneffekte auf der Ebene der dyadischen Beziehung zwischen Abnehmer und Lieferant repräsentieren wir durch eine Variable, die angibt, ob Abnehmer und Lieferant bereits früher geschäftlich in Beziehung standen (Vergangenheit). Solche früheren geschäftlichen Beziehungen sind in beinah allen Fällen mit positiven Erfahrungen des Abnehmers mit dem Lieferanten verbunden. Während bei etwa der Hälfte der Transaktionen (nämlich 635) bereits eine frühere geschäftliche Beziehung von Abnehmer und Lieferant vorliegt, gibt es lediglich 19 Transaktionen, bei denen der Abnehmer bereits früher Produkte vom Lieferanten bezogen hatte und dabei mit dem Lieferanten unzufrieden war – Unzufriedenheit mit dem Lieferanten führt offenbar zum Abbruch der Beziehung. Entsprechend unserer Hypothese H_2 über Lerneffekte erwarten wir mithin, dass eine frühere geschäftliche Beziehung des Abnehmers mit dem Lieferanten das ex ante-Management der fokalen Transaktion vermindert.

Im Zusammenhang mit Kontrolleffekten auf der Ebene der dyadischen Beziehung verwenden wir eine Variable, die angibt, in welchem Umfang und mit welcher Häufigkeit der Abnehmer zum Zeitpunkt der fokalen Transaktion zukünftige Transaktionen mit dem Lieferanten erwartete (Zukunft). Man beachte, dass der theoretisch zu erwartende Effekt dieser Variable auf das ex ante-Management der fokalen Transaktion vergleichsweise komplex ist. Einerseits nehmen die Kontrollmöglichkeiten des Abnehmers zu, wenn mit mehr und umfangreicheren zukünftigen Transaktionen gerechnet werden kann: Der „Schatten der Zukunft" kommt als Substitut für kostspieliges ex ante-Management in Betracht und dadurch kann das Vertrauen in den Lieferanten wachsen. Andererseits wird es aber auch attraktiver, in das ex ante-Management der fokalen Transaktion zu investieren, weil diese Investitionen in zukünftigen Transaktionen jedenfalls teilweise wiederverwendet werden können – man denke an die Wiederverwendung (von Teilen) eines schriftlichen Vertrags.[12] Durch diesen Investitionseffekt sollte das ex ante-Management der fokalen Transaktion zunehmen (vgl. in diesem Zusammenhang auch Williamsons 1985: 60–61 Argument, dass „Häufigkeit" von Transaktionen spezialisierte „governance structures" attraktiver macht). Ohne zusätzliche theoretische Argumente über die relative Stärke der beiden gegensätzlichen Effekte des Schattens der Zukunft können wir daher keine Hypothese über den Zusammenhang des Schattens der Zukunft mit dem ex ante-Management der fokalen Transaktion ableiten. Dennoch können wir eine Hypothese über Kontrolleffekte formulieren. Man vergleiche dazu fokale Transaktionen von Partnern, die bereits eine geschäftliche Beziehung hatten, mit fokalen Transaktionen ohne eine vorhergehende gemeinsame Beziehung von Abnehmer und Lieferant. Wenn bereits eine geschäftliche Beziehung vorlag, dann sind bereits früher Investitionen in das ex ante-Management von Transaktionen getätigt

12 Illustrativ ist in diesem Zusammenhang der Tatbestand, dass bei 217 der 635 Transaktionen, bei denen der Abnehmer bereits vor der fokalen Transaktion eine Geschäftsbeziehung mit dem Lieferanten unterhielt, der Vertrag für die fokale Transaktion eine mehr oder weniger angepasste Version eines früheren Vertrags war.

Tabelle 1: Regressionsanalyse der abhängigen Variable ex ante-Management, standardisierte Koeffizienten (kontrolliert für marginale Kosten des ex ante-Managements und für Grösse von Abnehmer und Lieferant)

Unabhängige Variable	Hypothese	Modell 1	Modell 2
Problempotenzial			
Volumen	+	.32**	.32**
Unsicherheit	+	.11**	.10**
Ersatzkosten	+	.15**	.14**
Lebensdauer	+	.09**	.10**
Profitabilität	+	.14**	.14**
Dyadische Einbettung			
Vergangenheit (1 = ja)	–	–.07**	–.07**
Zukunft	?	–.03	.03
Vergangenheit × Zukunft	–		–.09**
Erklärte Varianz (R^2)		.397**	.401**
Fallzahl (Transaktionen)		1142	1142

**, * und ~ für zweiseitige Signifikanzniveaus $p < 0.01$, $p < 0.05$ und $p < 0.10$; Huberkorrektur der Standardfehler (Huber 1967; Rogers 1993) für Clustering von Transaktionen beim Abnehmer.

worden und die Investitionen in das ex ante-Management der fokalen Transaktion werden einen geringeren Effekt für zukünftige Transaktionen haben als im Fall von Partnern, bei denen die fokale Transaktion zugleich die erste Transaktion ist. Wir können also erwarten, dass der mit dem Schatten der Zukunft verbundene Investitionseffekt in neuen geschäftlichen Beziehungen stärker sein wird als in Transaktionen zwischen Partnern, die bereits früher geschäftliche Beziehungen unterhielten. Ein Kontrolleffekt entsprechend unserer Hypothese H_3 sollte sich daher niederschlagen in einem negativen Interaktionseffekt vergangener und erwarteter zukünftiger geschäftlicher Beziehungen auf das ex ante-Management der fokalen Transaktion (Raub 1996; Raub und Snijders 2001).

Tabelle 1 präsentiert die empirischen Resultate (die Ergebnisse sind kontrolliert für Variablen, die marginale Kosten des ex ante-Managements beeinflussen – man denke an die juristische Expertise des Abnehmers – und für die Größe von Abnehmer und Lieferant; vgl. Batenburg et al. 2003; Buskens und Raub 2004 für Details). Die Resultate zeigen, dass zunehmendes Problempotenzial einhergeht mit umfangreicherem ex ante-Management, also geringerem Vertrauen. Mithin werden unsere Hypothesen H_1 und H_4 bestätigt. Frühere Erfahrungen mit dem Lieferanten reduzieren das ex ante-Management bzw. erhöhen das Vertrauen in den Lieferanten und das bestätigt unsere Hypothese H_2 über den Lerneffekt. Erwartungen zukünftiger Transaktionen reduzieren das ex ante-Management bzw. vergrößern Vertrauen, falls der Abnehmer bereits früher geschäftliche Beziehungen mit dem Lieferanten hatte. Dieses Ergebnis unterstützt unsere Hypothese H_3 über den Kontrolleffekt.

Wir präsentieren an dieser Stelle nur eine einzige repräsentative Analyse. Die Resultate sind jedoch robust für zahlreiche andere Modellspezifikationen. Resultate für eine Serie alternativer Modellspezifikationen berichten Batenburg et al. (2003). Ihre Analysen verwenden ausschließlich die Transaktionen aus der Studie von 1995. Die 1998

zusätzlich erhobenen Daten führen zu keinen wesentlichen Veränderungen der Ergebnisse.

V. Eine Vignettenstudie: Effekte der dyadischen Einbettung und der Netzwerkeinbettung

In diesem Beitrag möchten wir belegen, dass sich der Rational Choice-Ansatz in der Form spieltheoretischer Modellierungen und die quantitative Analyse sozialwissenschaftlicher Daten fruchtbar miteinander verbinden lassen. Goldthorpe (2000: 94) hat in seinem Plädoyer für die Allianz von RAT und QAD die quantitative Analyse sozialwissenschaftlicher Daten breit charakterisiert als „any analysis of extensive social data – though typically data collected via survey research – that involves the statistical investigation of relationships existing among variables". Auch wenn es bei QAD typischerweise um Surveydaten geht, so kann es doch nützlich sein, andere Arten von Daten und Forschungsdesigns nicht zu vernachlässigen. Die Verwendung verschiedenartiger Daten und Designs für die Überprüfung von Hypothesen trägt z.B. dazu bei, dass wir uns ein besseres Bild machen können von der Robustheit unserer empirischen Resultate. Auch sind typische Nachteile bei der Überprüfung von Hypothesen mit Surveydaten in Rechnung zu stellen, gerade auch Nachteile bei der Überprüfung vergleichsweise spezifischer Hypothesen, die wir aus theoretischen Modellen ableiten können. Die Messung von Variablen mit Surveydaten ist häufig nicht unproblematisch und Surveydaten können Annahmen verletzen, die dem zu prüfenden theoretischen Modell zugrunde liegen (vgl. Buskens 2002: Kap. 5 für eine ausführliche Diskussion einiger Probleme bei der Prüfung von Hypothesen über Einbettungseffekte auf Vertrauen mit unseren Surveydaten über IT-Transaktionen). Das Plädoyer für die Allianz von RAT und QAD wird also stärker, wenn wir unsere spieltheoretisch fundierten Hypothesen auch mit komplementären Daten und Designs überprüfen.

Wir wenden uns daher nun einer empirischen Überprüfung mittels einer Vignettenstudie (Rossi 1979; Rossi und Nock 1982) zu. Diese Studie hat außerdem den Vorzug, dass sie mit dem Ziel entworfen wurde, neben Hypothesen über Effekte dyadischer Einbettung gerade auch Hypothesen über Effekte der Netzwerkeinbettung zu überprüfen.

Mit *Netzwerkeinbettung* meinen wir, dass ein fokales Vertrauensproblem eingebettet ist in Beziehungen von Treugeber und Treuhänder mit dritten Parteien. Wir betrachten hier den Fall, dass der Treuhänder des fokalen Vertrauensspiels auch Treuhänder ist in Vertrauensspielen mit anderen Treugebern als Partnern und dass die Treugeber miteinander Kontakte unterhalten. Wir verzichten auf eine explizite spieltheoretische Modellierung derartiger Netzwerkeinbettung (vgl. Kreps 1990a; Raub und Weesie 1990; Buskens 2002 für solche Modellierungen und Analysen[13] und als Übersicht Buskens und Raub 2002) und belassen es bei einer Skizze von Hypothesen, die sich aus solchen Modellen gewinnen lassen. Dabei unterscheiden wir wiederum zwischen Lern- und

13 Ein verwandtes und soziologisch interessantes Modell verwendet Ziegler (1990) in seiner Analyse des zeremoniellen Gabentauschs.

Kontrolleffekten (vgl. Yamagishi und Yamagishi 1994: 138–139 für eine ähnliche Unterscheidung im Hinblick auf Effekte der Netzwerkeinbettung).

Im Kern geht es darum, dass das Netzwerk der Beziehungen des Treugebers mit anderen Treugebern des Treuhänders die Verbreitung von Informationen über das Verhalten des Treuhänders beeinflusst. Bei *Lerneffekten* kommt es auf die Informationen an, die der Treugeber im fokalen Vertrauensspiel von anderen Treugebern erhält. Unter der Voraussetzung, dass es sich wiederum um positive Informationen handelt, lässt sich insbesondere zeigen, dass Vertrauen zunimmt, wenn der Treugeber von mehr anderen Treugebern Informationen erhält, wenn also technisch gesprochen der Innengrad des Treugebers zunimmt (H_5), und wenn die Dichte des Netzwerks der Treugeber zunimmt (H_6). Für Lerneffekte auf der Ebene des Netzwerks kommt es darauf an, dass der Treugeber im fokalen Vertrauensproblem von möglichst vielen anderen Treugebern Informationen über den Treuhänder erhält. Für *Kontrolleffekte* auf der Ebene des Netzwerks kommt es demgegenüber darauf an, dass der Treugeber im fokalen Vertrauensproblem seinerseits möglichst viele andere Treugeber über das Verhalten des Treuhänders im fokalen Vertrauensproblem informieren kann. Dadurch können andere Treugeber ihr Verhalten in zukünftigen Vertrauensspielen vom Verhalten des Treuhänders im fokalen Vertrauensspiel abhängig machen. Kontrolleffekte auf der Ebene des Netzwerks hängen also technisch gesprochen vom Außengrad des Treugebers im fokalen Vertrauensspiel ab. Aus Modellen für Kontrolleffekte folgt, dass Vertrauen zunimmt, wenn der Außengrad zunimmt (H_7) und außerdem folgt erneut die Hypothese, dass Vertrauen zunimmt mit der Dichte des Netzwerks des Treugebers (H_6).[14]

Bemerkung: Man beachte, dass wir die Netzwerkeinbettung als exogen gegeben betrachten und uns hier nicht mit dem Problem beschäftigen, wie Netzwerke entstehen und sich verändern, wenn rationale Akteure ihre Partner wählen können (dieses Problem der Dynamik von Netzwerken ist komplex, wird aber immer mehr zum Gegenstand auch explizit spieltheoretischer Modellierungen, vgl. Dutta und Jackson 2003 für eine Sammlung zentraler Arbeiten und Kosfeld 2003 für eine Literaturübersicht, die auch erste experimentelle Arbeiten einschließt). Weiterhin vernachlässigen wir Probleme im Zusammenhang mit Informationen, die der Treugeber über den Treuhänder von dritten Parteien erhält und die der Treugeber über den Treuhänder an dritte Parteien weitergibt. In der Literatur (z.B. Lorenz 1988: 153–155; Raub und Weesie 1990: 648; Williamson 1996: 153–155; Blumberg 1998: 211–214; Buskens 2002: 18–20) wird gelegentlich bemerkt, dass derartige Informationen problematisch sein können. Informationen von dritten Parteien können etwa inkonsistent sein mit eigenen Erfahrungen. Die Verlässlichkeit von Informationen über den Treuhänder, die ein Treugeber von dritten Parteien erhält, kann zweifelhaft sein und zufälligen oder gar systematischen Verzerrungen unterliegen. Man kann annehmen, dass sowohl Lern- als auch Kontrolleffekte auf Netzwerkebene schwächer werden, wenn Information von dritten Parteien bzw. an dritte Parteien im angedeuteten Sinn problematischer wird.

Der Kauf von Gebrauchtwagen wird in der Literatur (Akerlof 1970; Dasgupta 1988) als paradigmatischer Fall eines Vertrauensproblems behandelt, mit dem (potenziellen) Käufer in der Rolle des Treugebers und dem Händler als Treuhänder, der Vertrauen durch den Verkauf einer „Zitrone" (lemon) zum Preis eines „Pfirsichs" (peach) missbrauchen kann. In der folgenden Vignettenstudie (vgl. Buskens und Weesie 2000 für

14 Auch hier verwenden wir Auszahlungsdominanz als Argument bei der Gleichgewichtsauswahl.

eine detaillierte Beschreibung) vergleichen Versuchspersonen verschiedene Situationen beim Gebrauchtwagenkauf. Die Versuchspersonen sind insgesamt 125 Studierende an Universitäten in den USA (University of Chicago) und in den Niederlanden (Utrecht und Tilburg). Die Versuchspersonen befinden sich also in der Rolle des Käufers bzw. Treugebers. Jeder Versuchsperson wurden zehn Paare von Vignetten mit unterschiedlichen Situationen vorgelegt und für jedes Paar konnte die Versuchsperson angeben, welche der beiden Situationen sie beim Gebrauchtwagenkauf bevorzugen würde. Die Bevorzugung einer Vignette interpretieren wir dahingehend, dass das Vertrauen der Versuchsperson in den Händler in der entsprechenden Situation größer ist als in der auf der alternativen Vignette beschriebenen Situation. In diesem Experiment variieren wir insgesamt sechs Merkmale der Situation (vgl. *Tabelle 2*). Unsere erste Variable und gleichzeitig unser einziges Merkmal der Transaktion ist der Preis des Gebrauchtwagens (1000 € bzw. 4000 €). Analog zu unserer Feldstudie nehmen wir an, dass der Umfang des Vertrauensproblems beim Gebrauchtwagenkauf positiv abhängt vom Preis. Innerhalb jedes Vignettenpaars halten wir den Preis konstant. Die Wahl zwischen zwei Vignetten kann daher nicht direkt vom Preis abhängen, aber es könnte der Fall sein, dass Einbettungseffekte bei preiswerten Gebrauchtwagen eine andere Rolle spielen als bei teuren. Wir beschränken uns mithin durch unser Design auf eine Analyse von Interaktionseffekten der Größe des Vertrauensproblems mit den Einbettungsvariablen.

