Herausgeber:
Zentralarchiv für Empirische Sozialforschung
Universität zu Köln

Das Zentralarchiv ist Mitglied der GESIS

Direktor: Prof. Dr. W. Jagodzinski
Geschäftsführer: E. Mochmann

Postanschrift:
Postfach 410 960
50869 Köln

Hausanschrift:
Bachemer Straße 40
50931 Köln

Telefon:

Zentrale 0221 / 4 76 94	- 0
Telefax	- 44
Redaktion	- 50

Redaktion:
Franz Bauske

E-mail: Name@ibm.za.uni-koeln.de
Internet: http://www.social-science-gesis.de

ISSN: 0723-5607

Die ZA-INFORMATION erscheint jeweils im Mai und November eines Jahres.

Sie wird kostenlos an Interessenten und Benutzer des Zentralarchivs abgegeben.

Inhaltsverzeichnis

Ankündigungen und Mitteilungen

Buchhinweise

Bei Beiträgen, die nicht von Mitarbeitern des Zentralarchivs verfaßt wurden, ist die Anschrift der Autoren beim jeweiligen Artikel angegeben. Die Inhalte der Beiträge entsprechen der Meinung der Autoren und geben nicht unbedingt die Ansicht der Redaktion wieder. Alle inhaltlichen Beiträge sind Gegenstand einer Beurteilung durch externe Gutachter.

Mitteilungen der Redaktion

Telefonische Befragungen erfreuen sich in der Bundesrepublik zunehmender Beliebtheit. Eine fast vollständige Versorgung aller Haushalte hat die Voraussetzung für ein preiswertes Erhebungsverfahren geschaffen, das mehr und mehr das persönliche Interview ersetzt. Nachdem die bundesdeutschen Telefonbücher auf CD-ROM erhältlich sind, ist zu überlegen, ob - und wenn ja wie - aus diesem Informationsbestand Befragungsteilnehmer sinnvoll ausgewählt werden können. ***Rainer Schnell*** schlägt ein Auswahlverfahren vor, das durch eine Programmierung auch weitgehend automatisiert werden kann.

Die Kompliziertheit log-linearer Modelle steht einer verbreiteten Anwendung in der Sozialforschung häufig im Wege, zumal der Eindruck besteht, daß sie nur schwer interpretierbar seien. ***Steffen M. Kühnel*** zeigt an einem Beispiel von ALLBUS-Daten auf, wie auch komplexe Modelle mit speziellen Design-Matrizen in SPSS-Prozeduren realisierbar sind.

Größere und komplexere Datensätze wie Paneldaten stellen eine besondere Herausforderung an die Datenanalyse. ***Björn Alecke*** stellt dar, wie man mit Hilfe der üblichen Statistik-Programmpakete eine Regressionsanalyse mit Panel-Daten durchführen kann. Er gibt eine Einführung in dieses Verfahren.

Petra Hartmann hat das Zentralarchiv in Richtung Kiel verlassen. Sie hat eine Stelle am Institut für Soziologie an der Christian-Albrechts-Universität angetreten. ***Markus Klein*** aus Mainz ist als wissenschaftlicher Mitarbeiter neu ins Zentralarchiv eingetreten. Sein Spezialgebiet ist die politische Soziologie.

Per Nielsen, Direktor des Dänischen Archivs, ist zur Jahreswende nach kurzer, schwerer Krankheit gestorben. Der in Europa und auch weltweit hoch angesehene Archivkollege bestach immer durch sein professionelles und ideenreiches Auftreten und war auf der anderen Seite ein stets gut aufgelegter Gesprächspartner: eine besondere Persönlichkeit. Seine Verdienste um das Archivwesen würdigt ***Ekkehard Mochmann.***

Franz Bauske

In memoriam

Per Nielsen

Per Nielsen, Director of the Dansk Data Arkiv (DDA) died on December 27, 1996 at the age of 49 years.

Per Nielsen started his career in the world of empirical social research at the Danish Institute for Social Research in 1973. He joined DDA when it began its operation in 1974. As junior partner of ***Ole Engberg*** he participated in international contacts within the emerging archival network. Since 1977 he has been Director of the DDA which was then a project under the Social Science Research Council. ***Per Nielsen*** managed to transfer DDA into an institute with national coverage under Odense University from 1978 onwards and to land it in the harbour of the Danish State Archives after 1990.

Already in the early seventies DDA and the Zentralarchiv co-operated in the development of standard instruments for international data documentation and data transfer. ***Per Nielsen*** undertook to finalise early developments of ***Hans Dieter Klingemann*** and ***Harm t´Hart*** for the study description scheme. Under his guidance DDA came up with the first guidebook of how to document data on the study level and laid the basis on which ***Sue Dodd*** could build her international standard. In these early years of the emerging international data movement ***Per Nielsen*** had a keen eye on detecting promising new developments and he pushed forward to getting those new tools implemented into daily work.

Per Nielsen was around when the International Association of Social Science Information Service and Technology (IASSIST) was founded in Toronto in 1974. DDA soon hosted the European Secretariat of IASSIST and he served as vice president of IASSIST. By his communicative skills he became a mediator between the male dominated European scene and the self conscious female colleagues from IASSIST. Later he also took over responsibilty for the Secretariat of the International Federation of Data Organisation for the Social Sciences (IFDO).

In the Council of European Social Science Data Archives (CESSDA) he soon became one of the strongest supporters of the idea to establish expert seminars, which would bring together the younger layer of the data movement to share their experience and to foster innovative ways of transboarder co-operation.

Given his political engagement and his active participation in political demonstrations, he also was one of the first to realise the importance of the emerging data protection laws and he was the reliable Danish source to compile an international overview in the international conferences 1978 and 1988 on this topic.

No wonder that his support was sought and appreciated when other nations were about to institutionalise their collective memory of public opinion. So ***Per Nielsen*** was one of the first to travel to South Africa to support the development of the South African Data Archive (SADA).

Per Nielsen was a highly appreciated professional and as a person he was a remarkable character. He has lived a colourful live. His joyful vivid open character also brought colour to official gatherings. It was fun to listen to him and rewarding to profit from his frank and honest arguments in serious discussions. It would not be sufficient to state that ***Per Nielsen*** has done his duty for the international scholarly community, it must be added that he also has given so much to his colleagues personally.

Per Nielsen knew that his time was running short to prepare the IASSIST / IFDO conference 1997 in Odense, but he wanted it to take place. His contribution to the development of the international social science infrastructure was gratefully acknowledged in this conference.

Ekkehard Mochmann

Erweiterungen im Datenangebot des Zentralarchivs

In den letzten Monaten ist eine Reihe von neuen Datensätzen archiviert worden. Neben der Archivnummer und dem Studientitel sind die Primärforscher bzw. die Erhebungsinstitute und das Erhebungsjahr aufgeführt. Mit einem Stern * sind diejenigen Datensätze versehen, die in *mehrfachgelochter* Form - *multi punch bzw. column binary* - vorliegen. Weitere Details zu den einzelnen Datensätzen sind auf Anfrage in Form von - Studienbeschreibungen erhältlich.
Der Datenbestands*katalog* enthält sämtliche Studienbeschreibungen und ist mit einem Informationsrückgewinnungssystem auf Diskette erhältlich oder über Internet ansprechbar. Das Datenbestands*verzeichnis* enthält eine Liste der Datensätze und wird auf Anfrage kostenlos zugeschickt. Es ist ebenfalls unter http://www.za.uni-koeln.de abrufbar.

2836 Polish General Social Survey 1995
Institute for Social Studies, University of Warsaw

2842 IAB-Beschäftigtenstichprobe 1975-1990 - Regionalfile[1]
2873 Berufsverläufe bei männlichen Erwerbspersonen (1970)
Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg

2843* Urlaub und Reisen '94
Forschungsgemeinschaft Urlaub und Reisen, Hamburg; GFM-GETAS, Hamburg

2844 Caritas-Armutsuntersuchung
Deutscher Caritasverband, Freiburg; R. Hauser, Johann Wolfgang Goethe-Universität, Frankfurt (Main), 1991

2845 Soll und Haben 4
SPIEGEL-Verlag, Hamburg; Infratest, München, 1995

2846 Meinungsbildner 1995
2847 Ausländer in Deutschland 1994
2848 Ausländer in Deutschland 1995
2849 Ausländer in Deutschland 1996
MARPLAN, Offenbach

1 Vgl. nachfolgenden Beitrag in dieser Ausgabe.

2850* OUTFIT 3
SPIEGEL-Verlag, Hamburg; MARPLAN, Offenbach, 1993

2851 Zukunftserwartungen und Zukunftsverhalten (1990)
2852 Zukunftserwartungen und Zukunftsverhalten (1992)
EMNID, Bielefeld

2857 Jugend in der Republik Polen, in Ost- und Westdeutschland
W. Melzer, Universität Bielefeld/TU Dresden, 1990

2858 Bürgerorientierungen gegenüber den soziopolitischen Vermittlungsstrukturen und dem politischen System der Bundesrepublik (Oktober 1994)
H. Kreikenbom, Jena; GFM-GETAS, Hamburg

2859 Studierende an hessischen Hochschulen zu Fragen der Politik
Institut für Sozialforschung, Johann Wolfgang Goethe-Universität, Frankfurt (Main), 1994

2861 Grundschule und Werbung (Klassenlehrerinnen und Klassenlehrer)
2862 Grundschule und Werbung (Schulleiterinnen und Schulleiter)
Deutsches Jugendinstitut, München, 1995

2870 Methodenstudie des Projekts "Determinanten der Ehescheidung"
H. Esser, Universität Mannheim, 1992

2871* Reiseanalyse 1992
Studienkreis für Tourismus, Starnberg; BASISRESEARCH, Frankfurt (Main); GFM-GETAS, Hamburg

2879 Die Landtagswahlen von 1946 in der Sowjetischen Besatzungszone
J. Falter, Johannes Gutenberg-Universität Mainz

2881 Potsdamer Elitestudie 1995
2882 Bevölkerungsbefragung zur Potsdamer Elitestudie 1995
W. Bürklin, Universität Potsdam; Infratest, München

2895 Lagebericht Mittelstand 1993/94
Westdeutsche Genossenschafts-Zentralbank, Düsseldorf

2913 Landtagswahl in Baden-Württemberg 1996
2914 Landtagswahl in Rheinland-Pfalz 1996
2915 Landtagswahl in Schleswig-Holstein 1996
Forschungsgruppe Wahlen, Mannheim

Ein neuer Datensatz der IAB-Beschäftigtenstichprobe jetzt im Zentralarchiv verfügbar: der Regionaldatenfile

von Anette Haas und Jürgen Hilzendegen[1]

Nachdem in Heft 38 der ZA-Information der Basisfile der IAB-Beschäftigtenstichprobe vorgestellt worden war, steht nun ein weiterer Datensatz aus dem IAB zur Verfügung.

Als Datenquelle für die Arbeitsmarktforschung ist die IAB-Beschäftigtenstichprobe von großer Bedeutung. Der Zeitraum von sechzehn Jahren (1975-1990), die hohen Fallzahlen und die Tagesgenauigkeit der Verlaufsangaben sind Eigenschaften, die besonders für erwerbsbiographische Forschungen unabdingbar sind. Dies zeigt sich auch in der großen Resonanz auf die erste verfügbare Stichprobe - *Basisfile* - (ZA-Nr. 2640), welche seit Februar 1996 über das Zentralarchiv zu beziehen ist.

Auf dem ersten Nutzerworkshop, der in Zusammenarbeit vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB) und dem Zentrum für Europäische Wirtschaftsforschung (ZEW) im Februar 1997 in Mannheim stattfand, lag der Schwerpunkt der bisherigen Forschungsausrichtungen auf Mobilitäts- sowie Lohnanalysen.

Der *Basisfile* enthält 1% der im Untersuchungszeitraum sozialversicherungspflichtig Beschäftigten. Bei den Merkmalen stehen Betriebsinformationen im Vordergrund, während nur eingeschränkte regionale Strukturmerkmale ausgewiesen werden. Dieses Manko des Basisfiles wurde durch Generierung eines eigenen *Regionalfiles* behoben, welcher ab sofort als Quelle zur wissenschaftlichen Nutzung zur Verfügung steht.

Die Gewinnung dieser Daten wurde mit Geldern der GESIS (Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen) sowie des WZB (Wissenschaftszentrum Berlin) finanziert und im IAB umgesetzt. Ziel des Projektes war es, eine neue 1% Stichprobe aus der Historikdatei der Beschäftigtenstatistik zu ziehen, die der Anforderung an möglichst uneingeschränkter Regionalinformation genügt. Bei der Stichprobenziehung wurde darauf geachtet, daß keine Person in beiden Stichproben vorhanden ist.

---

1 Anschrift der Autoren: ***Anette Haas*** (Dipl.-Vw.) ist wissenschaftliche Mitarbeiterin im IAB.
Jürgen Hilzendegen (Dipl.-Sozw.) ist Referent in der Hauptstelle der Bundesanstalt für Arbeit (Tätigkeitsschwerpunkt Statistik). Beide: Regensburger Str. 104, 90327 Nürnberg

Bei der Auswahl der Merkmale und der Stichprobenziehung wurde analog zum Basisfile vorgegangen. Somit stehen für den gleichen Zeitraum von 1975 bis 1990 jetzt zusätzlich Daten über 426.918 sozialversicherungspflichtig beschäftigte Personen zur Verfügung. Die Vorteile liegen wie schon beim Basisfile in den hohen Fallzahlen, den tagesgenauen Angaben und der Validität der Bruttoentgelte begründet[2].

Spezifika des Regionalfiles

Während im Basisfile der Schwerpunkt auf der Ausweisung einer systemfreien Betriebsnummer liegt, aber nur eingeschränkte regionale Strukturmerkmale enthalten sind, basiert der Regionalfile auf dem Konzept der Kreiskennziffer. Diese Schlüsselung ermöglicht Analysen auf administrativer Gebietsebene, wie kreisfreien Städten und Landkreisen, Regierungsbezirken oder Bundesländern.

Zusätzlich besteht die Möglichkeit einer funktionalen Gebietsgliederung, wie z. B. nach Arbeitsmarktregionen oder Verkehrsbezirken, die als Arbeitsgrundlage für räumliche Planung, Forschung und Politik dienen kann (vgl. StBA, 1996, 11, Wirtschaft und Statistik, S. 683-695).

Um jedoch den Anforderungen des Datenschutzes zu genügen, fehlen die Merkmale Staatsangehörigkeit und Betriebsgröße, die nur im Basisfile ausgewiesen sind. In Anlehnung an einen Vorschlag der Bundesforschungsanstalt für Landeskunde und Raumordnung (BfLR) wurden aus Anonymisierungsgründen Kreise, die weniger als 100.000 Einwohner aufweisen, zusammengelegt.

Im *Basisfile* sind folgende Angaben zu einer systemfreien Betriebsnummer verfügbar:

- Betriebsgröße
- Wirtschaftszweig
- Anzahl der sozialversicherungspflichtig Beschäftigten ohne abgeschlossene Berufsausbildung, mit abgeschlossener Berufsausbildung und mit Fachhochschul- bzw. Universitätsabschluß.

Im *Regionalfile* ist nur im TDA- File ein Betriebsnummernzähler vorhanden. Dieser gibt an, im wievielten Betrieb sich eine Person - ab dem 1.1.1975 gerechnet - befindet. Mit diesem Zähler ist gewährleistet, daß im Regionalfile betriebliche Mobilität (Betriebswechsel) untersucht werden kann.

2 In Zusammenarbeit mit der Universität Rostock wird zur Zeit ein weiteres Projekt durchgeführt, welches die DFG mitfinanziert. Die neue Stichprobenziehung umfaßt die Zeiträume 1975-1995 für Westdeutschland und 1992-1995 für Ostdeutschland.

In der folgenden Übersicht werden die Merkmale und Anonymisierungsmaßnahmen auf der Personenebene für beide Files gegenübergestellt.

Übersicht 1: Vergleich zwischen verfügbaren Merkmalen und angewendeten Anonymisierungsmaßnahmen von Basis- bzw. Regionalfile

Merkmale	Verfügbarkeit bzw. angewandte Anonymisierungsmaßnahme im	
	Basisfile	Regionalfile
Geschlecht	im Original vorhanden	
Familienstand	im Original vorhanden	
Ausbildung (B2-Schlüssel)	im Original vorhanden	
Stellung im Beruf (B1-Schlüssel)	im Original vorhanden	
Geburtsjahr	Aggregation - Bei Eintritt in die Datei unterhalb von 16 Jahren bzw. bei Austritt über 66 Jahren	
Beginn und Ende der Beschäftigung	Längsschnittverschiebung des gesamten Erwerbsverlaufs einer Person um eine Zufallskonstante	
Beendigungsgrund	Aggregation auf 8 Ausprägungen (z.B. Sperrzeiten, Anspruch erschöpft)	
Beruf (ausgeübte Tätigkeit)	Aggregation - von 334 Berufen behalten noch 234 ihre urspr. Klassifikation, die verbleibenden 100 werden zu 41 Berufsgruppen zusammengefaßt	Aggregation - 334 Berufe werden zu 119 Berufen bzw. Berufsgruppen zusammengefaßt
Sozialversicherungspfl. Bruttoentgelt	Rundung auf DM-Betrag	
Beginn und Ende von Leistung	Längsschnittverschiebung des gesamten Erwerbsverlaufs einer Person um eine Zufallskonstante	
Leistungsart	Aggregation auf ALG, ALHI, UHG*	
Staatsangehörigkeit	Aggregation auf 9 Nationalitäten und 7 Nationalitätengruppen	fehlt
Rentenversicherungsträger	im Original vorhanden	
Kreiskennziffer	nicht vorhanden	Zusammenfassung von Kreisen mit Einwohnern < 100.000.

* ALG: Arbeitslosengeld, ALHI: Arbeitslosenhilfe, UHG: Unterhaltsgeld

Hinweise für Nutzer zum Bezug der Daten über das Zentralarchiv

Die anonymisierte Regionalstichprobe, mit dem Titel *IAB-Beschäftigtenstichprobe 1975 - 1995 - Regionalfile*, kann über das Zentralarchiv unter der ZA-Nr. 2842 von den Nutzern angefordert werden. Die Antragsmodalitäten entsprechen dem Verfahren bei der Nutzung des Basisfiles (ZA-Nr. 2640). Als Unterlagen werden vom Antragsteller eine Projektbeschreibung mit Titel, die Angabe der Projektlaufzeit und eine Auflistung aller Projektmitarbeiter gefordert. Nach der Prüfung durch das Zentralarchiv werden die Anträge an das IAB zur formellen Genehmigung weitergeleitet. Der danach abzuschließende Nutzervertrag beinhaltet Projekttitel, Laufzeit und die Verpflichtung zur Rückgabe bzw. Löschung der Daten am Ende der Projektlaufzeit. Eine parallele Nutzung von Basis- und Regionalstichprobe ist möglich, allerdings dürfen beide Files nicht zusammengeführt werden, und es ist für jede Stichprobe ein gesonderter Projektantrag nötig.

Aufbau des Datensatzes und technische Aspekte

Dem Nutzer werden auf CD-ROM zwei sogenannte Rohdatensätze zur Verfügung gestellt. In der Beschäftigtendatei sind die Daten über die Versicherungsnehmer gespeichert. Dabei handelt es sich um eine Rechteckdatei, in der für jede Versicherungsnummer pro meldepflichtiges Ereignis ein Datensatz (record) enthalten ist. Informationen über Leistungen der Bundesanstalt für Arbeit, wie Arbeitslosengeld bzw. - hilfe und Unterhaltsgeld, sind in der Leistungsempfängerdatei enthalten.

Wie schon beim Basisfile konnte das Datenmanagement durch Komprimieren (ZOO-Archiv) erleichtert werden. Prof. Dr. ***Götz Rohwer*** (Max-Planck-Institut für Bildungsforschung, Berlin) hat eine spezielle Gesamtdatei erstellt und diverse Modifikationen der Ursprungsdaten vorgenommen. Die mit ZOO gepackten Daten können mit Hilfe des Programmes TDA 6.02 analysiert werden. Dieses wird separat auf einer Diskette mitgeliefert bzw. ist über das Internet[3] verfügbar.

Für die Datenanalyse empfiehlt sich die Dokumentation zur IAB-Beschäftigtenstichprobe 1975-1990, die als Beitragsband (BeitrAB 197, 1996) zur Arbeitsmarkt- und Berufsforschung erschienen ist. Die dort enthaltenen Beispielprogramme geben einen Einblick in die vielfältigen Analysemöglichkeiten.

3 Adresse: ftp: MTGOAT.MPIB-Berlin.MPG.DE (192.109.48.134), Login: TDA, Password: distribution1, Email für Anfragen etc.: rohwer@mpib-berlin.mpg.de

Übersicht 2: Dimension der au der CD-ROM enthaltenen Datensätze:

I) Rohdaten

1. Beschäftigtendatei (BSTREG.SDF)	
Datensatzlänge:	43
Personen (Versicherungsnr.):	426.918
Datensätze:	4.792.210
Bytes (nicht komprimiert):	210.857.240

2. Leistungsempfängerdatei (LEDREG.SDF)	
Datensatzlänge:	25
Personen (Versicherungsnr.):	140.778
Datensätze:	466.968
Bytes (nicht komprimiert):	12.141.161

II) Überarbeitete Datei von ***Götz Rohwer***

3. Gesamtdatei (BSTREG.DAT)	
Datensatzlänge:	52
Personen (Versicherungsnr.):	426.914
Datensätze:	5.776.242
Bytes (nicht komprimiert):	306.140.826
Bytes (komprimiert)	58.444.998

Literatur

Bender, Stefan ; Hilzendegen, Jürgen ; Rohwer, Götz; Rudolph, Helmut 1996:
Die IAB-Beschäftigtenstichprobe 1975-1990, BeitrAB 197, IAB, Nürnberg.

Haas Anette ; Hilzendegen, Jürgen: 1997:
IAB Info zur Beschäftigtenstichprobe, Nr. 5, Nürnberg.

Hilzendegen, Jürgen 1995:
Anonymisierung der Regionalstichprobe, Diskussionspapier, IAB, Nürnberg.

Hilzendegen, Jürgen 1996:
Datensatzbeschreibung und Codeplan (Kurzfassung) (Regionalfile), Manuskript, Nürnberg.

Zentralarchiv für Empirische Sozialforschung Universität zu Köln 1996:
IAB-Beschäftigtenstichprobe, ZA-Information 38, S. 15-19.

Arbeitsmarkt-Monitor für die neuen Bundesländer: ZA-Nr. 2142

Erhebungszeitraum: November 1990 bis November 1994

Primärforscher: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg

Datenerhebung: Infratest Sozialforschung, München; Infratest Burke, Berlin

Inhalt:
Achtwellige Panel-Befragung zur Entwicklung der Arbeitsmarktlage in den fünf neuen Bundesländern. Beschreibung der Arbeitssituation im Betrieb und des Arbeitsmarkts in der Region.
Themen: In allen Wellen wurde gefragt: Beurteilung der eigenen finanziellen Lage; derzeitige Arbeitsuche; Umkreis der Arbeitsuche in Kilometern; beim Arbeitsamt als arbeitslos oder Arbeitsuchende gemeldet; Zeitpunkt des Ausscheidens aus dem Betrieb und Grund für dieses Ausscheiden; Erwerbsstatus; Dauer der Beschäftigung in der derzeitigen Stelle; eigene Stellung innerhalb des Betriebes; Wirtschaftszweig des Betriebes; Zugehörigkeit zum öffentlichen Dienst; Betriebsgröße; Beschäftigung in Westdeutschland oder Westberlin; Teilnahme an berufsqualifizierenden Maßnahmen; Unterhaltsgeld durch das Arbeitsamt während der Maßnahme; Höhe des persönlichen Einkommens im Monat; Geburtsjahr; Erhebungsdatum.
In mindestens einer weiteren Erhebung wurde zusätzlich gefragt:
Derzeitige Lebenssituation: Einschätzung der Veränderung der wirtschaftlichen Lage gegenüber dem Vorjahr; allgemeine wirtschaftliche Lage der Gegend.
Erfahrungen mit Arbeitslosigkeit: Registrierung beim Arbeitsamt; einmalige oder mehrmalige Arbeitslosigkeit; Arbeitslosigkeitsdauer; Erhalt von Leistungen des Arbeitsamts; Erfahrungen mit dem Arbeitsamt.
Arbeitsuche: Suche nach Dauerstellung oder nach kurzfristiger Beschäftigung; Bereitschaft zur Annahme einer befristeten Stelle; Suche nach Vollzeit- oder Teilzeitstelle; Bewertung der Chancen auf einen Arbeitsplatz; berufsbezogene Arbeitsuche.
Ausbildungswünsche: Wunsch nach Berufsausbildung oder Studium; angestrebte Ausbildung; Aussichten auf einen Ausbildungsplatz oder eine Lehrstelle; Grund für die Auflösung eines Ausbildungsvertrages; Einschalten des Arbeitsamtes bei der Ausbildungsplatzsuche; Stellenangebote vom Arbeitsamt; Aussicht auf Studienbeginn; Alternative zur geplanten

Ausbildung; Interesse an einem Arbeitsplatz in Westdeutschland; präferierter Ausbildungsbereich.
Arbeitsbeschaffungsmaßnahmen: Vermittlung einer Arbeitsbeschaffungsmaßnahme durch das Arbeitsamt; Zeitpunkt dieser Vermittlung; Beschäftigung in einer Arbeitsbeschaffungsmaßnahme zum Befragungszeitpunkt; vorzeitiges Ausscheiden aus der Arbeitsbeschaffungsmaßnahme; Ende der Arbeitsbeschaffungsmaßnahme; Anschlußbeschäftigung; eigenes Bemühen um eine reguläre Arbeit während der Arbeitsbeschaffungsmaßnahme; Information des Arbeitsamtes über eine Beschäftigung außerhalb einer Arbeitsbeschaffungsmaßnahme.
Beruf und Familie: Zusammenleben mit Partner; Erwerbsstatus des Partners; Ausmaß der Erwerbstätigkeit des Partners; bevorzugte Möglichkeit des Geldverdienens.
Kinderbetreuung: Kinder unter 16 Jahren im Haushalt; Geburtsjahr des jüngsten Kindes; Betreuung der Kinder tagsüber; Möglichkeit einer Berufstätigkeit auch ohne Kinderkrippe, Kindergarten oder Hort; ganztägige Kinderbetreuungsmöglichkeit im Wohnviertel.
Die Situation in den letzten 6 Monaten: Ausscheiden aus einem Betrieb und Grund für dieses Ausscheiden.
Situation zum Befragungszeitpunkt: Wunsch nach einer neuen Berufstätigkeit; Gründe für die Nichtrückkehr in das Berufsleben.
Berufliche Tätigkeit zum Befragungszeitpunkt: Berufstätigkeit im gleichen Betrieb wie vor einem halben Jahr; Maßnahme zum Erhalt der Beschäftigung; befristetes Arbeitsverhältnis; Entsprechung von beruflicher Tätigkeit und erlerntem Beruf; erforderliche Ausbildung für die derzeitige Tätigkeit; Bewertung der technischen Ausstattung des Betriebes; Nutzung von computergesteuerten Arbeitsmitteln oder Anlagen; Innehaben einer Führungsposition; Kurzarbeit im Betrieb und eigene Betroffenheit; Aufstockung des Kurzarbeitgeldes des Arbeitsamtes durch den Betrieb; Ausmaß der Erwerbstätigkeit (Voll- bzw. Teilzeitstelle); Arbeitszeit pro Woche; präferierte Wochenstundenzahl; tatsächliche Wochenarbeitszeit; Gründe für Überstunden bzw. für Kurzarbeit; eigene Betroffenheit von der sogenannten Warteschleife; Entfernung zur Arbeitsstelle; Ausüben einer nebenberuflichen Tätigkeit; Art der Nebentätigkeit; Ausmaß der Nebentätigkeit; Dauer der Berufstätigkeit im letzten Jahr in Monaten.
Berufliche Veränderungen: Berufswechsel im letzten Jahr; Veränderung des Arbeitsplatzrisikos, der körperlichen Belastung, des Stresses und der Hektik seit dem letzten Jahr; Arbeitseinkommen; Änderung des Arbeitsplatzes; Arbeiten an neuen Maschinen bzw. Anlagen; Herstellen eines anderen Produktes; neue Kenntnisanforderungen; Reduzierung bzw. Wegfall der Arbeit; Zufriedenheit mit dem Arbeitsverdienst, mit der Möglichkeit, die eigenen beruflichen Kenntnisse und Fertigkeiten anzuwenden, mit den Aufstiegschancen, mit der Zusammenarbeit mit Kollegen, mit der Arbeitszeit, mit dem Arbeitsweg, mit dem Verhältnis zu den Vorgesetzten, Arbeitszufriedenheit (Skala); Einschätzung des Arbeitstempos im Betrieb.

Der Betrieb: Herstellungsprogramm des Betriebes; Entlassungen innerhalb des Betriebes; erwartete Veränderung der Beschäftigtenzahl in den nächsten 12 Monaten; Existenz des Betriebs vor einem Jahr; Verwaltung des Betriebs durch die Treuhandanstalt; Arbeiten in einer Beschäftigungs- bzw. Sanierungsgesellschaft; Grund für Arbeitsverhältnis im Westen; gewünschte Arbeitsdauer im Westen; Zugehörigkeit des Betriebs zu einem größeren Unternehmen; Personen aus dem Westen im Beschäftigungsbetrieb; Wohnort und gemeldeter Hauptwohnsitz; Art der Beendigung des Beschäftigungsverhältnisses im Westen; Grund für die Beendigung der Beschäftigung im Westen.
Die absehbare Zukunft: Geplante oder erwartete Veränderungen der beruflichen Situation.
Berufliche Qualifizierung: Qualifizierungsmaßnahme zum Erhebungszeitpunkt; Stundenaufwand pro Woche; Ziel der Qualifizierung; Dauer der Maßnahme; Veränderung der beruflichen Situation durch berufliche Weiterbildung; Interesse an Kursen zur beruflichen Qualifizierung; Ziele der Maßnahme; ausreichende Informiertheit über Qualifizierungsmaßnahmen.
Finanzielle Situation: Prozentualer Beitrag zum Haushaltseinkommen; Anzahl der Personen im Haushalt; Anzahl der Personen zwischen 16 und 65 Jahren im Haushalt; Berufstätigenzahl, Arbeitslosenzahl bzw. Anzahl der Rentner im Haushalt; Personen im Haushalt mit eigenem Einkommen; Höhe des monatlichen Nettohaushaltseinkommens; innerhäusliche Eigenproduktion.
Gesundheit: Selbsteinschätzung des Gesundheitszustandes.
Erwartungen: Erwartete wirtschaftliche Lage der Region in einem Jahr; eigene berufliche Zukunft (Skalometer).
Angaben zur Person: Höchster Schulabschluß; beruflicher Abschluß.

Grundgesamtheit und Auswahl: Untersuchungsgebiet: Mecklenburg-Vorpommern, Brandenburg, Sachsen, Sachsen-Anhalt, Thüringen.
Einfache Zufallsauswahl von Erwerbspersonen aus Melderegistern, die im September 1990 ihren Wohnsitz in der damaligen DDR hatten. Einbezogen wurden die Geburtsjahrgänge 1926 bis 1974. Ab Welle 4 wurde der Jahrgang 1975, ab Welle 6 der Jahrgang 1976 mit einbezogen.
Der Datensatz enthält die Informationen aus einer achtwelligen Panel-Erhebung. Die Bruttostichprobe umfaßte 15000 Personen. Die erste Welle wurde im November 1990 durchgeführt (10751 Befragte), die zweite im März 1991 (7929 Befragte), die dritte im Juli 1991 (7300 Befragte), die vierte im November 1991 (7956 Befragte), die fünfte im Mai 1992 (10956 Befragte), die sechste im November 1992 (9763 Befragte), die siebte im November 1993 (8351 Befragte), die achte im November 1994 (7549 Befragte). Bei den Angaben in der Klammer handelt es sich um Nettostichproben. Um Ausfälle auszugleichen, wurde ab der vierten Welle zusätzlich der Geburtsjahrgang 1975 und ab der sechsten Welle der Geburtsjahrgang 1976 aufgenommen. In der fünften Welle wurde die Stichprobe um brutto ca. 7000 Personen aufgestockt.

Erhebungsverfahren: Postalische Befragung

Datensatz: Anzahl der Einheiten: 14993
Anzahl der Variablen: 1229

Zugangsklasse: C

Veröffentlichung:

Bielenski, Harald; von Rosenbladt, Bernhard:
Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 11/90. Beiträge zur Arbeitsmarkt- und Berufsforschung. Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg 1991.

Bielenski, Harald; Enderle, Jovita; von Rosenbladt, Bernhard:
Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 3/91.
Beiträge zur Arbeitsmarkt- und Berufsforschung. Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg 1991.

Bielenski, Harald; Enderle, Jovita; von Rosenbladt, Bernhard:
Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 7/91. Beiträge zur Arbeitsmarkt- und Berufsforschung. Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg 1991.

o.V.: Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 11/91.Im Auftrag der Bundesanstalt für Arbeit, München, März 1992.

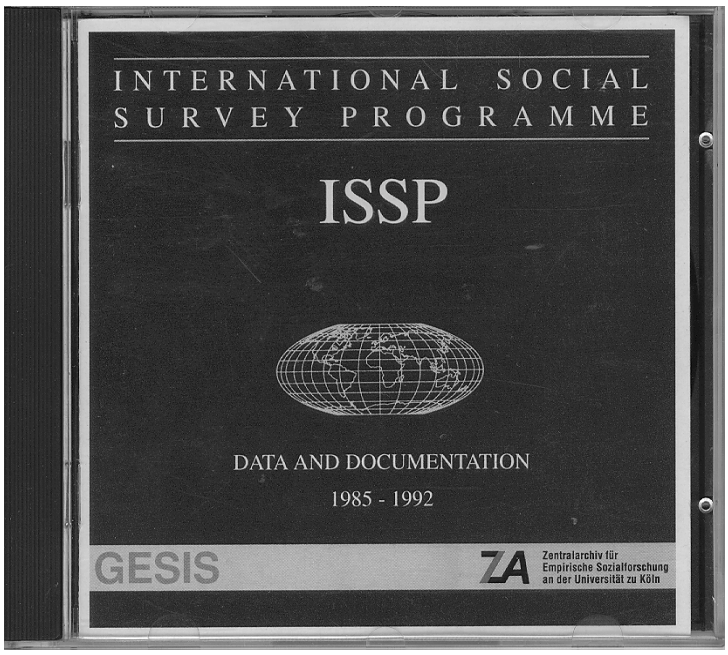
Enderle, Jovita; Bielenski, Harald; von Rosenbladt, Bernhard:
Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 5/92. Beiträge zur Arbeitsmarkt- und Berufsforschung. Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit, Nürnberg 1992.

o.V.: Arbeitsmarkt-Monitor für die neuen Bundesländer. Umfrage 11/92. Im Auftrag der Bundesanstalt für Arbeit. München, Mai 1993.

o.V.: Arbeitsmarkt-Monitor für die neuen Bundesländer, Umfrage 11/93. Im Auftrag der Bundesanstalt für Arbeit. München, April 1994.

ISSP CD-ROM 1985-1992

In der ZA-Information 36, Mai 1995, wurde die erste vom Zentralarchiv erstellte CD-ROM vorgestellt. Sie beinhaltet Daten und Dokumentation der Studien, die im International Social Survey Programme (ISSP) zwischen 1985 und 1992 erhoben wurden. Es handelt sich dabei um neun international vergleichende Datensätze aus bis zu siebzehn Ländern inklusive der maschinenlesbaren und recherchierbaren Codebücher sowie der Originalfragebögen aller beteiligten Länder als gescannte Image-Files. Die CD-ROM ist für DM 50,-- beim ZA erhältlich.



Im Frühjahr 1995, als die Planungen in die Endphase traten und die endgültigen Dateien mit der zugehörigen Software für die CD-ROM vorbereitet vorlagen, war zu entscheiden, in welcher Auflage produziert werden sollte. CD-Writer, die inzwischen bereits für unter DM 1000,-- im Handel erhältlich sind, waren vor zwei Jahren für einen erschwinglichen Preis noch nicht zu haben. Also mußte die CD bei einem Anbieter extern hergestellt werden. Das Zentralarchiv entschied sich dafür, die ISSP CD-ROM in einer Auflage von 500 Exemplaren zu bestellen, was zunächst als durchaus optimistische Einschätzung der erwarteten Nachfrage eingestuft werden konnte. Zwei Jahre später wissen wir, daß der Optimismus gerechtfertigt war. 90 % der Auflage ist bisher weltweit ausgeliefert worden und die Resonanz war durchweg positiv.

Bericht über das 26. Frühjahrsseminar des Zentralarchivs für Empirische Sozialforschung vom 17. Februar bis 7. März 1997

Analyse kategorialer Daten

Wie in den vergangenen Frühjahrsseminaren stand auch in diesem Jahr das Vermitteln von multivariaten Analyseverfahren im Mittelpunkt der Veranstaltungen mit drei Modulen, die weitgehend aufeinander aufgebaut waren: Einführung in die Logit- und Probitanalyse, log-lineare Modelle mit latenten Variablen und neuere Entwicklungen in der Analyse kategorialer Daten. Eine Gemeinsamkeit der in diesem Jahr behandelten Verfahren bestand darin, daß bei allen mit kategorialen Variablen gearbeitet wurde.

Dem Umstand, daß in der empirischen Sozialforschung die meisten der per Fragebogen erhobenen Daten kategoriales Skalenniveau haben, wird seit einigen Jahren auch in der statistischen Auswertung mehr und mehr Rechnung getragen. Anstatt einfacher multipler Regressionen werden logistische Regressionen gerechnet, wenn die zu erklärende Variable kategorial ist. Wenn alle Variablen kategorial skaliert sind, werden zunehmend log-lineare Modelle eingesetzt, um das Skalenniveau der Daten angemessen zu berücksichtigen. Auf der Ebene der Methodenforschung wurden in den letzten Jahren insbesondere im Bereich der log-linearen Modelle große Fortschritte gemacht, sei es durch den Einbezug von latenten Variablen oder durch das Berücksichtigen der ordinalen Ordnung einzelner Merkmale, wie z.B. bei den log-multiplikativen Modellen (letztere werden in der Literatur auch als RC-Modelle oder log-bilineare Modelle bezeichnet). Sowohl diese neueren Entwicklungen als auch neue Entwicklungen im Bereich der logistischen Regression wurden im diesjährigen Frühjahrsseminar vorgestellt und deren Anwendung anhand einer Vielzahl von sozialwissenschaftlichen Fragestellungen ausführlich diskutiert. Als Datensätze wurden sowohl der ALLBUS als auch der Schweizer Umweltsurvey verwendet.

In der ersten Woche wurden die Logit- und Probitmodelle behandelt; derartige Analyseverfahren werden in der aktuellen Forschungspraxis regelmäßig dann eingesetzt, wenn die zu erklärende Variable dichotom oder nominal skaliert ist - so z. B. bei der Frage nach Prädiktoren für die Wahlbeteiligung/Nichtbeteiligung oder zur Vorhersage bestimmter Verhaltensweisen. In der ersten Woche wurden folgende Gebiete behandelt: OLS-Regression, Schätzung des logistischen Regressionsmodells, Schätzmethoden und Modelltests, Probit-

modell und Parameterschätzung, Discrete Choice Modelle, Logit-/Probitmodell für polytome abhängige Variablen, konditionales Logitmodell, Modellerweiterungen. Als Programm wurde LIMDEP verwendet, eine ausführliche Einführung in dieses Analysepaket wurde im Rahmen der Vorlesungen gegeben. Die Dozenten waren Prof. Dr. ***Andreas Diekmann*** und Dr. ***Axel Franzen*** (beide Universität Bern).

In der zweiten Woche wurden log-lineare Modelle und ihre Erweiterung auf latente Variablen behandelt. Mit Hilfe von log-linearen Modellen kann u.a. bestimmt werden, wie Effekte in zwei- und mehrdimensionalen Kreuztabellen erklärt werden können - welche Haupt- bzw. Interaktionseffekte innerhalb der Tabelle signifikant sind und welchen Anteil die einzelnen Effekte an der Gesamtvariation der Daten haben. Anwendung findet dieses Verfahren überall dort, wo Kreuztabellen eingehender analysiert werden sollen, so. z. B. in der Mobilitätsforschung. Da die Schätzung von log-linearen Modellen bezüglich der Bestimmung und Interpretation von Fehlertermen problematisch ist, war eine logische Erweiterung die Einbeziehung von latenten Variablen. In der zweiten Woche wurden folgende Gebiete behandelt: log-linear model for two- and three-way tables, non-saturated hierarchical log-linear model, general log-linear model, logit model, log-linear path models, latent class model, restricted latent class model, latent trait model, latent class/latent trait models with covariates,
log-linear path model with latent variables. Dozent der zweiten Woche war Dr. ***Jeroen Vermunt*** (Tilburg University). Als Programm wurde LEM verwendet, welches vom Dozenten entwickelt wurde und mit dem die oben genannten Modelle und viele andere wie z.B. die Korrespondenzanalyse sehr effizient geschätzt werden können.

Die dritte Woche war den neueren Modellen zur Analyse von kategorialen Daten gewidmet. In dieser Woche wurden auch Querverbindungen dieser Modelle hin zu anderen Verfahren zur Analyse kategorialer Daten, z. B. der Korrespondenzanalyse, diskutiert. Eingesetzt werden diese Verfahren u.a., um in der Tabellenanalyse zu testen, ob einzelne Variablen ordinal skaliert sind und wie groß dann die Abstände zwischen den einzelnen Merkmalsausprägungen sind. Abgesehen von unterschiedlichen theoretischen Fragestellungen und unterschiedlichen Annahmen über die Struktur der Daten sind die Anwendungsgebiete dieser Verfahren letztlich gleich denen der zweiten Woche. In der dritten Woche wurden folgende Themen behandelt: Probabilistic vs. algebraic approaches to the analysis of nominal and ordinal data, canonical correlation and correspondence analysis, restricted correspondence analysis - one and multiple groups, linear-by-linear interaction models and the ANOAS approach, conditional and unconditional approaches, latent variable approach, RC models, multiple group RC(M) models, a latent variable framework for the analysis of categorical data, factor-analytic models for nominal and ordinal data. Dozent der dritten Woche war Prof. Dr. ***Ulf Böckenholt*** (University of Illinois at Urbana Champaign), als Programme wurden LEM und CDAS verwendet.

Zusätzlich zu den Vorlesungen wurden Informationen über die Dienstleistungen des Zentralarchivs, über GESIS und die internationale sozialwissenschaftliche Infrastruktur angeboten; Referenten waren ***Ekkehard Mochmann*** und ***Erwin Rose***. Die praktische Umsetzung des Lehrstoffs erfolgte in Arbeitsgruppen, die von den Zentralarchiv-Mitarbeitern ***Jörg Blasius***, ***Lorenz Gräf*** und ***Harald Rohlinger*** geleitet wurden.

Für das 27. Frühjahrsseminar 1998 (2. März bis 20. März) ist als Thema die Aufbereitung und Analyse komplexer Datensätze vorgesehen, als Wochenthemen werden voraussichtlich der "Aufbau von Datenbanken", "Probleme fehlender Werte und die Auswirkungen des Einsatzes unterschiedlicher Erhebungsverfahren" sowie die "Mehrebenenanalyse" behandelt. Interessenten können sich ab sofort beim Zentralarchiv anmelden. Eine nähere Beschreibung des Frühjahrsseminars und das vorläufige Programm wird in der Herbstausgabe der ZA-Information enthalten sein, die aktuellen Informationen stehen auf dem WWW-Server des Zentralarchivs zur Verfügung (http://www.za.uni-koeln.de/events/).

Das ZA-Frühjahrsseminar ist als Bildungsurlaub anerkannt, entsprechende "Mitteilungen an den Arbeitgeber" werden auf Anfrage zugeschickt.
Jörg Blasius

Einführung in das Internet: Arbeiten - Forschen - Experimentieren

Surfen. GESIS-Workshop 18. bis 20. Juni 1997

Bibliotheksrecherche vom PC-Arbeitsplatz aus, Teilhabe am schnellen Informationsaustausch und an Diskussionsgruppen; weltweites Suchen und Auffinden von Fachinformationen und eigenständiges Bereitstellen von Fachinformationen für die Profession im Netz. Das sind einige Möglichkeiten, wie heute der Zugang zum Internet genutzt werden kann: das Internet als weiteres Informationsmedium zum Arbeiten, Forschen und Experimentieren. Der Workshop richtet sich an Sozialwissenschaftlerinnen und Sozialwissenschaftler, die noch keine Erfahrungen mit diesem neuen Medium machen konnten bzw. dieses Medium künftig besser nutzen möchten. Während des Workshops werden sie beispielhaft in inhaltliche Anwendungen für die Profession eingeführt, wobei auch das Angebot der GESIS erläutert wird. Es wird auf Suchstrategien und die Nutzung von Internet-Suchmaschinen eingegangen. Daneben wird gezeigt, welche Software-Hilfsmittel (Email, WWW, News, FTP, Telnet) im Internet nutzbringend angewendet werden können. Der Workshop wird durchgeführt von ***Heiner Ritter*** (ZUMA) in Zusammenarbeit mit ***Lorenz Gräf*** (ZA), ***Uwe Jensen*** (ZA) und ***H. Peter Ohly*** (IZ) im Rahmen des GESIS-Internet-Gemeinschaftsprojektes. Für die praktischen Übungen stehen PCs zur Verfügung. Die Teilnehmerzahl ist auf 20 Personen begrenzt. Für die Teilnahme wird ein Beitrag von 100 Mark erhoben. Interessenten werden gebeten, sich beim Tagungssekretariat von ZUMA anzumelden (Email-Adresse: workshop@zuma-mannheim.de).

Rainer Metz erhielt Stinnes Award

Für herausragende wissenschaftliche Arbeiten auf den Gebieten Handel, Verkehr, Wirtschafts- und Stinnes-Unternehmensgeschichte vergibt die Stinnes-Stiftung jedes Jahr den Stinnes Award in Höhe von insgesamt DM 30.000,-- . Im Jahr 1996 wurde die Habilitationsschrift von ***Rainer Metz***: "Stochastische Trends und langfristige Wachstumsschwankungen: Neue Forschungsansätze und ihre theoretische und empirische Relevanz für die Wirtschaftsgeschichte" mit dem 1. Preis ausgezeichnet. Der Autor ist Bereichsleiter am Zentralarchiv, Abteilung Zentrum für Historische Sozialforschung (ZHSF), und Privatdozent für Wirtschaftsgeschichte und Methoden der empirischen Wirtschaftsforschung an der Universität St. Gallen.

Bei der systematischen Analyse des Wirtschaftsablaufs richtete sich das Bemühen der empirischen Konjunkturforschung von Anfang an auf die Entdeckung von Regelmäßigkeiten, mit denen man hoffte, die zukünftige Entwicklung prognostizieren zu können. Leider waren derartige Bemühungen nicht immer von Erfolg gekrönt. Häufig machten Unregelmäßigkeiten und historisch bedingte Zufälle den Prognostikern einen Strich durch die Rechnung. Seit einigen Jahren wird nun der "Zufall" auch bei der Analyse wirtschaftlicher Langfristent-wicklungen systematisch berücksichtigt, und zwar in Form sog. stochastischer Trendmodelle, die verstärkt auch bei der Finanzmarktanalyse und -prognose eingesetzt werden. Stochastische Trendmodelle stehen im Mittelpunkt der Habilitationsschrift von ***Rainer Metz***.

Dabei wird in mehrfacher Hinsicht methodisches und theoretisches Neuland betreten. Es ist die erste Untersuchung dieser Art, die unter Berücksichtigung dieser neu entwickelten ökonometrischen Modelle den Interpretationswert der traditionellen Erklärungshypothesen zur wirtschaftlichen Langfristentwicklung systematisch analysiert. Das Besondere dabei ist, daß der Zufall nicht, wie bislang üblich, als irreguläre Komponente aus der Betrachtung ausgeschlossen, sondern explizit in die Analyse einbezogen wird.

Neben einer ausführlichen Darstellung der Identifikation und Schätzung zufallsbedingter Entwicklungsprozesse, werden für die Reihe des deutschen Bruttoinlandsprodukts (BIP) von 1850-1990 folgende Fragen untersucht: Wie beeinflussen "große" und "kleine" Zufälle die langfristige Entwicklung? Sind Wachstumsschwankungen allein auf historisch bedingte Zufälle zurückzuführen? Sind die "Langen Wellen" wirtschaftlichen Wachstums das Produkt von exogen bedingten Einflüssen? Wie verläuft "störungsfreies" wirtschaftliches Wachstum?

Die Analyse ergibt eine stark durch Irregularitäten geprägte Kriegs- und Zwischenkriegszeit. Das "Wirtschaftswunder" erweist sich als eine historisch singuläre Rekonstruktionsphase. "Lange Wellen" mit einer Periodendauer von 30 bis 60 Jahren haben für das deutsche BIP keine Bedeutung. Die bislang für diese Reihe diskutierten ***Kondratieff***zyklen sind primär auf Zufallseinflüsse in der Zeit von 1914 bis 1960 zurückzuführen. Allerdings zeigt die "störungsfreie" Wachstumsrate von 1914-1950 einen von deutlichen Schwankungen geprägten nachhaltigen Anstieg, erreicht aber um 1980 wieder eine Höhe, wie sie für das Ende des 19. Jahrhunderts typisch gewesen ist. Das mag ein Hinweis darauf sein, daß auch in Zukunft mit einem eher bescheidenen Wachstum zu rechnen ist.

Die Arbeit steht in der Tradition einer aus den USA kommenden Forschungsrichtung, die man als Cliometrie bezeichnet, und die in Deutschland bislang kaum Fuß gefaßt hat. Der Preisträger wünscht sich, daß sich daran in Zukunft etwas ändern möge, besteht die Intention der Cliometrie doch auch darin, durch die Verbindung von Geschichte und Ökonomie zur Klärung von Gegenwartsfragen beizutragen. Erste Signale sind gesetzt. Im Jahr 1993 haben die Amerikaner ***Douglass North*** und ***Robert Fogel*** den Nobelpreis für Wirtschaftswissenschaften für ihre Verdienste um eben diese Forschungsrichtung erhalten.

Conference on "Empirical Investigation of Social Space"

The September 1997 conference announced in ZA-Information 38 (May 1996) has to be postponed because some of our key lecturers unexpectedly are not able to attend the conference as originally planned. We will inform about the new date via ZA-Information as well as via the WWW (http://www.za.uni-koeln.de/events).

The ICORE-Collection of Election Studies to National Parliaments in Europe

von Ekkehard Mochmann

Election research in Europe has proven quite successful in creating comprehensive databases on a national level. Most of the major election studies were processed and documented by the national social science data archives to facilitate further analysis. With respect to coverage across time and nations they belong to the most systematic and best documented collections of representative sample surveys. This relative data wealth for each nation was hardly visible and was not well reflected in research from a European perspective. In fact, for a number of studies data were not accessible and written documentation was only available in the original language for the respective country. Inspite of growing interest a systematic program for cross national comparative research so far did not exist. In 1989 the International Committee for Research into Elections and Representative Democracy (ICORE) was established with the objective to promote cross-national research into electoral behaviour and representative democracy. Its membership consisted of the directors of established surveys of national electorates in Europe.

To improve this situation ICORE decided to work towards the integration of the European election database by :

- supporting the translation of original questionnaires and documentation into English, where relevant studies of elections to national parliaments are not yet available in this form,
- the creation of an electronically searchable database of study descriptions for these election studies,
- the creation of an electronically searchable database of publications emanating from each national election study,
- the creation of an electronically searchable database of questions asked in the national election studies,
- the collection of all of the European national Election studies in a central archive.

In co-operation with the principle investigators and with support of the Council of European Social Science Data Archives (CESSDA) an initiative was started to make all election studies to national parliaments with the relevant material required for secondary analyses available in the Central Archive for Empirical Social Research in Cologne.

The election studies available at the Zentralarchiv so far have been documented in ***Mochmann***, ***Oedegaard***, ***Mauer*** (1997): "ICORE Inventory of National Election Studies: Belgium, Denmark, France, Germany, Great Britain, Hungary, The Netherlands, Norway and Sweden". The inventory gives an overview of all election dates and election studies carried out since 1945 and also includes standardised English study descriptions for almost all studies. Furthermore, detailed information is given about the studies which are available in the Zentralarchiv. The inventory informs about the format and the language of data sets, codebooks, questionnaires and study descriptions. At present 118 national election studies from the 9 countries, covered by the ICORE collection, are available for analysis at the Zentralarchiv.

This inventory of national election studies in Europe is a tangible result of the co-operation between principal investigators, research institutes, statistical offices and the social science data services, all working together in the effort of integrating the European data base. We are very grateful to all ICORE- and CESSDA- members, who provided us with the measurement instruments, documentation and data, that made it possible to compile this guide to resources for European election research. With respect to the accessibility of election studies within the ICORE-project national access regulations will be applied also in the Zentralarchiv, to the extent required by either principal investigators or national archives. In agreement with the donors this collection of data sets will be available for analysis within the Zentralarchiv only. It will not be distributed to external users.

European Funds Available to Participate in the Training and Mobility of Researchers (TMR) - Large Scale Facilities (LSF) Activity at the Zentralarchiv

TMR-LSF funds of the European Union are available for supporting access to large comparative databases in the Social Sciences at the University of Cologne and the University of Essex. The Large Scale Facility located at the Central Archive for Empirical Social Research (ZA) at the University of Cologne, Germany, invites researchers to work with comparative data sets in the research environment of the ZA.

Data available at the Zentralarchiv

The Zentralarchiv collection of comparative data consists of major collections such as the Eurobarometer surveys, the Central and Eastern Eurobarometer, the International Social Survey Programme (ISSP) and the ICORE-collection of election studies to national parliaments in Europe. Furthermore, the collections include the Civic Culture Study, the Political Action Surveys, the USIA Studies and the European- and World Values Survey. In addition, the Zentralarchiv holds an extensive collection of political manifestos provided by the "Comparative manifestos Projects" presenting the programmatic profiles of political parties in 20 countries from 1945-1988. Altogether, there are some 4000 data sets and collections available for secondary analysis.

What is provided by the Zentralarchiv

Access to the LSF will be provided free of charge and will include access to training seminars as well as infrastructural, logistical and scientific support. Within the TMR program financial support is given to cover international travel and subsistence expenses in accordance with travel expenditure regulations for researchers participating in LSF-activities. Researchers from the European Union or associated states may apply for financial support for a period ideally between one and six months. Researchers from Germany can only be supported to a limited extent.

Who Can Participate in the TMR-LSF Activity

All researchers from the EU member states and the Associated States Iceland, Norway, Liechtenstein and Israel may apply for participation. They must be entitled to publish the results of their work at the LSF in the open literature. An international board will select proposals on the basis of scientific merit through an independent peer review procedure.

How to Apply

If you want to apply please contact:
Ekkehard Mochmann,
Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln,
Bachemer Straße 40, D - 50931 Köln, Germany
Email: mochmann@za.uni-koeln.de
Fax: +49-221-47694-77

Data Frontiers in the Infoscape - Notizen von der IASSIST/IFDO Konferenz ‘97

von Brigitte Hausstein, Lorenz Gräf, Meinhard Moschner und Ekkehard Mochmann

Im Intervall von 4 Jahren veranstalten IFDO (International Federation of Data Organisations for the Social Sciences) und IASSISST (International Association of Social Science Information Service and Technology) gemeinsam ihre Kongresse in Europa. Diesmal wurde der Kongreß vom Dänischen Datenarchiv (DDA) in Odense organisiert. Etwa 130 Teilnehmer aus aller Welt nutzten das reichhaltige Angebot zum Informationsaustausch über neueste Entwicklungen des sozialwissenschaftlichen Datenservice in Plenarsitzungen, Workshops und Postersessions.

Beeindruckend war, wie selbstverständlich die statistischen Ämter Nordamerikas und der skandinavischen Länder am Dialog beteiligt sind. Gemeinsames Ziel ist es, Daten zu liefern für eine Sozialforschung, die moderne Gesellschaften verstehen hilft und den gestaltenden Politikern Rückmeldung über die Wirkungen ihrer Maßnahmen liefert. Am Beispiel einer Longitudinalstudie zur Auswirkung des dänischen Wohlfahrtssystems wurde das auf der Konferenz prototypisch vorgeführt. In Norwegen können solche Forschungen noch detaillierter durchgeführt werden. Dort haben Forscher nach Rücksprache mit den Datenschutzbeauftragten Zugang zu einer Registerdatenbank, in der für 10% der norwegischen Population alle Informationen aus den sozialen Sicherungssystemen zusammengetragen sind. Diese Datenbank steht ausschließlich Forschern zur Verfügung. Für administrative Zwecke kann sie nicht genutzt werden.

Gesellschaften korrekt zu beschreiben und Zusammenhänge zu prüfen setzt Zugang zu guten Daten voraus. Hierzu müssen Informationen angeboten werden, welche Daten existieren und wie sie zugänglich sind. Daten müssen durch Metadaten konsistent beschrieben und suchbar gemacht werden. Verschiedene Beiträge auf der Konferenz widmeten sich diesem Thema. Vielfältige, sich ergänzende Aktivitäten werden weltweit hierzu durchgeführt.

Originaldokumente - elektronisch verpackt und auf einen Blick verfügbar

Der Feldfragebogen und der Berichtsband in Papierform, das Codebuch in ASCII, der Feldbericht im Textverabeitungsformat, Graphiken als Bilddateien usw.: die vollständige Dokumentation zu einem Datensatz setzt sich häufig aus den unterschiedlichsten Formaten zusammen. Immer noch gilt die Papierform für Archive und Benutzer als das Austauschformat, das allgemeine Lesbarkeit und unverfälschtes Layout garantiert. Doch die Druck- und Kopierarbeiten sind aufwendig, der Versand ist zeitraubend, während die zugehörigen Datensätze über das Internet (FTP) in wenigen Minuten zur Verfügung gestellt werden können.

Eine vielversprechende Lösung des Problems wurde in Odense von einem Mitarbeiter des Inter-University Consortium for Political and Social Research (ICPSR) vorgestellt. Das amerikanische Archiv setzt für den Vertrieb seiner Datendokumentation auf das von Adobe Acrobat entwickelte PDF (Portable Document File) Format, welches in der Lage ist, die unterschiedlichsten maschinenlesbaren Dokumente unter einem einzigen und (derzeit) kostenlosen Browser (u.a. Acrobat Reader) auf nahezu allen Plattformen verfügbar zu machen. Lesezeichen, Indices oder Hyperlinks helfen dem Benutzer sich problemlos durch die verschiedenen Dokumente zu bewegen. Eine OCR (Optical Character Recognition)-Option erlaubt die Volltextsuche auch in ursprünglich gescannten Bildformaten (TIFF, GIF usw.). Daß der Text dabei nur im Hintergrund gehalten wird (Image+Text Datei) ist ein wesentlicher Vorteil gegenüber der immer noch wenig befriedigenden klassischen OCR-Bearbeitung: nurmehr wichtige Suchbegriffe müssen überarbeitet werden. Der Ausdruck ausgewählter Seiten im Original-Layout schließlich ist auf jedem Laser- oder graphikfähigen Matrixdrucker möglich (***Zack W. Allen***, ***John E. Gray*** , ICPSR, Ann Arbor).

PDF-Dateien lassen sich aber nicht nur auf CD-ROM oder über FTP vertreiben, sie sind mit einer entsprechenden „Plug-In“ Hilfs-Software auch über verschiedene Web-Browser (einschließlich Netscape) direkt ansprechbar. Daß das Internet der Informations- und Datenmarkt der Zukunft ist, daran ließen die Konferenzteilnehmer keinen Zweifel. Gearbeitet wird allerorts an Möglichkeiten der effektiven Suche nach relevanten Informationen und des leichten Zugriffs auf Daten und Dokumentation. PDF, so schien es zumindest in Odense, hat dabei alle Chancen ein Standard-Austauschformat zu werden, auch wenn die Archivhaltung selbst nicht auf die soft- und hardwareunabhängige Papierform, auf ASCII und TIFF (als Sicherungskopie) wird verzichten können und Sicherheitsprobleme in den neuen Entwicklungen von Adobe noch berücksichtigt werden müssen.

Eurobarometer-Fragebögen auf CD-ROM

Ebenfalls im PDF-Format werden das niederländische Steinmetzarchiv und das Zentralarchiv erstmalig den Gesamtbestand an Eurobarometer-Fragebögen zugänglich machen. Die

englisch- sowie französisch-sprachigen Basisfragebögen und (nahezu) alle Feldfragebögen in jeder der beteiligten Landessprachen aus bereits über 45 Umfragen werden zwei CD-ROMs füllen. Eine Betaversion wurde in Odense im Rahmen der traditionellen Poster-Sessions vorgestellt (***Cor van der Meer***, Steinmetz Archive, Amsterdam, ***Meinhard Moschner***, Zentralarchiv, Köln).

Internetzugriff und Datenextraktion auf Variablenebene

Über drei Dekaden bestimmte das OSIRIS Format den Standard für Datendokumentation auf der Variablenebene. Bereits Anfang der 90er Jahre zeichnete sich eine Neuorientierung ab. Unter Führung des ICPSR begann 1995 eine internationale Expertengruppe in der Data Documentation Initiative (DDI) die Entwicklung eines neuen Standards. Eine erste Fassung wurde verabschiedet. Sie wird mit Mitteln der National Science Foundation weiterentwikkelt und wurde inzwischen für Softwareentwicklungen freigegeben.

Auf diesem Hintergrund waren mehrere Initiativen zu sehen, die neue Metadatenmodelle und Programmpakete zur Datendokumentation vorstellten. Das von der Europäischen Union geförderte ILSES Projekt (Integrated Library and Survey - Data Extraction Service) entwickelt Instrumente, die Einzelnutzer in die Lage versetzen soll, auf bibliographische, methodische und technische Information, sowie auf empirische Daten aus großen komplexen Datenbasen, wie z.B. die Eurobarometer, direkt zuzugreifen. ILSES wird für Endnutzer wie auch Datenanbieter für Daten und Informationen aus unterschiedlichen Bibliotheks- und Datenbeständen entwickelt. Es basiert auf integrierten relationalen Datenbasen und soll 1998 über Internet zugreifbar sein (***David A. Schweizer***, iecProgamma, Groningen).

Auf Ebene der Studienbeschreibungen hat das NESSTAR Projekt inzwischen einen virtuell integrierten Datenbestandskatalog der europäischen Archive entwickelt. In Kooperation des englischen, dänischen und norwegischen Archivs wurde die Systementwicklung mit Mitteln der Europäischen Union geleistet (***Simon Musgrave***, UK Data Archive, Essex).

Der Einsatz computergestützter Interviewtechniken, wie CAPI war Ausgangspunkt für neue Techniken der Datendokumentation im Computer-assisted Survey Methods Program in Berkeley. Unter Beachtung des DDI Standards werden Datendokumentationen entwikkelt, die mit heute bereits im Internet verfügbaren Browsern recherchiert werden können. So können mit Browsern wie Netscape ohne weitere Software einfache Kreuztabellierungen oder Mittelwertvergleiche im Internet gerechnet werden (***Tom Piazza***, University of California, Berkeley).

Das U.S. Census Bureau widmet sich der Entwicklung eines Prototyps statistischer Metadatendokumentation für den Datenvertrieb über Internet und für die integrierte Verarbeitung von Umfragen. Das relationale Datenmodell ist konzipiert als elektronischer Katalog

von Information über Survey Design, Verarbeitung und Analyse der Daten (***Daniel W. Gillmann***, Washington).

Von den Teilnehmern der Sitzung Metadata Applications wurde mit Spannung die Demonstration der “Hyperlinked Eurotrends“ erwartet. ***Lorenz Gräf*** vom Zentralarchiv zeigte wie eine Auswahl von Eurobarometer-Studien auf Fragenebene mit Hyperlinks vernetzt wurden. Vergleichbare oder gar identische Frageformulierungen wurden in Form eines Continuity Guide abgebildet. Auf dieser Basis wurden die Codebücher zu den Studien untereinander verbunden. Über einen Index wurden Variablen zu gleichen Konzepten mit Hyperlinks vernetzt. So konnten vergleichbare Fragestellungen auf einem Hypertextpfad im Prototyp des Eurobarometer Trend Codebook Systems (EUTRECS) auf dem Bildschirm identifiziert werden. Interessenten können den Prototyp unter folgender Adresse testen: http:// gazza.za.uni-koeln.de/eutrecs/.

Die Zahl der unterschiedlichen Entwicklungen zeigt einen technischen Bereich im Umbruch. Erfreulicherweise waren alle Referenten sensibel für die Orientierung an gemeinsamen Standards. Bereits auf der Bootsfahrt nach dem Konferenzausflug zu Valdemars Schloß auf der Insel Taasinge wurden Optionen geprüft, die Produzenten der Systeme zu einer Planungskonferenz zusammenzubringen, um partiell erkennbare Parallelentwicklungen abzustimmen.

Datenschutz, historische Datenbasen, Neuentwicklungen von Archiven, Archivgesetze, langfristige Datenaufbewahrung bis zur Speicherung von Bildern waren weitere Themenkreise an der Schwelle zum Jahr 2000, nicht zu vergessen die engagierten kanadischen Kollegen und Kolleginnen mit ihren “Data Liberation Army“-T-Shirts.

Mit besonderer Freude wurde vermerkt, daß mit Unterstützung von IFDO und IASSIST erstmals Kollegen aus fünf osteuropäischen Ländern (Estland, Rußland, Polen, Slowenien und Ungarn) die Teilnahme an der Konferenz ermöglicht wurde. Diese Kollegen zeigten sich nicht zuletzt dank des Einsatzes der GESIS Außenstelle in Berlin bereits gut informiert über die internationalen Entwicklungen. Die Veränderung der politischen Verhältnisse in Osteuropa brachte zwar neue Möglichkeiten für die empirische Sozialforschung und die Entwicklung der Dateninfrastruktur, die begrenzten materiellen Ressourcen zwingen aber gerade die neu entstehenden Archive zur sehr sorgfältigen Auswahl effizientester Arbeitsinstrumente und Auswertung der reichhaltig in Odense angebotenen Erfahrungen.

Die hervorragende Konferenzorganisation durch die dänischen Kollegen und das informationsreiche Tagungsprogramm sorgten dafür, daß nicht nur die neuen Kollegen, sondern auch die nicht mehr ganz so jungen Pioniere des Data Movement mit der Überzeugung nach Hause fuhren, daß sich auch eine lange Anreise gelohnt hatte.

Data Archives and Their Functions in Social Research in Eastern Europe [1]

Vom 6. bis 7. Dezember 1996 fand an der Universität von Tartu (Estland) eine internationale Konferenz zum Thema „Data Archives and Their Functions in Social Research in Eastern Europe“ statt. Diese Konferenz wurde vom „Estonian Social Science Data Archive“ (ESSDA) des „Department of Social Sciences“ im Rahmen des vom „Higher Education Support Project“ (HESP) der „Open Estonia Foundation“ geförderten Projektes zur Schaffung eines nationalen Datenarchivs vorbereitet und ausgerichtet.

Ziel war es, über Erfahrungen und Probleme beim Aufbau und bei der Nutzung von Datenarchiven zu diskutieren und dabei sowohl westeuropäische als auch osteuropäische Erfahrungen einzubeziehen. Zugleich sollte mit der Konferenz ein Anstoß gegeben werden, um die Integration der osteuropäischen Archive in die internationale Archivwelt zu befördern.

Insgesamt nahmen 21 Wissenschaftler aus acht Ländern teil: Estland (13); Lettland (2); Litauen (1); Rußland (1); Ungarn (1); Schweden (1) Deutschland (1). Die „Open Society Archives“ in Budapest sandten eine Vertreterin.

Einführend wurde die Entwicklungsgeschichte des gastgebenden Datenarchivs umrissen. ***Rein Murakas*** (Leiter des ESSDA, Estland) berichtete darüber, daß bereits 1993 eine Gruppe von Soziologen, Psychologen, Politologen und Geographen der Tartu Universität mit den konzeptionellen Arbeiten für den Aufbau eines nationalen Datenarchivs begonnen hatten. 1994 wurden die Projektgelder bewilligt und inzwischen sind 214 Datensätze aus dem Zeitraum 1975 - 1994 im Archiv verfügbar. Im März 1996 beschloß der Wissenschaftliche Rat der Tartu Universität das Statut von ESSDA und im April wurde der wissenschaftliche Beirat des Archivs gewählt, bestehend aus Vertretern der Tartu Universität, von akademischen Einrichtungen außerhalb der Universität und von Markt- und Meinungsforschungsinstituten Estlands.

Iris Alfredson (Swedish Social Science Data Service, Schweden) berichtete über die Entstehung von Datenarchiven in den Ländern Westeuropas und stellte die Zusammenarbeit auf dem Gebiet der Dokumentation der Daten im Rahmen des „Council of European Social

1 Vgl.: ***Hausstein, B.***: Data Archives and Their Functions in Social Research in Eastern Europe. In: Sozialwissenschaften in Osteuropa. Newsletter Januar 1997. Hrsg.: InformationsZentrum Sozialwissenschaften, Bonn. Berlin 1997.

Data Archives“ (CESSDA) vor. Sie verwies auf die Notwendigkeit einer einheitlichen Dokumentation (Standard Study Description) und die in diesem Zusammenhang äußerst hilfreichen CESSDA-Expertenseminare.

Brigitte Hausstein (Zentralarchiv für Empirische Sozialforschung Köln) referierte in Vertretung des CESSDA-Präsidenten ***Ekkehard Mochmann*** zum Thema „European Data Resource Management for the Social Sciences“. Im Mittelpunkt dieses Beitrages stand die Forderung zur Integration der europäischen Datenbasis. Es wurde festgestellt, daß die Harmonisierung und Verbreitung von Standards für Meßinstrumente, Datenrepräsentation und Datendokumentation wesentliche Voraussetzungen für die Erstellung einer vergleichenden Datenbasis sind. Der Datenzugang und die Datennutzung wird durch koordinierte Netzwerke für den Datenservice wie z. B. CESSDA signifikant verbessert.

Im weiteren Verlauf stellten sich das ungarische, lettische und russische Datenarchiv vor. ***Josef Mészáros*** (Abteilungsleiter im TARKI, Ungarn) informierte über das Vorhaben TARKIs, die unterschiedlichen Arten von Informationen und Daten (Umfragedaten, amtliche Daten, historische, geographische Informationen) unter Nutzung entsprechender Software (ORACLE) für die Internet-Präsentation aufzubereiten. ***M. Pugacheva*** (Institut für Soziologie an der Russischen Akademie der Wissenschaften) referierte über die Probleme bei der Schaffung einer nationalen Datenbank für Rußland. ***A. Cesnavicius*** (Litauen) sprach über die Schwierigkeiten bei der Beschaffung empirischer Daten in seinem Land. ***A. Tabuns*** (Präsident des Wissenschaftsrates der Akademie der Wissenschaften, Lettland) stellte das „Latvian Social Science Data Archive“ (LSSDA) vor und informierte über die Internet- Darstellung des Archivs.

Abschließend präsentierte ***Jüri Saarniit*** (Sozialwissenschaftliche Fakultät der Tartu Universität, Estland) Ergebnisse eines Forschungsprojektes zum Wertewandel Jugendlicher in Estland, das auf Daten des ESSDA (Langzeitstudien seit 1977) und aktuellen Jugendstudien basiert.

Die Teilnehmer der Konferenz verabredeten einen umfangreichen Informationsaustausch bezüglich der Archivstandards (Archivierung, Aufbereitung und Dokumentation der Daten). Das ESSDA bemüht sich um die Aufnahme in CESSDA. Das Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln wird dabei Unterstützung leisten.

Im folgenden setzen wir die in der ZA-Information Heft 39 begonnene Vorstellung des Datenbestandes des ESSDA Tartu fort:

8 From lecture to examination 1977

Different aspects of studies at the Estonian Academy of Agriculture. Students' contentment with the arrangement of studies. The use of textbooks and making study notes. Passing the examinations. The ideas of about future jobs.

No. of respondents: 644 No. of variables: 519

9 "Sirp ja Vasar" (Sickle and Hammer) 1977-78

A survey carried out among the Estonian professionals, studying their opinions and expectations about the newspaper that features mainly cultural events. Which media are followed, what are the readers' interests.

No. of respondents: 616 No. of variables: 486

10 Rural life '78

The living and working conditions of people in Viljandi county: the needs and wishes of country people, their attitudes toward work, possibilities of spending spare time.

No. of respondents: 433 No. of variables: 232

11 Newspaper and life 1979

A survey conducted among the adult population in the southern and central part of Estonia about the role of local papers among the rest of the printed matter. Interests, orientations and living conditions.

No. of respondents: 934 No. of variables: 1023

12 INFO 1979

Which are the main sources of information about the vital events, which TV and radio broadcasts are watched and listened to? Which events are considered as most interesting and to what extent is the information available?

No. of respondents: 1017 No. of variables: 564

13 From desk to desk '79

The first national study of secondary school students' life style conducted by the research team headed by ***H. Liimets***.

No. of respondents: 1874 No. of variables: 518

14 Students' occupational choice 1979

The options of secondary school students in vocational choice. The ideas about one's future and job to be attained.

No. of respondents: 1012 No. of variables: 330

15 Newspaper and life 1981

A study conducted among the adult population in the counties of Tartu, Pärnu, Eastern Viru, Harju, and the city of Narva about the role of local papers among the rest of the printed matter. Interests, orientations, living conditions.

No. of respondents: 1280 No. of variables: 1034

16 Secondary school student 1981

Adolescents' value orientations and future expectations. Styles of spending leisure. Attitudes about the school and fellow students. Favourite broadcasts on the Estonian TV and Radio. Ideas about the social life.

No. of respondents: 995 No. of variables: 367

17 Spare time of secondary school students 1981

The usage of leisure by secondary school students. Adolescents' interests and orientations. Favourite books, TV and radio broadcasts. Preferred ways of spending leisure.

No. of respondents: 363 No. of variables: 134

18 Information service in the Estonian Agriculture 1981

Expert estimations about the availability of professional information. The availability and use of different sources of information. Contentment with the information. The use of information in daily work. The flow of information within agricultural enterprises. Participation in information systems.

No. of respondents: 590 No. of variables: 500

19 The life mode of city people 1981

A study of the life mode of city people. Relations with colleagues and different forms of spending leisure. Social and political participation. Material security and family income.

No. of respondents: 1110 No. of variables: 384

20 City problems 1981

The urban mode of life and its problems, work and educational attainment. Family incomes and the standard of living. The ethnic composition of city residents.

No. of respondents: 1449 No. of variables: 265

21 City problems, Rakvere 1982

Living conditions of city people, social-professional, social, and educational groups, Jobs, career outlooks, possibilities of spreading leisure, the standard of living, future plans.

No. of respondents: 1001 No. of variables: 265

22 What is there and what to do? 1983

A study conducted among the top executives and chief specialists of Estonian agricultural enterprises in order to obtain expert estimations about the labour supplies in Estonian agriculture.

No. of respondents: 720 No. of variables: 255

23 Some problems of our life and work 1984

A survey conducted in Saaremaa focused on respondents' contentment with the work and living conditions. Estimations about agricultural enterprises. The mode of life of rural people.