Unsere übrigen fünf Variablen sind sämtlich Dummyvariablen, mit denen wir Merkmale der sozialen Einbettung der Transaktion variieren. Auf der Ebene der dyadischen Beziehung von Käufer und Händler repräsentieren wir Lerneffekte wiederum

Tabelle 2: Kauf eines Gebrauchtwagens: Beschreibung der Variablen in der Vignettenstudie

Variable	Wert	Text
Volumen	0	Sie können ein Auto für ca. 1000 € kaufen.
	1	Sie können ein Auto für ca. 4000 € kaufen.
Vergangenheit	0	Sie haben bisher noch kein Auto bei diesem Gebrauchtwagenhändler erworben.
	1	Sie haben schon einmal ein Auto bei diesem Gebrauchtwagenhändler erworben und Sie waren mit diesem Auto zufrieden.
Zukunft	0	In einigen Wochen ziehen Sie um in einen anderen Landesteil.
	1	Sie werden Ihren Wohnsitz vorerst nicht wechseln.
Dichte	0	Der Gebrauchtwagenhändler ist in Ihrer Nachbarschaft unbekannt.
	1	Der Gebrauchtwagenhändler ist in Ihrer Nachbarschaft gut bekannt und hat dort viele Kunden.
Innengrad	0	Soweit Sie wissen, hat keiner Ihrer Freunde schon einmal ein Auto bei diesem Gebrauchtwagenhändler erworben.
	1	Sie haben Freunde, die schon einmal ein Auto bei diesem Gebrauchtwagenhändler erworben haben und damit zufrieden waren.
Außengrad	0	Sie haben keine anderweitigen sozialen Kontakte mit dem Gebrauchtwagenhändler.
	1	Der Gebrauchtwagenhändler und Sie spielen in der gleichen Fußballmannschaft.

durch eine Variable, die angibt, ob der Käufer bereits früher einen Gebrauchtwagen bei dem betreffenden Händler erworben hat, mit dem der Käufer zufrieden war, oder ob es sich um die erste Anschaffung bei diesem Händler handelt. Entsprechend Hypothese H_2 erwarten wir, dass positive frühere Erfahrungen mit dem Händler die Attraktivität der betreffenden Vignette erhöhen. Kontrolleffekte auf der Ebene der Dyade durch die Aussicht auf zukünftige Transaktionen operationalisieren wir mittels einer Variable, die angibt, ob der Käufer in naher Zukunft vor einem größeren Umzug steht. Ein derartiger Umzug würde zukünftige Interaktionen zwischen Käufer und Händler unwahrscheinlich machen und daher die Kontrollmöglichkeiten des Käufers verringern. Entsprechend Hypothese H_3 erwarten wir daher, dass eine Vignette attraktiver wird, wenn der Käufer nicht umzieht. Der Kontrolleffekt auf der Ebene der Dyade Käufer-Händler sollte sich in unserer Vignettenstudie also in einem Haupteffekt der entsprechenden Variablen niederschlagen, im Gegensatz zu dem Interaktionseffekt, auf den wir in diesem Zusammenhang bei der Feldstudie über IT-Transaktionen abgehoben haben. Das liegt daran, dass wir in der Feldstudie den Umfang des ex ante-Managements der fokalen Transaktion als Maß für (fehlendes) Vertrauen betrachtet haben und dass das ex ante-Management der fokalen Transaktion Effekte hat für die Anreizstruktur bei zukünftigen Interaktionen. Man beachte, dass von derartigen Effekten in der vorliegenden Vignettenstudie keine Rede ist: wenn der Käufer dem Gebrauchtwagenhändler vertraut und zum Kauf übergeht, dann werden dadurch die Anreize bei zukünftigen Interaktionen selbst nicht verändert. Daher können wir uns bei der Analyse von Kontrolleffekten auf der Ebene der dyadischen Beziehung von Treugeber und Treuhänder auf den Haupteffekt erwarteter zukünftiger Interaktionen mit dem Händler konzentrieren.[15]

Wenden wir uns nun den Variablen zu, die unterschiedliche Aspekte der Netzwerkeinbettung modellieren. Zunächst verwenden wir eine Variable, die angibt, ob der Händler in der Nachbarschaft des Käufers gut bekannt ist. Diese Variable repräsentiert die Dichte des Netzwerks. Es handelt sich um ein globales Netzwerkmerkmal im Gegensatz zu den individuellen Netzwerkpositionen von Händler und Käufer. Wenn der Händler besser bekannt ist, dann wird der Käufer mehr Kontakte mit anderen möglichen Kunden des Händlers unterhalten und es wird mehr Kontakte geben zwischen diesen anderen Parteien. Das verstärkt Lern- und Kontrolleffekte auf der Netzwerkebene. Auf Grund unserer Hypothese H_6 erwarten wir, dass die Attraktivität einer Vignette zunimmt, wenn der Händler besser bekannt ist.

Informationen des Käufers über den Händler von dritten Parteien operationalisieren wir als – wiederum positive – Informationen von Freunden des Käufers, die selbst

15 In unseren Studien schlägt sich Vertrauen entweder nieder in Merkmalen von Verträgen und Vertragsverhandlungen oder aber in der Wahl eines Geschäftspartners. Eine andere und ebenfalls interessante Möglichkeit wäre, dass das Vertrauen des Treugebers in den Treuhänder die Zahlungsbereitschaft des Treugebers beeinflusst: Vertrauenswürdige Treuhänder in der Rolle des Lieferanten oder Verkäufers erzielen höhere Preise, weniger vertrauenswürdige Treuhänder müssen ihre geringere Vertrauenswürdigkeit mit einem Preisabschlag kompensieren. Hypothesen über die Effekte sozialer Einbettung auf den Verkaufspreis in Analogie zu den in diesem Beitrag untersuchten Hypothesen über Einbettungseffekte liegen auf der Hand. Vgl. Diekmann und Wyder (2002) für eine empirische Studie zu solchen Einbettungseffekten auf den Preis bei Internet-Auktionen.

bereits Kunden des Händlers waren (Innengrad). Derartige Informationen repräsentieren Lerneffekte auf der Ebene des Netzwerks. Entsprechend unserer Hypothese H_5 erwarten wir, dass die Verfügbarkeit derartiger Informationen die Attraktivität einer Vignette erhöht.

Unsere letzte Variable verwenden wir zur Modellierung von Kontrolleffekten auf der Ebene des Netzwerks. Die Variable gibt an, ob Käufer und Händler gemeinsam Mitglied einer Fußballmannschaft sind. Diese Variable können wir als Maß für den Außengrad des Käufers betrachten, weil man erwarten kann, dass die Anzahl gemeinsamer Kontakte von Käufer und Händler im Fall einer solchen gemeinsamen Mitgliedschaft größer ist. Die gemeinsame Mitgliedschaft sorgt für zusätzliche Kontrollmöglichkeiten des Käufers dadurch, dass er den Ruf des Händlers als Geschäftsmann und als Mitspieler beeinflussen kann. Der Käufer kann im Fall der gemeinsamen Mitgliedschaft andere Mitspieler ermutigen, beim Händler zu kaufen oder aber das gerade nicht zu tun und er verfügt über Möglichkeiten positiver bzw. negativer „sozialer" Sanktionierung durch die Kontakte mit anderen Mitspielern. Ein rationaler Händler wird diese Sanktionsmöglichkeiten in Rechnung stellen. Auf Grund unserer Hypothese H_7 erwarten wir daher einen positiven Effekt der gemeinsamen Mitgliedschaft für die Attraktivität einer Vignette. Bei der Beschreibung der gemeinsamen Mitgliedschaft haben wir bewusst auf jeden Bezug auf früheres Verhalten des Händlers verzichtet. Damit wollen wir vermeiden, dass Versuchspersonen die gemeinsame Mitgliedschaft als Möglichkeit für Lerneffekte interpretieren. Es ist offensichtlich, dass die gemeinsame Mitgliedschaft Lerneffekte ermöglicht. Da der soziale Kontext einer gemeinsamen Mitgliedschaft in einer Fußballmannschaft vergleichsweise zahlreiche und umfangreiche soziale Sanktionen ermöglicht und mögliche Information über früheres Verhalten des Händlers durch die gemeinsame Mitgliedschaft sowohl positiv als negativ sein könnte, gehen wir davon aus, dass der mit der gemeinsamen Mitgliedschaft verbundene Kontrolleffekt dominant ist. Ein Vorteil der Operationalisierung des Außengrads durch gemeinsame Mitgliedschaft ist im übrigen, dass die implizit verwendete theoretische Annahme relativ realistisch bleibt, dass das Netzwerk von Käufer und Händler „common knowledge" ist: Käufer und Händler werden nämlich wissen, dass sie beide Mitglieder der Fußballmannschaft sind. In einer empirischen Studie zum Konsumentenverhalten haben DiMaggio und Louch (1998) gezeigt, dass Käufer es häufig bevorzugen, einen Gebrauchtwagen von einem Verwandten anstatt von einem Fremden zu erwerben, mit dem man keine gemeinsame soziale Beziehung unterhält. Dieses Resultat unterstützt in gewisser Weise unsere Argumentation.[16]

Das Wahlverhalten der Versuchspersonen in diesem Experiment analysieren wir mit einem Random Utility-Modell für multiple Entscheidungen (McFadden 1973). Wir nehmen also an, dass jede Vignette für eine Versuchsperson einen bestimmten Nutzen repräsentiert, der linear von Merkmalen der Vignette abhängt. Dieser Ansatz führt dazu, dass wir ein Probit Modell schätzen, bei dem die Differenzen zwischen den Variablen eines Vignettenpaares die unabhängigen Variablen sind, während unsere abhängige Variable die Wahrscheinlichkeit für die Wahl einer der beiden Vignetten ist. Die

16 Vgl. Buskens und Weesie (2000) für eine Diskussion verschiedener Probleme im Zusammenhang mit Operationalisierungen einiger der hier verwendeten unabhängigen Variablen.

Tabelle 3: Probitanalyse der Wahl von Vignetten

Unabhängige Variable	Hypothese	Alle Fälle	Chicago	Utrecht	Tilburg
Dyadische Einbettung					
Vergangenheit	+	1.09**	0.99**	1.19**	1.39**
Zukunft	+	0.57**	0.61**	0.61**	0.30
Netzwerkeinbettung					
Dichte	+	0.71**	0.67**	0.73**	0.73**
Innengrad	+	0.83**	0.77**	0.89**	0.86**
Außengrad	+	0.26**	0.18	0.28*	0.51*
Zahl der Versuchspersonen		125	40	72	13
Fallzahl		1249	400	720	129

Effektstärken sind proportional zu den Parameterschätzungen und werden daher nicht berichtet.

** und * für zweiseitige Signifikanzniveaus $p < 0.01$ und $p < 0.05$; Huberkorrektur der Standardfehler (Huber 1967; Rogers 1993).

Koeffizienten des Modells können wir als Regressionskoeffizienten interpretieren. Ein positiver Koeffizient impliziert, dass die betreffende Variable den Nutzen einer Vignette erhöht.

In *Tabelle 3* präsentieren wir Ergebnisse für die gepoolten Daten aller Versuchspersonen und separat für die Sitzungen in Chicago, Utrecht und Tilburg. Die Resultate aus den verschiedenen Sitzungen sind bemerkenswert stabil und es gibt keine signifikanten Unterschiede zwischen den Sitzungen. Wir konzentrieren uns daher auf die Ergebnisse für die gepoolten Daten. Alle Einbettungsvariablen haben einen positiven Effekt auf die Attraktivität einer Vignette, d.h. sie haben einen positiven Effekt auf die Wahrscheinlichkeit, mit der die Versuchperson erwartet, einen Gebrauchtwagen zu erhalten, der sein Geld wert ist. Alle Einbettungsvariablen haben demzufolge einen positiven Effekt für das Vertrauen des Käufers in den Händler. Es scheint so zu sein, dass die Lerneffekte auf der Ebene der dyadischen Beziehung und auf der Netzwerkebene am stärksten sind: positive Informationen über früheres Verhalten des Treuhänders fördern das Vertrauen des Treugebers.[17] Auch die Möglichkeit zukünftiger Interaktionen des Treugebers mit dem Treuhänder, der Außengrad des Treugebers und die Netzwerkdichte beeinflussen die Attraktivität einer Vignette positiv. Diese Resultate bestätigen, dass auch Kontrolleffekte – und zwar sowohl auf der Ebene der dyadischen Beziehung als auch auf der Ebene des Netzwerks – das Vertrauen des Käufers fördern.[18]

17 Bei der Interpretation der Stärke der verschiedenen Effekte ist Vorsicht geboten, da die Skalen, auf denen die Variablen gemessen wurden, nicht vergleichbar sind, und da die Stärke der Effekte von den spezifischen Formulierungen der verschiedenen Variablenausprägungen abhängen können.

18 Vgl. Rooks et al. (2000) für eine Vignettenstudie mit Einkaufsmanagern, die die Robustheit der Resultate der hier berichteten Studie indiziert und die Vermutung nahelegt, dass sich Studierende in Situationen der hier untersuchten Art nicht exzeptionell verhalten.

VI. Eine experimentelle Studie: Effekte institutioneller Einbettung am Beispiel der Kooperation durch „Pfänder"

Dyadische Einbettung und Netzwerkeinbettung repräsentieren wichtige soziale Bedingungen des Vertrauens. Diese Formen der Einbettung sind aber nicht immer ausreichende Bedingungen für Vertrauen unter rationalen Akteuren. Frühere Interaktionen des Treugebers mit dem Treuhänder und Informationen, die der Treugeber von anderen Treugebern über deren frühere Interaktionen mit dem Treuhänder erhält, können z.B. dazu führen, dass der Treugeber es nicht für ausreichend wahrscheinlich hält, dass der Treuhänder Vertrauen honorieren würde (technisch: die Wahrscheinlichkeit π ist so klein, dass (1) nicht erfüllt ist). Der Schatten der Zukunft, sei es durch zukünftige Interaktionen des Treuhänders mit dem Treugeber aus dem fokalen Vertrauensspiel selbst oder durch zukünftige Interaktionen mit anderen Treugebern, ist zu gering für individuell rationales bedingtes Vertrauen (technisch: die Wahrscheinlichkeit w ist so klein, dass (2) nicht erfüllt ist). Dies kann etwa gerade dann der Fall sein, wenn das fokale Vertrauensspiel dem Treuhänder eine „goldene Gelegenheit" für den Missbrauch von Vertrauen bietet (technisch: der Anreiz $T_2 - R_2$ für den Missbrauch von Vertrauen ist besonders groß).

In derartigen Situationen kann Vertrauen zustande kommen, wenn das Geben und Honorieren von Vertrauen durch andere soziale Bedingungen und Mechanismen „unterstützt" wird. Ein Beispiel für solche Mechanismen sind „Pfänder" und ähnliche „commitments" (im Sinn Schellings 1960; vgl. auch Williamson 1985: Kap. 7 und 8). Bleiben wir beim Problem des Gebrauchtwagenkaufs und nehmen wir an, dass die dyadische Einbettung und die Netzwerkeinbettung allein nicht ausreichen, um einen rationalen und eigeninteressierten Händler vom Verkauf der Zitrone abzuhalten. Dies droht den potenziellen Kunden, der dies absehen kann, vom Kauf abzuhalten, was wiederum zu dem für Kunde und Händler ungünstigen Ausgang „keine Transaktion" führen würde. Eine typische Maßnahme, die der Händler ergreifen kann, um diesem Problem zu begegnen, ist die Gewährung einer Garantie für den Gebrauchtwagen. Durch die Gewährung einer Garantie wird es für den Händler weniger attraktiv, eine Zitrone zu verkaufen, da der Händler die zum Zeitpunkt des Verkaufs für den Kunden noch verborgenen Mängel des Wagens später auf eigene Kosten beheben muss. Dies ist der *Bindungseffekt* des Pfandes „Garantie". Ebenso verringert sich für den Kunden der mögliche Schaden, den er durch Erwerb einer Zitrone erleiden könnte. Die Kosten der Reparatur verborgener Mängel braucht der Kunde nicht selbst zu tragen. Dies ist der *Kompensationseffekt.* Schließlich ist auch denkbar, dass die Gewährung einer Garantie einen *Signaleffekt* hat. Der Kunde kann aus der Gewährung der Garantie möglicherweise schließen, dass der Händler „ehrlich" ist, keine Zitronen verkauft oder vielleicht gar keine Zitronen hat.

Spieltheoretische Modelle kann man verwenden, um soziale Bedingungen im Sinn institutioneller Regeln für die Pfänder zu spezifizieren, unter denen ein rationaler Treuhänder ein Pfand gibt, welches einen rationalen Treugeber dazu veranlasst, Vertrauen zu geben, das durch den Treuhänder honoriert wird (vgl. Raub 1992; Weesie und Raub 1996; Snijders 1996; Raub und Weesie 2000b; Raub 2004 für derartige Modelle und deren Analyse). Man nimmt dabei einen institutionellen Kontext als gegeben an,

der dem Treugeber die Möglichkeit bietet, ein Pfand zu stellen, bevor das fokale Vertrauensspiel selbst gespielt wird. Die Stellung eines Pfands ist ein „strategischer Zug" im Sinn Schellings (1960), weil dadurch die späteren Anreize für das Geben und Honorieren von Vertrauen beeinflusst werden. Den Kontext, der die *Gelegenheit* für die Stellung des Pfands bietet, betrachten wir als exogen gegeben: das Vertrauensspiel ist *institutionell eingebettet*. Wenn der Treuhänder von dieser Gelegenheit Gebrauch macht und ein Pfand stellt, dann schafft er eine „private Institution" für die Interaktion mit dem Treugeber. In Colemans (1991: 54) Terminologie würde man sagen, dass der Treuhänder für ein „konstruiertes soziales Umfeld" sorgt, das Vertrauen fördert. Institutionelle Einbettung sorgt also für Gelegenheiten für die „private Ordnung" sozialer und ökonomischer Beziehungen (Macaulay 1986, Williamson 1985). Die private Institution selbst – das Pfand – ist endogen. Mit spieltheoretischen Modellen spezifiziert man also soziale Bedingungen, unter denen diese privaten Institutionen Resultat individuell rationalen Gleichgewichtsverhaltens sind (vgl. Schotter 1981 und Calvert 1995 für die Unterscheidung zwischen Institutionen als exogen gegebenen Restriktionen bzw. Opportunitäten einerseits und Institutionen als Resultaten von Gleichgewichtsverhalten andererseits). In der spieltheoretischen Analyse nimmt man also gerade nicht an, dass eine externe dritte Partei den Treuhänder dazu zwingt, ein Pfand zu geben. Man stellt vielmehr die „tiefere" Frage nach sozialen Bedingungen, unter denen der Treuhänder ein Pfand freiwillig und ohne externen Zwang stellt.