No. of respondents: 803 No. of variables: 503

24 Your household '84

Peoples engagement in housekeeping and attitude toward work in housekeeping. Private housekeeping, building activities, services, living conditions and the living standard. Dietary habits, the domestic division of work, ways of obtaining consumer goods. The age structure of households.
No. of respondents: 2033 No. of variables: 1035

25 From desk to desk 1984

A follow-up survey of secondary school students conducted by ***H. Liimets*** research team.
No. of respondents: 829 No. of variables: 478

26 Books and their readers 1984

The buying and reading of books by Estonians in Tallinn. Readers' wishes and proposals to the publishers. The role of books and other media channels, particularly that of magazines, in satisfying readers' interests and needs. Interests, orientations, living conditions and the mode of life.
No. of respondents: 1097 No. of variables: 1250

27 "Rahva Hääl" 1984 (People's Voice)

A study of the press reception, focusing on the mass-circulation national daily "Rahva Hääl". Explores how often the other papers and periodicals are read and to what extent TV and radio broadcasts are watched and listened to. Attitudes toward topics featured in RH are examined more specifically.
No. of respondents: 995 No. of variables: 298

28 You and your child (1) 1984

Topical problems for city people of different generations. Respondents' estimations from the position of city-dwellers as well as parents. Living conditions in different residential areas, the family, work and leisure-related items are considered separately. The problems of 8th grade students and their relations with the parents are examined.
No. of respondents: 679 No. of variables: 539

29 The Islanders (inhabitants of Saaremaa) 1984

The links of Saaremaa-origin youths who are studying on the mainland with their home island and their intentions of returning there. Students' future plans, the impact of their parents and home place on career choice.
No. of respondents: 519 No. of variables: 70

30 You and the literature year 1984

A study about the books published in 1984. Which books were published and what was their reception by the readers. Who were most popular and who least popular authors.
No. of respondents: 871 No. of variables: 1086

31 You and the literature year 1985

A study considering books published in 1985. Which books were published and what was their reception by the readers. Which authors were most and which least popular.

No. of respondents: 500 No. of variables: 1297

32 Me, the leader (in Estonian) 1985

Work and life problems of executives and managers. The work environment and other factors effecting the work of administrators. The habits of spending leisure and the living standard. Relations with the other people.

No. of respondents: 191 No. of variables: 435

33 Me, the leader (in Russian) 1985

Work and life problems of Russian-speaking executives and managers. The work environment and other factors effecting the work of administrators. The habits of spending leisure and the living standard. Relations with the other people.

No. of respondents: 81 No. of variables: 460

34 TV, Radio, and I. 1985 (children's questionnaire)

Children's exposure to the TV and radio. The role of TV and radio in children's spending of free time. Children's favourite broadcasts and their attitude toward these.

No. of respondents: 1069 No. of variables: 322

35 Public opinion (in Russian) 1985

The availability of information by Russian-speaking population; which media were followed. Satisfaction with life in the Soviet Estonia and opinions about social problems.

No. of respondents: 577 No. of variables: 666

36 Public opinion 1986

A mass media study examining the availability of information to the Estonian-speaking population and analysing which media were followed, from which radio and TV broadcasts and papers and periodicals information was received. The satisfaction with life in the Soviet Estonia and that time social problems were also considered.

No. of respondents: 1437 No. of variables: 392

37 The Chernobyl disaster 1986

The sources of information about the disaster. The adequacy of such information. T. Avikson's article "Chernobyl in July 1986."

No. of respondents: 465 No. of variables: 55

38 Cinema and film 1986

A study of cinema and film interests. Which films are popular and which are not. The sources of information about new films. The place of cinema among the other leisure

No. of respondents: 1470 No. of variables: 384

39 A pre-schooler and mass media 1986

How 5-7 years old home children watch the TV, listen to the radio and read (if read) books. The availability and impact of mass media on small children.

No. of respondents: 319 No. of variables: 163

40 I am a doctor 1986

A survey of doctors examines relations with the colleagues and patients and doctors' attitudes toward diseases and their treatment. Doctors' possibilities of spending leisure, their habits and life styles.

No. of respondents: 270 No. of variables: 559

41 From desk to desk 1986

A follow-up study of secondary school students conducted by disciples of ***H. Liimets***

No. of respondents: 1887 No. of variables: 609

42 You and the literature year 1986

A study considering the books published in 1986. Which books were published and what was their reception by the readers. Which authors were most and which least popular and which were the contacts with libraries.

No. of respondents: 725 No. of variables: 1176

43 Do you read magazines? '87

The survey embraced Estonian adult population and examined the degree of reading two magazines. Estimations of different magazines. Interests, orientations, life styles, mode of life.

No. of respondents: 1295 No. of variables: 647

44 Culture, radio, TV and we 1987

The role of culture-related broadcasts of the Estonian TV and Radio and culture in general in people's life. Cultural interests during the period of perestroika in Estonia. The viewers' and listeners' expectations in the field of culture.

No. of respondents: 990 No. of variables: 1384

45 Radio DJ-s 1987

Adolescents' musical taste. Favourite singers, groups, composers. Habits of listening to music. Which music broadcasts are favoured. What are the reasons for listening to music, what kind of audio-visual equipment is possessed? Other ways of spending leisure.

No. of respondents: 795 No. of variables: 330

46 Woman at work and in the family 1987

A study carried out in different counties looks at living, work, and recreation conditions during the Soviet era as well as relations with the other people, problems in these fields.

No. of respondents: 492 No. of variables: 818

47 The TV, radio, and I 1987

A study of young people examines youth's relationships with home, school and friends. Youth's favourite activities, leisure pursuits. How, when and why is the TV watched and radio listened to; which broadcasts were favoured, which not.

No. of respondents: 1065 No. of variables: 930

48 The university student and the books '87

A survey conducted in the scientific library of the Tallinn Technical University examined students' reading habits and interest in books, cultural interests, the reception of concrete literary texts.

No. of respondents: 402 No. of variables: 523

49 Me and the construction brigade 1987

The Estonian Secondary School Students' construction brigade as a summer-time life style. The reasons for joining the construction brigade. The brigade as a possibility for earning good money, feeling free, finding new friends, becoming familiar with real farm work, to see and hear new and interesting things.

No. of respondents: 664 No. of variables: 142

50 Expert estimations of "perestroika" 1988

The opinions of managers about the changes on social and economic life during the period of perestroika. Economic transformations. Difficulties in moving from the planned state system to a marked economy. The growth of the independence and responsibilities of enterprises. The transition of all-union enterprises to local administration.

No. of respondents: 221 No. of variables: 197

51 Work and work-mates 1988

The impact of work-mates' gender on work efficiency. Psychological atmosphere at workplace. Workers' assessment of work and colleagues.

No. of respondents: 207 No. of variables: 269

52 Me and my country 1988

Social processes occurring in different regions of Estonia. People's estimations of their home country. Cultural life in the countryside. Social and economic issues. The management in the countries.

No. of respondents: 946 No. of variables: 262

53 Public opinion 1988 (1)

Estonians' assessment of the social and political changes in the third year of "perestroika." The access to information from the mass media. Attitudes toward plans of mining phosphates in Virumaa and the state of environment in general.

No. of respondents: 1167 No. of variables: 447

54 Journalist and the press 1988

The study considers journalists' attitude toward their work. Journalists' living and work conditions, earnings and possibilities of spending leisure. The role of the press in society. The professional level of Estonian press. The press as the mirror of changes in Estonia. Self-evaluation of one's journalistic activities.

No. of respondents: 368 No. of variables: 1070

55 IME (self-managing Estonia) Panel for managers and executives 1988

A survey on different leadership styles and corporate development. Relationships within and outside the Estonian organisation. Work and pay conditions, the personnel potential of the organisation.

No. of respondents: 303 No. of variables: 118

56 The leader, economy, results 1988

The leadership and innovations in Estonian agricultural enterprises. The node of mind and attitudes toward the rural life, management, the family and general human problems of country people.

No. of respondents: 230 No. of variables: 508

57 Me and we in the changing rural life 1988

Estimations to the situation and development of agriculture. Life styles of rural people. Problems of farm work, auxiliary farms, possibilities of spending spare time.

No. of respondents: 549 No. of variables: 374

58 Music 1988

A youth study looking at the role of music among the interests of young people. Youth's attitude toward classical music, jazz, pop and rock music. Contentment with TV and radio music broadcasts. The reception and popularity of the TV and radio. The sources of information abut one's favourite music.

No. of respondents: 1342 No. of variables: 321

59 Tallinn 1988

Problems and possibilities of development related to the status of living environment in Tallinn. Migration in Tallinn. Work conditions and dependence of job on educational attainment. Problems of family planning and public health, living conditions.

No. of respondents: 1614 No. of variables: 579

60 Your opinion 1988

The opinions of Estonian adults about the social changes. The availability of information, the role of various media channels, attitudes toward topical issues.

No. of respondents: 888 No. of variables: 145

61 Rahva Hääl 1988 (People's Voice)

The opinions of Estonians about the perestroika-period press. The study focused on readers' attitudes toward the national daily "Rahva Hääl". Political changes connected with perestroika are also
examined.

No. of respondents: 942 No. of variables: 230

62 Public opinion '88 (December)

Social and political transformation in Estonia. The admission of the language law. The hoisting of national flag at the Tower of Pikk Hermann. The relations between Estonians and the Russians. The elections of the deputies to the Congress of Peoples Deputies in Moscow as a step toward democracy.

No. of respondents: 1149 No. of variables: 341

63 Public opinion 1989

The opinions of Estonian adult population about the social developments in 1988. People's attitudes toward the changes in social, economic, political and educational spheres. People's expectations about political independence. The exposure to different media channels.

No. of respondents: 969 No. of variables: 521

64 Family and home 1989

A study focusing on the listeners' opinions about the long-time Saturday-morning broadcast "Family and home". Which other family-specific broadcasts are listened to? Problems concerning the family and home were examined.

No. of respondents: 561 No. of variables: 200

65 Pre-schooler 1989

A survey carried out among the 5-6 year olds examines how children watch the TV, listen the radio and read books and what they think about them. When do pre-schoolers prefer drawing to playing, sporting to reading, how does a child's day look like in general.

No. of respondents: 1465 No. of variables: 426

66 Your opinion 1989

A public opinion survey conducted in April 1989 about the current political situation. Attitudes toward different political forces and problems. Ideas about Estonia's future trends. The exposure to different media channels.

No. of respondents: 924 No. of variables: 147

67 The rule of law I 1989

The political and economic development of society. The first perestroika-period steps toward the rule of law. Political preferences and most topical economic problems. Attitudes toward military service in the Soviet army.

No. of respondents: 851 No. of variables: 600

68 Life in Estonia (Mainor) 1990

Daily life in Estonia. Opinions, attitudes and contentment with work, family life and leisure. Interests, living conditions.

No. of respondents: 1426 No. of variables: 361

69 The life-style (in Russian) 1990

A study carried out among the Russian-speaking population considers the social, economic, work and everyday problems of the inhabitants of new residential districts. People's habits of spending spare time.

No. of respondents: 263 No. of variables: 678

70 Media channels 1990

The opinions of the Estonian TV watchers and listeners to the Estonian Radio about the content of the broadcasts. Which radio channels and to what degree are followed? Which are the most important broadcasts, what kind of information is needed?

No. of respondents: 912 No. of variables: 281

71 Manager I 1990

A survey conducted among the managers. The ideas of economic leaders about their enterprise as well as the economic and socio-political developments in Estonia in general. The current state of Estonian enterprises. Different styles of management.

No. of respondents: 912 No. of variables: 173

72 Manager II 1990

Managers' opinion of topical economic and political problems. Their legal income and social position. Attitudes toward the market economy and privatisation.

No. of respondents: 864 No. of variables: 184

73 From desk to desk 1990

A follow-up study of life styles of secondary school students

No. of respondents: 1753 No. of variables: 609

74 "Rahva Hääl" 1990 (People's Voice)

Opinions about the popular national papers "Rahva Hääl", "Päevaleht" (Daily News), "Maaleht" (Country Paper) and "Õhtuleht" (Evening News). To what degree are newspapers and magazines read and the TV and radio watched and listened to. The actuality of topics featured in the Estonian press.

No. of respondents: 915 No. of variables: 244

75 Your opinion (in Russian) 1990

The opinions of Russian-speaking population about some topical transition period problems, their attitudes and general mood. The social milieu, future expectations and relations with the Estonians of the Russian-speaking population.

No. of respondents: 610 No. of variables: 199

76 Hello, TV-fan! 1990

Attitudes about the TV children's broadcasts. The most popular children's broadcasts. Other possibilities of spending leisure in the afternoon.

No. of respondents: 1140 No. of variables: 49

77 The rule of law II 1990

Problems related to Estonia's political and socio-economic development. A survey conducted in Tallinn considers the enforcement of civil liberties, attitudes toward politics and religion. The identity option: Estonia versus the Soviet Union. Attitudes toward Presidential elections, unemployment, the freedom of press in Estonia.

No. of respondents: 833 No. of variables: 290

78 Manager III 1991

The survey of managers examines attitudes about the economy and enterprising in pre-kroon (currency) Estonia. Opinions about privatisation, the restitution of properties and the socio-political developments in Estonia. The state and development trends of Estonian enterprises.

No. of respondents: 764 No. of variables: 153

79 Pensioners 1991

The opinions of the TV serial "Prillitoos" (spectacle-case) of this broadcast. The actuality of subjects considered in this broadcast. Pensioners' life style in the early 1990s.

No. of respondents: 513 No. of variables: 203

80 A Radio and TV - Study 1991

A media study considering the reception and comprehensibility of Saturday and Sunday morning broadcasts "Peresaade" (Family Matters) and "Teleteater" (TV theatre).

No. of respondents: 371 No. of variables: 365

81 Your opinion 1991

A radio and TV study on the content and structure of the program. The popularity of different media channels, how they portray topical events. Why people turn to one or the other information channel? Attitudes toward transition period topical problems.

No. of respondents: 877 No. of variables: 479

82 You and your child (2) 1991

The ideas of different generations (children and their parents) about optimum living conditions and topical problems. Living conditions in Tallinn and its different residential districts. The family, work and leisure. The 9th grade students were studied more thoroughly as they faced their first choice in continuing the educational path.

No. of respondents: 484 No. of variables: 539

83 Culture '92

The reception and estimations about the culture-related broadcasts of the Estonian Radio and TV. Viewers' and listeners' opinions about different broadcasts featuring culture and entertainment. People's cultural orientations.

No. of respondents: 843 No. of variables: 505

84 New entrepreneurs in Estonia 1992

The situation of new entrepreneurs, mechanisms which help to cope in uncertain and difficult economic conditions. Different forms of entrepreneurship. The birth of firms. Turnover, revenue, investment strategies, personnel problems, Estonian socio-economic climate. Entrepreneur's social position.

No. of respondents: 256 No. of variables: 205

85 The TV and Radio in Estonia 1992

The viewing of TV channels in different hours. Reasons for this. The assessment of broadcasts.

No. of respondents: 362 No. of variables: 632

86 The TV and Radio in Estonia (in Russian) I 1992

The role of the Estonian TV and Radio among the non-Estonians. The popularity of the Estonian and Russian media channels. How much non-Estonians follow Estonian broadcasts and which ones?

No. of respondents: 458 No. of variables: 426

87 The TV and Radio in Estonia (in Russian) II 1993

The place of the Estonian TV and Radio among the Russian-speaking population. Potential viewers and listeners of Estonian broadcasts. Preferred channels and stations and their broadcasts. Attitudes toward the Estonian state and Estonians, interests and orientations (culture, sport, business, etc.)

No. of respondents: 606 No. of variables: 202

88 News 1993

Which channels provide information, what kind of news are needed most of all. Most popular news broadcasts and producers.

No. of respondents:752 No. of variables: 246

Ansprechpartner zum Datentransfer Osteuropa sind:
Brislinger, Evelyn Tel. 0221 47 69 4-67, e-mail: brislinger@za.uni-koeln.de
Riedel, Eberhard Tel. 0221 47 69 4-67, e-mail: riedel@za.uni-koeln.de
Hausstein, Brigitte (GESIS-Außenstelle Berlin) Tel. 030 308 742 49,
e-mail: hausstein@berlin.iz-soz.de

Praktische Ziehung von Zufallsstichproben für Telefon-Surveys

von Rainer Schnell [1]

Zusammenfassung

Es werden verschiedene Ansätze für Stichprobenpläne vorgestellt, mit denen aus Telefon-CD-ROMs Zufallsauswahlen von Telefonnummern durchgeführt werden können. Die dabei auftretenden praktischen Probleme und deren Lösungen werden diskutiert. Hierzu werden Hilfsprogramme in der String-Verarbeitungssprache AWK verwendet.

Abstract

Sample designs for telephone surveys using German CD-ROMs of telephone numbers are discussed. The practical problems during actual selection are solved by small programs using the string processing language AWK.

> "Well, last week we showed you how to become a gynaecologist. And this week on 'How to do it' we're going to show you how to play the flute (...) How to play the flute. (picking up a flute) Well here we are. You blow there and you move your fingers up and down there." (***Chapman*** 1990:63-64)

> "- Auswahlbasis auf der ersten Stufe sind sogenannte Knoten. Darunter versteht man Ortsvermittlungsbereiche bzw. Ortsnetze, die über keine Ortsvermittlungsbereiche verfügen. Davon gibt es etwa 6.000 in ganz Deutschland. Daraus wird eine Zufallsauswahl gezogen.
> - In einem weiteren Schritt wird dann für jeden ausgewählten Knoten eine Nummer ausgewählt." (***Fuchs*** 1994:162)

Da in der Bundesrepublik kein akademisch basiertes Erhebungsinstitut existiert, werden alle bundesweiten Surveys für face-to-face-Interviews durch die kommerziellen Datenerhebungsinstitute durchgeführt[2]. Telefoninterviews bieten hingegen Universitäten prinzipiell die Möglichkeit, lokale und auch bundesweite Erhebungen ohne Inanspruchnahme kom

1 Anschrift: Dr. ***Rainer Schnell*** ist Professor an der Fakultät für Verwaltungswissenschaft, Universität Konstanz, Postfach 5560, 78434 Konstanz, email: Rainer.Schnell@Uni-Konstanz.De

2 Zu den damit verbundenen Problemen vgl. u.a. ***Schnell*** (1997:253-254).

merzieller Institute durchzuführen. Der praktischen Umsetzung stehen vor allem zwei Probleme entgegen:

1. Es existiert anscheinend keine verfügbare CATI-Software in der Public Domain oder zu Preisen, die im akademischen Bereich noch finanziell tragbar wären[3].
2. Es wurden bislang kaum Stichprobenpläne für Telefonsurveys in der Bundesrepublik publiziert.

Das erste Problem läßt sich z.B. durch den Einsatz jeweils für den speziellen Zweck angefertigter Einzelprogramme umgehen[4]. Häufiger wird bei akademischen Surveys auf dieCATI-Technologie verzichtet und statt dessen traditionell mit Interviewern gearbeitet, die am Telefon einen gedruckten Fragebogen ausfüllen[5].

Das zweite Problem ist zumindest für bundesweite Stichproben schwieriger zu lösen. Ideen für Stichprobenpläne und Ansätze prinzipiell praktikabler Ziehungsverfahren wurden in der Bundesrepublik bisher nur für Auswahlen anhand der gedruckten Telefonbücher publiziert[6]. Es ist leicht einzusehen, daß die damit verbundenen praktischen Probleme aufgrund des hohen Arbeitsaufwandes beachtlich sind. Nach dem Erscheinen von CD-ROMs, die die Telefonnummern der BRD enthalten, wird die Durchführung eines Auswahlverfahrens für Telefonsurveys nunmehr häufig fälschlich als trivial angesehen. Praktische Probleme der Durchführung von Stichprobenziehungen zeigen sich aber erst bei ihrer tatsächlichen Realisierung. Einige der auftretenden Schwierigkeiten sollen hier dargestellt und Möglichkeiten zur ihrer Überwindung beschrieben werden.

Telefon-CDs als Auswahlgrundlage

Für die Bundesrepublik werden CD-ROMs mit den in den Telefonbüchern eingetragenen Nummern der Teilnehmer derzeit von drei verschiedenen Herstellern angeboten[7]. Um als Auswahlgrundlage brauchbar zu sein, muß die Datenbank die Möglichkeit bieten, alle Datensätze exportieren zu können[8]. Da sowohl die Telekom-CD als auch die Tele-Info-CD

3 Kommerzielle Systeme wie BLAISE, CI-3, Surveycraft oder in2itive liegen in der Anschaffung derzeit effektiv bei ca. 1000 bis 2000 DM pro Interviewerstation.

4 Leider fehlen diesen Programmen dann eine große Zahl derjenigen Features, die CATI-Interviews eigentlich interessant machen, wie z.B. das Call-Management.

5 Die wohl beste und detaillierteste Durchführungsanweisung für solche Surveys findet sich bei ***Lavrakas*** (1987).

6 Prinzipielle Möglichkeiten beschreibt ***Zeh*** (1987), technische Details in größerem Umfang finden sich bisher lediglich bei ***Frey, Kunz und Lüschen*** (1990:83-98).

7 DeTeMedien (1996): Telefonbuch für Deutschland (CD-ROM), Frankfurt.
Tele-Info Verlag (1996): Xi-Deutschland (CD-ROM), Garbsen.TopWare (1996): D-Info 3.0 (CD-ROM), Dorsten.

8 Dies ließe sich nur umgehen, wenn die Hersteller bereits in der Abfragesoftware der CDs die Möglichkeit einer Zufallsauswahl vorsehen würden.

ein Exportlimit von maximal einigen Hundert Nummern besitzt, kommt für die praktische Ziehung von Telefonstichproben nur die D-Info 3.0 CD in Frage[9].

Praktische Probleme der Dateigröße

Bei der Verwendung der D-Info 3.0 CD als Auswahlgrundlage entstehen eine Reihe technischer Probleme durch die Größe der resultierenden Dateien. Da die CD mit einem nicht dokumentierten Kompressionsalgorithmus beschrieben wurde, sind die Daten nicht direkt lesbar. Man kann Datenbestände allerdings exportieren. Die Software erlaubt als Fileformat Dbase- oder ASCII-Files. Dbase-Format besitzt den Vorteil, daß es direkt von anderen Programmen wie z.B. SPSS gelesen werden kann. Das Export-Format sieht aber eine fixe Recordgröße von 281 Byte vor. Bei 34.3 Millionen Einträgen ergäbe sich daher eine Dateigröße von 9.2 Gigabyte. DOS kann aber nur Platten bis ca. 2 Gigabyte verwalten. Das Dbase-Format eignet sich folglich nur für Files bis ca. 7.5 Millionen Einträgen. Werden mehr Einträge benötigt, muß auf das ASCII-Format ausgewichen werden. Verwendet man das ASCII-Format unter DOS, so sieht D-Info ein "embedded"-Stringformat vor: Die einzelnen Felder (Name, Straße, Ort, Nummer) sind jeweils in Anführungsstrichen eingeschlossen und durch Semikolon getrennt. Dieses Format führt für den gesamten Datenbestand zu einer Dateigröße über 2 Gigabyte. Unter DOS kann der gesamte Datenbestand daher nicht exportiert werden. Unter Windows erlaubt D-Info hingegen die Veränderung des ASCII-Exportformats. Verzichtet man auf die Einbettung in Anführungsstriche, so reduziert sich der Speicherbedarf um ca. 262 Megabyte auf 1.96 Gigabyte. Es empfiehlt sich hierbei einen Feldtrenner zu verwenden, der im normalen Datenbestand nicht vorkommt, so z.B. den senkrechten Strich (|). Diese Datei enthält ca. 34.324.400 Records[10]. Dateien dieser Größe lassen sich mit den meisten Standardprogrammen nicht mehr verarbeiten[11].

Hilfsprogramme für den Umgang mit dem Datenbestand

Das Exportprogramm der CD exportiert vier Felder (Name, Straße, Ort, Nummer); für die meisten Auswahlverfahren bei Telefonsurveys werden hiervon nur Nummer und Name benötigt. Da schon allein die kleinere Dateigröße den Umgang mit diesen Dateien wesentlich erleichtert,sollte man durch geeignete Filterprogramme die nicht benötigten Felder löschen.

---

9 Die Exportlimits lassen sich zwar prinzipiell durch verschiedene Möglichkeiten (z.B. Makros, Dekodierungsprogramme, Disassemblieren und Patchen der Retrievalsoftware) umgehen, allerdings wäre der Arbeitsaufwand trotzdem vergleichsweise hoch. Weiterhin würden solchen Verfahren klar gegen die Nutzungsbedingungen der CDs verstoßen.

10 Die genaue Zahl hängt davon ab, welches Kriterium man verwendet um fehlerhafte Einträge bereits im ersten Lesevorgang zu eliminieren.

11 Bemerkenswerterweise gilt dies sogar für eine Reihe von Compilern. So lassen sich z.B. Files über 500 Mbyte anscheinend in Turbo-Pascal 7.0 (***Borland*** 1992) nicht ansprechen. Die kompilierten PASCAL-Programme brechen mit irreführenden Fehlermeldungen ("file not found") ab.

Filterprogramme werden für eine Reihe weiterer Aufgaben bei der Durchführung von Auswahlverfahren benötigt.

Programmiert man einige Datenprüfungen mit, so lassen sich solche Filterprogramme in einer allgemeinen Programmiersprache wie FORTRAN, C++ oder PASCAL mit weniger als 100 Programmzeilen realisieren. Die Programmentwicklung ist allerdings wesentlich einfacher, wenn man eine auf Stringverarbeitung spezialisierte Programmiersprache wie Perl (***Wall, Schwartz*** 1993) oder AWK (***Aho, Kernighan*** und ***Weinberger*** 1988) verwendet[12]. Mit diesen Sprachen lassen sich solche Aufgaben mit wenigen Zeilen Programmcode erfüllen. Perl eignet sich auch für größere Programme, ist aber etwas schwerer zu erlernen als AWK. AWK ist eine ideale Sprache für kleine Gelegenheitsprogramme, die nur wenige Male benötigt werden und deren Effizienz keine Rolle spielt[13]. Die im folgenden berichteten Prozeduren wurden entsprechend zum größten Teil mit AWK-Programmen durchgeführt, die selten mehr als 10 Zeilen beanspruchen[14]. Beispiele für AWK-Programme bzw. "Oneliner" finden sich unten[15]. Falls in ASCII-Files nur einzelne Zeilen gelöscht oder gesucht werden sollen, empfiehlt sich neben AWK die Verwendung des Unix-Programms "grep", das auch in verschiedenen Varianten für DOS zur Verfügung steht. Weiterhin benötigt man für viele Aufgaben, wie z.B. der Bestimmung der Zahl der Doppeleinträge, ein leistungsfähiges Sortierprogramm, das unter DOS leider nicht zum Betriebssystem gehört[16].

Schließlich tritt ein besonderes Problem dadurch auf, daß nur wenige DOS und Windows-Editoren Files mit einer Größe von mehreren hundert Megabyte bearbeiten können[17]. Die für die Auswahl notwendigen Aufgaben lassen sich für den gesamten Datenbestand entwe-

---

12 Compiler bzw. Interpreter für beide Sprachen sind im Internet kostenlos erhältlich. AWK gehört bei UNIX-Systemen zu Lieferumfang. Eine DOS-Version von AWK (Autor: ***Rob Duff***) findet sich im SIMTEL-Archiv. SIMTEL ist z.B. über ftp.uni-heidelberg.de verfügbar. AWK liegt als AWK320.ZIP im Verzeichnis "simtelnet/msdos/txtutl". Weiterhin gibt es ein GNU AWK (GAWK), das auch für DOS compiliert wurde. Diese schnelle Version findet sich u.a. unter "http://www.leo.org/pub/comp/platforms/pc/gnuish/".

13 Den Hinweis auf die Möglichkeiten von AWK beim Datenmanagement verdanke ich Professor ***Wolfgang Sodeur***.

14 Da sich AWK aber bei Programmfehlern fast immer auf die Mitteilung "Syntax Error" ohne weitere Hinweise beschränkt, kann die Entwicklung der ersten eigenen Programme etwas länger beanspruchen als man ursprünglich glaubt.

15 Bei den Beispielprogrammen sind Zeilen, die mit "#" beginnen, Kommentarzeilen. In der ersten Kommentarzeile jedes Programms findet sich hier der Programmname. Die Beispielprogramme müssen in jeweils einem ASCII-File mit diesem Namen gespeichert werden, so z.B. "DUPS.AWK". Um das Programm auszuführen muß AWK installiert sein. Unter DOS erfolgt die Installation durch Kopieren eines Files (AWK.EXE) in ein beliebiges Verzeichnis im Suchpfad. Der Aufruf erfolgt dann mit "AWK DUPS inputfile > outputfile".

16 Im SIMTEL Verzeichnis "txtutil" (siehe oben) finden sich eine Reihe nützlicher DOS Hilfsprogramme, so z.B. auch das Sortierprogramm MSORT (MSORT115.ZIP) von ***Martin Katz***. MSORT sortiert auch Dateien, deren Größe über hundert Megabyte liegt, wenn auch - natürlich - mit entsprechender Dauer.

17 Ein solcher Editor ist z.B. MEL von American Cybernetics Inc. Eine Version findet sich im SIMTEL-Directory "simtelnet/msdos/editors" als "MELITE.ZIP".

der problemlos mit einer leistungsfähigen Unix-Maschine oder mit größeren Problemen mit einer Unix-Workstation bzw. einem PC mit den erwähnten Hilfsprogrammen durchführen. Für die Bearbeitung von Teilmengen des Datenbestandes, wie z.B. einzelner Gemeinden, reicht hingegen schon ein kleinerer PC mit Standardsoftware wie z.B. SPSS oder auch EXCEL aus.

Übersicht 1: Sechs verschiedene vollständige AWK-Programme zur Listenverarbeitung.

```
# DUPS.AWK: count the number of duplicate input lines
# output: only dups, first field: number of replications
BEGIN { old="_Z@1_Z@1Z_Z@1";c=1 }
{
if ($0==old) { c=c+1 }
else { if (c > 1) {print c,old; c=1}; old=$0 }
}
END { if (c > 1) {print c,old } }
 ———————————————————————

# UNIQUE.AWK: report only unique input lines, omit duplicates
BEGIN { old="_Z@1_Z@1Z_Z@1" }
{ if ($0 !~ old) { print $0; old=$0 } }
 ———————————————————————

# LINES.AWK: number of records
END {printf("%s%d\n","Number of Records: ",NR) }
 ———————————————————————

# FREQF1.AWK: frequencies of the first field
{
a=$1
count[a]++
}
END { for (i in count)
        print i ": ", count[i]
}
 ———————————————————————

# ONLYF1.AWK: read comma delimited, embedded strings.
# Kill ", keep only field 1
BEGIN {FS=","}
{ gsub("\"","",$0); print $1 }
 ———————————————————————

# PREFREQ.AWK: count dial prefix in a sorted file
BEGIN {FS="-";old="0000";c=1}
{
if ($1==old) { c=c+1 }
else { if (c > 1) {print old,c; c=1}; old=$1 }
}
END { print old,c }
```

Probleme der Datenbasis

Trivialerweise enthalten die Telefon-CDs ebenso wie die Telefonbücher lediglich die Einträge derjenigen Anschlüsse, deren Besitzer den Eintrag ins Telefonbuch wünschen und

deren Anschluß nicht kürzlich erfolgte. Da beide Mechanismen nicht zu zufälligen Ausfällen führen[18], sollte **keine** Auswahl aus der Liste (dem Telefonbuch oder der CD) erfolgen, sondern ein Eintrag in der Liste als Ausgangszahl für das "Randomized Last Digit"-Verfahren (RLD) verwendet werden, bei dem zur Ausgangszahl eine gleich verteilte Zufallszahl zwischen 0 und 9 addiert wird (ein Beispiel für ein entsprechendes AWK-Programm findet sich unten). In der Praxis wird hierauf allerdings häufig verzichtet: Zum einen aus Kostengründen (da viele der generierten Nummern keine Privatanschlüsse darstellen), zum anderen um sich Diskussionen mit aufgebrachten Zielpersonen, deren Nummer absichtlich nicht eingetragen wurde, zu ersparen.

Übersicht 2: AWK-Programm RLD.AWK: Addiert eine Zufallszahl zu einer Liste mit Telefonnummern

```
# RLD.AWK: add a random digit to a list of phone numbers
BEGIN {srand()}
{
i=index("-")
if (i==0)
 { print "no prefix in following input line: ", $0; exit }
pre=substr($0,1,i)
tel=substr($0,i+1)
gsub("-","",tel)
r=int(rand()*10)
printf("%s%d\n",pre,tel+r)
}
```

Die exportierten Dateien enthalten Sondernummern, deren Vorwahl erkennen läßt, daß sie kaum als Privatanschluß in Frage kommen, so z.B. 0130, 0180, 0190. Es empfiehlt sich, diese Nummern durch ein Filterprogramm zu löschen. Es stellt sich die Frage, ob dies auch für die Mobil-Telefonnummern (0161, 0171, 0172, 0177) gilt. Übliche Praxis in der BRD durch die kommerziellen Institute ist es, alle diese Nummern aus der Auswahl auszuschließen. Die Begründung hierfür liegt darin, daß vermutlich fast alle Besitzer von Mobil-Telefonen auch über einen stationären Anschluß erreicht werden können und sie daher eine höhere Auswahlwahrscheinlichkeit besäßen, wenn man sie nicht ausschlösse.

Dies ist vermutlich für die neuen Bundesländer zumindest zur Zeit nicht in gleichem Umfang gegeben wie für die alten Bundesländer. Eine Auszählung der entsprechenden Vorwahlen nach Bundesländern in der DInfo 3.0 CD zeigt so deutlich höhere Anteile von Sondervorwahlen (Mobilanschlüsse + 0130, 0180, 0190)in den neuen Bundesländern als in den alten Bundesländern (zwischen 2.8% und 3.8% in den neuen,zwischen 0.9% und 2.7%bzw. 3.7% mit dem Ausreißer Rheinland-Pfalz in den alten Bundesländern). Dieses Ungleichgewicht betrifft aber das generelle Problem der Anschlußdichte in den neuen Bundesländern

18 Eine Liste der amerikanischen Literatur zu diesem Problem findet sich bei ***Häder*** (1996).

und soll daher hier nicht weiter verfolgt werden. Generell ließe sich das Problem der Mobilanschlüsse nur durch die tatsächliche Berücksichtigung bei der Auswahl und der expliziten Frage nach weiteren stationären Anschlüssen klären. Abgesehen von der vermutlich höheren subjektiven Belastung der potentiellen Zielperson bei Mobilanschlüssen durch den Anruf und den daraus resultierenden Verweigerungen sowie des vermutlich doch geringen Anteils der nicht über stationäre Anschlüsse erreichbaren Zielpersonen sprechen gegen die Berücksichtigung der Mobilanschlüsse die weit höheren Gesprächsgebühren. Daher sollte das Filterprogramm auch die Mobiltelefonnummern vor der Auswahl löschen.

Die vom D-Info 3.0 Exportprogramm ausgegebenen Felder für Telefonnummern enthalten nicht nur Telefonnummern, sondern auch eine Reihe von Sondersymbolen wie z.B. "ISDN", "Q", "NEU" und "+". Diese Symbole sollten vom Filterprogramm gelöscht werden. Prüft man die verbleibenden Zeichenketten daraufhin, ob sie Telefonnummern darstellen, so entdeckt man eine Reihe von Fehlern. Hierzu gehören mehrere Telefonnummern in einem Feld, die durch Komma getrennt sein können oder auch nicht. In anderen Fällen stellen die Zeichen in diesem Feld keine Nummern dar, sondern Strings wie z.B. "Beratung". Insgesamt finden sich ca. 1500 solcher Fehler im gesamten Datenbestand. Dies ist zwar ein verschwindend geringer Anteil (ca. 0.0045%!), verhindert aber die problemlose Übernahme in automatische Wählprogramme[19].

Um die Erhebungskosten zu senken, wäre eine eindeutige Klassifikation der Telefonnummern nach Privat- und Geschäftsanschluß nützlich. Eine solche liegt im Datenbestand nicht vor.

Man kann natürlich die Strings im Feld mit dem Namen des Anschlußinhabers klassifizieren[20]. Hierzu benötigt man eine Liste von Strings, die auf einen Geschäftsanschluß hinde-uten. Einen vorläufigen Vorschlag für eine solche Liste zeigt die Abbildung 1. Legt man diese Liste zugrunde, so lassen sich je nach Bundesland zwischen 7.5% (NRW) und 12.2% (Sachsen) als Geschäftsanschluß klassifizieren[21]. Verwendet man keine "randomized last digit"-Stichproben, so können die entsprechenden Zeilen vom Filterprogramm gelöscht werden.

---

19 Ein Filterprogramm sollte daher prüfen, ob der Telefonnummernstring tatsächlich nur Ziffern und "-" enthält, wobei keine zwei "-" aufeinander folgen dürfen.

20 Hierbei muß beim Filterprogramm beachtet werden, daß Folgenummern bei Geschäftsanschlüssen häufig durch leere Namensfelder oder durch ein spezielles Symbol (z.B. "o") nach einem gültigen Namen gekennzeichnet sind.

21 Diese naheliegende Idee findet sich auch bei ***Marhenke*** (1996). Allerdings gibt ***Marhenke*** nicht die Liste an; er klassifiziert anhand seiner Kriterien für Kassel (und für die D-Info 2.0 CD) 1.6% der Anschlüsse als Geschäftsnummern (eigene Berechnung, R.S.); an anderer Stelle gibt er "mindestens 7%" an. Anhand der Kriterien der hier vorgeschlagenen Liste werden 9.2% der Anschlüsse als Geschäftsnummer klassifiziert.

Abbildung 1: Die verwendeten 75 Schlüsselworte zur Identifikation vermutlicher Geschäftsanschlüsse

'AG', ' e.V.', 'Fax', '&', 'Automobil', 'BTX', 'Btx', 'büro', 'Club', 'club', 'Center', 'center', 'Dienst', 'dienst', 'emeinschaft', 'ermietung', ' GdBR', 'Gesellschaft', 'Immobilien', 'Kfz', 'mbH', 'MBH', 'Mobiltelefon', 'Service', 'service', 'Studio', 'team', 'technik', 'Telefax', 'Verband', 'verband', 'verein', 'Verein', 'Versicherung', 'versicherung', 'Vertrieb', 'vertrieb', ' KG ', ' s. unter', 'o ', 'praxis', 'Praxis', 'Büro', 'Gebäude', 'Agentur', 'edaktion', 'deutschen ', 'Export', 'Antiqu', 'Industrie', 'Apotheke', 'Krankenhaus', 'Werkstatt', 'Stiftung', 'Videothek', 'nstitut', 'Autohaus', 'Arbeitskreis', 'agentur', 'betrieb', 'Betrieb', 'Verwaltung', 'verwaltung', ' Pension', 'direktion', 'Direktion', 'Zentrale', 'zentrale', 'Deutsche ', 'Deutscher ', 'Deutsches ', 'Aktiengesellschaft', 'achhandel', ' oHG', ' OHG'

Wählt man nur wenige Gemeinden aus, so zeigt sich rasch, daß die exportierten Dateien eine Reihe von Nummern enthalten, die in nicht ausgewählten Gemeinden liegen. Hierbei handelt es sich meistens um Querverweise, also z.B. Hauptniederlassungen. Solche Nummern lassen sich mit einem Filterprogramm problemlos beseitigen: Man schreibt nur die Nummern heraus, deren Vorwahl einer der ausgewählten Gemeinden entspricht.

Wählt man hingegen viele oder gar alle Gemeinden aus, so entsteht schon allein durch die Querverweise das Problem von Mehrfacheinträgen der gleichen Telefonnummer. Solche Nummern besäßen eine höhere Auswahlwahrscheinlichkeit. Die einzig saubere Lösung wäre daher eine Löschung der Mehrfacheinträge.

Hierzu benötigt man entweder nach Nummern sortierte Dateien oder eine Häufigkeitsauszählung aller Telefonnummern. Mit Standardsoftware sind beide Aufgaben zumindest für den gesamten Datenbestand mit PCs kaum zu schaffen[22]. Angesichts dieses Problems wird man eher zu der Annahme neigen wollen, daß die geringe Anzahl von Mehrfacheinträgen und die daraus resultierende Vergrößerung der Auswahlwahrscheinlichkeit dieser Nummern einen vernachlässigbaren Fehler darstellt. Dies gilt um so stärker, da diese Nummern vermutlich eher ausschließlich Geschäftsnummern darstellen, die später ohnehin nicht berücksichtigt werden.

Das Ausmaß der Mehrfacheinträge ist aber erstaunlich hoch, wie etwas mühselige Zählungen zeigen. Im gesamten Datensatz mit ca. 33.606.000 gefilterten Einträgen (ohne Mobil

22 Bei kleineren Datensätzen sind solche Prüfungen natürlich auch mit PCs möglich. Für diese Arbeit wurden die Sortierungen mit einem Mehrprozessor-Unix-System (SGI Power Challenge) durchgeführt; selbst hiermit wurden für 34 Millionen Records mehr als 10 CPU-Minuten benötigt. Die Bestimmung von Doppeleinträgen in der sortierten Datei ist z.B. durch das kurze AWK-Programm DUPS (siehe oben) möglich.

telefone und Datenfehler) sind ca. 32.753.000 Einträge Unikate. 853.000 Einträge finden sich mehrfach (ca. 2.5%). Hierbei handelt es sich um ca. 685.500 verschiedene Einträge. 87% aller Mehrfacheinträge sind Doppeleinträge; 9% kommen dreimal, 2% viermal vor. Die Anzahl der Einträge einer Nummer reicht bis zu 598; wobei die hohen Anzahlen immer Telekom-Service-Nummern entsprechen.

Betrachtet man das Problem der Mehrfacheinträge etwas genauer, so zeigen sich einige Probleme, die zu leicht verzerrten Auswahlen führen können. Nimmt man z.B. alle 928.104 Nummern, die von D-Info 3.0 als "Hamburger" Telefonnummern exportiert werden, so sind ca. 2% der Nummern mehr als einmal vorhanden. Einige Nummern kommen bis zu 30 mal in der Datei vor; allerdings sind 88% der Mehrfacheinträge lediglich Doppeleinträge. Filtert man anhand der oben beschriebenen Kriterien diejenigen Nummern heraus, die vermutlich Geschäftsanschlüsse darstellen, so beträgt der Anteil der mehrfach gelisteten Nummern immer noch 1.1%. Neben verbleibenden Geschäftsanschlüssen sind dies vor allem Paare, die unter zwei verschiedenen Namen eingetragen wurden (93% aller gefilterten Mehrfacheinträge sind Doppeleinträge). Daraus läßt sich der Schluß ziehen, daß z.B. der Anteil der zusammenlebenden, aber unverheirateten Paare in Telefon-Stichproben vermutlich höher geschätzt wird als bei anderen Verfahren.

Stichprobendesignvarianten

Für die Auswahl einer Stichprobe aus einer Telefon-CD gibt es eine Reihe verschiedener Möglichkeiten. Dies gilt in besonderem Maße für Stichproben aus dem gesamten Datenbestand. In diesem Fall liegt es nahe, mehrstufige Auswahlverfahren zu verwenden, da abgesehen von möglichen statistischen Vorteilen mehrstufige Verfahren die Größe der Einzeldateien erheblich reduzieren. Bei allen mehrstufigen Verfahren muß zunächst die Entscheidung getroffen werden, ob als Maß für die Größe einer Primäreinheit (z.B. einer Gemeinde) die Zahl der Einwohner (bzw. der Wahlberechtigten) oder die Zahl der Telefonanschlüsse (bzw. die Zahl der vermutlichen Privatanschlüsse) verwendet werden soll. Beides sche-int in der BRD üblich zu sein. Falls die Grundgesamtheit als "allgemeine Bevölkerung" definiert ist, so ist die Verwendung der Zahl der Einwohner bzw. die Zahl der Wahlberechtigten eher korrekt. Wird die Grundgesamtheit allerdings als die in Privathaushalten mit Telefonanschluß lebende Bevölkerung definiert, so ist die Verwendung der Zahl der Anschlüsse korrekt. Wird die Grundgesamtheit wie üblich als "allgemeine Bevölkerung" definiert und trotzdem die Zahl der Anschlüsse als Maß für die Größe der Primäreinheiten verwendet, so muß angenommen werden,daß die Anschlußquote über alle Primäreinheiten konstant ist. Dies ist definitiv falsch. Allerdings wurde bisher weder eine empirische Studie noch eine Simulation zu den tatsächlichen Konsequenzen dieser Vorgehensweise publiziert Hier könnte eine Ursache für mögliche Unterschiede in den Ergebnissen verschiedener Datenerhebungsinstitute liegen.Am einfachsten ist natürlich die Verwendung der Zahl der An-

schlüsse als Maß für die Größe der Primäreinheiten, da dann alle notwendigen Daten prinzipiell der CD entnommen werden können. Im folgenden werden daher - ungeachtet der statistischen und inhaltlichen Konsequenzen - zunächst Verfahren auf der Basis dieser Vorgehensweise erörtert.

Geschichtete Stichproben

Besonders naheliegend ist die Verwendung einer nach Bundesländern geschichteten Stichprobe. Die D-Info 3.0 CD erlaubt zwar indirekt den Export nach einzelnen Bundesländern[23], die resultierenden Dateien sind allerdings fehlerhaft. Vergleicht man die Zahl der Anschlüsse der einzelnen Bundesländerdateien mit der Gesamtzahl der Anschlüsse, so ergibt die zusammengeführte Bundesländerdatei 37.1 Millionen Anschlüsse, die Gesamtdatei enthält hingegen nur 34.3 Millionen (ungefilterte) Anschlüsse. Folglich werden 2.8 Millionen Anschlüsse mehrfach exportiert. Die Ursache liegt vor allem in der Zuordnung Hamburgs zu Schleswig-Holstein und zu Niedersachsen sowie Bremens zu Niedersachsen. Korrigiert man diese gravierenden Fehlzuordnungen, so verbleiben immer noch 553.000 überzählige Einträge. Zu ähnlichen Zahlen gelangt man über die Analyse der Vorwahlnumme-rn. Bereinigt man die exportierten Dateien der einzelnen Länder um die Sondernummern (0130, 0161, 0190 etc.) so zeigen sich 5247 verschiedene Vorwahlnummern, davon 5198 mit mehr als 10 Anschlüssen. 67 Vorwahlen mit mehr als 1000 Anschlüssen erscheinen mehrfach. Insgesamt finden sich fast 561.000 Einträge in mehr als einem Bundesland. Dies können Querverweise sein (wie z.B. auf Hauptniederlassungen), aber auch Vorwahlen in mehr als einem Bundesland. Beispiele für "berechtigte" Mehrfachzuordnungen sind einerseits Mannheim und Ludwigshafen, andererseits Ulm. Es ergeben sich aber auch schlicht falsche Mehrfachzuordnungen, z.B. bei 02224 (Bad Honnef) und 0241 (Düren). Zusammenfassend kann daher festgestellt werden, daß sich zwar aus den Daten der D-Info 3.0 CD eine nach Ländern geschichtete Stichprobe gewinnen ließe, allerdings nur mit sehr hohem Arbeitsaufwand. Für praktische Stichproben empfiehlt sich daher ein anderes Verfahren.

PPS-Stichproben

Eine andere Möglichkeit der Auswahl für eine bundesweite Studie besteht darin, die CD in gleicher Weise wie normale Telefonbücher als Auswahlgrundlage zu verwenden. In diesem Fall würde man vermutlich zu einem PPS-Design ("probability proportional to size", vgl. ***Schnell, Hill, Esser*** 1995:268-270) greifen: Auswahl der Primäreinheiten (z.B. Gemeinden oder Vorwahlen) entsprechend ihrer Größe, innerhalb der Primäreinheiten erfolgt dann die

23 Der Export einzelner Bundesländer ist nicht direkt möglich, sondern dadurch, daß man zunächst willkürlich einen Ort des jeweiligen Bundeslandes auswählt und dann im folgenden Exportdialog des Abfrageprogramms der D-Info CD die entsprechende Option selektiert.

Auswahl einer konstanten Zahl von Sekundäreinheiten (Haushalte) über eine einfache Zufallsstichprobe. Praktisch lassen sich PPS-Stichproben entweder mittels kommerzieller Software (***Frankel, Spencer*** 1990), Makros in Datenanalysesystemen (***Lehtonen, Pakhinen*** 1994:55-56) oder auch durch einige Zeilen AWK mittels des sogenannten Kumulationsverfahrens (vgl. ***Schnell, Hill, Esser*** 1995:445-447) ziehen[24]. Das Problem bei PPS-Stichproben liegt darin, daß man die Größe der Primäreinheiten benötigt. Die Zahl der zu einer Vorwahlnummer gehörenden Einheiten läßt sich aus der D-Info 3.0 CD auszählen, solange es sich um kleinere Datenbestände handelt. Wie oben geschildert, gibt es hier zwar das Problem der Mehrfacheinträge, daß man aber wohl bei den meisten Fragestellungen ohne große Folgen ignorieren kann. Zumindest beim gesamten Datensatz wirft aber die Häufigkeitsauszählung des gesamten Datenbestandes größere technische Probleme auf[25]. Falls man nicht die Zahl der Anschlüsse als Maß für die Größe der Primäreinheiten verwenden will, so kann man z.B. auf die Einwohnerzahlen der Gemeinden oder wie beim ADM-Design auf die Größe der Wahlbezirke rekurrieren[26]. Da das ADM-Ziehungsband nicht öffentlich zugänglich ist, müßten die Unterlagen über die Größe der Stimmbezirke beschafft werden. Eine Datei mit diesen Angaben für alle 80.000 Wahlbezirke ist beim Bundeswahlleiter im Statistischen Bundesamt gegen eine Gebühr von ca. 4.000 DM erhältlich. In einem weiteren Arbeitsschritt müßten kleinere Wahlbezirke zusammengefaßt und dann mit den Vorwahlen für die zusammengefaßten Einheiten zusammengeführt werden. Beides wäre mit den Aggregierungs- und Filemergeroutinen von Datenanalysesystemen wie SPSS oder STATA mit jeweils einem Kommando prinzipiell möglich[27]. Ähnliches gilt für die Einwohnerzahlen der Gemeinden, wobei nur für die größeren Gemeinden diese Zahlen problemlos beschaffbar sind.

Nachdem im ersten Schritt diese Vorwahlnummern mittels PPS ausgewählt wurden, muß im zweiten Schritt eine fixe Zahl von Einträgen für jede Vorwahl ausgewählt werden. Es empfiehlt sich, die Zahl der Einträge pro Vorwahl klein zu wählen, um den Klumpen-Effekt gering zu halten. Eine **realisierte** Zahl von weniger als 10 Einträgen bei ca. 250 verschiedenen Vorwahlen wäre eine übliche Größe.

---

24 Die letzte Methode hat den Vorteil, daß man mit nichtaggregierten Datenfiles beliebiger Größe arbeiten kann. Außerdem ist das Programm kostenneutral. Die PPS-AWK-Implementierung in DOS ist allerdings etwas mühsam, da AWK keine Sortierroutine enthält sondern auf das Betriebssystem zurückgreift.

25 Mit den Standardprogrammen ist diese Aufgabe nicht zu leisten. Mit Turbo-Pascal läßt sich die Gesamtdatei nicht bearbeiten. In AWK könnte man dieses Problem mit assoziativen Arrays bearbeiten; dies scheitert aber an einem Memory-Limit dieser Arrays in MSDOS-AWK. Am einfachsten läßt sich dieses Problem mit einem C oder FORTRAN-Compiler angehen.

26 Diese Anlehnung an das ADM-Design erwähnt auch ***Zeh*** (1987:344).

27 Da die Stimmbezirksdatei die Gemeindekennziffer enthält, müßten alle Vorwahlnummern für alle einzelnen Gemeindekennziffern in einer Datei vorliegen. Eine solche Datei scheint nicht allgemein verfügbar zu sein.

Liegt eine Datei vor, in der nach Vorwahlen sortierte Telefonnummern vorhanden sind, lassen sich die Einträge z.B. entweder durch eine einfache Zufallsstichprobe oder durch ein systematisches Auswahlverfahren (z.B. Ziehen jedes n-ten Eintrags) ziehen. Beide Methoden erfordern nur wenige AWK-Zeilen (siehe weiter unten). Liegt keine solche Datei vor, bleibt als praktisch gangbarer Weg zunächst nur der klassische Weg über die gedruckten Telefonbücher. Dann muß für jede PPS-ermittelte Vorwahl wieder manuell mit einem systematischen Verfahren aus dem Telefonbuch gezogen werden. Hierbei muß beachtet werden, daß das Druckbild der Telefonbücher sich unterscheidet, so daß die Anzahl der Eintr-äge pro Seite unterschiedlich ist. Korrekt wäre ein systematisches Verfahren also erst dann, wenn die Zahl der Einträge pro Vorwahl insgesamt bekannt oder jeweils geschätzt würde-[28]. Interessanterweise gibt aber die Telekom-CD die Zahl der Einträge pro Vorwahl in der Bildschirmmaske nach Auswahl eines Ortes aus. Damit kann diese CD direkt als Basis eines manuell durchgeführten systematischen Auswahlsverfahrens dienen; allerdings muß man für dieses Verfahren den Schreibtisch nicht mehr verlassen.

Übersicht 3: AWK-Programm SAMPLE.AWK: Zieht eine Zufallsstichprobe der Größe n aus N Eingabezeilen

```
# SAMPLE.AWK: select a random sample
BEGIN {
 if (ARGC != 4)  {
   print "Syntax:  AWK sample filename N n "
   print "where N=number of input  lines"
   print "      n=number of output lines"
   print "Example: AWK sample frame.lst 2000 10"
   exit }
 ni=0;srand();n=ARGV[2];m=ARGV[3];ARGC=2 }
{  if ( rand() < (m-ni)/(n-NR+1) ) { print $0 ; ni=ni+1 } }
END {  if ((NR != n) && (NR > 1)) \
 print "Number of units is ",NR," not ",n,\
               ". Selection invalid." }
```

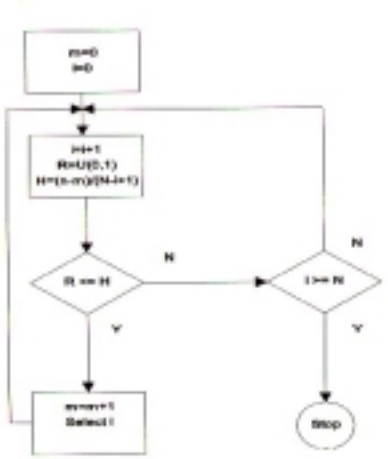
Sollten zukünftig dieCDs diese Bestimmung der Zahl der Einträge pro Vorwahl nicht mehr ermöglichen, wäre die Ziehung weiter erschwert und mit erheblichen Aufwand verbunden-[29]. Aufgrund der Kosten und des Arbeitsaufwandes scheiden alle PPS-Verfahren, die nicht auf der Auszählung der CD basieren,daher zumindest für "Lehrforschung" praktisch aus. Verwendet man aber die CD-Zählung als Basis der PPS-Stichprobe, so muß man (wie oben erwähnt) entweder die Grundgesamtheit ungewöhnlich definieren oder die unklaren Konsequenzen der falschen Annahme konstanter Anschlußquoten tragen. Aus diesem Grund

28 vgl. ***Frey, Kunz, Lüschen*** (1990:87).

29 Erfahrungsgemäß brauchen zwei studentische Hilfskräfte ca. 1 Woche, um auf diese Art eine Zufallsstichprobe von 2000 Zahlen zu ziehen. ***Frey, Kunz, Lüschen*** (1990:87) erwähnen ohne Seitenangabe den Vorschlag von ***Blankenship*** (1977), die Auswahl durch eine Schablone vorzunehmen. Bereitet man entsprechend für jedes der vorliegenden Telefonbücher eine eigene Schablone vor, so ist der Auswahlvorgang wesentlich rascher durchführbar.

scheint die einstufige Auswahl aus dem gesamten Datenbestand für bundesweite Surveys die einfachste, schnellste und am wenigsten problematische Lösung darzustellen. Für regional begrenzte Studien ist dies ohnehin fast immer die Methode der Wahl.

Abbildung 2: Sequentielle Ziehung ohne Zurücklegen (***Kennedy, Gentle*** 1980:238)



Einstufige Auswahlverfahren

Ein einfacher Algorithmus zur Auswahl von exakt n Elementen aus einer Liste von N Elementen findet sich bei ***Kennedy, Gentle*** (1980:238,vgl.Abbildung 1)[30]. Das entsprechende AWK-Programm SAMPLE.AWK findet sich oben[31]. Diese Methode setzt voraus, daß die Anzahl (N) der Elemente der Grundgesamtheit vor der Ziehung bekannt ist. Meist läßt sich diese Zahl recht einfach bestimmen (z.B. mit dem AWK-Oneliner "LINES.AWK", siehe T-ext 1). Bei großen Datenmengen wie beim gesamten Datenbestand der D-Info CD kann die Ermittlung von N aber mit Schwierigkeiten behaftet sein. In solchen Fällen empfiehlt sich die Verwendung eines Algorithmus, der nicht die Kenntnis von N voraussetzt[32]. Au-fgrund der unterschiedlichen Ausfallursachen bei Telefonsurveys (vgl. ***Schnell*** 1997:116-

30 Im Flußdiagramm ist i der Laufindex, m ist die Zahl der bisher ausgewählten Einheiten, n die Gesamtzahl der Einheiten in der Auswahlgrundlage. U(0,1) bezeichnet einen Pseudo-Zufallszahlengenerator, der gleich verteilte Zahlen im Intervall zwischen Null und Eins produziert. Dieser Algorithmus wird auch in einer Reihe kommerzieller Statistikpakete verwendet.

31 Da die in den üblichen Compilern eingebauten Pseudo-Zufallszahlengeneratoren meist kaum dokumentiert sind (vgl. hingegen ***Dohmann, Falk, Lessenich*** 1991) und auch die besten Generatoren mit Problemen behaftet zu sein scheinen (vgl. ***Ferrenberg, Landau, Wong*** 1992), empfiehlt sich im allgemeinen eher die Verwendung explizit programmierter Generatoren. In vielen Anwendungen hat sich bisher z.B. der portable Pseudo-Zufallszahlengenerator von ***Park, Miller*** (1988) bewährt. Das Programm findet sich z.B. bei ***Schnell, Hill, Esser*** (1995:443). Programmiert man den Generator unter AWK, so wird das Programm etwas unübersichtlich (Verwendung von Subroutinen) und langsam. Für praktische Zwecke würde man dieses Programm daher eher in einer allgemeinen Programmiersprache schreiben.

32 Ein Beispiel hierfür findet sich in der Arbeit von ***Pinkham*** (1987).

127) sollte eine wesentlich höhere Anzahl von Nummern (n) ausgewählt werden, als realisiert werden soll. Nach den Erfahrungen in einer Reihe von RLD-Surveys der "allgemeinen Bevölkerung" sollte diese Zahl bei ca. 300% der angestrebten Stichprobengröße liegen.

Die Zukunft der Telefonsurveys auf der Grundlage der Telefon-CDs

Das Erscheinen der Telefon-CD D-Info 3.0 hat die Möglichkeiten der Stichprobenziehung für Telefonsurveys vereinfacht. Durch den Export einer ASCII-Datei für das Untersuchungsgebiet läßt sich mit den hier angegebenen Programmen und Hilfsmitteln relativ problemlos eine Zufallsstichprobe der eingetragenen Telefonnummern ziehen bzw. eine RLD-Stichprobe erzeugen. Liegt einmal eine bereinigte ASCII-Datei der Telefonnummern vor, so kann eine Stichprobe in wenigen Minuten gezogen werden. Leider ist aufgrund von Datenschutzbemühungen zu erwarten, daß eine solche Auswahlgrundlage nicht dauernd zur Verfügung stehen wird. In diesem Fall bleiben nur wenige Möglichkeiten:

- Entweder man rekurriert zumindest für die Generierung der Ausgangsnummern des RLD-Verfahrens wieder auf die gedruckten Telefonbücher;
- man verwendet die alten CDs als Ausgangspunkt des RLD-Verfahrens
- oder man verwendet die Länge der eigentlichen Telefonnummer (ohne Vorwahl) als Ausgangspunkt eines "Random-Digit-Dialings" (RDD), bei dem alle Ziffern aus einem Pseudo-Zufallszahlengenerator gezogen werden.

Der Rückgriff auf die gedruckten Telefonbücher ist mühsam, aber sicher auch in Zukunft möglich[33]. Die weitere Verwendung der alten CDs müßte voraussetzen, daß sich die Häufigkeiten der Anfangsdigits des Telefonnetzes nicht verändert. Die Verwendung der Länge der Telefonnummern (ohne Vorwahl) als Ausgangsbasis für eine vollständige Zufallsgenerierung der Nummern setzt die Konstanz der relativen Häufigkeit unterschiedlicher Telefonnummernlängen voraus. Diese wird sich vermutlich langsamer ändern als die relative Häufigkeit der Anfangsdigits. Damit wäre RDD gegenüber RLD auf der Basis älterer CDs geraume Zeit verwendbar - bis entweder neue CDs verfügbar werden oder die Telekom (endlich) Daten über die Häufigkeit einzelner Anfangsdigits und die Länge der Telefonnummern zur Verfügung stellt. Allerdings wäre es auch möglich, daß ein immer größerer Anteil der Telefonbesitzer dazu übergeht, ihre Anrufbeantworter zum Aussieben der Anrufer zu verwenden - falls sie nicht ohnehin ihre intelligenten ISDN-Telefone so programmieren, daß sie keine Anrufe von unbekannten Anrufern akzeptieren. Dies wäre das definitive Ende der Telefonsurveys als Zugangsmethode zur "allgemeinen Bevölkerung".

33 Solange die Telekom-CD die Anzahl der Einträge pro Vorwahl ausgibt, kann diese CD zusammen mit der Stimmbezirksdatei als Grundlage für ein auf der letzten Stufe manuell durchgeführtes PPS-Verfahren verwendet werden (siehe oben).

Literatur

Aho, A.V., Kernighan, B.W., Weinberger, P.J. (1988):
The AWK Programming Language, Reading/Mass. (Addison-Wesley).

Blankenship, A.B. (1977):
Professional Telephone Surveys, New York (McGraw-Hill).

Borland GmBH (1992):
Turbo Pascal 7.0, München.

Chapman, G. u.a. (1990):
Monty Pythons Flying Circus: Just the Words, Vol.2, London.

Dohmann, B., Falk, M., Lessenich, K. (1991):
The Random Number Generators of the Turbo Pascal Family; in: Statistical Software Newsletter, 12, 1, S.129-132.

Ferrenberg, A.M., Landau, D.P., Wong, Y.J. (1992):
Monte Carlo Simulations: Hidden errors from "good" random number generators; in: Physical Review Letters, 69, 23, 3382-3384.

Frankel, M.R., Spencer, B.D. (1990):
Sample: A Supplementary Module for Systat and Sygraph, Evanston, Ill. (SYSTAT, Inc.).

Frey, J.H., Kunz, G., Lüschen, G. (1990):
Telefonumfragen in der Sozialforschung, Opladen.

Fuchs ,M. (1994):
Umfrageforschung mit Telefon und Computer, Weinheim (Beltz).

Häder, S. (1996):
Wer sind die Nonpubs? Zum Problem anonymer Anschlüsse bei Telefonumfragen; in: ZUMA-Nachrichten 39, S.45-68.

Hutchinson, H. (ed.)(1996):
Ci-3 User Manual, Version 1.1, SAWTOOTH Software, Sequim, WA.

Kennedy, W.J., Gentle, J.E. (1980):
Statistical Computing, New York (Marcel Dekker).

Lavrakas, P.J.(1987):
Telephone Survey Methods, Beverly Hills (Sage).

Lehtonen, R., Pahkinen, E.J. (1994):
Practical Methods for Design and Analysis of Complex Surveys, Chichester (Wiley).

Marhenke, W. (1996):
Telefonanschlußdaten als Auswahlgrundlage, Vortrag anläßlich des ZUMA-Symposiums "Vergleich von Stichprobenverfahren", Mannheim; wird 1997 in einer ZUMA-Buchpublikation erscheinen.

Park, S.K., Miller, K.W.(1988):
Random Number Generators: Good Ones are Hard to Find; in: Communications of the ACM, October 1988 (Volume 31, Number 10).

Pinkham, R.S. (1987):
An Efficient Algorithm for Drawing a Simple Random Sample; in: Applied Statistics, 36, 3, S.370-372.

Schnell, R.(1997):
Nonresponse in Bevölkerungsumfragen, Opladen (Leske+Budrich).

Schnell, R., Hill, P.B., Esser, E. (1995):
Methoden der empirischen Sozialforschung, 5. Auflage, München (Oldenboug).

Wall, L., Schwartz, R.L. (1993):
Programmieren in perl, München (Hanser).

Zeh, J.(1987): Stichprobenbildung bei Telefonumfragen; in: Angewandte Sozialforschung, 14, 4, S.337-347.

Benutzerdefinierte Design-Matrizen in log-linearen Analysen: Realisierungsmöglichkeiten in den SPSS-Prozeduren GENLOG und LOGLINEAR

von Steffen M. Kühnel [1]

Zusammenfassung

Der Anwendung log-linearer Modelle in der Sozialforschung steht oft die Vorstellung entgegen, daß diese Modelle recht kompliziert und daher kaum zu interpretieren seien. Das Verständnis für log-lineare Analysen wird erleichtert, wenn die Verwandtschaft zur multiplen Regression mit nominalskalierten Prädiktoren gesehen wird. Gleichzeitig kann so auch die Bedeutung der sogenannten Design-Matrix nahegebracht werden. Die volle Flexibilität log-linearer Modelle wird nämlich erst durch die Formulierung benutzerdefinierter Design-Matrizen erreicht. Anhand von Beispieldaten aus dem ALLBUS 1996 wird gezeigt, wie sich bei Anwendung der SPSS-Prozeduren GENLOG oder LOGLINEAR log-lineare Analysen mit benutzerdefinierten Design-Matrizen realisieren lassen.

Abstract

Applications of log-linear modelling are sometimes prevented by the impression that this technique is not user-friendly. Nevertheless, log-linear modelling is nothing more than multiple regression of the logarithms of cell counts on categorical predictors. Within this view the importance of the design matrix is easy to understand. The specification of user-defined design matrices within log-linear models allows for very flexible analyses of categorical data. It is shown how such analyses can be done using the SPSS procedures GENLOG or LOGLINEAR. An empirical example is given based on data from the ALLBUS 1996.

[1] Anschrift des Autors: Prof. Dr. ***Steffen M. Kühnel***, Justus-Liebig Universität Gießen, Fachbereich Gesellschaftswissenschaften, Institut für Politikwissenschaft, Karl-Glöcker-Str. 21E, 35394 Gießen

Log-lineare Modelle ermöglichen die multivariate Zusammenhangsanalyse bei kategorialen Daten. Die vielfältigen Möglichkeiten dieser Modellklasse lassen sich erst dann voll ausnutzen, wenn der Anwender die Möglichkeit hat, benutzerspezifische Design-Matrizen zu definieren. Es ist leider kaum bekannt, daß log-lineare Modelle mit benutzerdefinierten Design-Matrizen auch mit den SPSS-Prozeduren GENLOG oder LOGLINEAR geschätzt werden können. Im vorliegenden Beitrag möchte ich anhand eines einfachen Beispiels aus dem ALLBUS 1996 zeigen, wie hierbei vorzugehen ist. Für Leser, die mit der Logik log-linearer Modelle nicht so vertraut sind, will ich zunächst die Grundidee der log-linearen Analyse vorstellen.[2]

1. Die Logik log-linearer Tabellenanalysen

In der bekannten linearen Regression ergeben sich die Werte einer abhängigen Variable Y als lineare Funktion der Werte von erklärenden Variablen X_1, X_2, ..., und der Residual variable E:

$$Y = b_0 + b_1 \cong X_1 + b_2 \cong X_2 + ... + E \ .$$

Ganz analog kann auch die log-lineare Analyse einer mehrdimensionalen Tabelle als ein spezielles Regressionsmodell aufgefaßt werden. Dabei bilden die logarithmierten Häufigkeiten der Tabellenzellen die Werte der abhängigen Variable. Vorhergesagt werden diese Werte durch eine lineare Funktion von erklärenden (Design-) Variablen. Diese Sichtweise verdeutlicht auch die Bezeichnung "log-linear".

Über diese Analogie zur linearen Regression läßt sich die Logik der log-linearen Analyse und die Interpretation der Modellparameter relativ leicht nachvollziehen. Als Beispiel soll im folgenden der bivariate Zusammenhang zwischen der Wahlbeteiligung und der wahrgenommenen Bürgernähe von Politikern mit log-linearen Modellen untersucht werden. Im ALLBUS 1996 (ZA-Studiennummer 2800) finden sich hierzu zwei Fragen. Die perzipierte Bürgernähe von Politikern (V20) wird über ein Item aus der Anomia-Skala erfaßt, bei dem nach der Zustimmung zu der Meinung *"Die meisten Politiker interessieren sich in Wirklichkeit gar nicht für die Probleme der einfachen Leute"* gefragt wird. Als Antwortmöglichkeiten werden die Kategorien *"Bin derselben Meinung"* (Kode: 1), *"Bin anderer Meinung"* (Kode: 2) und *"Weiß nicht"* (Kode: 3) berücksichtigt. Die Wahlbeteiligung (V326) wird über die berichtete Wahlbeteiligung bei der letzten Bundestagswahl operationalisiert. Tabelle 1 gibt die in den ALLBUS-Daten vorzufindenden bivariaten Antworthäufigkeiten auf die beiden Fragen wieder. Neben den absoluten Häufigkeiten sind auch die logarith-

[2] Aus Platzgründen kann hier keine umfassende Einführung in die log-lineare Datenanalyse gegeben werden. Hierzu sei auf entsprechende Monographien verwiesen, etwa das kürzlich erschienene Lehrbuch von ***Andreß***, ***Hagenaars*** und ***Kühnel*** (1997) zur kategorialen Datenanalyse.

mierten Häufigkeiten aufgeführt. Der Wert 7.6917 der ersten Tabellenzelle ist beispielsweise der natürliche Logarithmus der Häufigkeit 2190 (ln(2190)≈7.6917).

Tabelle 1: Wahlbeteiligung und Bürgernähe von Politikern

Wahlbeteiligung bei letzter Bundestagswahl (V326)	Politiker nicht an Problemen interessiert (V20)		
	ja (1)	nein (2)	weiß nicht (3)
ja (1)	2190 *7.6917*	451 *6.1115*	180 *5.1930*
nein (2)	547 *6.3044*	75 *4.3175*	68 *4.2195*

Erster Wert: absolute Häufigkeit, zweiter Wert (kursiv): logarithmierte Häufigkeit
(Quelle: ALLBUS 1996)

In der log-linearen Analyse der Tabelle bilden die sechs logarithmierten Zellenhäufigkeiten die Werte der abhängigen Variable. Erklärt werden diese Häufigkeiten durch die Variablen, die die Tabelle definieren, hier also durch die Einschätzung der Desinteressiertheit von Politikern (V20) und durch die Wahlbeteiligung (V326). Beide Variablen werden zunächst als nominalskaliert aufgefaßt. In der linearen Regression kann eine nominalskalierte unabhängige Variable nicht direkt in die Modellgleichung aufgenommen werden. Die Berücksichtigung solcher Erklärungsgrößen erfolgt statt dessen indirekt über sogenannte Design-Variablen.[3] Gleiches gilt auch für log-lineare Modelle.

Die Umsetzung nominalskalierter Variablen in Design-Variablen kann nach unterschiedli-chen Regeln erfolgen. Gemeinsam ist allen Regeln, daß aus einer Variable mit insgesamt K verschiedenen Ausprägungen maximal K-1 Design-Variablen gebildet werden können. Im Beispiel werden also die drei Kategorien der Bürgernähe von Politikern in zwei DesignVariablen und die zwei Kategorien der berichteten Wahlbeteiligung in eine DesignVariable umgewandelt.[4] In der linearen Regression wird bei einer solchen Transformation häufig die *Dummykodierung* verwendet. Dabei werden den einzelnen Kategorien der Ausgangsvariable dichotome 0/1-kodierte Variablen zugeordnet, die immer dann den Wert '1' aufweisen, wenn

[3] Statt von 'Design-Variablen' wird meist von 'Dummy-Variablen' gesprochen. Da im folgenden von *Dummykodierung* bzw. *Effektkodierung* der Design-Variablen die Rede sein wird, verwende ich hier den neutraleren Ausdruck 'Design-Variable'.

[4] Da irgendeine der Ausprägungen einer Variable notwendigerweise realisiert wird, läßt sich das Auftreten einer Kategorie bei Kenntnis des Auftretens der übrigen Kategorien perfekt vorhersagen. Wird für jede Kategorie eine eigene Design-Variable erzeugt, enthält daher eine dieser Variablen keine zusätzlichen Informationen. Technisch gesprochen bestünde eine perfekte Multikollinearität unter der Prädiktoren.

die betreffende Kategorie bei einem Fall vorkommt. Die Kategorie, für die keine eigene 0/1-kodierte Design-Variable spezifiziert wird, wird als *Referenzkategorie* bezeichnet. Weist die nominalskalierte Ausgangsvariable bei einem Fall den Wert der Referenzkategorie auf, haben alle 0/1-kodierten Design-Variablen den Wert '0'. In der Varianzanalyse wird dagegen oft die *Effektkodierung* eingesetzt. Der Unterschied zur Dummykodierung liegt darin, daß die Referenzkategorie der Dummykodierung hier den Wert '-1' aufweist. Tabelle 2 zeigt die beiden Kodierungsmöglichkeiten für die Variablen aus Tabelle 1. Referenzkategorie bei der Einschätzung des Desinteresse von Politikern ist die Antwort "weiß nicht" (V20=3) und bei der Frage nach der Wahlbeteiligung die Antwort "nein" (V326=2).

Tabelle 2: Transformation der Modellvariablen in dummy- und effekt kodierte Design-Variablen

	Politiker sind desinteressiert (V20):			Wahlbeteiligung (V326)	
	ja (1)	nein (2)	weiß nicht (3)	ja (1)	nein (2)
Dummykodierung:					
Erste Design-Variable	1	0	0	1	0
Zweite Design-Variable	0	1	0		
Effektkodierung:					
Erste Design-Variable	1	0	-1	1	-1
Zweite Design-Variable	0	1	-1		

Mit Hilfe der Design-Variablen können nun die Vorhersagegleichungen für das log-lineare Modell formuliert werden. Zunächst soll ganz analog zur linearen Regression ein Modell geschätzt werden, bei dem die logarithmierten Häufigkeiten nur durch die Regressionskonstante und die Design-Variablen prognostiziert werden. Bei der Dummy-Kodierung ergeben sich dann folgende Vorhersagegleichungen:[5]

V326	V20	Y	≈	$\hat{\mathbf{Y}}$	=	$b_0 \cdot 1$	+	$b_1 \cdot D_1$	+	$b_2 \cdot D_2$	+	$b_3 \cdot D_3$
1	1	7.6917	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 1$	+	$b_2 \cdot 0$	+	$b_3 \cdot 1$
1	2	6.1115	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 1$	+	$b_3 \cdot 1$
1	3	5.1030	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 0$	+	$b_3 \cdot 1$
2	1	6.3044	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 1$	+	$b_2 \cdot 0$	+	$b_3 \cdot 0$
2	2	4.3175	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 1$	+	$b_3 \cdot 0$
2	3	4.2195	≈	?	=	$b_0 \cdot 1$	+	$b_1 \cdot 0$	+	$b_2 \cdot 0$	+	$b_3 \cdot 0$

[5] Um die Ähnlichkeit zur linearen Regression zu verdeutlichen, habe ich die Regressionskoeffizienten durch den Buchstaben "b" gekennzeichnet. In der üblichen auf ***L.A. Goodman*** zurückgehenden Notation werden für die Regressionsgewichte üblicherweise kleine Lambdas (λ) verwendet.