Bemerkung: Die Garantie für den Gebrauchtwagen kann man als Beispiel für ein vertragliches Pfand interpretieren. Während es in unserer Feldstudie über IT-Transaktionen um den *Umfang* der Investitionen in die vertragliche „Absicherung" von Vertrauensproblemen ging, richten wir uns nun also gewissermaßen auf den *Inhalt* von Verträgen. Dabei sollte man aber nicht übersehen, dass Vertrauensprobleme häufig auch durch nichtvertragliche und in diesem Sinn informelle Pfänder gelöst werden und dass sich die institutionelle Einbettung von Vertrauensproblemen keineswegs nur auf rechtliche Rahmenbedingungen bezieht (vgl. Weesie und Raub 1996: 207–212 und Snijders 1996: 1–3 für diverse Beispiele).

Die Verwendung von Pfändern als Mechanismen für die Lösung von Vertrauensproblemen ist Gegenstand spieltheoretischer Modelle (vgl. die obigen Literaturhinweise). Hypothesen, die aus diesen Modellen folgen, wurden inzwischen auch in verschiedenen Experimenten überprüft (Snijders 1996; Snijders und Buskens 2001). Wir runden unseren Beitrag jedoch ab durch einen kurzen Blick auf eine experimentelle Studie (Raub und Keren 1993), in der Pfänder als Mechanismen der Kooperation im *Gefangenendilemma* untersucht werden. Vertrauensspiele sind Modelle für Situationen mit einseitigen Anreizen (des Treuhänders) für opportunistisches Verhalten, während im Gefangenendilemma *beide* Akteure solche Anreize haben. Wir beschäftigen uns also abschließend mit dem in gewisser Weise komplexeren Fall wechselseitiger Opportunismusprobleme und der Lösung solcher Probleme durch den wechselseitigen Gebrauch von Pfändern. Als Beispiel für solche Situationen wollen wir kurz den Kauf und Verkauf von Wohneigentum betrachten. Dieser ist in den Niederlanden so organisiert, dass Käufer und Verkäufer zunächst einen Kaufvertrag („koopakte") abschließen. Die eigentliche Eigentumsübertragung („eigendomsoverdracht") erfolgt typischerweise erst einige Monate später. Dieses Arrangement bietet für beide Parteien Vorteile im Vergleich zur Transaktion „in einem Zug". Man gewinnt Flexibilität bei der Suche nach einem

geeigneten Objekt bzw. nach einem Käufer, beide Parteien gewinnen Zeit für die Vorbereitung ihrer Umzüge usw. Beiden Parteien können sich aber nach der Unterzeichnung des Kaufvertrags und vor der Eigentumsübertragung Gelegenheiten und Anreize für opportunistisches Verhalten bieten. Der Verkäufer stellt z.B. fest, dass sein eigener Neubau, den er zu beziehen gedenkt, nicht rechtzeitig fertig wird und er würde deshalb seine jetzige Wohnung gerne noch länger benutzen. Der Käufer findet zwischenzeitlich ein anderes Objekt, dass seinen Wünschen (noch) besser entspricht. Zur Lösung dieser Anreizprobleme stellen beide Parteien ein Pfand, das die Verwendung des Arrangements „Durchführung der Transaktion in zwei Schritten" erleichtert. Das Pfand des Käufers beträgt 10 Prozent des Kaufpreises, die er bei Unterzeichnung des Kaufvertrags bei einem Notar hinterlegt. Dieses Pfand erhält der Verkäufer, falls der Käufer später den Kaufvertrag verletzt. Der Verkäufer stellt dadurch ein Pfand, dass er sich mit dem Kaufvertrag verpflichtet, auf 10 Prozent des vereinbarten Preises der Immobilie zu verzichten, falls er seinerseits später den Kaufvertrag verletzt.

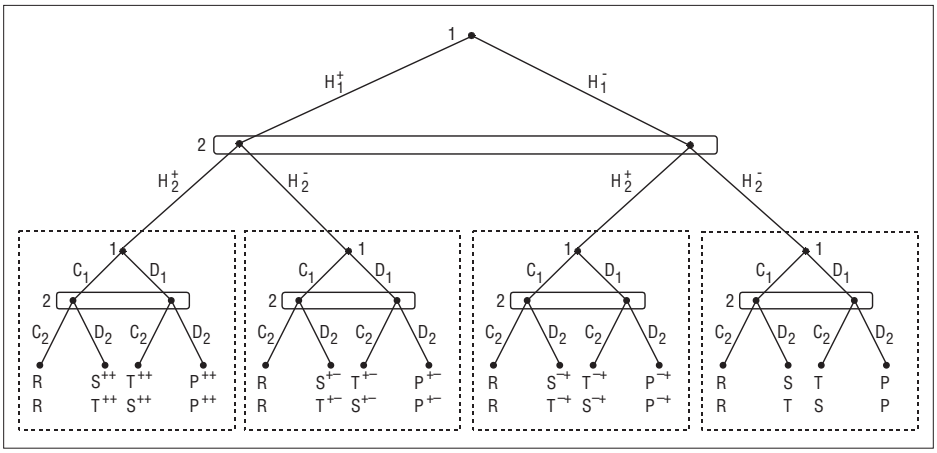
Abbildung 4: Das Gefangenendilemma ($S < P < R < T$)

		Akteur 2	
		Kooperation (C_2)	Defektion (D_2)
Akteur 1	Kooperation (C_1)	R,R	S,T
	Defektion (D_1)	T,S	P,P

Die bekannte Normalform des Gefangenendilemmas zeigt *Abbildung 4*. Man sieht, dass wechselseitige Defektion das eindeutige Gleichgewicht des Spiels ist. Defektion ist sogar *dominante Strategie*, d.h. einzige beste Antwort eines Akteurs gegen jede Strategie des anderen Akteurs. Wir betten jetzt das Gefangenendilemma in ein komplexeres Pfänderspiel ein, dessen extensive Form wir in *Abbildung 5* präsentieren. Das Gefangenendilemma ist, auch im technischen Sinn, ein Teilspiel dieses komplexeren Spiels. Wir wenden also das sogenannte „Nash-Programm" (Nash 1951) an, demzufolge alle ein- oder mehrseitigen Abmachungen und Verpflichtungen, einschließlich Pfänder und anderer strategischer Züge, explizit als Züge in einem erweiterten nichtkooperativen Spiel modelliert werden, mit dem Ziel, Bedingungen zu identifizieren, unter denen kooperatives Verhalten Gleichgewicht des erweiterten Spiels ist.

Das erweiterte Spiel ist ein Spiel mit vollständiger Information (technisch gesprochen: Es gibt am Anfang des Spiels keinen Zufallszug, dessen Ausgang durch mindestens einen Akteur nicht beobachtet werden kann). Wir abstrahieren damit von einer möglichen Signalfunktion von Pfändern und konzentrieren uns auf den Bindungs- und Kompensationseffekt (vgl. Raub und Weesie 2000b sowie Raub 2004 für komplexere spieltheoretische Modelle, in denen Pfänder auch Signalfunktion haben können). Das Spiel besteht aus zwei Phasen. In der ersten Phase entscheidet jeder der beiden Akteure, ob er ein Pfand stellt oder nicht. Einfachheitshalber nehmen wir an, dass die Stellung eines Pfandes kostenlos möglich ist (vgl. Weesie und Raub 1996 für ein Modell mit (Transaktions-)Kosten bei der Stellung eines Pfandes). Die Akteure entscheiden unabhängig voneinander über die Stellung eines Pfandes. Damit ist gemeint, dass ein

Abbildung 5: Einbettung des Gefangenendilemmas in ein Pfänderspiel ($S < P < R < T$, Superskripte indizieren mögliche Auszahlungsänderungen durch Pfänder)



Akteur bei seiner Entscheidung über die Stellung seines Pfandes nicht weiß, ob der andere Akteure ein Pfand stellt (man vgl. die Informationsmenge von Akteur 2 bei der Entscheidung über die Stellung seines Pfandes). In *Abbildung 5* notieren wir die Stellung eines Pfandes durch Akteur i mit H_i^+. Mit H_i^- deuten wir an, dass i kein Pfand stellt. Anschließend wird jeder Akteur über den ersten Zug des anderen Akteurs informiert, erfährt also, ob der andere Akteur ein Pfand gestellt hat. In der zweiten Phase des Spiels entscheiden die Akteure – wie im üblichen Gefangenendilemma – unabhängig voneinander, ob sie kooperieren oder defektieren. Danach endet das Spiel. In *Abbildung 5* verwenden wir für die Auszahlungen eine Notation, die die Züge beider Akteure im Verlauf des Spiels indiziert. Wir nehmen an, dass die Auszahlungen der Akteure additiv sind in a) ihren Auszahlungen, die sie, gegeben ihre Züge in der zweiten Phase des Spiels, im Gefangenendilemma erhalten würden, b) dem Wert des eigenen Pfandes, falls der Akteur dieses Pfand am Ende des Spiels verliert, und c) dem Wert des Pfandes des anderen Akteurs, falls man dessen Pfand am Ende des Spiels erhält. Im Hinblick auf institutionelle Regeln für die Pfänder wollen wir generell annehmen, dass ein Akteur ein Pfand *nur* dann verlieren kann, wenn er es in der ersten Phase des Spiels gestellt hat und dass er ein gestelltes Pfand jedenfalls dann *nicht* verliert, wenn er in der zweiten Phase des Spiels selbst kooperiert. Nehmen wir nun weiter an, dass die institutionellen Regeln für die Pfänder so beschaffen sind, dass ein Akteur i sein Pfand an den anderen Akteur j verliert, wenn i ein Pfand gestellt hat und danach in der zweiten Phase des Spiels einseitig defektiert. Nehmen wir auch an, dass beide Akteure ein Pfand stellen, dass danach Akteur 1 kooperiert und Akteur 2 defektiert. Die Auszahlung für Akteur 1 ist dann $S^{++} = S + K$, während Akteur 2 eine Auszahlung $T^{++} = T - K$ erhält, wobei K der Wert des Pfandes von Akteur 2 ist (die hier verwendete Annahme, dass der Wert des Pfandes von Akteur 2 für Akteur 2 dem Wert dieses Pfandes für Akteur 1 entspricht, ist dabei für die theoretische Analyse nicht wesentlich). Nehmen wir hingegen an, dass ein verlorenes Pfand *nicht* an den anderen Akteur

gegeben wird, dann wäre im angedeuteten Fall unter sonst unveränderten Annahmen die Auszahlung S^{++} für Akteur 1 nicht $S + K$, sondern S.[19]

Unter den beschriebenen Annahmen wird das Verhalten rationaler Akteure abhängen von den Auszahlungen T, R, P und S des Gefangenendilemmas, der Größe K, die den Wert des Pfandes repräsentiert, und den institutionellen Regeln, die festlegen, unter welchen Bedingungen ein Akteur sein Pfand verliert und ob ein verlorenes Pfand an den anderen Akteur gegeben wird oder aber „verfällt". Zentrale soziale Bedingungen sind also in diesem Modell institutionelle Regeln, die sich niederschlagen in der Größe K und die festlegen, unter welchen Umständen mit den Pfändern auf welche Weise verfahren wird. Die spieltheoretische Analyse (RAT) richtet sich auf die Ermittlung solcher sozialen Bedingungen, unter denen wechselseitige Stellung eines Pfandes und anschließende wechselseitige Kooperation Resultat individuell rationalen Verhaltens ist, also eines Verhaltens, dass jedenfalls durch ein (teilspielperfektes) Gleichgewicht unterstützt wird.

Wenden wir uns nun der Variante von QAD zu, die experimentelle Daten analysiert. Wir beschränken uns in diesem Beitrag auf die Skizze eines einzigen Experiments (vgl. Raub und Keren 1993 für eine genaue Beschreibung; Resultate mehrerer ergänzender Experimente werden in diesem Beitrag, in Keren und Raub 1993 und in Mlicki 1996 berichtet). Bei den Versuchspersonen handelt es sich um 73 Studierende verschiedener Studienrichtingen der Universität Utrecht. Jede Versuchsperson erhielt die Information, dass sie im Experiment gegen eine anonyme andere Versuchsperson spielen würde. De facto waren die Züge des Mitspielers vorprogrammiert und entsprachen rationalem (Gleichgewichts-)Verhalten. Im Hinblick auf die Auszahlungsfunktion der Versuchspersonen verwenden wir die vereinfachende Annahme „Nutzen = eigenes Geld". Jede Versuchsperson erhielt am Anfang des Experiments 5 Gulden (alle im Folgenden genannten Geldbeträge entsprechen Auszahlungen in niederländischen Gulden) und wurde informiert, dass sie im Lauf des Experiments, abhängig von ihren Entscheidungen und denen des Mitspielers, mehr Geld verdienen konnte. Das Experiment enthielt zwei Teile. Im ersten Teil spielte jede Versuchsperson ein Gefangenendilemma in der Form eines einfachen Kartenspiels. Eine Versuchsperson musste eine von zwei Karten unterschiedlicher Farbe wählen, wobei die vier möglichen Auszahlungen denen des Gefangenendilemmas entsprachen. Die Auszahlungen waren $T = 10$, $R = 8$, $P = 4$ und $S = 2$. Der fiktive Mitspieler wählte in diesem Teil des Experiments die Farbe, die der Defektion entsprach. Jede Versuchsperson wurde informiert über das Verhalten des Mitspielers und erhielt direkt danach eine monetäre Auszahlung (4 oder 2) entsprechend ihrem Verhalten in diesem Teil des Experiments. Dieser erste Teil des Experiments dient dazu, Versuchspersonen mit dem Experiment vertraut zu machen, liefert einen groben Indikator, ob die Nutzenfunktion der Versuchsperson unserer Annahme „Nutzen = eigenes Geld" entspricht (ein rationaler Akteur mit einer solchen Nutzenfunktion defektiert in diesem Teil des Experiments) und sorgt dafür, dass die Versuchsperson eine Information über den Mitspieler erhält, die konsistent ist mit der Annah-

19 In unserer Modellierung verwenden wir die radikale Annahme, dass jede Form von dyadischer Einbettung und von Netzwerkeinbettung fehlt. Vgl. Raub (1992) und Weesie et al. (1998) für komplexere Modelle, in denen mehrere Formen von Einbettung zugleich wirken.

me, dass dieser Mitspieler ein rationaler Akteur ist, für den „Nutzen = eigenes Geld" gilt (da der Mitspieler in diesem Teil des Experiments defektiert).

Im zweiten Teil des Experiments spielte jede Versuchsperson gegen den gleichen Mitspieler ein Pfänderspiel mit der extensiven Form aus *Abbildung 5* und $K = 4$ für den Wert des Pfandes. Die Versuchspersonen wurden explizit informiert über die Bedingungen, unter denen ein gestelltes Pfand verfallen oder gegebenenfalls an den Mitspieler gegeben würde. Versuchspersonen, die ein Pfand stellen wollten, übergaben den entsprechenden Betrag an den Versuchsleiter. Direkt nach ihrer eigenen Entscheidung über die Stellung eines Pfandes wurden Versuchspersonen über die entsprechende Entscheidung des Mitspielers informiert. Danach wurde das Kartenspiel aus dem ersten Teil des Experiments erneut gespielt und die Versuchspersonen erhielten eine Auszahlung entsprechend ihrer eigenen Farbwahl, der Farbwahl des Mitspielers, den vorhergehenden Entscheidungen im Hinblick auf die Stellung eines Pfandes und den Regeln für den Umgang mit gestellten Pfändern.

In diesem Experiment wurden drei verschiedene Versuchsbedingungen im Hinblick auf Eigenschaften der Pfänder verwendet. Eine *erste Gruppe* von Versuchspersonen spielte ein Pfänderspiel, in dem ein gestelltes Pfand genau dann an den Akteur, der es stellt, zurückgegeben wird, wenn dieser Akteur in der zweiten Phase des Spiels kooperiert. Ein verlorenes Pfand verfällt und wird nicht an den anderen Akteur gegeben. Unter diesen institutionellen Regeln hat das Pfand also einen bindenden, aber keinen kompensierenden Effekt. Spieltheoretische Analyse zeigt, dass das Pfänderspiel unter diesen Annahmen kein teilspielperfektes Gleichgewicht derart hat, dass beide Akteure ein Pfand geben, während das Spiel ein teilspielperfektes Gleichgewicht derart hat, dass beide Akteure auf die Stellung eines Pfandes verzichten und in der zweiten Phase des Spiels defektieren.

Eine *zweite Gruppe* von Versuchspersonen spielte ein Pfänderspiel, in dem ein gestelltes Pfand genau dann an den Akteur, der es stellt, zurückgegeben wird, wenn dieser Akteur in der zweiten Phase des Spiels kooperiert oder wenn beide Akteure in der zweiten Phase des Spiels defektieren. Ein verlorenes Pfand verfällt und wird nicht an den anderen Akteur gegeben. Wiederum hat das Pfand einen bindenden, aber keinen kompensierenden Effekt. Spieltheoretische Analyse zeigt, dass das Pfänderspiel unter diesen Annahmen ein teilspielperfektes Gleichgewicht derart hat, dass beide Akteure ein Pfand stellen. Die betreffenden Gleichgewichtsstrategien besagen, dass ein Akteur ein Pfand stellt und in der zweiten Phase des Spiels kooperiert, falls beide Akteure ein Pfand gestellt haben, während andernfalls in der zweiten Phase des Spiels defektiert wird. Dieses teilspielperfekte Gleichgewicht hat Auszahlungsdominanz im Vergleich zu allen anderen teilspielperfekten Gleichgewichten.

Für die *dritte Gruppe* von Versuchspersonen besagen die Regeln, dass ein gestelltes Pfand immer zurückgegeben wird, wenn der andere Akteur kein Pfand stellt. Wenn beide Akteure ein Pfand stellen, dann erhält ein Akteur sein Pfand außerdem dann zurück, wenn dieser Akteur in der zweiten Phase des Spiels kooperiert. Ein Akteur, der ein Pfand stellt, verliert dieses Pfand mithin genau dann, wenn dieser Akteur nach beiderseitiger Stellung eines Pfandes in der zweiten Phase des Spiels defektiert. Ein verlorenes Pfand wird an den anderen Akteur gegeben. Das Pfand ist unter diesen Regeln bindend und auch kompensierend. Für dieses Pfänderspiel zeigt die spieltheoretische

Analyse, dass es wiederum ein teilspielperfektes Gleichgewicht derart gibt, dass beide Akteure ein Pfand stellen. Die betreffenden Gleichgewichtsstrategien sind auch hier so beschaffen, dass ein Akteur ein Pfand gibt und in der zweiten Phase des Spiels kooperiert, falls beide Akteure ein Pfand gestellt haben, während andernfalls in der zweiten Phase des Spiels defektiert wird. Dieses teilspielperfekte Gleichgewicht hat Auszahlungsdominanz im Vergleich zu allen anderen teilspielperfekten Gleichgewichten.

Für die zweite und dritte Gruppe wird die Stellung eines Pfandes durch ein teilspielperfektes Gleichgewicht unterstützt. Die Gleichgewichtsstrategien sind *reaktiv* in dem Sinn, dass sie das Verhaltens des Akteurs abhängig machen vom Verhalten des Partners. Genauer gesagt enthalten die Gleichgewichtsstrategien das implizite Versprechen eigener Kooperation, falls der andere Akteur ebenfalls ein Pfand stellt, und die implizite Drohung eigener Defektion, falls der andere Akteur kein Pfand stellt. Teilspielperfektheit des Gleichgewichts stellt die Glaubwürdigkeit dieser Versprechungen und Drohungen in dem Sinn sicher, dass deren Implementation Teil eines Gleichgewichts in den entsprechenden Teilspielen ist. Man beachte übrigens, dass diese reaktiven Strategien auch für die erste Gruppe von Versuchspersonen ein Gleichgewicht bilden, dass dieses Gleichgewicht dort aber nicht teilspielperfekt ist: wenn etwa Akteur 1 ein Pfand gestellt hat, Akteur 2 aber nicht, dann ist im Folgenden Teilspiel Kooperation von Akteur 1 und Defektion von Akteur 2 das eindeutige Gleichgewicht und mithin ist unter den Bedingungen der ersten Gruppe die implizite Drohung von Akteur 1 nicht glaubwürdig, nach Verweigerung der Stellung eines Pfandes durch den Partner selbst zu defektieren.

Der Unterschied zwischen der zweiten und der dritten Gruppe in diesem Experiment wird deutlich, wenn wir uns die Frage stellen, ob die Stellung eines Pfandes nicht nur durch ein Gleichgewicht unterstützt wird, sondern auch durch Maximin-Verhalten. Das *Sicherheitsniveau* einer Strategie ist die minimale Auszahlung, die mit dieser Strategie erzielt wird. Eine *Maximin-Strategie* maximiert dieses Sicherheitsniveau. Eine Maximin-Strategie maximiert also sozusagen die Auszahlung eines Akteurs unter extrem pessimistischen Annahmen über das Verhalten des Partners. Die Maximin-Auszahlung im Pfänderspiel für das Gefangenendilemma entsprechend *Abbildung 5* ist offenkundig *P*. Man sieht leicht, dass das Sicherheitsniveau der Gleichgewichtsstrategie, die die Stellung eines Pfandes induziert, für die zweite Gruppe von Versuchspersonen kleiner ist als *P*. Die Gleichgewichtsstrategie führt nämlich zur Auszahlung $S^{++} = 2 < 4 = P$ für einen Akteur, der diese Gleichgewichtsstrategie gegen einen Partner spielt, der ein Pfand stellt, aber danach in der zweiten Phase des Spiels defektiert. Für die dritte Gruppe von Versuchspersonen ist demgegenüber die Gleichgewichtsstrategie, die die Stellung eines Pfandes induziert, zugleich auch Maximin-Strategie, und zwar deshalb, weil das Pfand unter den Versuchsbedingungen für diese Gruppe auch einen kompensierenden Effekt hat.

Für die dritte Gruppe von Versuchspersonen wird die Stellung eines Pfandes also nicht nur durch ein teilspielperfektes und auszahlungsdominantes Gleichgewicht unterstützt, sondern auch durch Maximin-Verhalten. Dies kann (vgl. Raub und Keren 1993; Weesie und Raub 1996) ein relevantes zusätzliches Argument dafür sein, dass die Stellung eines Pfandes Teil der Lösung des Spiels ist, wenn man z.B. annimmt, dass die Akteure unvollständig informiert sind über die Nutzenfunktion des Partners. Akteur 1

könnte etwa die Möglichkeit in Rechnung stellen, dass Akteur 2 Nutzen zieht aus der Minimierung der monetären Auszahlung von Akteur 1.

Aus der spieltheoretischen Analyse folgt damit, dass rationales Verhalten für die erste Gruppe von Versuchspersonen beinhaltet, dass kein Pfand gestellt und in der zweiten Phase des Pfänderspiels defektiert wird. Demgegenüber führt rationales Verhalten im Sinn von Verhalten entsprechend einem teilspielperfekten und auszahlungsdominanten Gleichgewicht für die zweite und dritte Gruppe von Versuchspersonen zur Stellung eines Pfandes und danach zur Kooperation, falls auch der andere Akteur ein Pfand gestellt hat. Für die dritte Gruppe von Versuchspersonen, nicht aber für die zweite, ist die Stellung eines Pfandes außerdem konsistent mit Maximin-Verhalten. Wir erwarten also (H_8), dass die Neigung, ein Pfand zu geben, bei den Versuchspersonen in der ersten Gruppe jedenfalls kleiner ist als bei den Versuchspersonen in der zweiten und dritten Gruppe.

Im Experiment erhielt die erste Gruppe von Versuchspersonen nach der eigenen Entscheidung über die Stellung eines Pfandes die Information, dass der „Mitspieler" kein Pfand gestellt hatte. In der folgenden Phase defektierte dieser Mitspieler. Die zweite und dritte Gruppe von Versuchspersonen spielte gegen einen Mitspieler, der ein Pfand stellte und, entsprechend dem auszahlungsdominanten teilspielperfekten Gleichgewicht, in der folgenden Phase kooperierte, falls auch die Versuchsperson selbst ein Pfand stellte, andernfalls aber defektierte. Ein Vergleich des Verhaltens von Versuchspersonen aus verschiedenen Gruppen in der zweiten Phase des Pfänderspiels ist nicht trivial, da sich die Gruppen unterscheiden im Hinblick auf den vorhergehenden Feedback über die Pfänderentscheidung des Mitspielers. Wir konzentrieren uns auf die Entscheidungen der Versuchspersonen über die Stellung eines Pfandes.

Tabelle 4: Zahl der Versuchspersonen, die nach Defektion im ersten Teil des Experiments im zweiten Teil ein Pfand stellen bzw. nicht stellen (Zahlen in Klammern schliessen Versuchspersonen ein, die im ersten Teil des Experiments kooperierten)

Gruppe	Versuchsbedingung: Stellung eines Pfandes ist ...	Pfand gestellt	kein Pfand gestellt
1	kein Teil eines teilspielperfekten Gleichgewichts	3[a] (4)[a]	15 (19)[b]
2	Teil eines teilspielperfekten Gleichgewichts, aber kein Maximin-Verhalten	9[a] (11)[a]	9 (11)
3	Teil eines teilspielperfekten Gleichgewichts und Maximin-Verhalten	18 (25)	1 (3)

[a] Eine Versuchsperson kooperierte nach der Stellung eines Pfandes *nicht*.
[b] Eine Versuchsperson defektierte nach Verweigerung der Stellung eines Pfandes *nicht*.

Aus *Tabelle 4* kann man ersehen, dass 74 Prozent aller Versuchspersonen im ersten Teil des Experiments, dem Gefangenendilemma, defektierten. Das Verhalten von Versuchspersonen, die im ersten Teil des Experiments kooperieren, scheint sich im übrigen im zweiten Teil des Experiments, also im Pfänderspiel selbst, kaum zu unterscheiden von dem Verhalten von Versuchspersonen, die im ersten Teil des Experiments defektieren.

Man könnte spekulieren, dass das Feedback über die Defektion des Mitspielers im ersten Teil des Experiments dabei eine wesentliche Rolle spielt. Man sieht, dass eine sehr große Mehrheit der Versuchspersonen in der ersten Gruppe kein Pfand stellt, entsprechend der spieltheoretischen Vorhersage. Ebenfalls entsprechend der spieltheoretischen Vorhersage stellt eine sehr große Mehrheit der Versuchspersonen in der dritten Gruppe ein Pfand. Von den Versuchspersonen in der zweiten Gruppe stellt die Hälfte ein Pfand. Der Anteil der Versuchspersonen, die in der zweiten Gruppe ein Pfand stellen ist signifikant größer als in der ersten Gruppe ($p < 0.05$, exakter Fisher-Test). Auffällig ist, dass sich die Versuchspersonen, von ganz wenigen Ausnahmen abgesehen, in der zweiten Phase des Pfänderspiels konsistent mit ihren Entscheidungen über die Stellung eines Pfandes in der ersten Phase verhalten. Insgesamt unterstützt das Experiment die Annahme, dass Pfänder eher dann als Kooperationsmechanismus verwendet werden, wenn der Gebrauch dieses Mechanismus konsistent ist mit individueller Rationalität. Insbesondere unterstützt das Experiment die Vermutung, dass das Konzept des teilspielperfekten Gleichgewichts in dem Sinn empirisch relevant ist, dass Pfänder nicht gestellt werden, wenn ihre Verwendung nicht zumindest durch ein teilspielperfektes Gleichgewicht unterstützt wird.

VII. Resümee

In diesem Beitrag haben wir einige Bausteine der Spieltheorie – zentrale Konzepte und Annahmen – skizziert und danach an Beispielen gezeigt, wie man Hypothesen über soziale Bedingungen individuellen Handelns und seiner kollektiven Folgen mittels spieltheoretischer Analysen generiert. Wir haben sodann gezeigt, wie man derartige Hypothesen empirisch überprüfen kann. Wir wollten damit Indizien liefern für die Vermutung, dass eine Allianz zwischen dem Rational Choice-Ansatz (RAT) und der quantitativen Analyse sozialwissenschaftlicher Daten (QAD) im Sinn von Goldthorpe möglich ist, und zwar auch dann, wenn man anknüpfend an Webers Definition der Soziologie die spieltheoretische Variante von RAT verwendet. Für eine solche Allianz kann man unterschiedliche und komplementäre Varianten von QAD heranziehen, d.h. unterschiedliche Daten und Forschungsdesigns.

Unser Beitrag ist umfangreich, aber nicht umfangreich genug für eine ausführliche Diskussion von Problemen bei der Anwendung spieltheoretischer Modelle in der empirisch orientierten Soziologie und für eine ausführliche Forschungsagenda. Als nur eines von vielen Problemen, die wir hier vernachlässigen, sei erwähnt, dass wir uns in unserem Beitrag – ähnlich vielen anderen empirischen Anwendungen der Spieltheorie – auf Hypothesen beschränkt haben, die im Kern etwas darüber aussagen, in welche Richtung sich Verhalten ändert, wenn sich soziale Bedingungen ändern. Wir haben uns in diesem Sinn auf „Marginalprognosen" konzentriert und spieltheoretische „Punktprognosen" beiseite geschoben, die mit der verfügbaren empirischen Evidenz häufig nicht oder nur unter sehr viel komplexeren Annahmen vereinbar sind (vgl. für die allgemeine Diskussion dieses Problems z.B. Green und Shapiro 1999: 55–56 und Übersichten wie etwa in Kagel und Roth 1995 oder Camerer 2003 zu empirischen Evidenzen, die mit spieltheoretischen Vorhersagen, vorsichtig ausgedrückt, nicht ohne weiteres in Einklang

gebracht werden können). Unsere Beispiele für Anwendungen sind sehr selektiv und wir haben zahlreiche für die Soziologie interessante spieltheoretische Modelle völlig außer Betracht gelassen. Mit den Hinweisen im Anhang wollen wir dem interessierten Leser den Zugang zu weiterführender Literatur erleichtern, in der diese Gesichtspunkte ausführlich erörtert werden.

Anhang: Literaturhinweise

Seit etwa 1990 sind zahlreiche ausgezeichnete *lehrbuchartige Darstellungen* der Spieltheorie erschienen. Stellvertretend für viele andere nennen wir Rasmusen (1994) als eine gründliche, aber technisch nicht zu anspruchsvolle Einführung in die *nichtkooperative Spieltheorie*. Ein deutschsprachiges Lehrbuch, das auch *kooperative Spiele* kurz behandelt, ist Güth (1992). Eine relativ neue Entwicklung in der Spieltheorie sind Modelle, die weitgehende Rationalitätsannahmen, die wir auch in diesem Beitrag verwendet haben, abschwächen oder aufgeben und etwa die Frage stellen, ob und unter welchen Bedingungen (und in welchem Sinn) „begrenzt rationales Verhalten" langfristig zu den Gleichgewichten der Spieltheorie führt (man beachte auch den Zusammenhang mit dem Gleichgewichtsauswahlproblem). Zur Modellierung *begrenzt rationalen Verhaltens* in Situationen strategischer Interdependenz und zur *evolutionären Spieltheorie* liegen inzwischen ebenfalls lehrbuchartige Darstellungen vor (z.B. Fudenberg und Levine 1998; Gintis 1999). Auch in der Soziologie wird zweifellos die sog. *„behavioral game theory"* einflussreich werden. Ziel dieser explizit an empirischen Anwendungen orientierten neueren Forschungsrichtung ist die Entwicklung von Modellen, die mit spieltheoretischen Instrumenten Verhalten – vor allem auch Verhalten in Experimenten – erklären, das inkonsistent ist mit „Standardannahmen" spieltheoretischer Analysen. Es geht dabei insbesondere um Verhalten in mittlerweile „klassischen" experimentellen Spielen, neben dem Gefangenendilemma und (Varianten des) Vertrauensspiel(s) etwa Diktator- und Ultimatumspiele oder auch verschiedenen Verhandlungsspielen. Man kann dabei z.B. an Versuche denken, die Eigennutzannahme durch komplexere Annahmen über die Nutzenfunktion der Akteure zu ersetzen. Dabei wird explizit angestrebt, mittels solcher komplexerer Nutzenannahmen systematisch neue Vorhersagen abzuleiten, so dass die empirische Prüfbarkeit der Modelle nicht leidet. Mit Camerer (2003) liegt auch zu diesem Programm bereits eine lehrbuchartige Darstellung vor. Im Zusammenhang mit empirischen Anwendungen spieltheoretischer Modelle muss schließlich die *experimentelle Ökonomie* erwähnt werden, die durch frühe Arbeiten von Selten im deutschsprachigen Raum mitbegründet wurde. Kagel und Roth (1995) ist das maßgebende Handbuch zu dieser Forschungsrichtung. Anwendungen der Spieltheorie in der Ökonomie sind inzwischen außerordentlich zahlreich und verzweigt. Wegen der Nähe zur Organisations- und Wirtschaftssoziologie verweisen wir lediglich auf spieltheoretisch orientierte lehrbuchartige Darstellungen wie Tirole (1988) und Milgrom und Roberts (1992) für das Teilgebiet der *Industrieökonomik*. Erwähnenswerte empirische Anwendungen der Spieltheorie sind sicherlich auch mehr „qualitative" Fallstudien von der Art der „analytic narratives" (Bates et al. 1998).

Ein informativer Beitrag zu *Grundlagenfragen im Zusammenhang mit der Anwendung der Spieltheorie in den empirisch orientierten Sozialwissenschaften* ist Kreps (1990b). Ein wichtiger kritischer Beitrag ist Green und Shapiro (1999). Zur *Geschichte der Spieltheorie* liegen inzwischen verschiedene Arbeiten vor. Eine konzise Übersicht mit Hinweisen auf weiterführende Literatur bietet Myerson (1999). Lesenswert ist in diesem Zusammenhang auch Nasars (1998) Nash-Biographie. Eine knappe *Geschichte spieltheoretischer Ideen in der Soziologie* bietet Swedberg (2001).