Die in den Gleichungen vorkommenden Werte der Design-Variablen ergeben sich aus der Position der entsprechenden Zelle in der ursprünglichen Häufigkeitstabelle (Tabelle 1). Um dies deutlich zu machen, sind links jeweils die Werte der Ausgangsvariablen V326 und V20 angegeben. Die Design-Variable D_1 hat immer dann den Wert '1', wenn die erste Kategorie der Bürgernähe von Politikern auftritt (V20=1). Ansonsten ist ihr Wert stets '0'. Die nächste Design-Variable D_2 bezieht sich auf die zweite Kategorie der Bürgernähe (V20=2) und die letzte Design-Variable D_3 auf die erste Kategorie der Wahlbeteiligung (V326=1). Die Werte der abhängigen Variablen (Y) sind die logarithmierten Besetzungen der Zellen von Tabelle 1. Sie werden durch Vorhersagewerte ($\hat{Y}$) prognostiziert, die sich mit Hilfe der Regressionskoeffizienten aus den Werten der Design-Variablen berechnen lassen. Die zunächst unbekannten Koeffizienten sind die Parameter des log-linearen Modells. Die Realisierungen der Design-Variablen bilden die *Design-Matrix*, die der Datenmatrix der erklärenden Variablen in der linearen Regression entspricht. Jede Zeile der Design-Matrix gibt die Werte aller Design-Variablen für eine Zelle der Ausgangstabelle an. Jede Spalte enthält alle Werte einer Design-Variable. Um auch die Regressionskonstante berücksichtigen zu können, enthält die erste Spalte der Design-Matrix eine Konstante mit dem Wert '1'.

Die vier Regressionskoeffizienten b_0 bis b_3 werden nun so bestimmt, daß die Vorhersagewerte den logarithmierten Häufigkeiten möglichst ähnlich sind. Anders als bei der im linearen Regressionsmodell üblichen Schätzung nach dem Kleinstquadrate-Kriterium wird bei log-linearen Modellen das Kriterium der Maximum-Likelihood-Methode verwendet (ML-Schätzung). Bei der ML-Schätzung werden die Koeffizienten so festgelegt, daß eine maximale Wahrscheinlichkeit besteht, daß die tatsächlich beobachteten Werte der abhängigen Variable - hier also die logarithmierten Zellenhäufigkeiten - als Funktion der Koeffizienten realisiert werden können.

Die Dummy-Kodierung der Design-Variablen wird bei der Prozedur GENLOG in SPSS verwendet, die in SPSS für Windows bei der Spezifikation eines log-linearen Modells über das Pull-Down-Menue aktiviert wird (***Norusis***/SPSS Inc., 1994). Alternativ kann die Schätzung der Modellparameter auch im Syntax-Fenster angefordert werden:

```
genlog V326 V20
   /print design freq estim /plot none
   /criteria = delta(0)
   /design V20 V326 .
```

Nach dem Prozedurnamen GENLOG werden zunächst die Variablen aufgeführt, die die zu analysierende Tabelle aufspannen. Im Beispiel sind dies die beiden ALLBUS-Variablen V20 und V326. Die hinter dem Schlüsselwort "/print" angegebenen Spezifikationen steuern die Ausgabe der Prozedur. Angefordert wird die Ausgabe der Design-Matrix ("design"), die beobachteten und vorhergesagten Zellenhäufigkeiten ("freq") und die Schätzungen der Regressionskoeffizienten ("estim"). Mit der Option "/plot none" wird die Ausgabe von

Abbildung 1: Ergebnisse der Parameterschätzung bei Dummykodierung (SPSS-Prozedur: GENLOG)

```
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
       1            Constant
       2            [V20 = 1]
       3            [V20 = 2]
       4       x    [V20 = 3]
       5            [V326 = 1]
       6       x    [V326 = 2]
Note: 'x' indicates an aliased (or a redundant) parameter.
      These parameters are set to zero.
Design Matrix
                          Cell          Parameter
Factor        Value  Structure     1     2     3     5
V326           ja
 V20            ja       1.000     1     1     0     1
 V20          nein       1.000     1     0     1     1
 V20    weiß nicht       1.000     1     0     0     1
V326          nein
 V20            ja       1.000     1     1     0     0
 V20          nein       1.000     1     0     1     0
 V20    weiß nicht       1.000     1     0     0     0
Table Information
                       Observed              Expected
Factor        Value       Count       %         Count       %
V326           ja
 V20            ja      2190.00 ( 62.38)      2199.11 ( 62.63)
 V20          nein       451.00 ( 12.85)       422.63 ( 12.04)
 V20    weiß nicht       180.00 (  5.13)       199.26 (  5.68)
V326          nein
 V20            ja       547.00 ( 15.58)       537.89 ( 15.32)
 V20          nein        75.00 (  2.14)       103.37 (  2.94)
 V20    weiß nicht        68.00 (  1.94)        48.74 (  1.39)
Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio       19.3674        2     6.E-05
         Pearson       19.3584        2     6.E-05
Parameter Estimates
                                                 Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     3.8865      .0721      53.92       3.75       4.03
        2     2.4012      .0663      36.21       2.27       2.53
        3      .7519      .0770       9.76        .60        .90
        4      .0000      .             .          .          .
        5     1.4082      .0425      33.16       1.32       1.49
        6      .0000      .             .          .          .
```

Grafiken zur Beurteilung der Modellanpassung unterdrückt. In SPSS wird standardmäßig der Wert 0.5 zu allen Zellenhäufigkeiten addiert.[6] Über die Option "criteria = delta (0)" wird diese Voreinstellung ausgeschaltet. Die zu analysierenden Tabellenhäufigkeiten bleiben so unverändert. In der letzten Option "/design" wird das log-lineare Modell spezifiziert, dessen Parameter geschätzt werden sollen. Die Angabe der beiden Modellvariablen V20 und V326 führt dazu, daß die Prozedur temporär für diese beiden Variablen dummykodierte Design-Variablen erzeugt, die zur Prognose der logarithmierten Häufigkeiten herangezogen werden.

[6] Für die Zahl null ist ein Logarithmus nicht definiert. Unbesetzte Tabellenzellen können daher bei log- linearen Analysen zu Problemen führen.

In Abbildung 1 ist leicht gekürzt die Ausgabe der Prozedur GENLOG dokumentiert. Zunächst werden Informationen zur Bedeutung der geschätzten Parameter gegeben. Der erste Modellparameter ist die Regressionskonstante ('Constant'). Es folgt das Regressionsgewicht der Design-Variable für die erste Ausprägung der Bürgernähe von Politikern ('V20=1'), anschließend das Gewicht für die Design-Variable der zweiten Ausprägung ('V20=2'). Die dritte Ausprägung von V20 ist die Referenzkategorie. Für sie wird keine eigene Design-Variable erzeugt. Dies wird durch ein 'x' in der mit "Aliased" überschriebenen Spalte gekennzeichnet. Anschließend folgt der Parameter der Design-Variable für die erste Ausprägung der Wahlbeteiligung ('V326=1'). Die zweite Ausprägung ist wiederum durch ein 'x' als Referenzkategorie gekennzeichnet.

Es folgt die Wiedergabe der bei der Schätzung verwendeten Design-Matrix. Zur leichteren Identifizierung der Tabellenzellen sind links die Ausprägungen der die Tabelle definierenden Variablen wiedergegeben. Die mit "Cell Structure" überschriebene Spalte bezieht sich darauf, daß den Tabellenzellen Gewichte zugeordnet werden können, um beispielsweise durch ein Gewicht von 0.0 unbesetzte Zellen von der Analyse auszuschließen. Voreinstellung ist das Gewicht 1.0 für jede Zelle. Schließlich folgen die eigentlichen Spalten der Design-Matrix. Durch die in der Spaltenüberschrift angegebene Parameternummer kann erschlossen werden, auf welchen Modellparameter sich eine Spalte der Design-Matrix bezieht.

Am Ende der Ausgabe werden die geschätzten Regressionskoeffizienten ('Estimate'), deren Standardfehler ('SE'), die Quotienten aus Koeffizienten und Standardfehler ('Z- value') und die Grenzen der asymptotischen 95%-Konfidenzintervalle ausgedruckt. Die ML-Schätzung der Koeffizienten ergibt bei den Daten des ALLBUS 1996 aus Tabelle 1 eine Regressionskonstante von 3.8865. Für das Regressionsgewicht der ersten Design- Variable D_1 wird der Wert 2.4012, für das der zweiten Design-Variable D_2 der Wert 0.7519 und für das der letzten Design-Variablen D_3 der Wert 1.4082 geschätzt. Setzt man diese Werte in die Vorhersagegleichung ein, ergeben sich die Prognosen für die logarithmierten Häufigkeiten. In Tabelle 3 ist die Berechnung der logarithmierten Häufigkeiten exemplarisch durchgeführt.

Die letzte Spalte der Tabelle enthält zusätzlich die vorhergesagten absoluten Häufigkeiten. Diese ergeben sich, wenn der Antilogarithmus das ist die Umkehrfunktion des Logarithmierens der Vorhersagewerte aus der ersten Spalte berechnet werden. 2199.31 ist beispielsweise der Antilogarithmus von 7.6959 ($2199.31 \approx e^{7.6959}$).[7] Das Modell sagt somit für die erste Tabellenzelle eine Häufigkeit von 2199.31 voraus. Dies ist der geschätzte Erwartungswert, der sich im Durchschnitt über alle möglichen Zufallsstichproben ergeben würde, falls die geschätzten Regressionskoeffizienten die tatsächlichen Populationswerte sind. Die

[7] Als Folge von Rundungsfehlern stimmen die in Tabelle 3 berechneten erwarteten Häufigkeiten nicht genau mit den von SPSS ausgedruckten Werten in Abbildung 1 überein.

prognostizierten Häufigkeiten werden daher auch als erwartete Häufigkeiten ('expected counts') bezeichnet.

Der Vergleich mit den beobachteten Häufigkeiten weist bei den Kategorien '2' und '3' von V20 (Bürgernähe) deutliche Abweichungen auf.[8] Das Modell scheint also nicht mit den Daten vereinbar zu sein. Im Unterschied zur linearen Regression, wo Abweichungen zwischen den Vorhersagewerten und den tatsächlichen Werten der abhängigen Variablen auf nichterfaßte Größen zurückgeführt werden können, die in der Residualvariable zusammengefaßt sind, wird in log-linearen Modellen grundsätzlich unterstellt, daß die Design-Variablen bei Kenntnis der "wahren" Modellparameter auch stets die "wahren" erwarteten Häufigkeiten wiedergeben. Abweichungen zwischen den Vorhersagen ($\hat{Y}$) und den beobachteten Werten (Y) können dann nur Folge von zufälligen Stichprobenschwankungen sein.[9]

Tabelle 3: Berechnung der vorhergesagten logarithmierten Häufigkeiten

$\hat{Y}$ =	$b_0 \cong 1$	$+ b_1 \cong D_1$	$+ b_2 \cong D_2$	$+ b_3 \cong D_3$	$e^{\hat{Y}}$
7.6959 =	3.8865 ≅ 1	+ 2.4012 ≅ 1	+ 0.7519 ≅ 0	+ 1.4082 ≅ 1	2199.31
6.0466 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 1	+ 1.4082 ≅ 1	422.67
5.2947 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 0	+ 1.4082 ≅ 1	199.28
6.2877 =	3.8865 ≅ 1	+ 2.4012 ≅ 1	+ 0.7519 ≅ 0	+ 1.4082 ≅ 0	537.91
4.6384 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 1	+ 1.4082 ≅ 0	103.38
3.8865 =	3.8865 ≅ 1	+ 2.4012 ≅ 0	+ 0.7519 ≅ 0	+ 1.4082 ≅ 0	48.74

Zur Beurteilung der Übereinstimmung von Modell und Daten können Goodness-of-Fit Teststatistiken herangezogen werden. In der SPSS-Ausgabe werden sowohl die Statistik für den Likelihood-Ratio-Test als auch die für Pearsons Anpassungstest ausgegeben. Unter der (Null-) Hypothese, daß Abweichungen zwischen beobachteten und erwarteten Häufigkeiten tatsächlich nur durch zufällige Stichprobenschwankungen hervorgerufen werden,sind beide Teststatistiken asymptotisch chiquadratverteilt. Die Freiheitsgrade ergeben sich dabei aus der Differenz der Zeilenzahl (i.a. die Anzahl der Zellen der zu analysierenden Tabelle) und der Spaltenzahl (Zahl der geschätzten Modellparameter) der Design-Matrix. Beide Test-

[8] Bezogen auf die relativen Häufigkeiten sind die Abweichungen allerdings nicht sehr groß. Dies ist eine Folge der starken Schiefe der Variable V20. Bei schiefen Verteilungen führen Prozentsatzdifferenzen leicht in die Irre.

[9] Etwas anders sieht es aus, wenn - wie z.B. in der latenten Klassenanalyse (LCA) - die Ausgangsvariablen, die die Tabelle generieren, auf inhaltlich relevante Größen und Meßfehler zurückgeführt werden. Aber auch dann gilt, daß *nach* der Berücksichtigung von Meßfehlern alle weiteren Abweichungen zwischen beobachteten und erwarteten Häufigkeiten auf Stichprobenschwankungen zurückgeführt werden.

statistiken weisen Werte um 19.4 auf; bei zwei Freiheitsgraden ergibt sich ein empirisches Signifikanzniveau ("Sig.") kleiner 0.001. Die Wahrscheinlichkeit rein zufälliger Abweichungen ist also nahezu null. Es gibt somit gute Gründe, an der Übereinstimmung von Modell und Daten zu zweifeln.

Obwohl das Modell offensichtlich nicht zutrifft, soll es doch als ein einfaches Beispiel dafür herangezogen werden, wie die Koeffizienten eines log-linearen Modells interpretiert werden können. Aus der Design-Matrix (bzw. Tabelle 3) läßt sich erkennen, daß die Regressionskonstante den Logarithmus der geschätzten Häufigkeit für die Tabellenzelle angibt, bei der alle Design-Variablen den Wert null aufweisen. Im Beispiel ist dies die letzte Tabellenzelle von Tabelle 1 bzw. die letzte Zeile von Tabelle 3, die der Ausprägungskombination V20=3 (keine Meinung zum Desinteresse von Politikern) und V326=2 (keine Wahlbeteiligung) entspricht. Die erwartete Häufigkeit von 48.74 ergibt sich wieder über den Antilogarithmus der prognostizierten logarithmierten Häufigkeit ($48.74.e^{3.8865}$).

Das Regressionsgewicht von 2.4012 der ersten Design-Variable bezieht sich auf die erste Ausprägung der Bürgernähe von Politikern. Aus dem Vergleich der Vorhersagegleichungen in Tabelle 3 läßt sich erkennen, daß dieser Koeffizient besagt, um welchen Wert die loga-rithmierten Zellenbesetzungen im Durchschnitt jeweils ansteigen, wenn statt der Referenzkategorie 'weiß nicht' (V20=3) die erste Ausprägung 'ja' der Variable betrachtet wird (V20=1). Da in dem Modell die Effekte der zweiten erklärenden Variable Wahlbeteiligung über den Koeffizienten der Design-Variable D_3 kontrolliert werden, handelt es sich um einen partiellen Effekt bei Kontrolle der Ausprägungen der Wahlbeteiligung. Der Antilogarithmus dieses Koeffizienten gibt dann an, um welchen Faktor die Zellenhäufigkeit im Durchschnitt über der Häufigkeit der Referenzkategorie liegt. Es gibt im (geometrischen) Mittel ungefähr 11 mal mehr Personen, die der Ansicht 'Politiker sind desinteressiert' zustimmen, als es Personen gibt, die mit 'weiß nicht' antworten ($e^{2.4012}$. 11.03).

Die Interpretation der übrigen Koeffizienten folgt der gleichen Logik. Für die zweite Kategorie der Bürgernähe (V20=2) ergibt sich ein Koeffizient von 0.7519. Der Antilogarithmus dieser Zahl, 2.12 ($.e^{0.7519}$), gibt wiederum den Faktor an, mit dem diese Kategorie im Durchschnitt die Häufigkeit der Referenzkategorie übersteigt. Nach dem spezifizierten log-linearen Modell gibt es im Durchschnitt 2.12 mal so viele Befragte, die Politiker für interessiert halten (V20=2), wie es Befragte gibt, die hierzu keine Meinung haben (V20=3). Der letzte Koeffizient bezieht sich auf die erste Kategorie der Wahlbeteiligung (V326=1). Der positive Wert von 1.4082 weist darauf hin, daß die Zahl der Wähler (V326=1) im Durchschnitt um den Faktor 4.09 ($.e^{1.4082}$) höher ist als die Zahl der Nichtwähler (V326=2).

Die Regressionskoeffizienten der dummykodierten Design-Variablen des log-linearen Modells besagen also, wie die (logarithmierten) Häufigkeiten der Zellenbesetzungen mit den Ausprägungen der erklärenden Variablen relativ zur Referenzkategorie variieren. Die Inter-

pretationslogik entspricht derjenigen der linearen Regression mit nominalskalierten unabhängigen Variablen. Ein Unterschied besteht allerdings zwischen der üblichen linearen Regression und der log-linearen Vorhersage von Tabellenzellen. In der linearen Regression wird die Beziehung zwischen einer inhaltlich interessierenden abhängigen Variable und deren Prädiktoren modelliert. Die logarithmierten Häufigkeiten der Tabellenzellen dürften aber in den seltensten Fälle von theoretischem Interesse sein. Sie sind nur eine Hilfsgröße bei der Untersuchung des eigentlich interessierenden Zusammenhangs.

Tatsächlich impliziert das spezifizierte log-lineare Modell auch Aussagen über den Zusammenhang zwischen der Wahlbeteiligung (V326) und der Bürgernähe von Politikern (V20). Das Modell behauptet nämlich, daß diese beiden Größen statistisch unabhängig voneinander sind. Um dies zu erkennen, sei daran erinnert, daß zwei Ereignisse genau dann statistisch unabhängig voneinander sind, wenn ihre gemeinsame Auftretenswahrscheinlichkeit das Produkt der Auftretenswahrscheinlichkeiten der einzelnen Ereignisse ist. In der einfachen Tabellenanalyse wird dies genutzt, um die bei Unabhängigkeit erwarteten Häufigkeiten zu berechnen. Diese sind nämlich gerade das Produkt der über die relativen Häufigkeiten der Randverteilungen geschätzten Wahrscheinlichkeiten der Ausprägungen von Spalten- und Zeilenvariable sowie der Fallzahl.

Auf gleiche Weise berechnen sich die vorhergesagten Häufigkeiten des log-linearen Modells aus Abbildung 1. Tabelle 3 ist zu entnehmen, daß die logarithmierte Häufigkeit der ersten Tabellenzelle die Summe der Regressionskonstante und der Regressionsgewichte der ersten und dritten Design-Variable ist: 7.6959 = 3.8865 + 2.4012 + 1.4082. Die absoluten Häufigkeiten ergeben sich über die Berechnung der Antilogarithmen. Aus der Summe wird dabei ein Produkt: $2199.31 = e^{7.6959} = e^{3.8865} A e^{2.4012} A e^{1.4082}$. Wie bei statistischer Unabhängigkeit gefordert, ist die vorhergesagte Häufigkeit der ersten Tabellenzelle proportional zum Produkt des Effektes der ersten Kategorie der Zeilenvariable von Tabelle 1 ($e^{2.4012}$) und des Effekts der ersten Kategorie der Spaltenvariable ($e^{1.4082}$). Der so berechnete Vorhersagewert entspricht daher auch bis auf Rundungsfehler den erwarteten Häufigkeiten, die man bei der üblichen Berechnung des Chiquadrattests auf statistische Unabhängigkeit benötigt. Gleiches gilt für die übrigen fünf Tabellenzellen.

Ein statistischer Zusammenhang zwischen den beiden Variablen V20 und V326 wird erst zugelassen, wenn zusätzlich zu den Regressionsgewichten der Design-Variablen *Interaktionseffekte* spezifiziert werden. Das log-lineare Modell ist dazu um weitere Prädiktoren zu erweitern. Diese ergeben sich als Produkte der Design-Variablen des ersten Modells. Im Beispiel werden also die Design-Variablen der Ausprägungen von V326 (D_1 und D_2) mit der Design-Variable für die Ausprägungen von V326 (D_3) multipliziert. Um das erweiterte Modell mit der SPSS-Prozedur GENLOG zu schätzen, können in der Option "/design" die Interaktionseffekte durch die Spezifikation "V326*V20" oder "V326 by V20" angefordert werden.

Abbildung 2: Parameterschätzung im Modell mit Interaktionseffekten (SPSS-Prozedur: GENLOG)

```
-> genlog  V326 V20
->   /print design estim /plot none
->   /criteria = delta(0) /design V20 V326 V20*V326 .
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            [V20 = 1]
        3            [V20 = 2]
        4       x    [V20 = 3]
        5            [V326 = 1]
        6       x    [V326 = 2]
        7            [V326 = 1]*[V20 = 1]
        8            [V326 = 1]*[V20 = 2]
        9       x    [V326 = 1]*[V20 = 3]
       10       x    [V326 = 2]*[V20 = 1]
       11       x    [V326 = 2]*[V20 = 2]
       12       x    [V326 = 2]*[V20 = 3]
Note: 'x' indicates an aliased (or a redundant) parameter.
      These parameters are set to zero.
Design Matrix
                          Cell                Parameter
Factor        Value  Structure     1     2     3     5     7     8
V326            ja
 V20            ja       1.000     1     1     0     1     1     0
 V20          nein       1.000     1     0     1     1     0     1
 V20    weiß nicht       1.000     1     0     0     1     0     0
V326         nein
 V20            ja       1.000     1     1     0     0     0     0
 V20          nein       1.000     1     0     1     0     0     0
 V20    weiß nicht       1.000     1     0     0     0     0     0
Goodness-of-fit Statistics
                   Chi-Square       DF       Sig.
Likelihood Ratio         .0000        0      .
         Pearson         .0000        0      .
Parameter Estimates
                                              Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     4.2195      .1213      34.79       3.98       4.46
        2     2.0849      .1286      16.21       1.83       2.34
        3      .0980      .1674        .59       -.23        .43
        4      .0000      .            .          .          .
        5      .9734      .1423       6.84        .69       1.25
        6      .0000      .            .          .          .
        7      .4138      .1502       2.76        .12        .71
        8      .8205      .1892       4.34        .45       1.19
        9      .0000      .            .          .          .
       10      .0000      .            .          .          .
       11      .0000      .            .          .          .
       12      .0000      .            .          .          .
```

Der Prozeduraufruf und Teile der SPSS-Ausgabe sind in Abbildung 2 wiedergegeben. Aus den Regressionskoeffizienten lassen sich wieder die Vorhersagen der logarithmierten Häufigkeiten berechnen (vgl. Tabelle 4). Der Vergleich der Regressionskonstante und der Koeffizienten für die Design-Variablen D_1, D_2 und D_3 mit den entsprechenden Koeffizienten des ersten Modells weist deutlich veränderte Werte auf. Änderungen der Koeffizientenschätzungen sind in der Regel zu beobachten, wenn die zusätzlich spezifizierten Interaktionseffekte die Vorhersagen der logarithmierten Häufigkeiten merklich verbessern können. Auch in der linearen Regression ändern sich oft die Regressionskoeffizienten, wenn zusätzliche erklärende Variablen in ein Modell aufgenommen werden.

Tabelle 4: Berechnung der vorhergesagten logarithmierten Häufigkeiten im Modell mit Interaktionseffekten

$\hat{Y}$ =	$b_0 \cong 1$	$+b_1 \cong D_1$	$+b_2 \cong D_2$	$+b_3 \cong D_3$	$+b_4 \cong (D_1 \cong D_3)$	$+b_5 \cong (D_2 \cong D_3)$
7.6116 =	4.2195 ≅ 1	+2.0849 ≅ 1	+0.0980 ≅ 0	+0.9734 ≅ 1	+0.4138 ≅ 1	+0.8205 ≅ 0
6.1114 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 1	+0.9734 ≅ 1	+0.4138 ≅ 0	+0.8205 ≅ 1
5.1929 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 0	+0.9734 ≅ 1	+0.4138 ≅ 0	+0.8205 ≅ 0
6.3044 =	4.2195 ≅ 1	+2.0849 ≅ 1	+0.0980 ≅ 0	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0
4.3175 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 1	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0
4.2195 =	4.2195 ≅ 1	+2.0849 ≅ 0	+0.0980 ≅ 0	+0.9734 ≅ 0	+0.4138 ≅ 0	+0.8205 ≅ 0

Die Goodness-of-Fit Teststatistiken weisen mit einem Wert von 0.0 (bei null Freiheitsgraden) darauf hin, daß eine perfekte Übereinstimmung zwischen den (nicht wiedergegebenen) erwarteten und beobachteten Häufigkeiten besteht. Dies ist jedoch kein empirischer Befund sondern eine logische Konsequenz der Tatsache, daß das Modell genau so viele Datenpunkte (=Zeilen der Design-Matrix) aufweist, wie unbekannte Modellparameter (=Spalten der Design-Matrix). Es ist bei solchen *saturierten Modellen* nicht möglich, anhand der Übereinstimmung der vorhergesagten mit den beobachteten Häufigkeiten die Güte des Modells zu beurteilen.

Welche Konsequenzen ergeben sich nun für die inhaltliche Interpretation? Diese Frage kann wiederum über die Berechnung der Vorhersagewerte beantwortet werden (Tabelle 4). Wie im ersten Modell besagt der positive Koeffizient der Design-Variable D_1, daß es verglichen mit der Referenzkategorien "weiß nicht" bei der wahrgenommenen Bürgernähe von Politikern im Durchschnitt deutlich mehr Personen gibt, die Politiker für desinteressiert halten. Der sehr geringe positive Wert von 0.098 für D_2 weist auf der anderen Seite darauf hin, daß verglichen mit den Meinungslosen bei Berücksichtigung von Interaktionseffekten nur mit geringfügig mehr Personen zu rechnen ist, die Politiker für interessiert halten. Dem positiven Effekt für die dritte Design-Variable ist schließlich zu entnehmen, daß es - wie im ersten Modell - im Durchschnitt deutlich mehr Wähler als Nichtwähler gibt.

Dieses Grundmuster wird nun durch die Interaktionseffekte modifiziert. Der erste Interaktionseffekt besagt, daß der Anstieg der Zellenhäufigkeit von der Referenzkategorie der Meinungslosen (V20=3) zu denjenigen, die Politiker für desinteressiert halten (V20=1), in der Gruppe der Wähler (V326=1) deutlich höher ist als in der Gruppe der Nichtwähler (V326=0). Bei den Wählern ist der Zuwachs der logarithmierten Häufigkeiten um den Wert des Interaktionseffekts höher als bei den Nichtwählern, also um 0.4138. Auf der Ebene der

absoluten Häufigkeiten übersteigt der Zuwachs der Wähler den der Nichtwähler daher um den Faktor 1.5126 (=$e^{0.4138}$). Alternativ kann der Interaktionseffekt auch so gelesen werden, daß der Anstieg der Zellenbesetzung beim Wechsel von den Nichtwählern (v326=2) zu den Wählern (V326=1) in der Gruppe derjenigen, die Politiker für desinteressiert halten (V20-=1), um den Faktor 1.5126 höher ist als in der Gruppe derjenigen, die keine Meinung haben (V20=3). Der zweite Interaktionseffekt ist noch höher: Die Besetzungszahlen steigen beim Wechsel von den Meinungslosen (V20=3) zu der Gruppe derjenigen, die Politiker für interessiert halten (V20=2), um den Faktor 2.2716 (=$e^{0.8205}$) stärker an, wenn es sich um Wähler (V326=1) und nicht um Nichtwähler (V326=2) handelt. Oder auf die Wahlbeteiligung bezogen: Beim Wechsel von Nichtwählern zu Wählern ist der Anstieg der Zellenbesetzung sehr viel stärker, wenn statt Meinungslosigkeit die Auffassung vertreten wird, daß Politiker interessiert sind. Zusammengenommen folgt also aus dem Modell, daß das Verhältnis von Wähler zu Nichtwählern bei den Meinungslosen am relativ geringsten und bei denjenigen, die Politiker für interessiert halten, am relativ höchsten ist.

Während die sogenannten *Haupteffekte* eines log-linearen Modells, das sind die Regressionsgewichte der Design-Variablen für die Kategorien der Modellvariablen, nur die unterschiedlichen Besetzungen der Kategorien dieser Variablen widerspiegeln, modellieren die Interaktionseffekte Zusammenhänge zwischen den inhaltlich interessierenden Variablen. Bei mehrdimensionalen Kreuztabellen können auch Interaktionseffekte höherer Ordnung spezifiziert werden. Dazu werden die Design-Variablen von drei oder mehr kategorialen Ausgangsvariablen miteinander multipliziert. Würde etwa bei den Daten aus Tabelle 1 zusätzlich zwischen Befragten aus den alten und den neuen Bundesländern unterschieden, könnte es sich zeigen, daß die Beziehung zwischen der Beurteilung der Interessiertheit von Politikern und der Wahlbeteiligung in den alten und neuen Bundesländern unterschiedlich ist. Im log-linearen Modell gäbe es dann deutliche Interaktionseffekte zwischen der Design-Variable der Wahlbeteiligung, den beiden Design-Variablen der Bürgernähe von Politikern und der Design-Variable für das Erhebungsgebiet. Die Interpretation eines log-linearen Modells mit solchen Interaktionseffekten zweiter oder noch höherer Ordnung wird allerdings schnell unübersichtlich. Wenn es die Daten zulassen, werden daher eher sparsame Modelle mit Interaktionseffekten möglichst geringer Ordnung bevorzugt.

2. Die Spezifikation benutzerdefinierter Design-Matrizen in SPSS

Die SPSS-Prozedur GENLOG verwendet grundsätzlich dummykodierte Design-Variablen bei der Spezifikation eines log-linearen Modells. Referenzkategorie ist die letzte Kategorie einer Ausgangsvariable, also deren numerisch höchster Wert. In der älteren SPSS-Prozedur LOGLINEAR wird dagegen als Voreinstellung Effektkodierung eingesetzt. Um die Unterschiede zwischen den beiden Kodierregeln zu verdeutlichen,soll das Modell aus Abbildung 2 auch mit effektkodierten Design-Variablen geschätzt werden. Um die Koeffizienten eines

log-linearen Modells mit Effektkodierung auch über die Prozedur GENLOG schätzen zu können, müssen die Spalten der Design-Matrix direkt vom Anwender generiert werden. Möglich wird dies durch die in der Prozedur implementierte Option der Schätzung von Effekten metrischer Prädiktoren, den sogenannten Kovariaten. Wird im Prozeduraufruf von GENLOG eine Variable als Kovariate aufgeführt, berechnet GENLOG für alle Zellen der zu analysierenden Tabelle die Mittelwerte der Kovariate über die jeweilige Anzahl der Fälle in den Zellen. Die resultierenden Mittelwerte können dann als zusätzliche Spalte in die Design-Matrix aufgenommen werden.[10] Über die Generierung geeigneter Kovariaten können somit benutzerdefinierte Spalten der Design-Matrix spezifiziert werden. Werden in der /DESIGN-Option ausschließlich benutzergenerierte Kovariaten aufgeführt, wird ein log-lineares Modell geschätzt, dessen Design-Matrix vom Benutzer frei gestaltbar ist. Die einzige Einschränkung besteht darin, daß die Regressionskonstante stets automatisch geschätzt wird.[11]

Für das Beispiel werden zunächst mit RECODE- und COMPUTE-Anweisungen aus den Ausgangsvariablen V20 (Bürgernähe) und V326 (Wahlbeteiligung) effektkodierte Design-Variablen für die Haupt- und Interaktionseffekte generiert:

```
recode V20 (1=1)(2=0)(3=-1) into D1.
recode V20 (1=0)(2=1)(3=-1) into D2.
recode V326 (1=1)(2=-1) into D3.
compute D1D3=D1*D3.
compute D2D3=D2*D3.
```

Die erste RECODE-Anweisung erzeugt für die erste Kategorie von V20 eine effektkodierte Design-Variable D1. Referenzkategorie ist die letzte Ausprägung der Ausgangsvariable, bei der die Design-Variable den Wert '-1' erhält. Auf analoge Weise werden für die zweite Kategorie von V20 und für die erste Kategorie von V326 effektkodierte Design-Variablen D2 und D3 gebildet. Schließlich werden mit den beiden COMPUTEAnweisungen durch einfaches Multiplizieren der gerade gebildeten Design-Variablen zusätzliche Produktvariablen D1D3 und D2D3 für die Schätzung von Interaktionseffekten generiert.

In der Modellspezifikation werden die so erzeugten Design-Variablen als Kovariaten hinter dem Schlüsselwort "WITH" direkt nach der Nennung der die Tabelle definierenden Variablen aufgeführt und in der /DESIGN-Option als Effekte angegeben. Der Prozeduraufruf und die Resultate der Modellschätzung sind in Abbildung 3 wiedergegeben. Der Vergleich mit den entsprechenden Werten des log-linearen Modells bei Dummykodierung (Abb. 2)

10 Eine mögliche Variation zwischen den Werten der Kovariaten in einer Tabellenzelle wird also ignoriert. Die Vorgehensweise ist daher keine Alternative zu Modellen, die - wie die logistische Regression - auf Individualdatenebene kategoriale abhängige Variablen durch metrische Prädiktoren erklären.

11 Außerdem werden redundante Spalten einer Designmatrix automatisch entfernt. Mathematisch ausgedrückt muß die Design-Matrix vollen Spaltenrang haben. Als eine technische Einschränkung ist schließlich die Zahl der möglichen Kovariaten auf 200 beschränkt.

weist auf gänzlich verschiedene Werte hin. Die (im Ausdruck nicht aufgeführten erwarteten Häufigkeiten) und die Goodness-of-Fit Teststatistiken sind jedoch identisch. Tatsächlich handelt es sich bei den beiden Modellen um *Reparametrisierungen* der gleichen Aussagen zur Struktur der analysierten Tabelle. Unterschiede bei Vorhersagewerten können erst dann auftreten, wenn zwei Modelle unterschiedliche Aussagen beinhalten, wie dies z.B. bei den beiden Modellen mit bzw. ohne Interaktionseffekten der Fall ist.

Abbildung 3: Parameterschätzung im Modell mit Dummykodierung

```
-> genlog V326 V20 with D1 D2 D3 D1D3 D2D3
->   /print des est /plot non /crit delta(0)
->   /des D1 D2 D3 D1D3 D2D3 .
Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            D1
        3            D2
        4            D3
        5            D1D3
        6            D2D3
Design Matrix
                          Cell                    Parameter
Factor        Value  Structure     1        2        3        4        5
V326           ja
 V20             ja       1.000     1    1.000     .000    1.000    1.000
 V20           nein       1.000     1     .000    1.000    1.000     .000
 V20    weiß nicht        1.000     1   -1.000   -1.000    1.000   -1.000
V326          nein
 V20             ja       1.000     1    1.000     .000   -1.000   -1.000
 V20           nein       1.000     1     .000    1.000   -1.000     .000
 V20    weiß nicht        1.000     1   -1.000   -1.000   -1.000    1.000
Design Matrix (continued)
                          Cell    Parameter
Factor        Value  Structure        6
V326           ja
 V20             ja       1.000     .000
 V20           nein       1.000    1.000
 V20    weiß nicht        1.000   -1.000
V326          nein
 V20             ja       1.000     .000
 V20           nein       1.000   -1.000
 V20    weiß nicht        1.000    1.000
Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio         .0000        0      .
         Pearson         .0000        0      .

Parameter Estimates                              Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     5.6396      .0325     173.36       5.58       5.70
        2     1.3585      .0353      38.44       1.29       1.43
        3     -.4251      .0485      -8.76       -.52       -.33
        4      .6924      .0325      21.29        .63        .76
        5      .0012      .0353        .03       -.07        .07
        6      .2046      .0485       4.22        .11        .30
```

Die Regressionskonstante gibt bei Effektkodierung den Durchschnittswert der logaritmierten erwarteten Häufigkeiten wieder. Tabelle 5 verdeutlicht, warum dies so ist. Werden nämlich die Vorhersagegleichungen für alle Tabellenzellen aufsummiert, bleibt auf der rechten Seiten der Gleichung nur die Summe der Regressionskonstanten übrig. Alle anderen Spalten addieren sich stets zum Wert null. Da die Summe der sechs Vorhersagewerte also das sechsfache der Regressionskonstante ergibt, muß diese 1/6 dieser Summe sein.Aus der Tat-

sache, daß die Summe der Werte einer Design-Variable null ist, folgt weiter für die Interpretation des zugehörigen Koeffizienten, daß dieser die Abweichung vom Durchschnittswert erfaßt. Der Koeffizient 1.3585 des Effektes der Design-Variable D_1 für die erste Ausprägung von V20 besagt also, daß die logarithmierten Häufigkeiten der Tabellenzellen, bei der die Variable V20 den Wert '1' aufweist, bei Kontrolle von V326 und der Berücksichtigung der Interaktionseffekte im Mittel um den Wert 1.3585 vom Durchschnittswert aller Tabellenzellen abweichen. Entsprechend weist der nächste Koeffizient -0.4251 darauf hin, daß die Tabellenzellen, die sich auf die zweite Kategorie von V20 beziehen, im Schnitt eine um 0.4251 geringere logarithmierte Häufigkeit aufweisen. Da die Summe aller Abweichungen vom Durchschnitt null ergibt, ist die durchschnittliche Abweichung in der Referenzkategorie das Negative der Summe der übrigen Abweichungen. Die durchschnittliche Abweichung der 'Meinungslosen' bei der Bewertung der Bürgernähe (V20=3) ist also $-0.9334 = 1.3585 \cdot (-1) + -0.4251 \cdot (-1)$.