Anwendungen spieltheoretischer Modelle in der Soziologie finden sich inzwischen auch in den „Mainstream"-Zeitschriften des Fachs und regelmäßig in Zeitschriften wie etwa *Rationality and Society, Journal of Mathematical Sociology* und *Journal of Conflict Resolution.*

In diesem Beitrag haben wir uns auf *Anwendungen auf Vertrauens- und Kooperationsprobleme* konzentriert. Die hier skizzierten Studien zum Vertrauen beim wirtschaftlichen Handeln sind Teil

eines Forschungsprogramms, das zu mehreren komplementären Studien mit den auch hier verwendeten und anderen, verwandten Datensätzen geführt hat (vgl. außer den bereits oben erwähnten Arbeiten auch Prosch 1999 und Gautschi 2002). Auf die Beziehung Arbeitgeber-Arbeitnehmer werden ähnliche Überlegungen angewendet in Abraham (1996) und Abraham und Prosch (2000). Anwendungen auf die Interaktion von Partnern im Haushalt bieten Kalmijn et al. (1999) und Abraham (2002). In der Soziologie und ihren sozialwissenschaftlichen Nachbardisziplinen gibt es inzwischen eine große Zahl von Arbeiten mit Anwendungen spieltheoretischer Analysen auf Kooperationsprobleme. In der Soziologie scheint Voss (1982; 1985) derjenige zu sein, der als erster systematisch den Gedanken ausgearbeitet hat, dass soziale Ordnung darauf beruhen kann, dass rationale und eigeninteressierte Akteure bei Kooperationsproblemen ihre langfristigen Interessen berücksichtigen. In neueren Arbeiten (vgl. etwa Voss 2001) hat er aufgezeigt, wie solche Modelle angewendet werden können für die Erklärung der Entstehung und Stabilisierung sozialer Normen. Spieltheoretische Analysen kollektiven Handelns und der Produktion kollektiver Güter bieten z.B. Bendor und Mookherjee (1987) und auch Diekmanns (z.B. 1985) Freiwilligendilemma gehört in diese Gruppe von Beiträgen.

Bei Kooperationsproblemen geht es um die Frage, wie groß der Kuchen ist, der zwischen den Akteuren verteilt werden kann. *Verteilungsfragen* und *Fragen sozialer Ungleichheit* betreffen das Problem, wie ein Kuchen gegebener Größe zwischen den Akteuren aufgeteilt wird und ob bzw. unter welchen Bedingungen manche Akteure einen größeren Teil des Kuchens bekommen als andere. Auch Verteilungsprobleme werden mit spieltheoretischen Mittels untersucht. Anwendungen in der Soziologie findet man z.B. in der Literatur zum „Netzwerktausch" (vgl. Braun 2004, auch für weitere Literaturhinweise), aber auch Boudons (1974) einflussreichen Arbeiten über ungleiche Bildungschancen und soziale Ungleichheit in modernen Industriegesellschaften liegt im Kern ein spieltheoretisches Modell zugrunde (Boudon 1977: Kap. V). Becker (1991) und viele an Becker anschließende Arbeiten wenden spieltheoretische Modelle bei der soziologischen Analyse der Interaktion von Partnern im Haushalt an.

Ein passender Abschluss für einen Beitrag über spieltheoretische Modelle und empirische Anwendungen in einem Band zu Methoden der empirischen Sozialforschung scheint uns ein Hinweis auf die *Integration theoretischer und statistischer Modelle*. Bei Anwendungen des Rational Choice-Ansatzes auf parametrische Entscheidungssituationen ohne strategische Interdependenz zwischen Akteuren (bzw. bei Anwendungen, in denen von solchen Interdependenzen abstrahiert wird), kann man dazu bekanntlich von Random Utility-Modellen (McFadden 1973) Gebrauch machen. Weesie (2000) zeigt Wege für die Anwendung derartiger Modelle auch in strategischen Situationen.

Literatur

Abraham, Martin, 1996: Betriebliche Sozialleistungen und die Regulierung individueller Arbeitsverhältnisse. Endogene Kooperation durch private Institutionen. Frankfurt a.M.: Peter Lang.

Abraham, Martin, 2002: Die Haushalts- und Partnerschaftsorganisation selbständig Erwerbstätiger. Habilschrift Leipzig.

Abraham, Martin, und *Bernhard Prosch*, 2000: Long-Term Employment Relationships by Credible Commitments. Rationality and Society 12: 283–306.

Akerlof, George A., 1970: The Market for ‚Lemons': Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics 89: 488–500.

Arrow, Kenneth. J., 1974: The Limits of Organization. New York: Norton.

Axelrod, Robert, 1987: Die Evolution der Kooperation. München: Oldenbourg (Original 1984).

Batenburg, Ronald S., und *Werner Raub*, 1995: Het Externe Management van Automatisering 1995 (Das externe Management der Automatisierung 1995). [Datensatz] Universität Utrecht.

Batenburg, Ronald S., *Werner Raub* und *Chris Snijders*, 2003: Contacts and Contracts: Temporal Embeddedness and the Contractual Behavior of Firms. Research in the Sociology of Organizations 20: 135–188.

Bates, Robert H., Avner Greif, Margaret Levi, Jean-Laurent Rosenthal und *Barry R. Weingast*, 1998: Analytic Narratives. Princeton, NJ: Princeton University Press.

Becker, Gary S., 1991 (1981): A Treatise on the Family. Erw. Aufl., Cambridge, MA: Harvard University Press.

Bendor, Jonathan, und *Dilip Mookherjee*, 1987: Institutional Structure and the Logic of Ongoing Collective Action. American Political Science Review 81: 129–154.

Berg, Joyce, John Dickhaut und *Kevin McCabe*, 1995: Trust, Reciprocity, and Social History. Games and Economic Behavior 10: 122–142.

Blau, Peter M., 1964: Exchange and Power in Social Life. New York: Wiley.

Blossfeld, Hans-Peter, und *Gerald Prein* (Hg.), 1998: Rational Choice Theory and Large-Scale Data Analysis. Boulder, CO: Westview.

Blumberg, Boris F., 1998: Das Management von Technologiekooperationen: Partnersuche und vertragliche Planung. Wiesbaden: Gabler.

Boudon, Raymond, 1974: Education, Opportunity, and Social Inequality. New York: Wiley.

Boudon, Raymond, 1977: Effets Pervers et Ordre Social. Paris: PUF.

Braun, Norman, 2004: Tausch in Netzwerken. S. 129–141 in: *Andreas Diekmann* und *Thomas Voss* (Hg.), Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme. Rolf Ziegler zu Ehren. München: Oldenbourg.

Burt, Ronald S., und *Marc Knez*, 1995: Kinds of Third-Party Effects on Trust. Rationality and Society 7: 255–292.

Buskens, Vincent, 2002: Social Networks and Trust. Boston, MA: Kluwer.

Buskens, Vincent, und *Werner Raub*, 2002: Embedded Trust: Control and Learning. Advances in Group Processes 19: 167–202.

Buskens, Vincent, und *Werner Raub*, 2004: Soziale Mechanismen rationalen Vertrauens: Eine theoretische Skizze und Resultate aus empirischen Studien. S. 183–216 in: *Andreas Diekmann* und *Thomas Voss* (Hg.), Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme. Rolf Ziegler zu Ehren. München: Oldenbourg.

Buskens, Vincent, und *Jeroen Weesie*, 2000: An Experiment on the Effects of Embeddedness in Trust Situations: Buying a Used Car. Rationality and Society 12: 227–253.

Calvert, Randall L., 1995: Rational Actors, Equilibrium, and Social Institutions. S. 57–94 in: *Jack Knight* und *Itai Sened* (Hg.), Explaining Social Institutions. Ann Arbor, MI: University of Michigan Press.

Camerer, Colin F., 2003: Behavioral Game Theory. Experiments in Strategic Interaction. New York: Russell Sage.

Camerer, Colin F., und *Keith Weigelt*, 1988: Experimental Tests of a Sequential Equilibrium Reputation Model. Econometrica 56: 1–36.

Coleman, James S., 1991: Grundlagen der Sozialtheorie, Bd. 1: Handlungen und Handlungssysteme. München: Oldenbourg (Original 1990).

Coleman, James S., 1992: Grundlagen der Sozialtheorie, Bd. 2: Körperschaften und die moderne Gesellschaft. München: Oldenbourg (Original 1990).

Dasgupta, Partha, 1988: Trust as a Commodity. S. 49–72 in: *Diego Gambetta* (Hg.), Trust: Making and Breaking Cooperative Relations. Oxford: Blackwell.

Diekmann, Andreas, 1985: Volunteer's Dilemma. Journal of Conflict Resolution 29: 605–610.

Diekmann, Andreas, und *Thomas Voss*, 2004: Die Theorie rationalen Handelns. Stand und Perspektiven. S. 13–29 in: *Andreas Diekmann* und *Thomas Voss* (Hg.), Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme. Rolf Ziegler zu Ehren. München: Oldenbourg.

Diekmann, Andreas, und *David Wyder*, 2002: Reputation und Vertrauen bei Internet-Auktionen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 674–693.

DiMaggio, Paul, und *Hugh Louch*, 1998: Socially Embedded Consumer Transactions: For What Kinds of Purchases Do People Most Often Use Networks? American Sociological Review 63: 619–637.

Dutta, Bkaskar, und *Matthew O. Jackson* (Hg.), 2003: Networks and Groups. Models of Strategic Formation. Berlin: Springer.

Esser, *Hartmut*, 1993: Soziologie. Allgemeine Grundlagen. Frankfurt a.M.: Campus.

Friedman, Milton, 1953: The Methodology of Positive Economics. S. 3–43 in: *Milton Friedman*, Essays in Positive Economics. Chicago, IL: University of Chicago Press.

Fudenberg, *Drew*, und *David K. Levine*, 1998: The Theory of Learning in Games. Cambridge, MA: MIT Press.

Gautschi, *Thomas*, 2002: Trust and Exchange. Effects of Temporal Embeddedness and Network Embeddedness on Providing and Dividing a Surplus. Amsterdam: Thela Thesis.

Gintis, *Herbert*, 1999: Game Theory Evolving. Princeton, NJ: Princeton University Press.

Goldthorpe, *John H.*, 2000: On Sociology. Numbers, Narratives, and the Integration of Research and Theory. Oxford: Oxford University Press.

Gouldner, *Alvin W.*, 1960: The Norm of Reciprocity: A Preliminary Statement. American Sociological Review 25: 161–178.

Granovetter, *Mark S.*, 1985: Economic Action and Social Structure: The Problem of Embeddedness. American Journal of Sociology 91: 481–510.

Green, *Donald P.*, und *Ian Shapiro*, 1999: Rational Choice. Eine Kritik am Beispiel von Anwendungen in der Politischen Wissenschaft. München: Oldenbourg (Original 1994).

Güth, *Werner*, 1992: Spieltheorie und ökonomische (Bei)Spiele. Berlin: Springer.

Harsanyi, *John C.*, 1967/68: Games with Incomplete Information Played by ‚Bayesian' Players I-III. Management Science 14: 159–182, 320–334, 486–502.

Hedström, *Peter*, und *Richard Swedberg*, 1998: Social Mechanisms: An Introductory Essay. S. 1–31 in: *Peter Hedström* und *Richard Swedberg* (Hg.), Social Mechanisms. An Analytical Approach to Social Theory. Cambridge: Cambridge University Press.

Huber, *Peter J.*, 1967: The Behavior of Maximum Likelihood Estimates under Non-Standard Conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 221–233.

Kagel, *John*, und *Alvin E. Roth* (Hg.), 1995: Handbook of Experimental Economics. Princeton, NJ: Princeton University Press.

Kalmijn, *Matthijs*, *Wim Bernasco* und *Jeroen Weesie* (Hg.), 1999: Huwelijks- en Samenwoonrelaties in Nederland. De Organisatie van Afhankelijkheid (Haushalte in den Niederlanden. Die Organisation von Interdependenz). Assen: Van Gorcum.

Keren, *Gideon*, und *Werner Raub*, 1993: Resolving Social Conflicts through Hostage Posting: Theoretical and Empirical Considerations. Journal of Experimental Psychology: General 122: 429–448.

Kosfeld, Michael, 2003: Network Experiments. S. 131–152 in: *René van den Brink* und *Robert P. Gilles* (Hg.), Wiskundig Economische Perspectieven. Opstellen aangeboden aan prof. dr. Pieter H.M. Ruys (Perspektiven der mathematischen Ökonomie. Festschrift für Pieter Ruys). Tilburg: VAET.

Kreps, *David M.*, 1990a: Corporate Culture and Economic Theory. S. 90–143 in: *James E. Alt* und *Kenneth A. Shepsle* (Hg.), Perspectives on Positive Political Economy. Cambridge: Cambridge University Press.

Kreps, *David M.*, 1990b: Game Theory and Economic Modelling. Oxford: Clarendon Press.

Lindenberg, *Siegwart*, 1977: Individuelle Effekte, kollektive Phänomene und das Problem der Transformation. S. 46–84 in: *Klaus Eichner* und *Werner Habermehl* (Hg.), Probleme der Erklärung sozialen Verhaltens. Meisenheim am Glan: Hain.

Lindenberg, *Siegwart*, 1981: Erklärung als Modellbau. S. 20–35 in: *Werner Schulte* (Hg.), Soziologie in der Gesellschaft. Bremen: Zentraldruckerei der Universität.

Lindenberg, *Siegwart*, 1992: The Method of Decreasing Abstraction. S. 3–20 in: *James S. Coleman* und *Thomas J. Fararo* (Hg.), Rational Choice Theory: Advocay and Critique. Newbury Park, CA: Sage.

Lorenz, *Edward H.*, 1988: Neither Friends nor Strangers: Informal Networks of Subcontracting in French Industry. S. 94–107 in: *Diego Gambetta* (Hg.), Trust: Making and Breaking Cooperative Relations. Oxford: Blackwell.

Macaulay, *Stewart*, 1963: Non-Contractual Relations in Business: A Preliminary Study. American Sociological Review 28: 55–67.

Macaulay, Stewart, 1986: Private Government. S. 445–518 in: *Leon Lipson* und *Stanton Wheeler* (Hg.), Law and the Social Sciences. New York: Russell Sage.

McFadden, Daniel, 1973: Conditional Logit Analysis of Qualitative Choice Behavior. S. 105–142 in: *Paul Zarembka* (Hg.), Frontiers in Econometrics. New York: Academic Press.

Milgrom, Paul, und *John Roberts*, 1992: Economics, Organization and Management. Englewood Cliffs, NJ: Prentice Hall.

Mlicki, Pawel P., 1996: Hostage Posting as a Mechanism of Co-operation in the Prisoner's Dilemma Game. S. 164–183 in: *Wim B.G. Liebrand* und *David M. Messick* (Hg.), Frontiers in Social Dilemmas Research. Berlin: Springer.

Myerson, Roger B., 1999: Nash Equilibrium and the History of Economic Theory. Journal of Economic Literature 37: 1067–1082.

Nasar, Sylvia, 1998: A Beautiful Mind. New York: Simon & Schuster.

Nash, John F., 1951: Non-Cooperative Games. Annals of Mathematics 54: 286–295.

Nohria, Nitin, und *Robert G. Eccles* (Hg.), 1992: Networks and Organizations: Structure, Form, and Action. Boston: Harvard Business School Press.

Prosch, Bernhard, 1999: Die Absicherung von Lieferbeziehungen. Frankfurt a.M.: Lang.

Rapoport, Anatol, 1974: Introduction. S. 1–14 in: *Anatol Rapoport* (Hg.), Game Theory as a Theory of Conflict Resolution. Dordrecht: Reidel.

Rasmusen, Eric, 1994: Games and Information: An Introduction to Game Theory. 2. Aufl. Oxford: Blackwell.

Raub, Werner, 1992: Eine Notiz über die Stabilisierung von Vertrauen durch eine Mischung von wiederholten Interaktionen und glaubwürdigen Festlegungen. Analyse und Kritik 14: 187–194.

Raub, Werner, 1996: Effects of Temporal Embeddedness on Ex Ante Planning under Incomplete Information. ISCORE paper 87: Universität Utrecht.

Raub, Werner, 2004: Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling. Rationality and Society 16: 319–366.

Raub, Werner, und *Gideon Keren*, 1993: Hostages as a Commitment Device: A Game-theoretic Model and an Empirical Test of Some Scenarios. Journal of Economic Behavior and Organization 21: 43–67.

Raub, Werner, und *Chris Snijders*, 2001: A Reluctant Match: Models for the Analysis of Trust in Durable Two Party Relations. S. 195–200 in: *Cristiano Castelfranchi* und *Yao-Hua Tan* (Hg.), Trust and Deception in Virtual Societies. Dordrecht: Kluwer.

Raub, Werner, und *Jeroen Weesie*, 1990: Reputation and Efficiency in Social Interactions: An Example of Network Effects. American Journal of Sociology 96: 626–654.

Raub, Werner, und *Jeroen Weesie*, 2000a: The Management of Durable Relations. S. 1–32 in: *Jeroen Weesie* und *Werner Raub* (Hg.), The Management of Durable Relations: Theoretical Models and Empirical Studies of Households and Organizations. Amsterdam: Thela Thesis.

Raub, Werner, und *Jeroen Weesie*, 2000b: Cooperation via Hostages. Analyse und Kritik 22: 19–43.

Rogers, William M., 1993: Regression Standard Errors in Clustered Samples. Stata Technical Bulletin 13: 19–23.

Rooks, Gerrit, 2002: Contract en Conflict: Strategisch Management van Inkooptransacties (Vertrag und Konflikt. Strategisches Einkaufsmanagement). Amsterdam: Thela Thesis.

Rooks, Gerrit, Ronald Batenburg, Werner Raub und *Chris Snijders*, 1998: Het Extern Management van Automatisering 1998 (Das externe Management der Automatisierung 1998). [Datensatz] Universität Utrecht.

Rooks, Gerrit, Werner Raub, Robert Selten und *Frits Tazelaar*, 2000: Cooperation between Buyer and Supplier: Effects of Social Embeddedness on Negotiation Effort. Acta Sociologica 43: 123–137.

Rossi, Peter H., 1979: Vignette Analysis: Uncovering the Normative Structure of Complexity. S. 176–186 in: *Robert K. Merton, James S. Coleman* und *Peter H. Rossi* (Hg.), Qualitative and Quantitative Social Research: Papers in Honor of Paul F. Lazarsfeld. New York: Free Press.

Rossi, Peter H., und *Steven L. Nock* (Hg.), 1982: Measuring Social Judgments: The Factorial Survey Approach. Beverly Hills, CA: Sage.

Schelling, Thomas C., 1960: The Strategy of Conflict. London: Oxford University Press.

Schotter, *Andrew*, 1981: The Economic Theory of Social Institutions. Cambridge: Cambridge University Press.

Selten, *Reinhard*, 1965: Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. Zeitschrift für die gesamte Staatswissenschaft 121: 301–324, 667–689.

Snijders, *Chris*, 1996: Trust and Commitments. Amsterdam: Thesis Publishers.

Snijders, *Chris*, und *Vincent Buskens*, 2001: How to Convince Someone that You Can Be Trusted? The Role of ‚Hostages'. Journal of Mathematical Sociology 25: 355–383.

Swedberg, *Richard* (Hg.), 1993: Explorations in Economic Sociology. New York: Russell Sage.

Swedberg, *Richard*, 2001: Sociology and Game Theory: Contemporary and Historical Perspectives. Theory and Society 30: 301–335.

Taylor, *Michael*, 1987: The Possibility of Cooperation. Cambridge: Cambridge University Press (rev. ed. of Anarchy and Cooperation. London: Wiley 1976).

Tirole, *Jean*, 1988: The Theory of Industrial Organization. Cambridge, MA: MIT Press.

Voss, *Thomas*, 1982: Rational Actors and Social Institutions: The Case of the Organic Emergence of Norms. S. 76–100 in: *Werner Raub* (Hg.), Theoretical Models and Empirical Analyses. Contributions to the Explanation of Individual Actions and Collective Phenomena. Utrecht: ESP.

Voss, *Thomas*, 1985: Rationale Akteure und soziale Institutionen. München: Oldenbourg.

Voss, *Thomas*, 2001: Game-Theoretical Perspectives on the Emergence of Social Norms. S. 105–136 in: *Michael Hechter* und *Karl-Dieter Opp* (Hg.), Social Norms. New York: Russell Sage.

Weber, *Max*, 1976 (1921): Wirtschaft und Gesellschaft. 5. Aufl. Tübingen: Mohr 1976.

Weesie, *Jeroen*, 2000: Statistical Models for Dyadic Decision Making. In: *Jeroen Weesie* und *Werner Raub* (Hg.), The Management of Durable Relations: Theoretical Models and Empirical Studies of Households and Organizations. Amsterdam: Thela Thesis.

Weesie, *Jeroen*, *Vincent Buskens* und *Werner Raub*, 1998: The Management of Trust Relations via Institutional and Structural Embeddedness. S. 113–138 in: *Patrick Doreian* und *Thomas Fararo* (Hg.), The Problem of Solidarity: Theories and Models. Amsterdam: Gordon and Breach.

Weesie, *Jeroen*, und *Werner Raub*, 1996: Private Ordering: A Comparative Institutional Analysis of Hostage Games. Journal of Mathematical Sociology 21: 201–240.

Williamson, *Oliver E.*, 1985: The Economic Institutions of Capitalism. New York: Free Press.

Williamson, *Oliver E.*, 1996: The Mechanisms of Governance. Oxford: Oxford University Press.

Yamagishi, *Toshio*, und *Midori Yamagishi*, 1994: Trust and Commitment in the United States and Japan. Motivation and Emotion 18: 129–166.

Ziegler, *Rolf*, 1990: The Kula: Social Order, Barter and Ceremonial Exchange. S. 141–168 in: *Michael Hechter*, *Karl-Dieter Opp* und *Reinhard Wippler* (Hg.), Social Institutions: Their Emergence, Maintenance, and Effects. New York: Aldine.

Die Autorinnen und Autoren

Beck-Bornholdt, Hans-Peter, 1950, Prof. Dr. rer. nat., Universität Hamburg, Institut für Allgemeinmedizin. Forschungsgebiete: Forschungsmethodik, Evidence Assessment klinischer Studien, Nonsense. Veröffentlichungen: Der Hund der Eier legt – Erkennen von Fehlinformation durch Querdenken, Reinbek 2001 (mit H.-H. Dubben); Der Schein der Weisen – Irrtümer und Fehlurteile im täglichen Denken, Reinbek 2003 (mit H.-H. Dubben); Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall, Reinbek 2005 (mit H.-H. Dubben).

Brüderl, Josef, 1960, Dr. rer. pol., Professor für Statistik und sozialwissenschaftliche Methodenlehre an der Fakultät für Sozialwissenschaften der Universität Mannheim. Forschungsgebiete: Methoden, Wirtschaftssoziologie, Familiensoziologie. Veröffentlichungen: Der Erfolg neugegründeter Betriebe, Berlin 1998 (mit P. Preisendörfer und R. Ziegler); The Dissolution of Marriages, Journal of Mathematical Sociology 25, 2001 (mit F. Kalter); The Survival of Newly Founded Firms, Journal of the Royal Statistical Society 168, 2005 (mit G. Kauermann und G. Tutz).

Buskens, Vincent, 1968, Dr. disc. soc., Associate Professor für Soziologie und Methoden der Sozialforschung an der Universität Utrecht und am Interuniversity Center for Social Science Theory and Methodology (ICS). Forschungsgebiete: Soziale Netzwerke, mathematische Soziologie, Computersimulation, sozialwissenschaftliche Methoden, Rational Choice Theorie. Veröffentlichungen: Social Networks and Trust, Boston 2002; Embedded Trust: Control and Learning, Advances in Group Processes 19, 2002 (mit V. Buskens) (der Beitrag erhielt 2003 den James S. Coleman Award der Section on Rationality and Society der American Sociological Association); The Governance of Relations in Markets and Organizations, Research in the Sociology of Organizations, Vol. 20, Oxford 2003 (mit W. Raub und C. Snijders).

Couper, Mick P., 1957, Ph.D., Professor für Soziologie am Survey Research Center, University of Michigan. Forschungsgebiete: Methoden der empirischen Sozialforschung, Online-Forschung, technische Möglichkeiten der Befragung. Veröffentlichungen: Picture This! An Analysis of Visual Effects in Web Surveys, Public Opinion Quarterly 68, 2004 (mit K. Kenyon und R. Tourangeau); Does Voice Matter? An Interactive Voice Response (IVR) Experiment, Journal of Official Statistics 20, 2004 (mit E. Singer und R. Tourangeau).

Coutts, Elisabeth, 1967, Dipl.-Psych., wissenschaftliche Mitarbeiterin am Institut für Soziologie, Eidgenössische Technische Hochschule (ETH) Zürich. Forschungsgebiete: wissenschaftliches Fehlverhalten, Methoden der empirischen Sozialforschung, Urteils- und Entscheidungsforschung. Veröffentlichung: Das DEFECT-Projekt. Sampling-Errors und Non-sampling-Errors in komplexen Bevölkerungsstichproben, Opladen (im Druck) (mit R. Schnell, F. Kreuter und D. Thume).

Diekmann, Andreas, Professor für Soziologie an der ETH Zürich. Forschungsgebiete: Umweltsoziologie, Bevölkerung und Arbeitsmarkt, Methoden und Modelle, Untersuchungen zur sozialen Kooperation. Veröffentlichungen: Umweltsoziologie, Reinbek 2001 (mit P. Preisendörfer); Handbuch der Demographie (Hg.), Springer 2001 (mit U. Müller und B. Nauck); Reputation und Vertrauen bei Internet-Auktionen, Kölner Zeitschrift für Soziologie und Sozialpsychologie 54, 2002; Green and Greenback. The Behavioral Effects of Environmental Attitudes in Low-Cost and High-Cost Situation, Rationality and Society 15, 2003 (mit P. Preisendörfer); Do Parents of Girls Have a Higher Risk of Divorce? An Eighteen-Country

Study, Journal of Marriage and the Family 66, 2004 (mit K. Schmidheiny); The Power of Reciprocity. Fairness, Reciprocity, and Stakes in Variants of the Dictator Game, Journal of Conflict Resolution 48, 2004; The Human Dimensions of Global Environmental Change (Hg. mit T. Dietz, C.C. Jäger und E. A. Rosa), MIT-Press (im Druck).

DiPrete, *Thomas A.*, 1950, PH.D., Professor für Soziologie, Department of Sociology, Columbia University, New York. Forschungsgebiete: Soziale Schichtung und Mobilität, Arbeit und Arbeitsmarkt, Organisation, ökonomische Soziologie, quantitative Methoden. Veröffentlichungen: The Bureaucratic Labor Market: The Case of the Federal Civil Service, New York 1989; Discipline, Order and Student Behavior in American High Schools, Washington 1982; Losers and Winners: the Financial Consequences of Separation and Divorce for Men, American Sociological Review 66, 2001 (mit P. A. McManus); Life Course Risks, Mobility Regimes, and Mobility Consequences: A Comparison of Sweden, Germany, and the U.S., American Journal of Sociology 108, 2002; Do Cross-National Differences in Household Standard-of-Living Mobility Parallel Cross-National Differences in Occupational Mobility? A Comparison of Germany, Sweden, and the USA, Current Sociology 51, 2003.

Dubben, *Hans-Hermann*, 1955, PD Dr. rer. nat., Universität Hamburg, Institut für Allgemeinmedizin. Forschungsgebiete: Irrtümer und Trugschlüsse in der biomedizinischen Forschung. Veröffentlichungen: Systematic Review of Publication Bias in Studies on Publication Bias, British Medical Journal 2005: 433–434 (mit H.-P. Beck-Bornholdt); Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall, Reinbek 2005 (mit H.-P. Beck-Bornholdt); Der Schein der Weisen – Irrtümer und Fehlurteile im täglichen Denken, Reinbek 2003 (mit H.-P. Beck-Bornholdt); Der Hund der Eier legt – Erkennen von Fehlinformation durch Querdenken, Reinbek 2001 (mit H.-P. Beck-Bornholdt).

Engel, *Uwe*, 1954, Dr. phil., Professor für Soziologie mit dem Schwerpunkt Statistik und empirische Sozialforschung, Universität Bremen. Forschungsgebiete: Survey Methodologie, Mehrebenen- und Panelanalyse, Schulwirkungsforschung, strukturelle Ungleichheitsanalyse. Veröffentlichungen: Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre, Frankfurt a.M. 2001; Nonresponse und Stichprobenqualität. Ausschöpfung in Umfragen der Markt- und Sozialforschung, Frankfurt a.M. 2004 (mit M. Pötschke, C. Schnabel und J. Simonson); Bildung und soziale Ungleichheit. Methodologische und strukturelle Analysen, Bonn 2005.

Engelhardt, Henriette, 1968, PD Dr. rer.soc., wissenschaftliche Mitarbeiterin am Vienna Institute for Demographie der Österreichischen Akademie der Wissenschaften. Forschungsgebiete: statistische Methoden, Demographie, Familien- und Bevölkerungssoziologie, Veröffentlichungen: Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation, Sociological Methods & Research 32, 2004 (mit T. A. DiPrete); Fertility and Women's Employment Reconsidered: A Macro-level Time-series Analysis in Developed Countries, 1960–2000, Population Studies 58, 2004 (mit T. Kögel und A. Prskawetz); On the Changing Correlation Between Fertility and Female Employment over Space and Time, European Journal of Population 20, 2004 (mit A. Prskawetz).

Flache, *Andreas*, 1963, Dr., Dipl.-Informatiker, Assistenzprofessor für Soziologie an der Universität Groningen. Forschungsgebiete: Soziale Integration, soziale Netzwerke, Kooperation, formale Modellierung sozialer Prozesse. Veröffentlichungen: Stochastic Collusion and the Power Law of Learning. A General Reinforcement Learning Model of Cooperation, Journal of Conflict Resolution 46, 2002 (mit M. W. Macy); How May Virtual Communication Shape Cooperation in a Work Team? A Formal Model Based on Social Exchange Theory, Analyse & Kritik 26, 2004. Weitere neuere Arbeiten erschienen u.a. in Rationality and Society, Proceedings of the National Academy of Sciences U.S.A. und Annual Review of Sociology.

Gabler, Siegfried, 1951, PD Dr. sc. math., Chief Statistician beim Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim. Forschungsgebiete: Stichprobentheorie, Designeffekte. Veröffentlichungen: Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland, Münster 2002 (mit S. Häder); Combining Random Sampling and Census Strategies – Justification of Inclusion Probabilities Equal to 1, Metrika 2005 (mit H. Stenger); Approximations to b* in the Prediction of Design Effects due to Clustering, Survey Methodology 2005 (mit P. Lynn).

Gadenne, Volker, 1948, Dr. phil., Professor für Philosophie und Wissenschaftstheorie an der Johannes Kepler Universität Linz. Forschungsgebiete: Methodologie der Sozialwissenschaften, Theorien des Bewusstseins, Körper-Geist-Problem. Veröffentlichungen: Wirklichkeit, Bewusstsein und Erkenntnis, Rostock 2003; Philosophie der Psychologie, Bern 2004.

Gangl, Markus, 1972, Dr. rer. pol., Professor für Methoden der empirischen Sozialforschung und angewandte Soziologie an der Universität Mannheim. Forschungsgebiete: Arbeitsmarktsoziologie, Sozialstrukturanalyse, quantitative Verfahren. Veröffentlichungen: Transitions from Education to Work in Europe, Oxford 2003 (Hg. mit W. Müller); Welfare States and the Scar Effects of Unemployment, American Journal of Sociology 109, 2004; Institutions and the Structure of Labour Market Matching in the United States and West Germany, European Sociological Review 20, 2004.

Glemser, Axel, 1971, M.A., Leiter der Stichprobenabteilung bei TNS Infratest München. Veröffentlichung: ADM-Telefonstichproben in der Praxis, in: Siegfried Gabler und Sabine Häder: Telefonstichproben, Münster 2002.

Häder, Sabine, 1961, Dr. oec., Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim. Forschungsgebiete: Stichproben, insbesondere Telefonstichproben und Stichproben für interkulturelle Umfragen. Veröffentlichung: Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland, Münster 2002 (mit S. Gabler).

Jann, Ben, 1972, Lic. rer. soc., Assistent an der Professur für Soziologie der ETH Zürich. Forschungsgebiete: Arbeitsmarkt, Ungleichheit, Umwelt, Methoden. Veröffentlichungen: Einführung in die Statistik, München 2005; Earnings Returns to Education in Urban China: A Note on Testing Differences among Groups, American Sociological Review 2005; Tabulation of Multiple Responses, Stata Journal 2005.

Knäuper, Bärbel, 1965, Dr. phil., Department of Psychology, McGill University. Forschungsgebiete: Angewandte Soziale Kognition mit Schwerpunkt auf Gesundheit (Selbstregulation von Gesundheitsverhalten), Konstruktion psychometrischer Skalen, Validität von Selbstberichten, Fragebogenkonstruktion. Veröffentlichung: Cognition, Aging, and Self-reports, Philadelphia 1999 (mit N. Schwarz, D.C. Park und S. Sudman).

Krämer, Walter, 1948, Dr. rer. pol., Professor für Wirtschafts- und Sozialstatistik, Universität Dortmund, Institut für Wirtschafts- und Sozialstatistik. Träger des „Deutschen Sprachpreises" 1999. Forschungsgebiete: Statistik, quantitative Methoden, Methodologie, Wissenschaftstheorie. Veröffentlichungen: Die Krankheit des Gesundheitswesens, Frankfurt 1989; So lügt man mit Statistik, Frankfurt a.M. 1991; Lexikon der populären Irrtümer, Frankfurt 1996 (mit G. Trenkler); The Robustness of the F-test to Spatial Autocorrelation Among Regression Disturbances, Statistica 63, 2003; Datenanalyse mit SAS – Statistische Verfahren und ihre grafischen Aspekte, Berlin 2004 (mit O. Schoffer und L. Tschiersch).

Kuckartz, Udo, 1951, Professor für empirische Pädagogik und Forschungsmethoden an der Philipps-Universität Marburg. Forschungsgebiete: Methoden, insbes. computergestützte Analyseverfahren für qualitative Daten, sozialwissenschaftliche Umweltforschung, Innovationsforschung. Veröffentlichungen: Umweltbewusstsein im Wandel, Opladen 2003 (mit H. Grunen-

berg); Qualitative Datenanalyse: computergestützt. Methodische Hintergründe und Beispiele aus der Forschungspraxis, Wiesbaden 2004 (Hg. mit H. Grunenberg und A. Lauterbach); Einführung in die computergestützte Analyse qualitativer Daten, Wiesbaden 2005.

Macy, Michael W., 1948, Ph. D. Sociology, Professor of Sociology, Cornell University. Forschungsgebiete: Kollektives Handeln, Normentstehung, selbstorganisierende Gruppendynamik, Kaskaden in Netzwerken, kognitive Spieltheorie, agentenbasierte Computermodellierung, Laborexperimente. Veröffentlichungen: Learning Dynamics in Social Dilemmas, Proceedings of the National Academy of Sciences U.S.A., May 14, 2002 (mit A. Flache); The Evolution of Trust and Cooperation between Strangers: A Computational Model, American Journal of Sociology 110, 2005 (mit D. Centola und R. Willer). Weitere neuere Arbeiten erschienen u.a. in American Sociological Review, Annual Review of Sociology und Journal of Conflict Resolution.