Tabelle 5: Berechnung der vorhergesagten logarithmierten Häufigkeiten im Modell mit Interaktionseffekten bei Effektkodierung

$\hat{y}$ =	$b_0 \cdot 1$	$+b_1 \cdot D_1$	$+b_2 \cdot D_2$	$+b_3 \cdot D_3$	$+b_4 \cdot (D_1 \cdot D_3)$	$+b_5 \cdot (D_2 \cdot D_3)$
7.6117=	5.6396 · 1	+1.3585 · 1	-0.4251 · 0	+0.6924 · 1	+0.0012 · 1	+0.2046 · 0
6.1115=	5.6396 · 1	+1.3585 · 0	-0.4251 · 1	+0.6924 · 1	+0.0012 · 0	+0.2046 · 1
5.1928=	5.6396 · 1	+1.3585 · -1	-0.4251 · -1	+0.6924 · 1	+0.0012 · -1	+0.2046 · -1
6.3045=	5.6396 · 1	+1.3585 · 1	-0.4251 · 0	+0.6924 · -1	+0.0012 · -1	+0.2046 · 0
4.3175=	5.6396 · 1	+1.3585 · 0	-0.4251 · 1	+0.6924 · -1	+0.0012 · 0	+0.2046 · -1
4.2196=	5.6396 · 1	+1.3585 · -1	-0.4251 · -1	+0.6924 · -1	+0.0012 · 1	+0.2046 · 1

Die beiden Haupteffekte der Bürgernähe von Politikern besagen also, daß es überdurchschnittliche viele Personen gibt, die Politiker für nicht interessiert halten, daß es dagegen unterdurchschnittlich viele Personen gibt, die Politiker für interessiert halten, und daß es noch weniger Personen gibt, die gar keine Meinung zu diesem Thema haben. Analog ergibt sich als Interpretation des Koeffizienten für die Wahlbeteiligung, daß in der analysierten Tabelle die logarithmierten Zellenhäufigkeiten um 0.6924 über dem Durchschnitt liegen, wenn es sich um Wähler handelt (V326=1). Umgekehrt liegen die logarithmierten Häufigkeiten bei Nichtwählern im Durchschnitt um -0.6924 unter dem Durchschnitt aller Zellen.

Die Interaktionseffekte modifizieren wiederum diese Grundaussage.Der sehr geringe Koeffizient von 0.0012 besagt, daß bei Zustimmung zur These der Desinteressiertheit von Politikern (V20=1) die Zahl der Wähler nur ganz geringfügig über der durchschnittlichen Zahl

von Wählern in den Tabellenzellen liegt und entsprechend die Zahl der Nichtwähler nur geringfügig unter dem Durchschnitt der Nichtwähler. Der letzte Koeffizient von 0.2046 besagt entsprechend, daß die Zahl der Wähler, die Politiker eher für interessiert halten (V20=2), überdurchschnittlich hoch ist und die Zahl der Nichtwähler, die Politiker für desinteressiert halten, dann deutlich unter dem Durchschnitt liegt. Für die Gruppe der Meinungslosen berechnet sich die Abweichung vom Durchschnitt der Wähler durch Aufsummieren der negativen Werte der beiden Interaktionseffekte: der negative Wert von -0.2058 (=-0.012 + -0.2046) weist darauf hin, daß es in dieser Gruppe besonders wenige Wähler und dafür umgekehrt besonders viele Nichtwähler gibt.

Bei Effektkodierung werden die Koeffizienten also ähnlich interpretiert wie bei der Dummykodierung. Der Unterschied liegt allein darin, daß bei der Effektkodierung jeweils Abweichungen vom Durchschnittswert gemessen werden, bei der Dummykodierung dagegen Abweichungen von einer Referenzgruppe. Das Endergebnis ist aber dasselbe. Beide Modelle besagen, daß es unter denjenigen, die die Politiker für nicht desinteressiert halten, besonders viele Wähler gibt und bei den Meinungslosen besonders wenige. Oder auch umgekehrt: unter den Wählern gibt es besonders viele Personen, die Politiker für nicht desinteressiert halten und besonders wenige, die keine Meinung zu diesem Thema haben.

Trotz der gleichen empirischen Behauptungen der Modelle gibt es bei den geschätzten Koeffizienten erhebliche Unterschiede. Im Modell mit Dummykodierung (Abbildung 2) sind beide Interaktionseffekte deutlich von null verschieden, der zweite Haupteffekt der Einschätzung des Desinteresses von Politikern aber praktisch vernachlässigbar. Beim Modell mit Effektkodierung sind dagegen alle Haupteffekte deutlich von null verschieden; dafür ist aber der erste Interaktionseffekt sehr klein. Ursache dieser Differenzen ist wiederum der unterschiedliche Bezugspunkt. Relativ zur Gruppe der Meinungslosen (Dummykodierung) gibt es bei Kontrolle der übrigen Effekte im Durchschnitt kaum mehr Personen, die Politiker nicht für desinteressiert halten. Verglichen zum Gesamtdurchschnitt (Effektkodierung) ist diese Zahl dagegen deutlich geringer. Umgekehrt ist es bei den Interaktionseffekten. Relativ zur Zahl der Wähler unter den Meinungslosen gibt es relativ mehr Wähler in der Gruppe, die Politiker für desinteressiert halten. Die Abweichung vom Durchschnitt aller Wähler ist dagegen nur gering. Inhaltliche Bedeutung bekommen die unterschiedlichen Sichtweisen erst, wenn versucht wird, sparsamere log-lineare Modelle zu schätzen, bei denen nicht signifikante Koeffizienten ausgelassen werden.

Ich erwähnte bereits, daß der erste Interaktionseffekt mit einem Wert von 0.0012 sehr klein ist. Der Koeffizient ist auch bei einer Irrtumswahrscheinlichkeit von 5% nicht signifikant von null verschieden. Sichtbar wird letzteres an den kleinen Z-Werten oder daran, daß die von SPSS ausgedruckten asymptotischen Konfidenzintervalle den Wert null einschließen. Bei der ML-Schätzung sind die geschätzten Koeffizienten asymptotisch normalverteilt. Der

Quotient aus Schätzung und Standardfehler (Z-Wert) kann daher als Teststatistik der Nullhypothese herangezogen werden, daß ein Koeffizient in der Grundgesamtheit null ist. Ist die Nullhypothese richtig, ist der Z-Wert bei größeren Fallzahlen in etwa normalverteilt. Ein Koeffizient ist also in einem zweiseitigen Test mit einer Irrtumswahrscheinlichkeit von etwa 5% signifikant von null verschieden, wenn sein Z-Wert größer +2 oder kleiner -2 ist.

Es liegt nun nahe, ein log-lineares Modell zu spezifizieren, bei dem nur die signifikanten Parameter berücksichtigt werden. Dazu wird einfach in der /DESIGN-Option die entsprechende Kovariate ausgelassen. Aufruf und Ergebnis der Modellschätzung sind in Abbildung 4 festgehalten. Die geringen Chiquadrat-Werte weisen darauf hin, daß die (nicht ausgedruckten) beobachteten und erwarteten Häufigkeiten sehr dicht beieinander liegen. Das Modell kann also die wesentlichen Aspekte der tabellierten Daten gut wiedergeben. Der Vergleich der Koeffizienten in den Abbildungen 3 und 4 zeigt weiter, daß sich die jeweils entsprechenden Koeffizienten kaum unterscheiden. Damit bleibt auch die Interpretation im wesentlich unverändert: Es gibt überdurchschnittlich viele Befragte (b_1=1.359), die Politiker für desinteressiert halten (V20=1) und unterdurchschnittlich viele Personen (b_2=-0.496), die Politiker für interessiert halten (V20=2). Die Zahl der Meinungslosen (V20=3) weicht im Durchschnitt noch stärker nach unten vom Gesamtmittel ab (- 0.863=-b_1-b_2). Bei der Wahlbeteiligung ist die Zahl der Wähler (V326=1) überdurchschnittlich (b_3=0.693), die Zahl der Nichtwähler (V326=2) entsprechend unterdurchschnittlich (-0.693=-b_3) hoch. Schließlich gibt es unter den Wählern überdurchschnittlich viele Befragte, die Politiker nicht für desinteressiert halten (b_5=0.205) und unterdurchschnittlich viele, die keine Meinung haben (-0.205=-b_5).

Das Interessante an dem Modell ist, daß der überdurchschnittliche Anstieg von Wählern in der Gruppe derjenigen, die Politiker nicht für desinteressiert halten, gerade genauso groß ist wie der überdurchschnittliche Rückgang bei den Meinungslosen. In gewisser Hinsicht wird dadurch für den Zusammenhang von Bürgernähe und Wahlbeteiligung eine Metrik definiert, bei der die Meinungslosen den einen Pol bilden und Personen, die Politiker für interessiert halten, den anderen Pol. Die mittlere Position wird von Personen eingenommen, die Politiker für desinteressiert halten.[12]

Die Rangordnung bzw. Metrik der Bürgernähe (V20) kann auch bei Effektkodierung modelliert werden. Dazu wird eine zusätzliche Design-Variable INT gebildet:

```
recode V20 (1=1)(2=2)(3=0) into INT.
if(V326=2) INT=0
```

12 Diese Art von Beziehung wird nach ***L. A. Goodmann*** als log-lineares Assoziationsmodell bezeichnet und für Modelle vorgeschlagen, bei denen eine Variable ordinales oder metrisches Skalenniveau aufweist. Das Modell kann gleichzeitig auch als log-lineare Darstellung des ordinalen Logitmodells der benachbarten Kategorien aufgefaßt werden.

Abbildung 4: Schätzung des Modells mit nur einem Interaktionseffekt bei Effektkodierung

```
-> genlog V326 V20 with D1 D2 D3 D1D3 D2D3
->   /print des est /plot non /crit delta(0)
->   /des D1 D2 D3 D2D3.

Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            D1
        3            D2
        4            D3
        5            D2D3

Design Matrix
                          Cell                    Parameter
Factor        Value  Structure     1        2        3        4        5
V326           ja
 V20            ja       1.000     1    1.000     .000    1.000     .000
 V20           nein      1.000     1     .000    1.000    1.000    1.000
 V20    weiß nicht       1.000     1   -1.000   -1.000    1.000   -1.000
V326          nein
 V20            ja       1.000     1    1.000     .000   -1.000     .000
 V20           nein      1.000     1     .000    1.000   -1.000   -1.000
 V20    weiß nicht       1.000     1   -1.000   -1.000   -1.000    1.000

Goodness-of-fit Statistics
                    Chi-Square       DF       Sig.
Likelihood Ratio         .0011        1      .9737
         Pearson         .0011        1      .9737

Parameter Estimates
                                                  Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     5.6391      .0295     191.32       5.58       5.70
        2     1.3591      .0290      46.93       1.30       1.42
        3     -.4255      .0469      -9.07       -.52       -.33
        4      .6932      .0213      32.50        .65        .74
        5      .2049      .0470       4.36        .11        .30
```

Den drei Kategorien der Bürgernähe (V20) werden hier entsprechend ihrer Rangfolge bei der Höhe der Wahlbeteiligung (V326) die Werte '1' (V20=1), '2' (V20=2) und '0' (V20=3) zugeordnet. Da es sich um einen Interaktionseffekt handelt, gelten diese Scores nur in der Gruppe der Wähler. Bei Nichtwählern weist die Variable INT den Wert '0' auf.

Die Modellspezifikation und Teile der SPSS-Ausgabe sind in Abbildung 5 wiedergegeben. Die Haupteffekte basieren auf dummykodierten Design-Variablen. Daher ist es bei diesem Modell nicht notwendig, die Design-Variablen für die Haupteffekte als Kovariaten zu spezifizieren. Die Gleichheit der Chiquadrat-Werte bzw. der (in der Abbildung nicht wiederge-gebenen) erwarteten Häufigkeiten weisen darauf hin, daß es sich wie bei den saturierten Modellen aus Abbildung 2 und 3 um eine Reparametrisierung des Modells mit Effektkodierung handelt. Tatsächlich läßt sich aus der Design-Matrix ablesen, daß das Modell postuliert, daß der (logarithmierte) Anstieg der Wähler beim Wechsel von den Meinungslosen zu denjenigen, die Politiker für desinteressiert halten, halb so groß ist, wie der Anstieg von den Meinungslosen zu denjenigen, die Politiker für nicht desinteressiert halten.Ausgehend von denjenigen, die Politiker für desinteressiert halten, entspricht also der Anstieg zu denjenigen,die Politiker nicht für desinteressiert halten,dem Rückgang bei den Meinungslosen.

Abbildung 5: Modellierung metrischer Interaktion bei Dummykodierung

```
-> genlog V326 V20 with INT
->   /print des est /plot non /crit delta(0)
->   /des V20 V326 INT.

Correspondence Between Parameters and Terms of the Design
Parameter   Aliased  Term
        1            Constant
        2            [V20 = 1.00]
        3            [V20 = 2.00]
        4       x    [V20 = 3.00]
        5            [V326 = 1.00]
        6       x    [V326 = 2.00]
        7            INT

Design Matrix
                          Cell                Parameter
Factor        Value  Structure     1     2     3     5        7
V326           ja
 V20            ja       1.000     1     1     0     1    1.000
 V20          nein       1.000     1     0     1     1    2.000
 V20    weiß nicht       1.000     1     0     0     1     .000
V326         nein
 V20            ja       1.000     1     1     0     0     .000
 V20          nein       1.000     1     0     1     0     .000
 V20    weiß nicht       1.000     1     0     0     0     .000

Goodness-of-fit Statistics
                   Chi-Square       DF       Sig.
Likelihood Ratio         .0011        1      .9737
         Pearson         .0011        1      .9737

Parameter Estimates
                                                 Asymptotic 95% CI
Parameter   Estimate         SE    Z-value      Lower      Upper
        1     4.2172      .0996      42.36       4.02       4.41
        2     2.0878      .0951      21.95       1.90       2.27
        3      .0982      .1675        .59       -.23        .43
        4      .0000          .          .          .          .
        5      .9766      .1056       9.25        .77       1.18
        6      .0000          .          .          .          .
        7      .4099      .0940       4.36        .23        .59
```

Die mögliche Einsparung eines Koeffizienten läßt sich bereits in Abbildung 2 erkennen. Der zweite Interaktionseffekt ist dort nämlich mit einem Wert von 0.8205 ziemlich genau das Doppelte des Wertes 0.4138 des ersten Interaktionseffekts. Wird nun in der Design-Matrix für das Modell aus Abbildung 2 das zweifache der letzten Spalte zur vorletzten Spalte addiert und die letzte Spalte anschließend gelöscht, ergibt sich gerade die Design-Matrix für das Modell aus Abbildung 5. Zur Verdeutlichung sind in Tabelle 6 noch einmal beide Design-Matrizen direkt untereinander geschrieben.

Die Design-Matrix für das Modell aus Abbildung 5 läßt sich daher auch als Spezifikation einer linearen Restriktion über die Koeffizienten der Design-Matrix aus Abbildung 2 verstehen, bei der der letzte Parameter aus dem Modell in Abbildung 2 auf das Doppelte des vorletzten Parameters festgesetzt wird. Tatsächlich lassen sich mit benutzerdefinierten Design-Matrizen beliebige lineare Restriktionen über die Parameter eines log-linearen Modells spezifizieren und schätzen. Die Logik entspricht der an anderer Stelle vorgestellten Logik der Spezifikation sparsamer logistischer Modelle (***Kühnel***, 1992).

Tabelle 6: Zusammenfassung von Spalten der Design-Matrix

a. Designmatrix des saturierten Ausgangsmodells (Abbildung 2)

V326	V20	'1'	D_1	D_2	D_3	$D_1 \cdot D_3$	$D_2 \cdot D_3$
1	1	1	1	0	1	1	0
1	2	1	0	1	1	0	1
1	3	1	0	0	1	0	0
2	1	1	1	0	0	0	0
2	2	1	0	1	0	0	0
2	3	1	0	0	0	0	0

b. Designmatrix des nicht saturierten Modells (Abbildung 5)

V326	V20	'1'	D_1	D_2	D_3	$(D_1 \cdot D_3) + 2 \cdot (D_2 \cdot D_3)$
1	1	1	1	0	1	1
1	2	1	0	1	1	2
1	3	1	0	0	1	0
0	1	1	1	0	0	0
0	2	1	0	1	0	0
0	3	1	0	0	0	0

Ausgehend von den Parameterschätzungen (Abbildung 5) läßt sich das Modell noch weiter vereinfachen. Der zweite Haupteffekt für V20 (D_2) ist nämlich wie bereits im saturierten Modell aus Abbildung 2 bei einer Irrtumswahrscheinlichkeit von 5% nicht signifikant von null verschieden. Es sollte daher möglich sein, auch diesen Koeffizienten auf null zu setzen. Dies kann über die Spezifikation einer benutzerdefinierten Design-Matrix erfolgen, bei der auch die Design-Variablen für die Haupteffekte als Kovariaten eingehen. Der verbleibende erste Haupteffekt von V20 läßt sich aber auch in der /DESIGN-Option durch die Angabe der Kategoriennummer in einer Klammer nach dem Variablennamen ansprechen:

```
genlog V326 V20 with INT
    /print freq des est /plot none /crit delta (0)
    /design V20(1) V326 INT .
```

Das Modell postuliert gegenüber dem Modell aus Abbildung 5 zusätzlich, daß im Durchschnitt bei Kontrolle der unterschiedlichen Häufigkeiten der Wähler und Nichtwähler und des Zusammenhangs zwischen Wahlbeteiligung und Bürgernähe von Politikern die Häufigkeiten der Meinungslosen (V20=3) sich nicht von den Häufigkeiten derjenigen unterscheiden, die Politiker nicht für desinteressiert halten (V20=2). Abweichungen gegenüber diesen beiden Kategorien weist nur die erste Kategorie auf, für die daher eine dummykodierte De-

sign-Variable spezifiziert ist. Die Schätzung des Modells gibt bei zwei Freiheitsgraden einen Chiquadrat-Wert von 0.34 für die Likelihood-Ratio Teststatistik und einen Wert von 0.35 für Pearsons Teststatistik. Auch dieses Modell paßt also gut zu den Daten. Da sich die zusätzliche Einsparung eines Koeffizienten aber nur auf die Randverteilung einer Variable (V20) bezieht und nicht auf die Beziehung zwischen den Variablen, ist das Einsparen eines Parameters inhaltlich kaum interessant. Auf eine Wiedergabe der geschätzten Koeffizienten sei daher hier verzichtet.

3. Schlußbemerkung

Oftmals steht der Anwendung log-linearer Modelle bei der Analyse kategorialer Daten die Vorstellung entgegen, daß diese Modelle recht kompliziert und kaum zu interpretieren seien. Die Konzeption log-linearer Modelle als Regressionsmodelle für logarithmierte Häufigkeiten kann m.E. die Interpretation erleichtern. Der wesentliche Unterschied zur linearen Regression mit nominalskalierten unabhängigen Variablen bzw. zur Varianzanalyse besteht dann allein darin, daß Zusammenhänge zwischen zwei Variablen nicht durch Haupteffekte, sondern durch Interaktionseffekte modelliert werden. Die Sichtweise als Regressionsmodell erleichtert auch das Verständnis der Design-Matrix eines log-linearen Modells, die dann der üblichen Datenmatrix der Prädiktoren in der linearen Regression entspricht.

Durch die Verwendung benutzerdefinierter Design-Matrizen ergibt sich eine hohe Flexibilität bei der Modellspezifikation. Es ist z.B. möglich, sehr sparsame Modelle zu schätzen, die keine "überflüssigen" Parameter enthalten. Inferenzstatistisch gesehen erhöht dies die Teststärke bei Hypothesenprüfungen. Wichtiger ist aber, daß die Möglichkeiten benutzerdefinierter Design-Matrizen dazu genutzt werden können, Modelle zu spezifizieren, die für die Untersuchung der jeweils interessierenden inhaltlichen Fragestellung angemessener sind als Standardmodelle. Als einfaches Beispiel wurde oben die Beziehung zwischen der Bürgernähe von Politikern und der Wahlbeteiligung als eine quasi-metrische Beziehung aufgefaßt, die über einen einzigen Koeffizienten (Interaktionseffekt) erfaßt werden kann.

Voraussetzung für die Nutzung dieser Möglichkeiten log-linearer Analysen ist eine Software, die den Anwendern den direkten Zugriff auf die Design-Matrix erlaubt. In der SPSS-Prozedur GENLOG ist dies über die Spezifikation von Kovariaten möglich. Die gleiche Technik läßt sich aber auch bei Nutzung der älteren Prozedur LOGLINEAR anwenden, bei der außerdem benutzerspezifische Design-Matrizen direkt spezifiziert werden können (siehe Anhang). Die einzige Einschränkung in SPSS besteht darin, daß es in beiden Prozeduren nicht möglich ist, Modelle ohne Konstante zu schätzen. Ansonsten stehen aber auch mit diesen Standardprozeduren alle Möglichkeiten der log-linearen Analyse mit benutzerdefinierten Design-Matrizen zur Verfügung.

Literatur:

Andreß, ***H.-J***., ***Hagenaars***, ***J***. und ***Kühnel***, ***S***.***M***. (1997)
Analyse von Tabellen und kategorialen Daten. Berlin u.a.: Springer.

Kühnel, ***S.M.*** (1992)
Sparsame Modellierung mit logistischen Zufallsnutzenmodellen. ZA- Information 31, S. 70-92.

Norusis, ***M.J***. und SPSS Inc. (1994)
SPSS Advanced Statistics 6.1. Chicago: SPSS Inc.

Zentralarchiv (1996): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ALLBUS 1996. Codebuch, ZA-Nr. 2800. Köln, ZA.

ANHANG:

Das Einlesen von Tabellendaten in SPSS und die Spezifikation benutzerdefinierter Design-Matrizen in der Prozedur LOGLINEAR

Log-lineare Modelle basieren auf den Besetzungen von Tabellenzellen. Solche Informationen können ohne großen Aufwand in SPSS eingelesen werden. Um mit der Prozedur GENLOG oder LOGLINEAR die Daten aus Tabelle 1 zu analysieren, können folgende SPSS-Anweisungen verwendet werden:

```
data list free / V326 V20 N.
begin data
1 1 2190   1 2 451   1 3 180   2 1 547   2 2 75   2 3 68
end data.
formats V20 V326 (f1.0).
var lab V20 'Politiker sind nicht an Problemen interessiert'
  /V326'Wahlbeteiligung letzte Bundestagswahl'.
val lab V20 V326 1'ja' 2'nein' 3'weiß nicht'.
weight by N.
```

In der Anweisung "data list" werden die beiden Modellvariablen V326 und V20 aufgeführt, die die zu analysierende Tabelle aufspannen. Die zusätzliche Variable N wird für die Eingabe der Besetzungszahlen benötigt. Die Dateneingabe erfolgt zwischen den Schlüsselwörtern "begin data" und "end data.". Für jede Tabellenzelle werden die jeweiligen Werte der Modellvariablen sowie die Besetzungzahl aufgeführt. Die drei nachfolgenden Anweisungen "formats", "var lab" und "val lab" definieren das Ausgabeformat für die Modellvariablen und Variablen- und Wertetiketten. Sie sind nicht notwendig, sondern dienen allein der Übersichtlichkeit der späteren SPSS-Ausgaben. Wichtig ist die Anweisung "weight by N.", die dafür sorgt, daß die Datenanalyse tatsächlich auf den korrekten Fallzahlen beruht.

Für die Schätzung der log-linearen Modelle muß nicht notwendigerweise die Prozedur GENLOG verwendet werden. Alle Analysen können auch mit der Prozedur LOGLINEAR ausgeführt werden, die bereits in älteren SPSS-Versionen und in SPSS/PC verfügbar ist. In

den neueren Versionen von SPSS für Windows kann die Prozedur LOGLINEAR ausschließlich über ein Syntax-Fenster angefordert werden. Der wesentliche Unterschied zu GENLOG besteht darin, daß LOGLINEAR als Voreinstellung effektkodierte Design-Variablen verwendet. Soll beispielsweise das in Abbildung 3 wiedergegebene saturierte Modell mit LOGLINEAR geschätzt werden, kann dies folgendermaßen realisiert werden:

```
loglinear V326(1 2) V20(1 3)
  /print design freq est
  /crit delta(0)
  /design V20 V326 V326*V20 .
```

Wie bei der Prozedur GENLOG folgt nach dem Prozedurnamen die Angabe der Variablen, die die zu analysierende mehrdimensionale Tabelle definieren. Im Unterschied zur neueren Prozedur muß hinter den Variablennamen der kleinste und der größte zu berücksichtigende Ausprägungswert angegeben werden. Die Optionen "/PRINT", "/CRIT" und "/DESIGN" entsprechen den gleichnamigen Optionen von GENLOG. Die Druckerausgabe ist etwas unübersichtlicher, enthält aber im wesentlichen die gleichen Informationen wie die Ausgabe von GENLOG. Um die Zuordnung der geschätzten Koeffizienten zu den Design-Variablen zu erleichtern, empfiehlt sich stets die Angabe des Schlüsselworts "design" in der /PRINT-Option.

Es ist zu beachten, daß die Prozedur LOGLINEAR sowohl bei der Wiedergabe der Design-Matrix wie auch bei der Ausgabe der geschätzten Koeffizienten die Konstante ignoriert, obwohl sie tatsächlich aus den Daten berechnet wird. Die Ursache für das Fehlen der Regressionskonstante liegt darin, daß die ML-Schätzung in der Prozedur LOGLINEAR von einer Multinomialverteilung der Häufigkeiten in den Zellen der Ausgangstabelle ausgeht. Dies hat die Konsequenz, daß die Fallzahl der zu analysierenden Tabelle ein fest vorgegebener Modellparameter ist, der nicht mehr aus den Daten geschätzt werden braucht. Dieser vorgegebene Modellparameter wird bei der Effektkodierung des log-linearen Modells durch die Regressionskonstante modelliert. Die Prozedur GENLOG unterstellt dagegen als Voreinstellung Poisson-Verteilungen für die einzelnen Tabellenzellen und betrachtet daher die Regressionskonstante nicht als vorgegebene Größe, sondern als zu schätzenden Parameter.

Wird bei der Nutzung von LOGLINEAR der Wert der Konstanten benötigt, muß dieser per Hand aus den erwarteten Häufigkeiten berechnet werden. Über die Design-Matrix läßt sich die jeweilige Rechenformel ableiten. Sind alle Design-Variablen effektkodiert, ist der Wert der Regressionskonstante stets das arithmetische Mittel der logarithmierten erwarteten Häufigkeiten. Bei dummykodierten Design-Variablen ist die Konstante der Logarithmus der erwarteten Häufigkeit der Zelle, bei der alle übrigen Design-Variablen den Wert null aufweisen.

Sollen benutzerdefinierte Design-Matrizen spezifiziert werden, ist die gleiche Vorgehensweise möglich, die auch bei der Prozedur GENLOG angewendet wird. Die gewünschten Design-Variablen sind explizit über COMPUTE-, RECODE- und/oder IF-Anweisungen zu generieren und anschließend als Kovariaten in das Modell aufzunehmen. Als Beispiel habe ich im folgenden die SPSS-Anweisungen für die Schätzung der log-linearen Modelle mit Dummykodierung aus den Abbildungen 1, 2 und 5 über die Prozedur LOGLINEAR aufgelistet. Mit den ersten drei RECODE-Anweisungen werden zunächst die drei dummykodierten Design-Variablen D1, D2 und D3 für die Haupteffekte von V20 und V326 gebildet. Die beiden nachfolgenden COMPUTE-Anweisungen generieren die Interaktionseffekte. Die letzte COMPUTE-Anweisung erzeugt die Variable INT für den Interaktionseffekt aus dem in Abbildung 5 wiedergegebenen Modell. Anschließend folgt die Spezifikation der einzelnen Modelle:

```
recode V20(1=1)(2,3=0) into D1.
recode V20(1,3=0)(2=1) into D2.
recode V326(1=1)(2=0) into D3.
compute D1D3=D1*D3.
compute D2D3=D2*D3.
compute INT=D1D3+2*D2D3.
* Modell aus Abbildung 1.
loglinear V326(1 2)  V20( 1 3) with D1 D2 D3
  /print des est /crit delta(0)
  /des D1 D2 D3.
* Modell aus Abbildung 2.
loglinear V326(1 2)  V20( 1 3) with D1 D2 D3 D1D3 D2D3
  /print des est /crit delta(0)
  /des D1 D2 D3 D1D3 D2D3.
* Modell aus Abbildung  5.
loglinear V326(1 2) V20(1 3) with D1 D2 D3 INT
  /print des est /crit delta(0)
  /des D1 D2 D3 INT.
```

Neben der Realisierung über Kovariaten erlaubt die Prozedur LOGLINEAR auch die direkte Spezifikation von Design-Matrizen in der in GENLOG nicht verfügbaren Option "/contrast". Mit dieser Option können u.a. für jede Modellvariable benutzerspezifizierte Design-Variablen gebildet werden. Ausgangspunkt ist jeweils eine quadratische Matrix, die so viele Zeilen und Spalten enthält,wie die Variable,für die DesignVariablen spezifiziert werden, Ausprägungen hat. Die Einschätzung des Desinteresses von Politikern hat drei Ausprägungen. Die Matrix für benutzerspezifizierte Design-Variablen hat entsprechend drei Zeilen mit jeweils drei Spalten. Jede Zeile steht für eine zu spezifizierende Design-Variable und jede Spalte für eine Ausprägung der Ausgangsvariable. Der Aufbau der Matrix entspricht also der Definition von Design-Variablen in Tabelle 2. Es wird allerdings eine Zeile mehr spezifiziert. Da bei einer Variable mit K Ausprägungen aber nur maximale K-1 unabhängige Design-Variablen berücksichtigt werden, sind tatsächlich nur die letzten K-1 Zeilen der Matrix relevant. Die erste Zeile steht gewissermaßen für die Regressionskonstante und

sollte daher die Werte '1' aufweisen. Sollen also für die drei Kategorien der Variable V20 zwei effektkodierte Design-Variablen gebildet werden, geschieht das über die Option "/contrast" durch folgende Spezifikation:

```
/contrast (V20)=special ( 1  1  1
                          1  0  0
                          0  1  0  )
```

Nach dem Schlüsselwort "/CONTRAST" folgt in Klammern der Name der Variable, für die benutzerspezifizierte Design-Variablen erzeugt werden sollen. Nach einem Gleichheitszeichen wird durch das Schlüsselwort 'SPECIAL' angezeigt, daß - wiederum in Klammern - die Angabe einer quadratischen Matrix mit der Definition der Design-Variablen folgt. Die erste Zeile der Matrix enthält die drei Einsen für die Regressionskonstante. Es folgen die Werte der ersten Design-Variable. Da das erste Element der Zeile den Wert '1' aufweist und die beiden übrigen Elemente jeweils den Wert '0', bedeutet dies, daß die Design-Variable bei der ersten Ausprägung von V20 den Wert '1' hat und bei den übrigen Ausprägungen den Wert null. Die Zeile definiert also eine dummykodierte Designvariable für V20=1. Die letzte Zeile enthält die Definition für die zweite Design-Variable. Das Muster "0 1 0" bewirkt, daß die Design-Variable bei der zweiten Ausprägung von V20 den Wert '1' aufweist, ansonsten den Wert '0'.

Die Elemente der quadratischen Matrix können auch hintereinander in eine Reihe geschrieben werden. Außerdem können aufeinanderfolgende Wiederholungen abgekürzt werden. Anstelle der Zeichenfolge '1 1 1' kann also die Abkürzung '3*1' geschrieben werden. Wie in SPSS üblich, können die Schlüsselwörter auch bis auf drei Zeichen abgekürzt werden. Für jede Variable kann ein eigenes /CONTRAST-Statement spezifiziert werden. Soll also das ursprünglich mit der Prozedur GENLOG geschätzte log-lineare Modell ohne Interaktionseffekte aus Abbildung 1 mit der Prozedur LOGLINEAR und der Definition von Kontrasten geschätzt werden, kann dies im Syntax- Fenster mit folgender Anweisung realisiert werden:

```
loglinear V326(1 2) V20(1 3)
  /print freq des est /crit delta(0)
  /cont (V20)=spec(1 1 1  1 0 0  0 1 0)
  /cont (V326)=spec(1 1  1 0)
  /design V20 V326 .
```

Die /CONTRAST-Option der Prozedur LOGLINEAR erlaubt zunächst nur die Definition von Design-Matrizen jeweils einer einzigen Variable, aber nicht die Definition von Interaktionseffekten zwischen mehreren Variablen. Mit einem kleinen Trick kann diese Einschränkung aber umgangen werden. Der Trick besteht darin, alle Ausprägungskombinationen der zu analysierenden Tabelle als Kategorien einer einzigen Modellvariablen zu definieren. Dann lassen sich alle Effekte einschließlich der Interaktionseffekte in einer einzigen /CONTRAST-

Anweisung spezifizieren. Um die sechs Zellen von Tabelle 1 als Ausprägungen einer neuen Variable X zu definieren, können folgende SPSS-Anweisungen formuliert werden:

```
compute X=V326*10+V20.
recode X(11=1)(12=2)(13=3)(21=4)(22=5)(23=6).
var lab X'Kombination von V326 und V20'.
val lab x 1'1 1' 2'1 2' 3'1 3' 4'2 1' 5'2 2' 6'2 3'.
```

Es ist anschließend möglich, mit der SPSS-Prozedur LOGLINEAR Modelle für die Variable X zu schätzen. Mit der Option "/CONTRAST" können dabei für X beliebige Design-Variablen spezifiziert werden. Da X sechs Ausprägungen hat, hat die benutzerspezifizierte quadratische Matrix 6 Zeilen mit je sechs Spalten. Diese Matrix ist dann aber gerade die transponierte, d.h. um 90° gekippte, Design-Matrix für ein saturiertes log-lineares Modell. Durch Nullsetzen von Zeilen dieser Matrix können sparsamere Modelle spezifiziert werden. Um beispielsweise für X das Modell aus Abbildung 1 zu schätzen, wird folgende /CONTRAST-Anweisung benutzt:

```
/contrast (X)=special   ( 1  1  1  1  1  1
                          1  0  0  1  0  0
                          0  1  0  0  1  0
                          1  1  1  0  0  0
                          0  0  0  0  0  0
                          0  0  0  0  0  0)
```

Die erste Zeile der Matrix steht für die Regressionskonstante. In den nächsten drei Zeilen werden die Design-Variablen spezifiziert. Die letzten beiden Zeilen sind auf null gesetzt. Wird diese /CONTRAST-Option beim Aufruf von LOGLINEAR verwendet, wird wiederum das Modell aus Abbildung 1 geschätzt:

```
loglinear X(1 6)
  /print freq des est
  /cont (X)=spec(6*1  1 0 0 1 0 0  0 1 0 0 1 0  1 1 1 0 0 0 12*0 )
  /crit delta(0) /design X .
```

Die letzten beiden Spalten der Design-Matrix enthalten nur Nullen. Die diesen Spalten zugeordneten Koeffizienten erhalten dann automatisch ebenfalls den Wert null. Deren Standardfehler und darauf aufbauende Statistiken werden nicht berechnet. Außerdem werden die Spalten bei der Berechnung der Freiheitsgrade für die Goodness-of-Fit Statistiken nicht berücksichtigt.

Regressionsanalyse mit Panel-Daten: Eine Einführung

von Björn Alecke[1]

Zusammenfassung

Dieser Beitrag gibt eine Einführung in die verschiedenen Verfahren zur regressionsanalytischen Behandlung von Panel-Daten, in der auf Ableitungen und eine extensive Benutzung von Matrix-Algebra verzichtet wird. Allerdings kann die Darstellung nicht ganz ohne eine formale Schreibweise auskommen, wobei jedoch im letzten Abschnitt anhand eines konkreten Rechenbeispiels die benutzten Formeln näher erläutert werden. Auf diese Weise möchte der Beitrag einerseits aufzeigen, daß die Regressionsanalyse mit Panel-Daten im wesentlichen nur auf rechentechnisch einfach durchzuführende Transformationen der Daten hinausläuft und mit Hilfe der üblichen Statistik-Programmpakete (z.B. SPSS) durchgeführt werden kann, und andererseits einen leichteren Zugang zur bestehenden Lehrbuchliteratur ermöglichen.

Abstract

This article gives a short introduction into existing methods of investigating panel data by means of regression analysis without relying on extensive use of matrix algebra and formal derivatives. Although a formal presentation can not be completely avoided, a simple example is given in the final section to illustrate the main formulas. By doing this, the article is intended on the one hand, to show that regression analysis of panel data requires in essence only straightforward transformations of data, and on the other hand, to permit a more readily accessability to the textbook literature for interested readers.

[1] ***Björn Alecke*** (Dipl.-Vw.) ist wissenschaftlicher Mitarbeiter an der Universität Münster, Institut für Wirtschafts- und Sozialgeschichte, Hüfferstr. 1a, 48149 Münster.

I. Einleitung

Mit Panel- oder auch Longitudinaldaten bezeichnet man einen Datensatz, der bei $i = 1,2,\ldots N$ Untersuchungseinheiten (Merkmalsträgern) die beobachteten Werte einer oder mehrerer Variablen (die Merkmale) für $t = 1,2,\ldots T$ verschiedene Zeitpunkte erfaßt. Als Beispiel für einen Panel-Datensatz könnte man etwa die Erfassung von Stimmenanteilen für politische Parteien und bestimmte sozio-ökonomischen Variablen bei *N* Wahlkreisen über *T* Wahlperioden anführen. Möchte man mit Hilfe dieses Datensatzes beispielsweise die Hypothese überprüfen, daß der Anteil der SPD-Wähler vom Ausmaß der Arbeitslosigkeit beeinflußt wird, so könnte man folgende Regressionsfunktion $SPD = \alpha + \beta \cdot ALQ$ schätzen, wobei die Verwendung von Paneldaten gegenüber reinen Querschnitts- bzw. Zeitreihendaten eine Reihe von Vorteilen bietet: unmittelbar offensichtlich ist die gestiegene Zahl von Freiheitsgraden, da die Stichprobengröße hier *NT* beträgt. Dies wird die Genauigkeit (Effizienz) der Schätzung erhöhen. Ebenso vermindert sich die Gefahr von Multikollinearität, da im allgemeinen bei Paneldaten die Streuung zwischen den erklärenden Variablen größer sein wird. Ein wesentlicher Vorteil liegt darin, daß erst Paneldaten die Beantwortung bestimmter ökonomischer Fragestellungen ermöglichen. Dazu sei ein von ***Baltagi*** (1995, S.5) angeführtes Beispiel wiedergegeben:

"Suppose that we have a cross-section of women with a 50% average yearly labour force participation rate. This might be due to (a) each woman having a 50% chance of being in the labour force, in any given year, or (b) 50 % of the women work all the time and 50% do not. Case (a) has high turnover, while case (b) has no turnover. Only panel data could discriminate between these cases".