Müller, Walter, 1942, Dr. rer. soc., Ph.D. h.c., Professor für Methoden der empirischen Sozialforschung und angewandte Soziologie, Universität Mannheim, und Forscher am Mannheimer Zentrum für Europäische Sozialforschung. Forschungsgebiete: Vergleichende Analyse der Sozialstruktur fortgeschrittener Gesellschaften, insb. Bildung, Arbeit, soziale Ungleichheit und Mobilität. Veröffentlichungen: Klassenstruktur und Parteiensystem: Zum Wandel der Klassenspaltung im Wahlverhalten, Kölner Zeitschrift für Soziologie und Sozialpsychologie 50, 1998; Transitions from Education to Work in Europe, Oxford 2003 (mit M. Gangl); The Reemergence of Self-Employment: A Comparative Study of Self-Employment Dynamics and Social Inequality, Princeton 2004 (mit R. Arum).

Pötter, Ulrich, 1956, Dr. rer. pol., Ruhr-Universität Bochum, Fakultät für Sozialwissenschaft, Sektion Sozialwissenschaftliche Methodenlehre und Statistik. Forschungsgebiete: Statistische Methoden, Analyse von unvollständigen Daten, Ereignisanalyse. Veröffentlichungen: Grundzüge der sozialwissenschaftlichen Statistik, Weinheim 2001 (mit G. Rohwer); Methoden sozialwissenschaftlicher Datenkonstruktion, Weinheim 2002 (mit G. Rohwer); Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung, Weinheim 2002 (mit G. Rohwer).

Raub, Werner, 1953, Dr. disc. soc., Professor für theoretische Soziologie an der Universität Utrecht und am Interuniversity Center for Social Science Theory and Methodology (ICS). Forschungsgebiete: theoretische Soziologie, Organisations- und Wirtschaftssoziologie, mathematische Soziologie und Anwendungen spieltheoretischer Modelle in den Sozialwissenschaften. Veröffentlichungen: The Management of Durable Relations. Theoretical Models and Empirical Studies of Households and Organizations, Amsterdam 2000 (mit J. Weesie); Embedded Trust: Control and Learning, Advances in Group Processes 19, 2002 (mit V. Buskens) (der Beitrag erhielt 2003 den James S. Coleman Award der Section on Rationality and Society der American Sociological Association); The Governance of Relations in Markets and Organizations, Research in the Sociology of Organizations, Vol. 20, Oxford 2003 (mit V. Buskens und C. Snijders); Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling, Rationality and Society 16, 2004.

Rohwer, Goetz, 1947, Dr. phil., Professor für Methoden der empirischen Sozialforschung und Statistik an der Ruhr-Universität Bochum. Forschungsgebiete: Methoden der empirischen Sozialforschung. Veröffentlichungen: Grundzüge der sozialwissenschaftlichen Statistik, Weinheim 2001 (mit U. Pötter); Methoden sozialwissenschaftlicher Datenkonstruktion, Weinheim 2002 (mit U. Pötter); Wahrscheinlichkeit, Begriff und Rhetorik in der Sozialforschung, Weinheim 2002 (mit U. Pötter).

Rost, Jürgen, 1952, Prof. Dr., Dipl.-Psych., Universität Kiel, Professur am IPN-Leibniz Institut für die Pädagogik der Naturwissenschaften. Forschungsgebiete: quantitative Methoden, Modellbildung, Skalierung. Veröffentlichungen: Gedächtnispsychologische Grundlagen naturwis-

senschaftlichen Wissens, Weinheim 1980; Rating Scale Analysis with Latent Class Models, Psychometrika 53, 1988; Die Erfassung transsituativer Copingstile durch Stimulus Response Inventare, Diagnostica 42, 1996 (mit M. v. Davier); Applications of Latent Trait and Latent Class Models in the Social Sciences, Münster 1997 (Hg. mit R. Langeheine); Handeln für die Umwelt – Anwendung einer Theorie, Münster 2001 (mit C. Gresele und T. Martens); Zeitgeist und Moden empirischer Analysemethoden, ZUMA-Nachrichten Spezial, Bd. 8, 2002; Lehrbuch Testtheorie – Testkonstruktion, Bern 2004 (2. Aufl.).

Scherer, Stefani, 1970, Dr. phil., Università Milano-Bicocca, Dipartimento di Sociologia e Ricerca Sociale. Forschungsgebiete: Arbeitsmarktsoziologie, Bildungssoziologie, statistiche Methoden. Veröffentlichungen: Mehr Risiken – Mehr Ungleichheit? Abbau von Wohlfahrtsstaat, Flexibilisierung von Arbeit und die Folgen, Frankfurt a.M. 2003 (mit W. Müller); Steppingstones or Traps? Work Employment and Society 18, 2004; Patterns of Labour Market Entry, European Sociological Review 21, 5, 2005.

Schwarz, Norbert, 1953, Dr. phil. habil., Professor für Psychologie an der University of Michigan und Research Professor am Institute for Social Research. Forschungsgebiete: Kognition und Urteilsbildung. Veröffentlichungen: Cognitive Aging, Philadelphia, 2000 (Hg. mit D. C. Park); Well-being: The Foundations of Hedonic Psychology, New York 2003 (Hg. mit D. Kahneman und E. Diener); A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method, Science 306, 2004 (mit D. Kahneman, A. Krueger, D. Schkade und A. Stone); Implicit Measures of Attitudes: Procedures and Controversies, New York (im Druck) (Hg. mit B. Wittenbrink).

Simonson, Julia, 1974, Dr. rer. pol., wiss. Assistentin an der Universität Bremen, Arbeitsgebiet Statistik und empirische Sozialforschung. Forschungsgebiete: Methoden empirischer Sozialforschung, Sozialstrukturanalyse, soziale Integration. Veröffentlichungen: Individualisierung und soziale Integration. Zur Entwicklung der Sozialstruktur und ihrer Integrationsleistungen, Wiesbaden 2004; Nonresponse und Stichprobenqualität. Ausschöpfung in Umfragen der Markt- und Sozialforschung, hg. vom ADM e.V., Frankfurt a.M. 2004 (mit U. Engel, M. Pötschke und C. Schnabel).

Stokman, Frans N., 1941, Professor für Sozialwissenschaftliche Forschungsmethoden an der Universität Groningen, Niederlande. Forschungsgebiete: soziale Netzwerke (und ihre Evolution), Modelle kollektiver Entscheidungsfindung. Georg Simmel Preis des International Network for Social Network Analysis, 2004. Veröffentlichungen: Strategic Decision Making, Advances in Group Processes 17, 2000 (mit M. A. L. M. van Assen, J. van der Knoop und R. C. H. van Oosten); Conflict Measures in Cooperative Exchange Models of Collective Decision Making, Rationality and Society 15, 2003 (mit M. A. L. M. van Assen und R. van Oosten); Winners and Losers in the European Union. Special Issue of European Union Politics 5, 2004 (Hg. mit R. Thomson); The European Union Decides, Cambridge (im Erscheinen) (Hg. mit R. Thomson, C. H. Achen und T. König).

Troitzsch, Klaus G., 1946, Dr. phil. habil., Professor für sozialwissenschaftliche Informatik am Institut für Wirtschafts- und Verwaltungsinformatik, Fachbereich Informatik, Universität Koblenz-Landau, Campus Koblenz. Forschungsgebiete: Quantitative Methoden, stochastische Prozesse, Computersimulation, speziell Multiagentensimulation. Veröffentlichungen: Simulation for the Social Scientist, Maidenhead/New York 1999 (mit N. Gilbert); Dynamik interagierender Bevölkerungen, in: Ulrich Mueller (Hg.), Handbuch der Demographie 1: Modelle und Methoden, Berlin 2000.

Vieth, Manuela, 1977, wissenschaftliche Assistentin (AIO) an der Universität Utrecht, Niederlande. Forschungsgebiete: soziale Normen, Kooperation, Fairness, Reziprozität, Spieltheorie und Rational Choice Theorie, Computersimulation, (faktorielle) Online-Surveys. Veröffentli-

chungen: Fairness im Ultimatumspiel. Eine spieltheoretische Modellierung, Zeitschrift für Soziologie 32, 2003 (auch in: Andreas Diekmann und Ben Jann (Hg.), Modelle Sozialer Evolution, Wiesbaden 2004).

Wagner, Michael, 1955, Dr. phil., Professor für Soziologie, Forschungsinstitut für Soziologie, Universität zu Köln. Forschungsgebiete: Familiensoziologie, Bildungssoziologie, Lebenslauf und Alter, quantitative Methoden. Veröffentlichungen: Bilanz der deutschen Scheidungsforschung. Versuch einer Meta-Analyse, Zeitschrift für Soziologie 32, 2003 (mit B. Weiß); Angewandte Soziologie, Wiesbaden 2004 (Hg. mit R. Kecskes und C. Wolf); Schulverweigerung. Empirische Analysen zum abweichenden Verhalten von Schülern, Kölner Zeitschrift für Soziologie und Sozialpsychologie 56, 2004 (mit I. Dunkake und B. Weiß).

Weiß, Bernd, 1974, M.A., wiss. Mitarbeiter am Forschungsinstitut für Soziologie, Universität zu Köln. Forschungsgebiete: Familiensoziologie, Jugenddelinquenz, quantitative Methoden. Veröffentlichungen: Bilanz der deutschen Scheidungsforschung. Versuch einer Meta-Analyse, Zeitschrift für Soziologie 32, 2003 (mit M. Wagner); Schulverweigerung. Empirische Analysen zum abweichenden Verhalten von Schülern, Kölner Zeitschrift für Soziologie und Sozialpsychologie 56, 2004 (mit M. Wagner und I. Dunkake).

Wirth, Heike, 1962, Dr. phil., Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim. Forschungsgebiete: Soziale Ungleichheit, Familiensoziologie, Datenschutz bei amtlichen Mikrodaten. Veröffentlichungen: Bildung, Klassenlage und Partnerwahl, Opladen 2000; Zur Weitergabe von älteren amtlichen Mikrodaten an die Wissenschaft, Allgemeines Statistisches Archiv, 2004 (mit P. Lüttinger); Kinderlosigkeit von Frauen im Spiegel des Mikrozensus. Eine Kohortenanalyse der Mikrozensen 1987 bis 2003, Wirtschaft und Statistik, 2005 (mit K.-J. Duschek).

Wolf, Christof, 1963, PD Dr. rer. pol., Wissenschaftlicher Leiter am Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim. Forschungsgebiete: Sozialstrukturanalyse, quantitative Methoden, Gesundheitssoziologie, Religionssoziologie. Veröffentlichungen: Struktur und sozialer Wandel einer Mittelstadt, Opladen 2002 (mit J. Friedrichs und R. Kecskes); Angewandte Soziologie, Wiesbaden 2003 (Hg. mit R. Kecskes und M. Wagner); Advances in Cross-National Comparison, New York 2002 (Hg. mit J. Hoffmeyer-Zlotnik).

English Summaries

Andreas Diekmann: **Current Issues in Empirical Research Methodology,** pp. 8–32.

This article reviews the basic problems of research methodology, as well as the issues involved in new data collection modes, data analytical methods and model development in the social sciences. This overview refers primarily to the contributions in this volume. We conclude by arguing in favour of greater professionalism in social science research. Specifically, the quality of such research should be raised through greater attention to and integration of previous research, clearer standards, more frequent replication of empirical findings and steady improvement in the methods employed.

Keywords: social research, research methodology, data collection, data analysis, model development

Stichworte: Sozialforschung, Forschungsmethodik, Datenerhebungsverfahren, Datenanalyse, Modellbildung

Volker Gadenne: **Empirical Research and Normative Theory of Science. Critical Rationalism under Scrutiny,** pp. 33–50.

Is critical rationalism still relevant for the methodology of social research? Does empirical science at all need a normative philosophical foundation? In order to answer these questions Popper's rules of falsification and his rejection of induction and justification are analysed. It is argued that some (though not all) rules of falsification seem to be useful heuristic principles. As to justification, it is shown that empirical science cannot dispense with a methodological principle that allows for the tentative acceptance of hypotheses as true. Finally, some functions of methodological rules and normative principles for the empirical sciences are pointed out and discussed.

Keywords: critical rationalism, theory of science, logic of social research, Popper, falsification

Stichworte: kritischer Rationalismus, Wissenschaftstheorie, Logik der Forschung, Methodologie der Sozialforschung, Popper, Falsifizierung

Walter Krämer: **Statistics – Helping or Hindering Progress in the Social Sciences?,** pp. 51–60.

It is argued that descriptive statistics and statistical inference help the social sciences in different degrees. While there is still much need for further improvements in basic issues of definition and measurement, the marginal utility of conventional mathematical statistics might well by now be negative.

Keywords: measurement, statistical fallacies, statistical inference, descriptive statistics

Stichworte: Meßprobleme, schließende Statistik, statistische Trugschlüsse, deskriptive Statistik

Hans-Hermann Dubben and *Hans-Peter Beck-Bornholdt:* **The Meaning of Statistical Significance,** pp. 61–74.

First the good news: there is reasonable hope, that in the near future the devastating and usually unread flood of more than 5,000 biomedical publications per day with alleged "significant" results will shrink to an easy to survey streamlet. The bad news: the prevailing interpretation of p-values is incorrect, making it easy to obtain "statistically significant" but false findings. Correcting this false interpretation is the key step towards the good news. Imagine the results of your clinical study (that is methodologically perfect) show higher survival rate in the experimental arm as compared to standard treatment. The observed difference might be a chance result due to random variations in treatment outcome. Therefore you consult your statistician raising the clinically relevant question: "How likely is it that I am wrong, when I consider the experimental therapy to be superior to the standard therapy?" After some calculations he replies: "The difference of your results is statistically significant (p=0.03)." With this, the most important criterion for publication is fulfilled, but: What does that sentence mean? The statistician's reply does not mean what most clinicians believe: "When I regard the experimental therapy as superior, the probability of being wrong is only 3 percent." The p-value is not the answer to your question! In fact it barely means: "If both therapy arms are actually equivalent, the probability to measure a difference as observed or bigger simply by chance is 3 percent." The difference between this two sentences is not hair-splitting but perhaps the most frequent and most severe error in modern biomedical research. The basic problem is elucidated with everyday examples and "statistically significant" results. It will be shown that p-values alone are a meaningless measure of evidence.

Keywords: statistical significance, significance test, hypothesis test, p-value, error probability, predictive value, Bayesian statistics, Bayes

Stichworte: Statistische Signifikanz, Signifikanztest, Hypothesentest, p-Wert, Irrtumswahrscheinlichkeit, prädiktiver Wert, bayesianische Statistik, Bayes

Jürgen Rost: **Counting or Scaling?,** pp. 75–92.

In the empirical social sciences, indicators usually are counted, item responses are added, and the numbers obtained are taken as scale values of variables. The Rasch-model is the only model for questionnaire data, in which the non weighted sums of item responses represent the complete statistical information in the data with respect to the measures of the persons. In the present paper, the process of analysing test or questionnaire data according the Rasch-model and its generalisations is described. This is done under the condition that no precise a-priori-hypothesis of the psychometric structure of the test is given. In the course of this analysis a number of models is used, i.e. the Rasch-model, partial credit model, rating scale model, latent class analysis, latent class models for ordinal data, the mixed Rasch-model and hybrid-models. In order to illustrate the methods of scaling analysis, including quantitative and classifying models, nine items on behaviour evaluation have been chosen from the ALLBUS 2002. This scaling analysis has figured out four groups, in which counting the item responses is a meaningful operation of scaling. Each of these groups has its own specific profile of mean item responses, serving as a frame of references for the interpretation of the sum scores.

Keywords: data analysis, scaling, scaling analysis, Rasch-model, latent class analysis, hybrid-model

Stichworte: Datenanalyse, Skalierung, Skalenanalyse, Rasch-Modell, Analyse latenter Klassen, Hybrid-Modell

Heike Wirth and *Walter Müller:* **Microdata from Official Sources. Its Research Potential for Empirical Social Research,** pp. 93–127.

Official microdata are next to science based databases an essential source for empirical social research. However, the research potential of official microdata is far from being fully exploited in Germany. This is partly due to previous restrictions in the access to such data, but partly also to lacking knowledge about its usefulness in many parts of the scientific community. Recently, with support of the Federal Ministry of Education and Research great efforts have been made to improve the access to official microdata. The present paper addresses the current potential of official microdata for empirical social research in Germany. It starts with a brief overview on the development of the legal conditions with respect to data access and then describes in general terms the strengths and weaknesses of various official microdata bases available in Germany. Their research potential is analysed by reviewing examples of recent major studies based on such data. Finally, the paper explores the future use of it in empirical social research.

Keywords: official microdata, official statistics, database, data access, public use file, secondary analysis, research potential, historical data analysis

Stichworte: amtliche Mikrodaten, amtliche Statistik, Datenbasis, Datenzugang, öffentlicher Datensatz, Sekundäranalyse, historische Datenanalyse

Siegfried Gabler: **Weighting Problems in Data Analysis,** pp. 128–147.

In the analysis of empirical data it is often assumed that the underlying sampling procedure is simple random sampling. But this is true only very seldom. Usually complex sample designs are used. The danger of drawing wrong conclusions is given not only for point estimators but also in computing variance estimators and confidence intervals. Weights should be assigned to respondent records to incorporate the sample design and adjustments for non-response and noncoverage. Extreme weights should be avoided. In the last years model-assisted approaches are developed in data analysis where the model specifies the parameter of interest, but all inference is based on the sample design. Fortunately, more and more statistical packages allow the user to take weights into consideration and include a model-assisted approach as part of the design-based approach.