Der zusätzliche Nutzen von Paneldaten ist jedoch auch mit Kosten verbunden, die in der Form höherer Anforderungen bei der Durchführung der Regressionsanalyse bestehen. Deutlich wird dies in der Vielzahl der in der Ökonometrie entwickelten Verfahren zur Behandlung von Paneldaten, die in der untenstehenden Tabelle 1 aufgeführt werden.

Die Verfahren unterscheiden sich dabei hinsichtlich der getroffenen Annahmen über den deterministischen und stochastischen Teil des (linearen) Regressionsmodells. Unterschiedliche Annahmen über den stochastischen Teil sind auch aus der herkömmlichen Regressionsanalyse bekannt. Zumeist wird davon ausgegangen, daß bei Zeitreihenuntersuchungen eine Autokorrelation der Residuen bestehen kann, während bei Querschnittsuntersuchungen oftmals Heteroskedastizität der Residuen zu beobachten ist. Bei Paneldaten als kombinierte Querschnitts - und Zeitreihenuntersuchung können deshalb in vielen Fällen Residuen vermutet werden, die sowohl durch Autokorrelation als auch durch Heteroskedastizität gekennzeichnet sind.

Tabelle 1: Taxonomie von Regressionsmodellen mit kombinierten Zeitreihen- und Querschnittsdaten

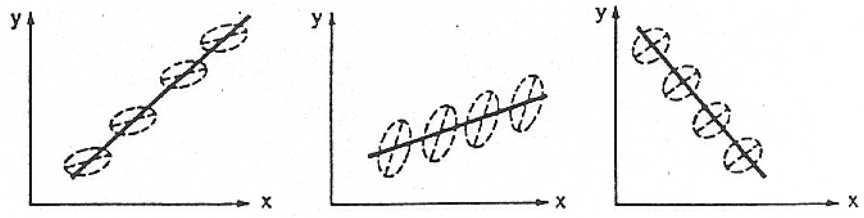
Annahmen über			
Regressionskonstante	Regressionsgewichte	Residuen	Modellbezeichnung
gemeinsam über alle *i* und *t*	gemeinsam über alle *i* und *t*	$E(\mathbf{ee}') = \sigma_e^2 \mathbf{I}_{NT}$	Classical Pooling
gemeinsam über alle *i* und *t*	gemeinsam über alle *i* und *t*	$E(\mathbf{ee}') = \mathbf{V}$	Kmenta-Model
verschieden über alle *i*	gemeinsam über alle *i* und *t*	Fixed effects	Least Squares Dummy Variable-Model
verschieden über alle *i*	gemeinsam über alle *i* und *t*	Random effects	Error Components Model
verschieden über alle *i* und *t*	gemeinsam über alle *i* und *t*	Fixed effects	Least Squares Dummy Variable-Model
verschieden über alle *i* und *t*	gemeinsam über alle *i* und *t*	Random effects	Error Components Model
verschieden über alle *i*	verschieden über alle *i*	Fixed effects	SURE Model
verschieden über alle *i*	verschieden über alle *i*	Random effects	Swamy Random Coefficient Model
verschieden über alle *i* und *t*	verschieden über alle *i* und *t*	Random effects	Hsiao Random Coefficient Model

Quelle: vgl. Johnston (1986), S. 397 und Judge (1985), S. 517.

Die verschiedenen Annahmen über den deterministischen Teil, also divergierende Parameterwerte für verschiedene Individuen und Zeitpunkte, lassen sich mit einer eventuell bestehenden Heterogenität von Paneldaten begründen, deren Nichterkennen zu Verzerrungen bei der Koeffizientenschätzung führen kann (Heterogenitätsbias). In Anlehnung an ***Hsiao*** (1986, S.6) sei folgendes Beispiel für die Schätzung der obigen Regressionsfunk-tion zwischen Stimmenanteil und Arbeitslosigkeit gegeben. Angenommen, man habe für den gesamten Datensatz, d.h. über alle Wahlkreise und -perioden, die Regressionsfunktion ge-

schätzt und dabei unterstellt, die Parameterwerte für α und β seien für alle Wahlkreise gleich, so könnten sich die in den Abbildungen 1 und 2 dargestellten Fälle ergeben.

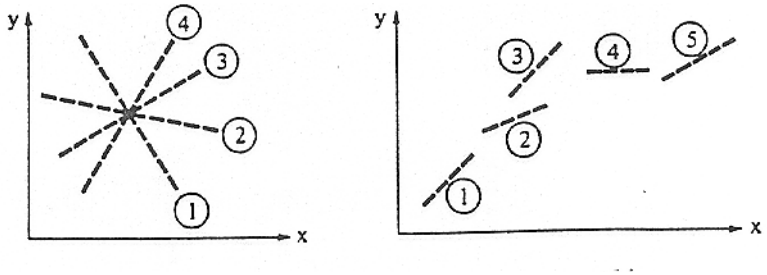
Abbildung 1



1a) 1b) 1c)

Quelle:Hsiao (1986), S.7

In Abbildung 1 wurde angenommen, daß jeder der (hier zur Vereinfachung nur 4 dargestellten) Wahlkreise einen anderen Wert für die Regressionskonstante α aufweise, so daß eigentlich die gestrichelten Linien die jeweiligen für die Wahlkreise gültigen Regressionsfunktionen angeben sollten. Vernachlässigt man den individuellen Unterschied in den Regressionskonstanten, so wird die durchgezogene Linie geschätzt, die offensichtlich sowohl für die Regressionskonstante als auch für das Regressionsgewicht β einen falschen Schätzwert liefert. Dabei zeigt sich, daß je nach Situation die Richtung der Verzerrungen unterschiedlich sein kann (Fall 1a) Überschätzung von β, 1b) Unterschätzung, 1c) falsches Vorzeichen). **Abbildung 2**



2a) 2b)

Quelle: Hsiao (1986), S.7

Abbildung 2 zeigt den Fall, daß sowohl α und β für die Wahlkreise verschieden sind. Im Fall 2a) würde die (nicht gezeigte) Regressionsgerade für den kompletten Datensatz eine Art Mittelwert aus den individuellen Regressionsgeraden bilden, während für den Fall 2b) sich eine logarithmische Form ergäbe. Folglich muß man bei Paneldaten in Erwägung ziehen, daß die zu schätzenden Parameter über die Individuen und/oder über die Zeit variieren.

Grundsätzlich lassen sich hier vier Fälle unterscheiden:

1. Die Regressionsgewichte sind konstant, aber die Regressionskonstante variiert über die Individuen:

$$y_{it} = \beta_{1,i} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \ldots + \beta_K x_{K,it} + e_{it}$$

2. Die Regressionsgewichte sind konstant, aber die Regressionskonstante variiert über die Individuen und die Zeit:

$$y_{it} = \beta_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \ldots + \beta_K x_{K,it} + e_{it}$$

3. Alle Koeffizienten, also Regressionsgewichte und Regressionskonstante, variieren über die Individuen:

$$y_{it} = \beta_{1,i} + \beta_{2,i} x_{2,it} + \beta_{3,i} x_{3,it} + \ldots + \beta_{K,i} x_{K,it} + e_{it}$$

4. Alle Koeffizienten, also Regressionsgewichte und Regressionskonstante, variieren über die Individuen und die Zeit:

$$y_{it} = \beta_{1,it} + \beta_{2,it} x_{2,it} + \beta_{3,it} x_{3,it} + \ldots + \beta_{K,it} x_{K,it} + e_{it}$$

Diese Fallunterscheidungen spiegeln sich in Tabelle 1 in den unterschiedlichen Annahmen über den deterministischen Teil der zu schätzenden Regressionsfunktion wider. In den folgenden Abschnitten werden die angeführten Modelle näher vorgestellt, wobei im letzten Abschnitt ein simplifiziertes Rechenbeispiel gegeben wird.

II. Modelle mit konstantem Parametervektor

1. Das "Classical Pooling"-Modell

Dieses Modell ist eine einfache Erweiterung des klassischen (linearen) Regressionsverfahrens auf einen Paneldatensatz. Neben dem konstanten Parametervektor für jedes Individuum und über alle Zeitpunkte wird für die Residuen Homoskedastiziät und fehlende Autokorrelation sowie auch eine fehlende Korrelation ihrer Ausprägungen zwischen den Individuen unterstellt.

Für den Fall von *K* erklärenden Variablen lauten die Beobachtungsgleichungen für *N* Individuen ($i=1,2,..N$) und bei *T* Zeitpunkten ($t=1,2...T$):

$$
\begin{aligned}
y_{11} &= \beta_1 + \beta_2 x_{2,11} + \beta_3 x_{3,11} + \ldots + \beta_K x_{K,11} + e_{11} \\
y_{12} &= \beta_1 + \beta_2 x_{2,12} + \beta_3 x_{3,12} + \ldots + \beta_K x_{K,11} + e_{12} \\
&\vdots \\
y_{1T} &= \beta_1 + \beta_2 x_{2,1T} + \beta_3 x_{3,1T} + \ldots + \beta_K x_{K,1T} + e_{1T} \\
y_{21} &= \beta_1 + \beta_2 x_{2,21} + \beta_3 x_{3,21} + \ldots + \beta_K x_{K,21} + e_{21} \\
y_{22} &= \beta_1 + \beta_2 x_{2,22} + \beta_3 x_{3,22} + \ldots + \beta_K x_{K,22} + e_{22} \\
&\vdots \\
y_{2T} &= \beta_1 + \beta_2 x_{2,2T} + \beta_3 x_{3,2T} + \ldots + \beta_K x_{K,2T} + e_{2T} \\
&\vdots \\
y_{N1} &= \beta_1 + \beta_2 x_{2,N1} + \beta_3 x_{3,N1} + \ldots + \beta_K x_{K,N1} + e_{N1} \\
y_{N2} &= \beta_1 + \beta_2 x_{2,N2} + \beta_3 x_{3,N2} + \ldots + \beta_K x_{K,N2} + e_{N2} \\
&\vdots \\
y_{NT} &= \beta_1 + \beta_2 x_{2,NT} + \beta_3 x_{3,NT} + \ldots + \beta_K x_{K,NT} + e_{NT}
\end{aligned}
$$

Das Gleichungssystem, bestehend aus insgesamt *NT* einzelnen Beobachungsgleichungen, läßt sich kompakt in Matrixnotation schreiben

$$\mathbf{y} = \mathbf{X}ß + \mathbf{e}$$

wobei

$$
\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{NT} \end{bmatrix}_{NT \times 1}
\qquad
\mathbf{X} = \begin{bmatrix} 1 & x_{2,11} & x_{3,11} & \cdots & x_{K,11} \\ 1 & x_{2,12} & x_{3,12} & \cdots & x_{K,12} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,1T} & x_{3,1T} & \cdots & x_{K,1T} \\ 1 & x_{2,21} & x_{3,21} & \cdots & x_{K,21} \\ 1 & x_{2,22} & x_{3,22} & \cdots & x_{K,22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,NT} & x_{3,NT} & \cdots & x_{K,NT} \end{bmatrix}_{NT \times K}
\qquad
\mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1T} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{NT} \end{bmatrix}_{NT \times 1}
$$

und

$$ß = (\beta_1 \beta_2 \cdots \beta_K)'$$

Die Varianz-Kovarianzmatrix der Residuen **V** läßt sich allgemein darstellen als

$$\mathbf{V} = \begin{bmatrix} E(e_{11}^2) & E(e_{11}e_{12}) & \cdots & E(e_{11}e_{1T}) & E(e_{11}e_{21}) & E(e_{11}e_{22}) & \cdots & E(e_{11}e_{NT}) \\ E(e_{12}e_{11}) & E(e_{12}^2) & \cdots & E(e_{12}e_{1T}) & E(e_{12}e_{21}) & E(e_{12}e_{22}) & \cdots & E(e_{12}e_{NT}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_{1T}e_{11}) & E(e_{1T}e_{12}) & \cdots & E(e_{1T}^2) & E(e_{1T}e_{21}) & E(e_{1T}e_{22}) & \cdots & E(e_{1T}e_{NT}) \\ E(e_{21}e_{11}) & E(e_{21}e_{12}) & \cdots & E(e_{21}e_{1T}) & E(e_{21}^2) & E(e_{21}e_{22}) & \cdots & E(e_{21}e_{NT}) \\ E(e_{22}e_{11}) & E(e_{22}e_{12}) & \cdots & E(e_{22}e_{1T}) & E(e_{22}e_{21}) & E(e_{22}^2) & \cdots & E(e_{22}e_{NT}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_{NT}e_{11}) & E(e_{NT}e_{12}) & \cdots & E(e_{NT}e_{1T}) & E(e_{NT}e_{21}) & E(e_{NT}e_{22}) & \cdots & E(e_{NT}^2) \end{bmatrix}_{NT\times NT}$$

wobei die Blöcke entlang der Hauptdiagonalen die Varianz-Kovarianzmatrix für das jeweilige Individuum angeben, während abseits der Diagonalen die Kovarianzen abgebildet werden, die zwischen den Individuen bestehen. Die Matrix hat also folgende Struktur

$$\begin{bmatrix} [\mathbf{V_1}] & [\mathbf{V_{1,2}}] & \cdots & [\mathbf{V_{1,N}}] \\ [\mathbf{V_{2,1}}] & [\mathbf{V_2}] & \cdots & [\mathbf{V_{2,N}}] \\ \vdots & \vdots & \vdots & \vdots \\ [\mathbf{V_{N,1}}] & [\mathbf{V_{N,2}}] & \cdots & [\mathbf{V_N}] \end{bmatrix}$$

Unter den Annahmen des klassischen Regressionsmodells von Homoskedastizität

$E(e_{it}^2) = \sigma_e^2$ für $i = 1,2,\ldots N$

und fehlender Korrelation der Residuen zu verschiedenen Zeitpunkten und zwischen den Individuen

$E(e_{it}e_{is}) = 0 \quad (t \neq s)$

$E(e_{it}e_{js}) = 0 \quad (i \neq j)$ für $t,s = 1,2,\ldots T$

läßt sie sich einfacher als

$$\mathbf{V} = \begin{bmatrix} \sigma_e^2 & 0 & \cdots & 0 \\ 0 & \sigma_e^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_e^2 \end{bmatrix}_{NT\times NT} = \sigma_e^2 \mathbf{I}_{NT}$$

mit $\mathbf{I}_{NT}$ als Einheitsmatrix der Ordnung *NT* schreiben. Weil zwischen den Individuen keine Korrelation besteht, weisen die Blöcke $\mathbf{V_{i,j}}$ abseits der Hauptdiagonalen den Wert 0 auf. Wegen der Unabhängigkeitsannahme der Residuen zu verschiedenen Zeitpunkten für jedes

einzelne Individuum sind innerhalb der Blöcke $\mathbf{V_1}, \mathbf{V_2}$ ... $\mathbf{V_N}$ nur die Elemente auf der Diagonalen (d.h. zu gleichen Zeitpunkten) besetzt. Da für jedes Individuum diese Varianz σ_e^2 als gleich angenommen wurde, lassen sich diese Blöcke als N Einheitsmatrizen der Ordnung T interpretieren, so daß sich insgesamt die Matrix wie oben aufgezeigt ergibt.

Vorausgesetzt die getroffenen Annahmen stimmen mit dem datengenerierenden Prozeß überein, kann dieses Modell mit der üblichen Kleinste Quadrate Methode (Ordinary Least Squares, im folgenden OLS) geschätzt werden. Unterstellt man weiterhin, wie üblich, normalverteilte Residuen, können die bekannten Verfahren zur Hypothesenüberprüfung und zur Bildung von Konfidenzintervallen verwendet werden.

2. Das "Kmenta"-Modell

Wie weiter oben schon erwähnt kann nicht immer davon ausgegangen werden, daß die restriktiven Annahmen über die Residuen erfüllt sind. ***Kmenta*** (1986, S.618ff.) hat deshalb die Schätzung des "Classical Pooling"-Modells dahingehend modifiziert, daß Heteroskedastizität und Autokorrelation berücksichtigt werden. Es ändert sich folglich die Struktur der Varianz-Kovarianzmatrix der Residuen.

Es besteht jetzt die Möglichkeit zu einer individuell verschiedenen Varianz der Residuen:

$$E(e_{it}^2) = \sigma_i^2$$

Die Residuen für jedes Individuum sind autokorreliert:

$$e_{it} = \rho_i e_{i,t-1} + u_{it}$$

wobei $u_{it} \sim N(0, \sigma_{ui}^2)$, $e_{it} \sim N(0, \frac{\sigma_{ui}^2}{1-\rho_i^2})$ und $E(e_{i,t-1}u_{jt}) = 0$ für alle i,j

Zu beachten ist, daß hier die Möglichkeit verschiedener Autokorrelationskoeffizienten für die einzelnen Individuen gegeben ist. Aus obigen Annahmen läßt sich die Kovarianz der Residuen des jeweiligen Individuums zu verschiedenen Zeitpunkten ableiten:

$$\mathrm{E}\left(\mathrm{e_{it}}\,\mathrm{e_{is}}\right) = \rho_{\mathrm{i}}^{\mathrm{t-s}}\sigma_{\mathrm{i}}^2 \quad (\mathrm{t} \neq \mathrm{s})$$

Weiterhin wird jedoch angenommen, daß zwischen den Individuen keine Korrelation besteht,

$$E(e_{it}e_{js}) = 0 \quad (i \neq j) \text{ für } t,s = 1,2,\ldots T$$

so daß sich durch Einsetzen dieser Werte in die allgemeine Form der Varianz-Kovarianzmatrix der Residuen **V**, eine blockdiagonale Darstellung für diese ergibt

$$\mathbf{V} = \begin{bmatrix} \mathbf{V_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V_N} \end{bmatrix}_{NT \times NT}$$

wobei

$$\mathbf{V_i} = \sigma_i^2 \begin{bmatrix} 1 & \rho_i & \rho_i^2 & \cdots & \rho_i^{T-1} \\ \rho_i & 1 & \rho_i & \cdots & \rho_i^{T-2} \\ \rho_i^2 & \rho_i & 1 & \cdots & \rho_i^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1 \end{bmatrix}_{T \times T}$$

und jede **0** für eine Matrix der Ordnung [*T*x*T*] steht.

Da die Annahmen des klassischen Regressionsmodells verletzt sind, d.h. die Varianz-Kovarianzmatrix der Residuen besitzt nicht mehr die Form $\mathbf{V} = \sigma_e^2 \mathbf{I}_{NT}$, würde eine Schätzung dieses Modells mit Hilfe der OLS Methode zwar zu konsistenten, aber nicht mehr effizienten Schätzern führen. Hinzu kommt, daß die OLS-Formel für die Berechnung der Standardfehler des Parametervektors verzerrt ist, so daß im folgenden das Verallgemeinerte KQ-Verfahren (Generalized Least Squares, GLS) anzuwenden ist. Allerdings wird theoretisch davon ausgegangen, die Werte der Matrix **V** seien bekannt, man hätte also die nötigen Informationen über die Varianzen und Kovarianzen der nicht beobachtbaren Residuen. Der Parametervektor läßt sich beim GLS-Verfahren über die Formel

$$\tilde{\boldsymbol{ß}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

schätzen. Die Verwendung der Formel ist aber nicht operational, da die meisten Computer-Programme eine Schätzung des Parametervektors über die OLS-Formel vornehmen und die GLS-Formel nicht implementiert ist. Nun läßt sich aber zeigen, daß unter allgemeinen Voraussetzungen die Matrix **V** sich folgendermaßen zerlegen läßt.

$$\mathbf{PVP}' = \sigma_e^2 \mathbf{I}.$$

Multipliziert man die Beobachtungsgleichung auf beiden Seiten mit der Matrix **P,** so ergibt sich

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{ß} + \mathbf{Pe} \text{ bzw.} \mathbf{y}^* = \mathbf{X}^*\boldsymbol{ß} + \mathbf{e}^*$$

Für die Varianz-Kovarianzmatrix der Residuen **e*** des transformierten Modells sind nun die klassischen Annahmen wieder erfüllt, wie aus

$$E\left[\mathbf{e}^{*}\mathbf{e}^{*\prime}\right] = E\left[\mathbf{P}\mathbf{e}\mathbf{e}'\mathbf{P}'\right] = \mathbf{P}E\left[\mathbf{e}\mathbf{e}'\right]\mathbf{P}' = \mathbf{P}\mathbf{V}\mathbf{P}' = \sigma_e^{\,2}\mathbf{I}$$

ersichtlich ist. Wendet man das OLS-Verfahren auf das transformierte Modell an, so ergibt sich der gesuchte GLS-Schätzer $\tilde{\boldsymbol{ß}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

In der Praxis sind allerdings die Varianzen und Kovarianzen der nicht beobachtbaren Residuen, also die Werte von **V**, nicht bekannt. Die Lösung dieses Problems liegt in einer Schätzung dieser Werte und der entsprechenden Transformationsmatrix, die dann zur Berechnung von **ß** verwandt werden. Dieses Verfahren bezeichnet man deshalb mit Estimated Generalized Least Squares (EGLS) mit der entsprechenden Formel

$$\hat{\tilde{\boldsymbol{ß}}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

Dabei läßt sich zeigen, daß eine konsistente Schätzung der Werte in $\hat{\mathbf{V}}$ zu einem Schätzwert von **ß** führt, der in großen Stichproben dieselben optimalen Eigenschaften des GLS-Schätzers aufweist. Ein bekanntes Beispiel für eine EGLS-Schätzung ist das sogenannte Cochrane-Orcutt Verfahren zur Beseitigung von Autokorrelation. Genau dieses Verfahren wird auch von ***Kmenta*** zur Ermittlung einer Transformationsmatrix übernommen, mit der eine Beseitigung der Autokorrelation der Residuen im ersten Schritt erfolgt. Hierauf folgt dann in einem zweiten Schritt eine weitere Transformation der beobachteten Werte zur Beseitigung der Heteroskedastizität. Im einzelnen lauten die Schritte:

- Durchführung einer OLS-Schätzung für alle NT Beobachtungswerte und anschließende Berechnung des Autokorrelationskoeffizienten für jedes Individuum über folgende Formel:

$$\hat{\rho}_i = \frac{\sum_{t=2}^{T} e_{it}e_{i,t-1}}{\sum_{t=2}^{T} e_{i,t-1}^2}$$

- Transformation der Beobachtungswerte von erklärter und erklärenden Variablen gemäß

$$\begin{aligned}
\mathrm{y_{it}}^{*} &= \sqrt{1-\hat{\rho}_\mathrm{i}^2}\ \mathrm{y_{it}} && \textbf{für } \mathrm{t}=1\\
\mathrm{y_{it}}^{*} &= \mathrm{y_{it}} - \hat{\rho}_\mathrm{i}\mathrm{y_{i,t-1}} && \textbf{für } \mathrm{t}=2,3\ldots\mathrm{T}\\
\mathrm{x_{k,it}}^{*} &= \sqrt{1-\hat{\rho}_\mathrm{i}^2}\ \mathrm{x_{k,it}} && \textbf{für } \mathrm{t}=1\\
\mathrm{x_{k,it}}^{*} &= \mathrm{x_{k,it}} - \hat{\rho}_\mathrm{i}\mathrm{x_{k,i,t-1}} && \textbf{für } \mathrm{t}=2,3\ldots\mathrm{T}
\end{aligned}$$

OLS-Schätzung über das folgende transformierte Modell

$$y_{it}^{*} = \beta_1^{*} + \beta_2 x_{2,it}^{*} + \beta_3 x_{3,it}^{*} + \ldots + \beta_K x_{K,it}^{*} + e_{it}^{*}$$

- Anschließende Schätzung der Varianz der Residuen e_{it} * für jedes Individuum über

$$\hat{\sigma}_i^2 * = \frac{1}{T-K} \sum_{t=1}^{T} \hat{e}_{it}^{*2}$$

wobei diese Varianzschätzung zur weiteren Transformation verwendet wird, um die noch im Ansatz befindliche Heteroskedastiztät zu eliminieren.

- Nochmalige Transformation von y_{it} * und $x_{k,it}$ * gemäß

$$y_{it} ** = \frac{y_{it}*}{\hat{\sigma}_i *} \text{ und } x_{k,it} ** = \frac{x_{k,it}*}{\hat{\sigma}_i *}$$

und OLS-Schätzung folgender Modellgleichung

$$y_{it} ** = \beta_1 ** + \beta_2 x_{2,it} ** + \beta_3 x_{3,it} ** + \ldots + \beta_K x_{K,it} ** + e_{it} **$$

Da die Residuen e_{it} ** nach diesen Transformationen wieder die klassischen Annahmen erfüllen, also weder Autokorrelation noch Heteroskedastizität aufweisen, besitzt die OLS-Schätzung von **ß** in großen Stichproben wieder wünschenswerte Eigenschaften und die üblichen Verfahren zur Hypothesenüberprüfung und Bildung von Konfidenzintervallen können angewandt werden.

Eine Vereinfachung des obigen Modells besteht in der Annahme eines für alle Individuen gleichen Autokorrelationskoeffizienten, der über die Formel

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=2}^{T} e_{i,t-1}^2}$$

konsistent geschätzt werden kann. Diese Annahme kann insbesondere bei Datensätzen mit wenigen Zeitreihenbeobachtungen angebracht sein, da selbst bei voneinander abweichenden Autokorrelationskoeffizienten der Individuen durch diese Vereinfachung ein Effizienzgewinn erzielt werden kann, der sich aufgrund des größeren Unsicherheitsbereichs einer Schätzung individuell verschiedener Autokorrelationskoeffizienten auf der Basis nur weniger Zeitreihenbeobachtungen ergibt.

Bevor jedoch die obigen Transformationen im Rahmen des "***Kmenta***"-Modells vorgenommen werden, empfiehlt sich auf der Grundlage der berechneten Werte für die Autokorrelation und Varianz der Residuen eine Überprüfung der Hypothese,daß tatsächlich eineEGLS-Schätzung vonnöten ist. Hierzu können die verschiedenen in der ökonometrischen Literatur vorgeschlagenen formalen Testprozeduren auf Autokorrelation (beispielsweise der ***Durbin***-

Watson Test) oder Heteroskedastizität (beispielsweise der ***Goldfeld-Quandt*** Test) herangezogen werden (siehe hierzu etwa ***Judge*** et al. (1989), Kap.9).

Eine wesentliche Erweiterung des "***Kmenta***"-Modells besteht in der Annahme einer möglichen Korrelation der Residuen zwischen den Individuen jeweils zu gleichen Zeitpunkten ("contemporaneous correlation"),

$$E(e_{it}e_{jt}) = \sigma_{ij} \neq 0 \quad (i \neq j)$$

so daß die Matrix **V** nicht mehr blockdiagonal ist. Auch hier kann das EGLS-Verfahren über eine konsistente Schätzung der Elemente von **V** durchgeführt werden (siehe hierzu ***Kmenta*** (1986), S. 622ff.).

III. Modelle mit variablen Regressionskonstanten

1. Individual- und Zeiteffekte

Bisher wurde davon ausgegangen, daß für alle Individuen zu jedem Zeitpunkt der Parametervektor konstant sei. Weiter oben wurde darauf hingewiesen, daß eine solche Annahme in vielen Fällen nicht gerechtfertigt ist und zu verzerrten Parameterschätzungen führen kann. Im folgenden sollen Modelle vorgestellt werden, die von unterschiedlichen Regressionskonstanten ausgehen. Zur Begründung dieser Annahme sei noch einmal das Beispiel der Wahlfunktion betrachtet, deren zu schätzende Regressionsgleichung lautet:

$$SPD_{it} = \alpha + \beta \cdot ALQ_{it} + e_{it}$$

In der Regressionsanalyse wird angenommen, daß die Residuen e_{it} den Einfluß von sogenannten latenten bzw. impliziten Variablen wiedergeben. Diese Variablen üben zwar einen Einfluß auf die zu erklärende Variable aus, werden aber nicht in die Regressionsfunktion aufgenommen, da man ihren Einflüssen nur geringe Bedeutung zuschreibt oder aber weil diese Variablen nicht meßbar bzw. nicht beobachtbar sind. Die Residuen als Summe dieser Einflüsse werden als zufällig interpretiert, und ihre einzelnen Ausprägungen "springen" zwischen den einzelnen Beobachtungspunkten in einem nicht zu beobachteten Wertebereich hin und her. Diese Variation wird bei einem Panel-Datensatz sowohl unabhängig von dem Individuum als auch der Zeit angenommen. Allerdings könnten die Residuen auch den Einfluß latenter Variablen beinhalten, die zwar von Individuum zu Individuum verschieden, aber zeitkonstant sind, oder umgekehrt für jedes Individuum gleich, aber von Zeitpunkt zu Zeitpunkt verschieden.

Im Rahmen der Schätzung obiger Wahlfunktion wäre das mit der geographischen Lage eines Wahlkreises verbundene "traditionelle Wählerverhalten" ein Beispiel für eine latente

Variable W_i, die offensichtlich individuen-spezifische, aber zeitkonstante Ausprägungen annimmt. Umgekehrt wäre die "politische Stimmung" ein Beispiel für eine latente Variable Z_t, die zu verschiedenen Zeitpunkten divergierende, jedoch für alle Individuen gleiche Werte annimmt. Bei Berücksichtigung dieser latenten Variablen in den Residuen läßt sich dieser schreiben als

$$e_{it} = \mu W_i + \lambda Z_t + v_{it}$$

Allerdings liegen für die latenten Variablen keine Beobachtungen vor und somit ist eine Schätzung ihrer Parameter μ und λ nicht möglich. (Natürlich könnte man versuchen, die latenten Variablen explizit zu modellieren, indem man etwa den Anteil an Wählern mit einer bestimmten Konfessionszugehörigkeit eines Wahlkreises, die im Zeitablauf (nahezu) konstant sein, sich aber zwischen den Wahlkreisen unterscheiden dürfte, als Indikator für das "traditionelle Wählerverhalten" oder aber die Konjunkturlage als Ausdruck der "politischen Stimmung" bestimmt, aber es sei angenommen, daß letztendlich auch mit diesen Variablen die Einflüsse nur unzureichend quantifiziert und gemessen werden können.) Faßt man die Produkte $\mu_i = \mu W_i$ bzw. $\lambda_t = \lambda Z_t$ nun zu sogenannten Individual- und Zeiteffekten zusammen und ordnet die zu schätzende Regressionsfunktion um,

$$SPD_{it} = \alpha + \beta ALQ_{it} + \mu W_i + \lambda Z_t + v_{it} = \alpha + \mu_i + \lambda_t + \beta ALQ_{it} + v_{it} = \alpha_{it} + \beta ALQ_{it} + v_{it}$$

so erhält man ein Modell mit unterschiedlichen Regressionskonstanten. Je nachdem, ob man die Individual- und Zeiteffekte als feste Größen, und somit die unterschiedlichen Regressionskonstanten als schätzbare Parameter ansieht, oder aber als Zufallsvariablen betrachtet, ergibt sich eine weitere Unterteilung in Fixed Effects- und Random Effects-Modelle, die im folgenden näher vorgestellt werden sollen. Dabei soll vereinfachend zuerst einmal nur von dem Vorliegen von Individualeffekten μ_i ausgegangen werden.

2. Das "Least Squares Dummy Variable"-Modell

a) Einführung von Dummy-Variablen

Eine in der Regressionsanalyse übliche Methode zur Erfassung von unterschiedlichen Regressionskonstanten besteht in der Einführung von sogenannten Dummyvariablen, die jeweils den Wert 1 bei Vorliegen einer bestimmten qualitativen Ausprägung und sonst den Wert 0 annehmen. In unserem Fall ist die qualitative Ausprägung jeweils durch das einzelne Individuum gegeben, und somit sind *N* Dummy Variablen mit der folgenden Definition einzuführen

$$D_{jt} = \begin{cases} 1 & \textbf{falls } j = i \\ 0 & \textbf{falls } j \neq i \end{cases}$$

Die zu schätzende Regressionsgleichung nimmt folgende Gestalt an

$$y_{it} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k x_{k,it} + v_{it}$$

Gilt für das betrachtete *i*te Individuum *i=j*, so nimmt die Dummy-Variable zu jedem Zeitpunkt den Wert 1 an, während alle anderen Dummy-Variablen den Wert 0 aufweisen, und der geschätzte Koeffizient β_{1j} gibt die Regressionskonstante des betreffenden Individuums an. Um dies zu verdeutlichen, sei das entstehende Gleichungssystem wiedergegeben:

$$\begin{aligned}
y_{11} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,11}+\beta_3 x_{3,11}+\ldots+\beta_K x_{K,11}+e_{11}\\
y_{12} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,12}+\beta_3 x_{3,12}+\ldots+\beta_K x_{K,11}+e_{12}\\
&\vdots\\
y_{1T} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,1T}+\beta_3 x_{3,1T}+\ldots+\beta_K x_{K,1T}+e_{1T}\\
y_{21} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,21}+\beta_3 x_{3,21}+\ldots+\beta_K x_{K,21}+e_{21}\\
y_{22} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,22}+\beta_3 x_{3,22}+\ldots+\beta_K x_{K,22}+e_{22}\\
&\vdots\\
y_{2T} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,2T}+\beta_3 x_{3,2T}+\ldots+\beta_K x_{K,2T}+e_{2T}\\
&\vdots\\
y_{N1} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,N1}+\beta_3 x_{3,N1}+\ldots+\beta_K x_{K,N1}+e_{N1}\\
y_{N2} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,N2}+\beta_3 x_{3,N2}+\ldots+\beta_K x_{K,N2}+e_{N2}\\
&\vdots\\
y_{NT} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,NT}+\beta_3 x_{3,NT}+\ldots+\beta_K x_{K,NT}+e_{NT}
\end{aligned}$$

Nimmt man für die Residuen die klassischen Annahmen als gegeben an, so kann dieses Modell wie ein gewöhnlicher Gleichungssatz mit Hilfe des OLS-Verfahrens geschätzt werden. Die sich ergebenden Parameterwerte für die Dummy-Variablen können wie alle anderen Parameter behandelt werden, sind also den üblichen Testmethoden zugänglich.

b) Bildung von Durchschnitten

Die Schätzung des Modells bei Einführung von *N* Dummy-Variablen ist mit der Inversion einer Matrix der Ordnung (*N+K-1*) verbunden, was insbesondere bei großem *N*, also einer Stichprobe mit vielen Untersuchungseinheiten, zu numerischen Problemen führen kann. Es läßt sich jedoch eine andere Form der Darstellung finden, die diesen Nachteil vermei-det, indem der Vektor der Regressionsgewichte getrennt von dem der Regressionskonstante geschätzt wird. Dabei macht man sich die spezielle,sich nach Einführung von Dummy-Variablen ergebende Struktur der Matrix der erklärenden Variablen zunutze, indem man die

Formel für die sogenannte partionierte OLS-Schätzung verwendet. Zur Veranschaulichung sei noch einmal von der Gleichung ausgegangen

$$y_{it} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k x_{k,it} + v_{it}$$

Bildet man für jedes Individuum einen zeitlichen Durchschnitt der Beobachtungswerte, so erhält man folgenden Ausdruck

$$\overline{y}_{i.} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.} + \overline{v}_{i.}$$

mit

$$\overline{y}_{i.} = \frac{1}{T} \sum_{t=1}^{T} y_{it} \quad \overline{x}_{k,i.} = \frac{1}{T} \sum_{t=1}^{T} x_{k,it} \quad \overline{v}_{i.} = \frac{1}{T} \sum_{t=1}^{T} v_{it}$$

Da D_{jt} für (i=j) zu jedem Zeitpunkt den Wert 1 annimmt, gilt $D_{jt} = \overline{D}_{j.}$. Subtrahiert man nun die beiden Gleichungen voneinander, so ergibt sich

$$y_{it} - \overline{y}_{i.} = \sum_{k=2}^{K} \beta_k (x_{k,it} - \overline{x}_{k,i.}) + (v_{it} - \overline{v}_{i.})$$

Wendet man auf diese Gleichung eine OLS-Schätzung ohne Regressionskonstante an, so ist die Schätzung des Parametervektors $\mathbf{ß} = (\beta_2\ \beta_3 \ldots \beta_K)'$ identisch mit der Schätzung der Regressionsgewichte, die sich bei Einführung von Dummy-Variablen ergeben hätte. Diese Schätzung der Regressionsgewichte, die die interindividuelle Streuung durch Bildung der Durchschnitte eliminiert und nur die Streuung innerhalb eines Individuums berücksichtigt, wird als "within"-Schätzung bezeichnet. Jede der Variablen wird hier als Abweichung von individuell verschiedenen Mittelwerten ausgedrückt. Nach der Schätzung der Regressionsgewichte können die individuellen Regressionskonstanten über folgende Formel geschätzt werden.

$$\beta_{1i} = \overline{y}_{i.} - \beta_2 \overline{x}_{2,i.} - \beta_3 \overline{x}_{3,i.} - \ldots - \beta_K \overline{x}_{K,i.} = \overline{y}_{i.} - \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.}$$

Eine Schätzung der Varianz der Residuen kann über die residuale Abweichungsquadratsumme vorgenommen werden, die sich unabhängig von dem gewählten Schätzansatz (Einführung von Dummy-Variablen oder Bildung von Durchschnitten) ergibt.