Keywords: random sampling, sample design, design weight, cell weighting, calibration, model-assisted approach

Stichworte: Zufallsstichprobe, Stichprobenplanung, Designgewichtung, Zellgewichtung, Kalibrierung, Modell gestützte Analyse

Sabine Häder and *Axel Glemser:* **Designs for Telephone Survey Sampling in Germany,** pp. 148–171.

Currently, in Germany more than 40 percent of all interviews in market research are conducted by telephone. A prerequisite for a high quality of the telephone surveys is an appropriate generation of sample. Therefore, there is a need for an adequate sampling frame. The telephone book cannot be used due to under representation, too many households are not listed. Also unlisted households have different socio-demographic characteristics to listed households. The advantages and disadvantages of alternative sample frames will be discussed, as well as designs for telephone survey sampling. The Gabler-Häder-design for the academic world and the ADM-design for the professional market researcher are established as "state of the art" in Germany in the last couple of years. The ongoing development of mobile phones is a new challenge for both sampling systems.

Keywords: sampling, sampling design, telephone interviewing, response rate, Gabler-Haeder-design, ADM-master sample

Stichworte: Stichprobenziehung, Stichprobenplanung, Telefoninterview, Antwortrate, Gabler-Häder-Stichprobenplan, ADM-Stichprobenplan

Ulrich Pötter: **Probabilistic Selection Models,** pp. 172–202.

Missing or incomplete data are a pertinent to any social survey. Interviewees do not answer some questions or refuse to participate at all. Nowadays, response rates in surveys rarely reach 40 percent. Can the answers of 40 percent of the selected individuals be treated as if one had answers from all of them? Probabilistic selection models can be used to speculate on the connexion between the attained information and the possible answers of the non-responders. These models can be extended to cover incomplete data such as grouped or censored data as well as processes of self-selection and the evaluation of social programmes. A brief review of some of these applications is given. Next, it is shown that the assumptions used in probabilistic selection models necessarily refer to concepts transcendenting empirically accessible social facts. However, the amount and impact of these assumptions can be reduced if possibly available additional incomplete information is fully used. Therefore, probabilistic selection models are formulated in a form that allows for grouped, censored and general coarsened data. Finally, several proposals for the sensitivity analysis of probabilistic selection models are presented and it is argued that approaches restricted to too small neighbourhoods of a given model can be quite misleading. Throughout, responses on income questions obtained in the German ALLBUS 1996 are used as an illustration.

Keywords: sampling theory, response rate, incomplete data, probabilistic selection model, sensitivity analysis

Stichworte: Stichprobentheorie, Antwortrate, unvollständige Daten, probabilistisches Selektionsmodell, Sensitivitätsanalyse

Norbert Schwarz and *Bärbel Knäuper:* **Cognition and Survey Methodology: Age-sensitive Context Effects,** pp. 203–216.

The collaboration of cognitive psychologists and survey methodologists resulted in an increased understanding of the cognitive and communicative processes underlying context effects in self-reports. This understanding set the stage for a theoretical analysis of differential context effects, as we illustrate for age-sensitive context effects. Specifically, we find that question order effects decrease with age, whereas response order effects increase with age. Both observations can be traced to age-related changes in working memory function. Similarly, context effects on behavioral frequency reports increase with age, unless the behavior is of high personal relevance. The size of the observed effects is often sufficient to reverse the ordinal placement of age groups. Any observed age difference in reported attitudes and behavior may reflect (a) a difference in the phenomenon under study, (b) a difference in the response process, or (c) an unknown mix of both.

Keywords: survey methodology, age effects, context effects, cognition, attitude measurement, behavioral frequency reports

Stichworte: Umfragemethodologie, Alterseffekte, Kontexteffekte, Kognition, Einstellungsmessung, Verhaltensberichte

Mick P. Couper and *Elisabeth Coutts:* **Online Surveys: An Updated Review of Issues and Approaches,** pp. 217–243.

In 2000, Couper reviewed research and practices on web surveys. Since then, the popularity of online surveys has grown, along with the body of research on sources of error in web surveys and development of a set of best practices. This article builds on Couper's (2000) review article in providing an overview of the main sources of survey error and their impact on online studies. These include variations in internet use among different population subgroups, issues involved in sampling the internet population, nonresponse and measurement error. A typology is then presented of the kinds of web surveys in common use, distinguishing between non-probability and probability surveys. The possibilities and limitations of each of those approaches are discussed in light of recent findings.

Keywords: online survey, web survey, survey error, coverage, response rate, nonresponse

Stichworte: Online-Befragung, Internet-Umfrage, Fehlerquellen bei Umfragen, Abdeckung, Antwortquote, Nonresponse

Christof Wolf: **Egocentric Networks. Data Collection and Data Quality,** pp. 244–273.

This article focuses on the techniques for the collection of egocentric network data in standardized mass surveys. After a short presentation of the different traditions of network research a detailed description of the different data collection methods for egocentric networks is given. The emphasize of the discussion lies on techniques that employ name generators, i.e. questions asking for names of contact persons. The second part of the paper is concerned with the quality of egocentric network data. Results related to the validity and reliability of name generator questions, the quality of proxi-measures for characteristics of network persons and relationships are discussed. Similarly, information regarding the quality effects of the mode of data collection and of interviewer behavior are investigated. The article closes with recommendations for surveying egocentric networks.

Keywords: egocentric network, survey research, data quality, data collection, interview, name generator question, validity, reliability

Stichworte: egozentriertes Netzwerk, Umfrageforschung, Datenqualität, Datenerhebung, Interview, Namensgeneratorfrage, Zuverlässigkeit, Gültigkeit

Frans N. Stokman and *Manuela Vieth:* **What Binds Us When With Whom? Content and Structure in Social Network Analysis,** pp. 274–302.

In the increasingly popular field of social network analysis, the concept of social capital is often used to indicate the opportunities social networks give to groups and individuals for the realization of their goals. In this article it is argued that most network studies concentrate too much on network structure and neglect that their findings are context specific, depending on which goals are at stake and which types of relationships are studied. The effect is a fragmentation of network approaches without a specification under which conditions they are appropriate. In this article a heuristic is proposed that enables an integration of network approaches. This is done by giving analytic primacy to actors' goals that cannot be realized in isolation and thus lead to joint production and sharing. The heuristic explicitly incorporate material gain goals, hedonic goals ("feeling good") and normative goals ("behaving appropriately"). The heuristic guides the selection of the combination of processes and corresponding social network relationships that matter in a given setting. It leads to theories and hypotheses on what is produced by which network under which structural conditions and why. The analytic power of the heuristic is illustrated in two completely different contexts, namely policy networks and friendship networks.

Keywords: social capital, network exchange theory, norms, policy networks, networks in organizations, friendship networks, collective decision making, joint production

Stichworte: Soziales Kapital, Netzwerk-Austauschtheorie, Normen, Politiknetzwerke, Netzwerke in Organisationen, Freundschaftsnetzwerke, kollektive Entscheidungsfindung, gemeinschaftliche Produktion

Uwe Engel and *Julia Simonson:* **Social Context in Multi-Level-Analysis,** pp. 303–329.

The paper deals with contextual analysis and shows how to employ multilevel modelling techniques for purposes of carrying out such analyses. First, basic ideas for the analysis of hierarchically structured data are introduced along with a review of areas of applied multi-level modelling. Then part II presents some basics of the hierarchical linear regression approach, followed by a third part on structural analysis and compositional effects. Part IV presents models for cross-sectional and panel data. Finally, part V addresses a multilevel approach for the analysis of categorical dependent variables. For illustrative purposes, up to four waves of the German Socio-economic Panel were used to estimate multilevel models for the effects of individual status inconsistency and household-level status incongruency on a couple of social integration indicators.

Keywords: contextual analysis, multilevel analysis, compositional effects, structural analysis, panel data, growth curves, status inconsistency, GSOEP

Stichworte: Kontextanalyse, Mehrebenenanalyse, Kompositionseffekte, strukturelle Analyse, Paneldaten, Wachstumskurven, Statusinkonsistenz, SOEP

Josef Brüderl and *Stefani Scherer:* **Methods for Analyzing Sequence Data,** pp. 330–347.

This article surveys methods for analyzing sequence data. It deals with the representation of life-courses as sequences, with optimal matching analysis, and sociological applications of sequence data methods. Two applications on the transition from school to work and on the diversity of living arrangements give a sense for the power of these methods. The final section discusses the analytical potential of optimal matching analysis.

Keywords: data analysis, life course, sequence data, optimal matching analysis

Stichworte: Datenanalyse, Lebensverlauf, Sequenzdaten, Optimal-Matching-Analyse

Götz Rohwer: **Durations and Rates in Competing Risk Situations,** pp. 348–367.

The paper discusses the definition and estimation of durations if episodes can end in several different destination states. It is shown that the Kaplan-Meier procedure leads to pseudo survivor functions which have no meaningful interpretation in terms of durations. As an alternative, the papers proposes to use time-dependent functions which directly capture the proportion of objects moving into the different destination states. To illustrate the argument, the papers uses a simulation model for the duration of marriages which can end in divorce or widowhood.

Keywords: duration of episodes, survival analysis, Kaplan-Meier procedure, competing risk situations

Stichworte: Episoden-Verweildauer, Survival-Analyse, Kaplan-Meier-Verfahren, konkurrierende Risiken

Henriette Engelhardt: **Causal Analysis with Single and Pooled Time Series**, pp. 368–395.

Time series are gaining increasing importance as databases for empirical research in the social sciences. Often it is not only diverse variables of time series data which are cumulated, but also time series of different research units. The structure of the data which have been combined into these pooled time series is very similar to the structure of panel data. While typical panel data are characterised by many research units but only few points in time, pooled time series are characterised by only a few research units and many time points. This review article examines multiple regression analyses for single and pooled time series and various versions of these schemes. Specific estimation problems are discussed which stem primarily from specific attributes of the time series data. We illustrate the analysis of pooled time series using data on the effect of women's employment on fertility in 22 OECD countries (1960-2000).

Keywords: time series, pooled and single time series, causal analysis, regression analysis, estimation problems, women's employment, fertility

Stichwörter: Zeitreihen, zusammengefasste und separate Zeitreihen, Kausalananlyse, Regressionsanalyse, Schätzprobleme, Frauenerwerbstätigkeit, Fertilität

Markus Gangl and *Thomas A. DiPrete:* **Matching Methods for the Causal Analysis of Observational Data,** pp. 396–420.

Having close linkages with the counterfactual concept of causality, nonparametric matching estimators have recently gained in popularity in the statistical and econometric literature on causal analysis. Introducing key concepts of the Rubin causal model (RCM), the paper discusses the implementation of counterfactual analyses by propensity score matching methods. We emphasize the suitability of the counterfactual framework for sociological questions as well as the assumptions underlying matching methods relative to standard regression analysis. We then illustrate the application of matching estimators in an analysis of the causal effect of unemployment on workers' subsequent careers.

Keywords: matching methods, causal analysis, nonparametric estimators, observational data, Rubin causal model, propensity score matching, regression analysis

Stichworte: Matchingverfahren, Kausalanalyse, nichtparametrische Verfahren, Beobachtungsdaten, Rubin Causal Model, Propensity Score Matching, Regressionsanalyse

Ben Jann: **Regression Diagnostics in Small Samples,** pp. 421–453.

Like all other statistical procedures, regression methods focus on the analysis of only select aspects of the data material at hand. Accordingly, a given set of regression results may result from many different data constellations, not all of which lend themselves to straightforward interpretation. Particularly in small samples, regression estimates may, for example, depend critically on a few extreme data points, a situation in which the reliability of the conclusions derived from the regression analysis is greatly reduced. In this contribution, I will therefore introduce a few simple graphical and formal instruments for the diagnosis of influential data. These tools should be applied to data analysis as a matter of course. The steps to be taken after the identification of "atypical" cases are also discussed.

Keywords: data analysis, regression analysis, reliability, extreme data points, small samples, robust regression

Stichworte: Datenanalyse, Regressionsanalyse, Zuverlässigkeit, Extremwerte, kleine Stichproben, robuste Regression

Udo Kuckartz: **Computer Aided Analysis of Qualitative Data,** pp. 454–478.

More and more computer software is used for the analysis of qualitative data (QDA-software). The article presents the basic techniques of QDA-analysis that are centred around systematic categorisation of the textual material. In the beginning QDA software offered primarily code-and-retrieve procedures. Chunks of text were coded and later on a synopsis of segments coded with the same category could be done. Modern QDA-software allows much more, e.g. searching for complex patterns of codes and visualising relationships between codes. Finally the article reviews the discourse on the methodological questions related to these new techniques of computer-assisted analysis, particularly the effect on the research process and the question of quality of qualitative data. It will be argued that the new techniques offer a chance for more accuracy and for a more systematic analysis of qualitative data, which overcomes the anecdotic evidence that has often be criticised in the past.

Keywords: computer aided data analysis, QDA-software, qualitative data, quality of data

Stichworte: Computerunterstützte Datenanalyse, QDA-software, qualitative Forschung, Datenqualität

Michael Wagner and *Bernd Weiss:* **Meta-Analysis as a Method of Social Research,** pp. 479–504.

In empirical social research, we usually distinguish between primary and secondary analyses. In the last two decades, a third type of empirical social research, called meta-analysis (Glass 1976), has been developed. Meta-analysis is the statistical analysis of the empirical results of independent studies. The aim of meta-analysis is to summarize the empirical results from primary or secondary analyses and to explain (possible) discrepancies between these results. The article focuses on three goals: first, we demonstrate that meta-analysis covers the whole research process starting with the phrasing of a problem, followed by data collection and statistical analyses, and ending with the interpretation of results. Second, we show various methods used for the synthesis of results, including examples. Finally, we discuss methodological problems and the potential use of meta-analysis in sociology.

Keywords: meta-analysis, primary analysis, secondary analysis, research process, data collection, statistical analysis,

Stichworte: Meta-Analyse, Primärforschung, Sekundäranalyse, Forschungsprozess, Datenerhebung, Datenanalyse

Klaus G. Troitzsch: **Dynamic Models of Systems,** pp. 505–535.

After a short survey of the history of formal models in the social sciences and of different approaches of computer simulation used during the past four decades, this chapter discusses two fundamentally different types of formal process modelling. On the one hand, simulation is not more than the numerical solution of systems of mathematical equations. The most interesting of these do not have analytical solutions such that without numerical simulation no conclusions could be drawn from the mathematical model. One of the representatives of this type is system dynamics which became famous 30 years ago when the Club of Rome sponsored research into the dynamics of growth in a finite world. On the other hand, one should not talk about computer simulation in a strict sense unless one makes uses of the possibilities of modern, object-oriented software and represents the real-world entitites of one's target system with more or less autonomous software objects which depend on each other in the same manner as do the respective real-world entities. Multilevel simulation, the classical microanalytic simulation, cellular automata and, particularly, multi-agent models belong to this latter class of simulation models.

Keywords: computer simulation, formal process modelling, system dynamics, multilevel simulation, microanalytic simulation, cellular automata, multi-agent models

Stichworte: Computersimulation, formale Modellierung, System-Dynamics-Ansatz, Mehrebenenmodell, Mikrosimulation, zellulare Automaten, Multiagentenmodell

Andreas Flache and *Michael W. Macy:* **"Bottom-up" Models of Social Dynamics. Agent-based Computer Models and Methodological Individualism,** pp. 536–559.

Agent-based computer modeling is a new approach that aims to model theoretically how complex social macrodynamics emerge from the interactions of autonomous, yet interdependent individual actors ("agents"). Our contribution describes the main principles of the agent-based computer modeling and discusses its relationship to methodological individualism. We argue that agent-based computer modeling offers to methodological individualists a tool that allows combining the rigor of formal modeling with more freedom in the choice of behavioral assumptions than formal game theory does. We illustrate this with a comparison of game theoretical and agent-based models of emergent social order. We discuss in particular models that address the effect of relational stability, network structure and behavioral assumptions on spontaneous social order. We conclude our contribution with a discussion of methodological principles that agent-based modelers should adhere to in order to fully exploit the potential of this method for social theory.

Keywords: agent-based modeling, computational modeling, computer simulation, game theory, social order, methodological individualism

Stichworte: agentenbasierte Modellierung, Computermodellierung, Computersimulation, Spieltheorie, soziale Ordnung, methodologischer Individualismus

Werner Raub and *Vincent Buskens:* **Game-theoretic Models and Empirical Applications in Sociology,** pp. 560–598.

Game theory provides a toolbox for describing and analysing social situations with interdependent actors. We sketch concepts and assumptions of game theory and subsequently offer some examples of applications in sociology. Our applications focus on a core topic of sociological theory and fields such as the sociology of organisations, economic sociology, and the sociology of law, namely, how social conditions affect trust and cooperation in social and economic life. We show how game-theoretic models can be used to generate testable hypotheses. We offer empirical evidence from a series of studies employing different designs: a survey, a vignette study, and an experiment. We wish to show by way of example that game-theoretic modelling can contribute to an integration of systematic theory construction based on the rational choice approach and empirical research in sociology.

Keywords: theory construction, rational choice approach, hypotheses testing, game-theoretic modelling, trust, cooperation, exchange

Stichworte: Theoriekonstruktion, Rational-Choice-Ansatz, Hypothesentest, spieltheoretische Modellierung, Vertrauen, Kooperation, Austausch