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{v}_{it}^2}{NT - (N + K - 1)}$$

Im Ergebnis bleibt festzuhalten, daß im "LSDV"-Modell über die Einführung von Dummy-Variablen Individualeffekte berücksichtigt werden können. Ist die Zahl der Individuen groß, empfiehlt sich eine getrennte Schätzung von Regressionskonstanten und Regressionsgewichten. Dies kann über eine einfache Transformation der Beobachtungswerte erreicht werden, bei der jede Variable als Abweichung von individuell verschiedenene Mittelwerten ausgedrückt wird. Da so jeweils nur die zeitliche Streuung innerhalb der Individuen berücksichtigt wird, bezeichnet man diesen Schätzer auch als "within"-Schätzer.[2]

c) Test auf Vorliegen von Individualeffekten

Die naheliegende Frage, ob überhaupt von dem Vorliegen von Individualeffekten bzw. von unterschiedlichen Regressionskonstanten ausgegangen werden kann, läßt sich mit Hilfe eines F-Tests entscheiden, der die nichterklärten Abweichungsquadratsummen der Residuen miteinander vergleicht, die sich jeweils bei Verwendung des "Classical-Pooling"-Modells bzw. des "LSDV"-Modells ergeben. Die Nullhypothese lautet dementsprechend

$$H_0 = \beta_{11} = \beta_{12} = \ldots = \beta_{1N}$$

während die Alternativhypothese sich als

H_1: nicht alle der β_{1i} sind gleich

formulieren läßt. Die zugehörige Teststatistik

$$F = \frac{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2 - \sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(N-1)}}{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(NT-(N+K-1))}}$$

ist bei Gültigkeit der Nullhypothese F-verteilt mit $(N-1,\ NT-(N+K-1))$ Freiheitsgraden, wobei $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2$ die Abweichungsquadratsumme der Residuen des "Classical Pooling" -Modells ist. Dieses wird auch als das restringierte Modell bezeichnet, da hier die Restriktion gleicher Regressionskonstanten für alle Individuen vorausgesetzt wird. Die Freiheitsgrade im Zähler (*N* - 1) entsprechen dabei der Zahl der eingeführten Restriktionen. Die Abweichungsquadratsumme des unrestringierten Modells wird durch die Quadratsumme der

[2] Voraussetzung ist hierbei natürlich, daß eine zeitliche Streuung gegeben ist, die nicht durch Meßfehler hervorgerufen wird.

Residuen des "LSDV"-Modells $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2$ gegeben. Die Freiheitsgrade im Nenner $(NT-(N+K-1))$ entsprechen den Freiheitsgraden des unrestringierten Modells (Stichprobengröße minus der Zahl der zu schätzenden Parameter).

Übersteigt der empirische F-Wert bei einem vorgegebenen Signifikanzniveau $(1-\alpha)$ den kritischen F-Wert, so ist die Nullhypothese gleicher Regressionskonstanten abzulehnen und dementsprechend das "LSDV"-Modell vorzuziehen. ***Judge*** et al. (1988, S.476) weisen darauf hin, die Signifikanz der Dummyvariablen nicht über die jeweiligen t-Werte, sondern mit dem obigen F-Test, der dementsprechend auch als mehrfach partieller F-Test bezeichnet wird, zu überprüfen, da sich sonst unter bestimmten Umständen unterschiedliche Empfehlungen bezüglich der Signifikanz der individuellen Dummy-Variablen ergeben.

2. "Error-Components"-Modell

a) Herleitung der GLS-Schätzung

Bei Verwendung des "LSDV"-Modells konnte für jedes Individuum eine eigene Regressionskonstante geschätzt werden, weil die Individualeffekte als feste Größen interpretiert wurden.

$$y_{it}=\beta_1+\mu_i+\sum_{k=2}^{K}\beta_k x_{k,it}+v_{it}=\beta_{1i}+\sum_{k=2}^{K}\beta_k x_{k,it}+v_{it}$$

Im "Error Components"-Modell ("EC"-Modell) werden demgegenüber die Individualeffekte als zufällig betrachtet, so daß auch die individuellen Regressionskonstanten $\beta_{1i}=\overline{\beta}_1+\mu_i$ als Zufallsvariable zu interpretieren sind. Über die zufälligen Individualeffekte werden folgende Annahmen gemacht

$$E(\mu_i)=0 \quad E(\mu_i^2)=\sigma_\mu^2 \quad E(\mu_i\mu_j)=0$$

und weiterhin, daß die zufälligen Individualeffekte mit den Residuen unkorreliert sind. Die Individuen werden als eine Stichprobe aus einer größeren Grundgesamtheit betrachtet und das Ziel ist nun eine Schätzung des für diese Grundgesamtheit gültigen Parametervektors $\boldsymbol{ß}=(\overline{\beta}_1\ \beta_2\ldots\beta_K)'$.

Die Schätzgleichung lautet

$$y_{it}=\overline{\beta}_1+\sum_{k=2}^{K}\beta_k x_{k,it}+\mu_i+v_{it}=\overline{\beta}_1+\sum_{k=2}^{K}\beta_k x_{k,it}+w_{it}$$

wobei sich nun für den Mittelwert und die Varianzen bzw. Kovarianzen der Residuen w_{it} folgende Ausdrücke ergeben

$$E(w_{it}) = E(\mu_i) + E(v_{it}) = 0$$

$$E(w_{it}^2) = E(\mu_i^2) + E(v_{it}^2) + 2E(\mu_i v_{it}) = \sigma_\mu^2 + \sigma_v^2$$

wegen $E(\mu_i v_{it}) = 0$

$$E(w_{it} w_{is}) = E(\mu_i^2) + E(v_{it} v_{is}) + E(\mu_i v_{it}) + E(\mu_i v_{is}) = \sigma_\mu^2$$

wegen $E(v_{it} v_{is}) = 0$ für $t \neq s$

Ferner ist $E(w_{it} w_{js}) = E(\mu_i \mu_j) + E(v_{it} v_{is}) + E(\mu_j v_{it}) + E(\mu_i v_{js}) = 0$ für $t \neq s, i \neq j$

Die Varianz-Kovarianzmatrix der Residuen besitzt für jedes einzelne Individuum die gleiche Darstellung

$$\mathbf{V_i} = \begin{bmatrix} \sigma_\mu^2 + \sigma_v^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{bmatrix}_{T \times T}$$

aus der hervorgeht, daß sie die Annahme der Homoskedastizität (konstante Varianzen) erfüllen, allerdings über verschiedene Beobachtungszeitpunkte miteinander korreliert sind. Im Gegensatz zu dem von ***Kmenta*** betrachteten Fall eines autoregressiven Prozesses bleibt die Größenordnung dieser Korrelation jedoch über die Zeit konstant. Die Varianz-Kovarianz-matrix der Residuen für alle Individuen ist dementsprechend zwar blockdiagonal mit den über $\mathbf{V_i}$ gegebenen Blöcken, allerdings besitzt sie nicht die Darstellung einer den klassischen Annahmen entsprechenden Diagonalmatrix $\sigma_e^2 \mathbf{I}_{NT}$.

Deswegen ist für die Schätzung des Parametervektors das schon bei den ***Kmenta***-Modellen gezeigte GLS-bzw. EGLS-Verfahren anzuwenden. Auch hier besteht die Aufgabe in der Formulierung einer geeigneten Transformationsmatrix **P** und der Schätzung ihrer Elemente. Es läßt sich nun zeigen, daß tatsächlich die Varianz-Kovarianzmatrix der Residuen w_{it} für das "EC"-Modell eine Darstellung besitzt, die zu einer Transformation der Beobachtungswerte gemäß

$$\mathbf{Py} = \mathbf{PX}ß + \mathbf{Pe} \text{ bzw. } \mathbf{y}^* = \mathbf{X}^* ß + \mathbf{e}^*$$

führt, so daß eine anschließende OLS-Schätzung auf **y*** und **X*** der GLS-Schätzung entspricht. Ohne auf die Herleitung dieser Matrix einzugehen, seien nur die notwendigen Transformationen angeführt

$y_{it}^{ec} = y_{it} - \theta \bar{y}_{i.}$ und $x_{k,it}^{ec} = x_{k,it} - \theta \bar{x}_{k,i.}$

wobei $\theta = 1 - \frac{\sigma_v}{\sigma_1}$ mit $\sigma_1 = \sqrt{T\sigma_\mu^2 + \sigma_v^2}$

Dementsprechend ergibt sich der GLS-Schätzer über eine OLS-Schätzung folgender Gleichung

$$y_{it} - \theta \bar{y}_{i.} = (1-\theta)\bar{\beta}_1 + \sum_{k=2}^{K} \beta_k (x_{k,it} - \theta \bar{x}_{k,i.}) + w_{it}$$

Stellt man diese Gleichung der Schätzgleichung des "within"-Schätzers

$$y_{it} - \bar{y}_{i.} = \sum_{k=2}^{K} \beta_k (x_{k,it} - \bar{x}_{k,i.}) + (v_{it} - \bar{v}_{i.})$$

gegenüber, so wird die Beziehung zwischen diesen beiden Ansätzen und die Rolle der einzelnen Komponenten der Residuen ("Error Components") über den Faktor θ deutlich. Auch beim "EC"-Modell sind die transformierten Beobachtungswerte als Abweichungen vom individuellen Mittelwert aufzufassen, allerdings wird dieser vorher mit einem Faktor gewichtet, der sich als das Verhältnis von zwei Varianzen ergibt. Es läßt sich zeigen, daß der Ausdruck für den GLS-Schätzer einen gewichteten Durchschnitt aus zwei OLS-Schätzern darstellt, wobei sich die Gewichte aus der Größenordnung der jeweiligen Varianzkomponenten ergeben. Dabei ist der eine OLS-Schätzer der bekannte "within" Schätzer, der die Informationen über die Streuung innerhalb der Individuen ausnutzt, während der andere OLS-Schätzer als "between"-Schätzer bezeichnet wird, der die Streuung zwischen den Individuen ausnutzt. Dieser ergibt sich aus einer OLS-Schätzung über die *N* individuellen Mittelwerte

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{ß} + \bar{\mathbf{w}}$$

wobei $\bar{\mathbf{y}}, \bar{\mathbf{X}}$ und $\bar{\mathbf{w}}$ jeweils die typischen Elemente $\bar{y}_{i.}, \bar{x}_{k,i.}$ und $\bar{w}_{i.}$ enthalten.

Der Faktor θ entscheidet nun über den Einfluß dieser Streuungsarten, der bei der GLS-Schätzung berücksichtigt wird. Aus der Definition ergibt sich, daß θ bei großem *T* oder großem σ_μ relativ zu σ_v gegen den Wert 1 strebt. Dies ist gleichbedeutend mit einem großen Einfluß des "within"-Schätzers, im Grenzfall *T* gegen ∞ stimmen die Schätzwerte für **ß** im "LSDV"-Modell mit denen des "EC"-Modells überein. Umgekehrt, bei relativ großem σ_v gegenüber σ_μ, tendiert θ gegen 0 und der GLS-Schätzer stimmt mit dem OLS-Schätzer des "Classical Pooling"-Modells überein.

b) Die EGLS-Schätzung

In der Praxis jedoch sind die Varianzen der Residuenkomponenten nicht bekannt, so daß man vom GLS- zum EGLS-Verfahren übergeht, in dem eine Schätzung der Varianz-Kovarianzmatrix **V** bzw. der Transformationsmatrix **P** vorgenommen wird. Dies ist gleichbedeutend mit einer Schätzung der Varianzkomponenten bzw. ihrer Quadratwurzeln σ_v und σ_μ zur Bildung der gesuchten Größe θ und anschließender Transformation der Beobachtungswerte, wie oben gezeigt. Dabei macht man sich nun die schon oben erwähnten "within"- und "between"- Schätzer zunutze:

Es kann gezeigt werden, daß die beim "LSDV"-Modell aus den Residuen erhaltene ("within") Schätzung der Varianz $\hat{\sigma}_v^2$ der Residuen eine unverzerrte Schätzung für σ_v^2 darstellt. Um eine Schätzung für den Term σ_μ^2 zu erhalten , bildet man den "between"-Schätzer, der einer OLS-Schätzung über die *N* individuellen, über die Zeit gemittelten Werte der Beobachtungsvariablen entspricht.

$$\overline{y}_{i.} = \overline{\beta}_1 + \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.} + \overline{w}_{i.}$$

Bildet man die Varianz der Residuen $\overline{w}_{i.} = \mu_i + \overline{v}_{i.}$ so ergibt sich aufgrund der postulierten Unabhängigkeitsannahme

$$\text{var}(\mu_i + \overline{v}_{i.}) = \sigma_\mu^2 + \frac{\sigma_v^2}{T} = \frac{\sigma_1^2}{T}$$

Zur Schätzung der Varianz können die sich bei Durchführung der "between"-Schätzung ergebenden Residuen verwendet werden. Der so erhaltene Varianzschätzer mit *T* multipliziert, ergibt dann den gesuchten Ausdruck für die zweite Varianzkomponente in θ. Transformiert man die Beobachtungswerte von **y** und **X** gemäß

$$y_{it}^{ec} = y_{it} - \hat{\theta}\overline{y}_{i.} \text{ und } x_{k,it}^{ec} = x_{k,it} - \hat{\theta}\overline{x}_{k,i.}$$

$$\text{wobei } \hat{\theta} = 1 - \frac{\hat{\sigma}_v}{\hat{\sigma}_1} \text{ mit } \hat{\sigma}_1 = \sqrt{T\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2}$$

und führt mit den transformierten Variablen eine OLS-Schätzung durch, erhält man den gesuchten EGLS-Schätzer. Dabei ist allerdings zu beachten, daß auch die Regressionskonstante entsprechend transformiert wird. Anstelle der Einsen in der Matrix der erklärenden Variablen sind also die Werte $1-\hat{\theta}$ einzugeben. Die Vorgehensweise zur Ermittlung des EGLS-Schätzers des "EC"-Modells läßt sich zusammengefaßt darstellen:

- Berechnung des "LSDV"-Modells, entweder über Einführung von Dummy-Variablen oder über die transformierten Beobachtungswerte, ausgedrückt als Abweichung vom in-

dividuellen Mittelwert, wobei letztere Variante rechentechnische Vorteile bezüglich der später durchzuführenden Transformationen besitzt. Ermittlung der Quadratsumme der Residuen des "LSDV"-Modells, die bei beiden Vorgehensweisen identisch ist, und Schätzung der Varianzkomponente $\hat{\sigma}_v^2$ über

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(NT-(N+K-1))}$$

- Berechnung des "between"-Schätzers über eine OLS-Regression bei Verwendung der *N* individuellen Mittelwerte von erklärter und erklärenden Variablen zur Ermittlung der Quadratsumme der Residuen dieses Modells. Anschließende Schätzung der Varianzkomponente $\hat{\sigma}_1^2 = T\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2$ mit Hilfe dieser Residuen

$$\frac{\hat{\sigma}_1^2}{T} = \frac{\sum_{i=1}^{N}\hat{\bar{w}}_{i.}^2}{N-K}$$

- Berechnung von $\hat{\theta} = 1 - \frac{\hat{\sigma}_v}{\hat{\sigma}_1}$

- Bildung der transformierten Beobachtungswerte

$y_{it}^{ec} = y_{it} - \hat{\theta}\bar{y}_{i.}$ und $x_{k,it}^{ec} = x_{k,it} - \hat{\theta}\bar{x}_{k,i.}$

- Eine OLS-Regression über die entsprechend transformierten Beobachtungswerte ist gleichbedeutend mit der gesuchten EGLS-Schätzung

c) Test auf Vorliegen von Individualeffekten

Wie beim "LSDV"-Modell kann auch für das "EC"-Modell die Frage, ob Individualeffekte vorliegen, mit Hilfe des F-Tests entschieden werden, der die Quadratsummen der Residuen des restringierten Modells ("Classical Pooling") mit denen des unrestringierten Modells ("LSDV") vergleicht. Die Bildung des "LSDV"-Modells zur Erfassung von Individualeffekten reicht für diese Entscheidung, unabhängig davon, ob man die Individualeffekte als fest oder zufällig betrachtet. Da zur praktischen Ermittlung des EGLS-Schätzers die "within"-Schätzung, also das "LSDV"-Modell, ohnehin benötigt wird, bedeutet dies keinen zusätzlichen Rechenaufwand.

Eine andere Teststatistik, die von ***Breusch*** und ***Pagan*** vorgeschlagen wurde und als ein Lagrange-Multiplikator-Test anzusehen ist, benötigt demgegenüber nur die Residuen des restringierten Modells. Die Nullhypothese lautet

$$H_0: \quad \sigma_\mu^2 = 0$$

und unter ihr ist die Teststatistik

$$LM = \frac{NT}{2(T-1)} \left(\frac{\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{e}_{it} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2} - 1 \right)^2$$

asymptotisch $\chi^2_{(1)}$ -verteilt. Dabei sind die $\hat{e}_{it}$'s die Residuen des "Classical-Pooling"-Modells. Übersteigt der empirische $\tilde{\chi}^2$-Wert bei vorgegebenem Signifikanzniveau $(1-\alpha)$ den kritischen $\chi^2_{(1)}$-Wert, ist die Nullhypothese abzulehnen, und es kann auf Vorliegen von Individualeffekten geschlossen werden.

4. Fixed versus Random Effects

Eine offensichtliche Frage ist, ob bei Vorliegen von Individualeffekten diese als fest oder als zufällig anzusehen sind, ob also das "LSDV"-Modell oder das "EC"-Modell Anwendung finden sollte. Ein erster Gesichtspunkt betrifft die Größenordnung der Stichprobe. Wenn *T* bei kleinem *N* sehr groß wird, so gehen die beiden Modelle ineinander über, und das rechentechnisch einfachere "LSDV"-Modell dürfte vorzuziehen sein. Trifft dieser Fall jedoch nicht zu, so können erhebliche Unterschiede zwischen den Modellen erwartet werden und andere Aspekte müssen bei der Entscheidung berücksichtigt werden.

Bei der Herleitung dieser Modelle wurden Individualeffekte als Komponente der ursprünglichen Residuen aufgefaßt, die den Einfluß latenter Variablen auf die zu erklärende Variable wiedergeben. Während man den Einfluß der latenten Variablen, die nicht nur über die Zeit, sondern auch über die Individuen variieren, weiterhin im "LSDV"-Modell in einem als zufällige Größe zu interpretierenden Residuen v_{it} erfaßt, werden die zeitinvarianten und für jedes Individuum unterschiedlichen latenten Variablen als feste Individualeffekte angesehen, deren Einfluß über die Parameter der Dummy-Variablen geschätzt wird. Allerdings liefern diese Dummy-Variablen keine Informationen über die Ursachen, welche zu einer Niveauverschiebung der Regressionsfunktion für jedes Individuum führen, ihre Parameter messen lediglich deren Einfluß auf Kosten eines Verlustes an Freiheitsgraden. Insofern scheint es angebrachter, die bestehende Unkenntnis über die Individualeffekte ähnlich der über die anderen latenten Variablen zu behandeln, das heißt, diese als ebenfalls zufällig zu betrachten.

In der ökonometrischen Literatur wird jedoch darauf hingewiesen, daß die Individualeffekte immer als zufällige Größen interpretiert werden können, da sie erst nach der Ziehung der

Stichprobe bekannt sind bzw. geschätzt werden können. Es ändert sich lediglich die Sichtweise, mit der Schlußfolgerungen über diese Stichprobe gezogen werden können: Bei dem Fixed Effects-Modell ist die statistische Inferenz bedingt (abhängig) von den jeweiligen Individuen in der Stichprobe, während bei der Annahme von Random Effects eine Inferenz unbedingt (unabhängig) von den Individuen dieser Stichprobe erfolgt. Damit werden Aussagen über die dahinterstehende Grundgesamtheit ermöglicht. Der Anwender muß folglich über die Art der zu treffenden Aussagen entscheiden. Bedingte Inferenz sollte dann getroffen werden, wenn die Individuen nicht als eine Stichprobe aus einer übergeordneten Grundgesamtheit zu betrachten sind oder wenn es insbesondere die in der Stichprobe enthaltenen Individuen sind, über die er Aussagen treffen will. Sollen demgegenüber die Schlußfolgerungen die Grundgesamtheit betreffen und können die Individuen als eine Stichprobe hieraus angesehen werden, dann empfiehlt sich die unbedingte Inferenz mit Hilfe der Annahme von Random Effects.

Im Gegensatz zum Fixed Effects-Modell müssen hierzu jedoch restriktive Annahmen über die Verteilung der Zufallsvariablen μ gemacht werden, so daß nur bei deren Gültigkeit diese zusätzliche Information zu einem Effizienzgewinn des "EC"-Modells führen kann.

Deshalb sollte die Eignung der Annahme überprüft werden, daß die Individualeffekte identisch und unabhängig verteilte Zufallsvariablen mit einem Mittelwert von Null und konstanter Varianz sind. So kann zum Beispiel gezeigt werden, daß bei einer Korrelation zwischen den erklärenden Variablen und den Individualeffekten der EGLS-Schätzer des "EC"-Modells verzerrt und inkonsistent ist, während gerade in dieser Situation sich der "LSDV"-Schätzer als effizient erweist.

So könnte man sich im Rahmen der Schätzung einer Wahlfunktion zum Beispiel vorstellen, daß die mit einer bestimmten geographischen Lage verbundenen Individualeffekte eines Wahlkreises in Form des "traditionellen Wählerverhaltens" über die vorherrschende Wirtschaftsstruktur in diesem Gebiet mit der Arbeitslosigkeit korreliert sind, da in einem landwirtschaftlich strukturierten Gebiet einerseits eher konservativ gewählt wird, während die Arbeitslosigkeit tendenziell geringer als in industriellen Ballungszentren ausfällt.

Um die Hypothese einer bestehenden Korrelation zwischen erklärenden Variablen und Individualeffekten zu überprüfen, kann ein von ***Hausman*** entwickeltes Testverfahren angewandt werden, wobei man sich die unterschiedlichen Eigenschaften des "LSDV"- und des "EC"-Schätzers zunutze macht.Unter der Nullhypothese fehlender Korrelation ist der"EC"-Schätzer unverzerrt und effizient hingegen bei Vorliegen einer Korrelation verzerrt, während der "LSDV"-Schätzer sowohl bei Gültigkeit der Nullhypothese wie auch bei bestehender Korrelation konsistent ist. Deswegen kann bei Gültigkeit der Nullhypothese erwartet werden, daß, zumindest asymptotisch, die beiden Schätzer nur zufällig voneinander abweichen werden, während diese Abweichung bei bestehender Korrelation weitaus größer

sein dürfte. Im wesentlichen wird beim ***Hausman***-Test also der sich empirisch ergebende Unterschied beider Schätzer auf Signifikanz geprüft.

Zur Durchführung des ***Hausman***-Tests eignet sich besonders die F-Test Version, bei der wiederum die Quadratsummen von zwei Modellen miteinander verglichen werden. Das eine Modell ist dabei das "EC"-Modell, während das andere eine Kombination des "EC"- und des "LSDV"-Modells darstellt,

$$y_{it} - \theta\,\bar{y}_{i.} = (1-\theta)\bar{\beta}_1 + \sum_{k=2}^{K} \beta_k^{EC}(x_{k,it} - \theta\,\bar{x}_{k,i.}) + \sum_{k=2}^{K} \beta_k^{LSDV}(x_{k,it} - \bar{x}_{k,i.}) + w^*_{it}$$

indem man zusätzlich zu den bereits im Ansatz befindlichen Regressoren des "EC"-Modells die des "LSDV"-Modells einbezieht.

Die Nullhypothese lautet

$$H_0 : \beta_k^{LSDV} = 0 \quad \text{gegen} \quad H_1 : \beta_k^{LSDV} \neq 0$$

Man prüft nun die Signifikanz der zusätzlich eingeführten Parameter des "LSDV"-Modells über die Teststatistik

$$F = \frac{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2 - \sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2 *}{(K-1)}}{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2 *}{(NT-2K+1)}}$$

dabei entspricht $\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2$ der Quadratsumme der Residuen des restringierten Modells, also des "EC"-Modells, mit $(K-1)$ als der Zahl der eingeführten Restriktionen, während $\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2 *$ die Quadratsumme der Residuen des unrestringierten Modells, also des kombinierten "EC"- und "LSDV"-Modells, und $(NT-2K+1)$ die Zahl der Freiheitsgrade des unrestringierten Modells angeben.

Unter der Nullhypothese ist die Teststatistik Fverteilt mit den angegebenen Freiheitsgraden von Zähler und Nenner. Übersteigt der empirische F-Wert den bei einem vorgegebenen Signifikanzniveau ermittelten kritischen F-Wert, so ist die Nullhypothese abzulehnen. Dies bedeutet, daß man von einer Korrelation zwischen erklärenden Variablen und Individual-

effekten ausgehen kann, und somit das "LSDV"-Modell dem "EC"-Modell in dieser Situation vorzuziehen ist.

Zusammenfassend schlagen ***Judge*** et al. (1985, S.527) vor, daß "a reasonable prescription is to use the error components model if the μ_i ~ i.i.d. $(0, \sigma_\mu^2)$ assumption is a reasonable one and *N* is sufficiently large for reliable estimation of σ_μ^2; otherwise, particularly when μ_i and $\mathbf{X}_i$ are correlated, or *N* is small, use the dummy variable (within) estimator." Auf die Frage, wie groß *N* sein sollte, führen sie an anderer Stelle (1989, S.490) eine Untersuchung von ***Taylor*** (1980) an, der gezeigt hat, daß selbst bei relativ geringem Stichprobenumfang $\left[T \geq 3, N-K \geq 9\,;\, T \geq 2, N-K \geq 10\right]$ die Effizienzvorteile des "EC"-Schätzers zum Tragen kommen.

5. Berücksichtigung von Zeiteffekten

a) Zeiteffekte im "LSDV"-Modell

Bisher wurde zur Vereinfachung von möglichen Zeiteffekten λ_t abstrahiert, deren Einführung jedoch im Rahmen der obigen Modelle eine naheliegende Erweiterung darstellt.

Im "LSDV"-Modell können entweder *T* zusätzliche Dummy-Variablen zur Erfassung der Zeiteffekte mit der folgenden Definition

$$\mathrm{D}_{\mathrm{is}} = \begin{cases} 1 & \textbf{falls } \mathrm{t} = \mathrm{s} \\ 0 & \textbf{sonst} \end{cases}$$

eingeführt werden oder die Beobachtungswerte von **y** und **X** werden wie folgt transformiert

$$y_{it}^{LSDV} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{y}_{..}$$

$$x_{k,it}^{LSDV} = x_{k,it} - \bar{x}_{k,i.} - \bar{x}_{k,.t} + \bar{x}_{k,..}$$

wobei

$$\bar{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad \bar{y}_{.t} = \frac{1}{N}\sum_{i=1}^{N} y_{it} \qquad \bar{y}_{..} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$$

$$\bar{x}_{k,i.} = \frac{1}{T}\sum_{t=1}^{T} x_{k,it} \qquad \bar{x}_{k,.t} = \frac{1}{N}\sum_{i=1}^{N} x_{k,it} \qquad \bar{x}_{k,..} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{k,it}$$

Dabei wird ein einzelner Beobachtungspunkt nicht mehr nur als Abweichung vom jeweiligen individuellen Durchschnitt über alle Zeitpunkte, sondern auch als Abweichung vom jeweili-

gen zeitlichen Durchschnitt über alle Individuen ausgedrückt. Führt man eine OLS-Regression mit Individual- und Zeitdummies oder mit den transformierten Beobachtungswerten durch, so erhält man den "LSDV"-Schätzer für feste Individual- und Zeiteffekte. Auch hier kann die Einführung von Zeiteffekten mit Hilfe eines F-Tests und entsprechend definierten Quadratsummen der Residuen überprüft werden.

b) Zeiteffekte im "EC"-Modell

Für die Durchführung einer Schätzung des "EC"-Modells müssen die Beobachtungswerte ebenfalls transformiert werden

$$y_{it}^{ec} = y_{it} - \theta_1 \overline{y}_{i.} - \theta_2 \overline{y}_{.t} + \theta_3 \overline{y}_{..}$$

$$x_{k,it}{}^{*} = x_{k,it} - \theta_1 \overline{x}_{k,i.} - \theta_2 \overline{x}_{k,.t} + \theta_3 \overline{x}_{k,..}$$

Dabei drücken die Koeffizienten θ wieder den Einfluß der Varianzkomponenten aus

$$\theta_1 = 1 - \frac{\sigma_v}{\sigma_1}, \quad \theta_2 = 1 - \frac{\sigma_v}{\sigma_2}, \quad \theta_3 = \theta_1 + \theta_2 - 1 + \frac{\sigma_v}{\sigma_3}$$

mit $\sigma_1^2 = \sigma_v^2 + T\sigma_\mu^2$, $\sigma_2^2 = \sigma_v^2 + N\sigma_\lambda^2$

und $\sigma_3^2 = \sigma_v^2 + T\sigma_\mu^2 + N\sigma_\lambda^2$

Eine OLS-Schätzung über die Werte **y*** und **X***, wobei zur Transformation die entsprechenden Varianzkomponenten vorher geschätzt werden müssen, entspricht der EGLS-Schätzung des "EC"-Modells mit Individual- und Zeiteffekten.

Neben der Überprüfung auf Vorliegen von Individual-und Zeiteffekten durch den schon unter a) genannten F-Test, kann der von ***Breusch*** und ***Pagan*** vorgeschlagene LM-Test mit entsprechend erweiterter Teststatistik

$$LM = \frac{NT}{2}\left\{\frac{1}{T-1}\left(\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right)^2 + \frac{1}{N-1}\left(\frac{\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right)^2\right\} \sim \chi_2^2$$

verwandt werden.

Bei den obigen Betrachtungen wurde bisher unterstellt, daß sich die Individual- und/oder Zeiteffekte allein auf die Regressionskonstante auswirkten, während die Regressionsgewichte weiterhin als konstant angenommen wurden. In der ökonometrischen Literatur werden jedoch auch Modelle behandelt, in denen Individual- und/oder Zeiteffekte für den gesamten Parametervektor zugelassen werden (vgl. Tabelle 1). Deren Erläuterung würde allerdings den Rahmen dieser Einführung sprengen, so daß auf die unten angeführte Literatur verwiesen wird. Statt dessen soll im nächsten Abschnitt ein simplifiziertes Rechenbeispiel gegeben werden, um die in den vorangegangenen Abschnitten dargestellten Modelle zu veranschaulichen. Gleichzeitig soll anhand dieses Beispiels aufgezeigt werden, daß zur praktischen Umsetzung einer Regressionsanalyse mit Panel-Daten sich die Benutzung eines Tabellenkalkulationsprogrammes (z.B. Excel) in Kombination mit einem Statistik-Programmpaket empfiehlt, welches das Einlesen der Daten aus dem Tabellenkalkulationsprogramm unterstützt (z.B. SPSS für Windows, das explizit einen Befehl zur Eingabe von Excel-Dateien beinhaltet).

IV. Rechenbeispiel: Die Schätzung einer Wahlfunktion

Angenommen man habe in 4 Wahlkreisen für 5 Wahlperioden die Werte des jeweiligen Anteils an SPD-Wählern und die jeweilige Arbeitslosenquote beobachtet, die in Tabelle 2 wiedergegeben sind:

Tabelle 2: Anteil der SPD-Wähler und Arbeitslosenquote für 5 Wahlperioden und 4 Wahlkreise

	Wahlkreis 1		Wahlkreis 2		Wahlkreis 3		Wahlkreis 4	
Wahlperiode	SPD	ALQ	SPD	ALQ	SPD	ALQ	SPD	ALQ
1	38,46	4,32	45,52	5,13	22,86	2,39	41,86	6,49
2	35,32	5,86	48,71	4,77	18,52	1,77	44,33	6,18
3	30,78	3,85	47,01	4,68	22,93	2,86	43,21	6,05
4	35,34	4,08	50,32	7,36	25,02	3,15	46,69	7,77
5	30,83	3,99	40,05	4,28	35,13	4,01	49,7	8,12
Jahres∅	34,15	4,42	46,32	5,24	24,89	2,84	45,16	6,92

Ziel sei es, zu überprüfen, ob der Anteil der SPD-Wähler von einer Zunahme der Arbeitslosigkeit beeinflußt wird, wobei angenommen wird, daß die beiden Variablen in linearer Abhängigkeit stehen: $SPD_{it} = \alpha_i + \beta \cdot ALQ_{it} + e_{it}$

Für die Regressionsgewichte β wird für alle Kreise ein gleicher Wert erwartet; über die Regressionskonstante α_i bestehen dagegen keinerlei Vorabinformationen, so daß die Möglichkeit von individuell verschiedenen Regressionskonstanten der Wahlkreise hier in Betracht zu ziehen ist.

Das "Classical Pooling"-Modell erfordert nun nichts weiter als eine Dateneingabe, in der die entsprechenden Beobachtungswerte von abhängiger und unabhängiger Variable untereinandergeschrieben werden. Die Spalte (1) in Tabelle 3 gibt den Vektor **y** für die abhängige Variable (den Anteil der SPD-Wähler) wieder, während die Spalten (2) und (3) die Matrix **X** der unabhängigen Variablen (Konstante und Arbeitslosenquote) darstellen. Anzumerken ist, daß die Eingabe des konstanten Terms, also der mit lauter Einsen besetzten Spalte (2) von den Statistik-Programmen übernommen wird.

Für das "Classical Pooling"-Modell ergibt sich folgende Regressionsgleichung, wobei RSS die Residual Sum of Squares, also die Abweichungsquadratsumme der Residuen, angibt:

$$SPD_{it} = 14,56 + 4,75 \cdot ALQ_{it} + \hat{e}_{it} \qquad RSS = 508,26$$

Wie erwähnt liegen dem "Classical Pooling"-Modell sehr restriktive Annahmen über die Residuen zugrunde, so daß man im Rahmen des "***Kmenta***"-Modells für die 4 Wahlkreise jeweils eine verschieden große Varianz und Autokorrelation der Residuen zulassen kann. Den ersten Schritt des "***Kmenta***"-Modells haben wir mit der obigen Schätzung bereits unternommen, so daß nun aus den Residuen dieser Schätzgleichung, die in Spalte (4) aufgeführt sind, die zur Transformation der Variablen erforderlichen Werte gemäß der in Abschnitt II.2 angegebenen Formeln errechnet werden müssen. Zuerst erfolgt eine Berechnung der Autokorrelationskoeffizienten $\hat{\rho}_i$, wobei sich für den ersten Haushalt der Wert wie folgt ergibt:

$$\hat{\rho}_1 = \frac{(-7,08) \cdot 3,37 + (-2,07) \cdot (-7,08) + 1,4 \cdot (-2,07) + (-2,69) \cdot 1,4}{3,37^2 + (-7,08)^2 + (-2,07)^2 + 1,4^2} = \frac{(-15,87)}{67,73} = (-0,23)$$

Die Werte für die 3 anderen Wahlkreise lauten $\hat{\rho}_2 = 0,73$, $\hat{\rho}_3 = 0,7$ und $\hat{\rho}_4 = 0,43$.

Im nächsten Schritt sind die ursprünglichen Beobachtungswerte gemäß der Formeln in II.2 zu transformieren, wobei hier nur die Berechnung für den ersten Wahlkreis exemplarisch veranschaulicht werden soll:

$y_{11}^{*} = \sqrt{1-(-0,23)^2} \cdot 38,46 = 37,43$ $\quad x_{11}^{*} = \sqrt{1-(-0,23)^2} \cdot 4,32 = 4,2$

$y_{12}^{*} = 35,32 - (-0,23) \cdot 38,46 = 44,17$ $\quad x_{12}^{*} = 5,86 - (-0,23) \cdot 4,32 = 6,85$

$y_{13}^{*} = 30,78 - (-0,23) \cdot 35,32 = 38,9$ $\quad x_{13}^{*} = 3,85 - (-0,23) \cdot 5,86 = 5,2$

$y_{14}^{*} = 35,34 - (-0,23) \cdot 30,78 = 42,42$ $\quad x_{14}^{*} = 4,08 - (-0,23) \cdot 3,85 = 4,97$

$y_{15}^{*} = 30,83 - (-0,23) \cdot 35,34 = 38,96$ $\quad x_{15}^{*} = 3,99 - (-0,23) \cdot 4,08 = 4,93$

Die Ergebnisse für die Berechnung auch der 3 anderen Wahlkreise finden sich in den Spalten (5), (6) und (7) der Tabelle 3. Zur Spalte (6) ist eine Bemerkung angebracht. Diese ersetzt den konstanten Term in Spalte (2) des vorhergehenden Modells. Auch die mit Einsen besetzte Spalte (2) ist gemäß der angegebenen Formeln zu transformieren, so daß sich die entsprechenden Werte in Spalte (6) ergeben. Eine nochmalige Schätzung des Regressionsmodells mit den transformierten Beobachtungswerten führt zu dem folgenden Ergebnis,

$$SPD_{it}^{*} = 19,22 + 3,54 \cdot ALQ_{it}^{*} + \hat{e}_{it}^{*} \qquad RSS = 212,88$$

wobei zu beachten ist, daß bei der Schätzung die üblicherweise von den Statistikprogrammen automatisch mitgeschätzte Konstante unterdrückt wird, und an ihrer Stelle der entsprechend transformierte konstante Term $x_{1,it}^{*}$ eingegeben wird (SPSS bietet unter dem Menü-Punkt "Optionen..." mit dem Befehl "Konstante in Gleichung" die Möglichkeit, die Schätzung der Regressionskonstante zu unterdrücken). Da wir im Ansatz aber noch Heteroskedastizität vermuten, muß aus dem Residualvektor $\hat{\mathbf{e}}^{*}$, der in Spalte (8) angegeben wird, noch die für jeden Wahlkreis als unterschiedlich angenommene Varianz der Residuen $\hat{\sigma}_i^2$ bestimmt werden, für den ersten Wahlkreis lautet diese:

$$\hat{\sigma}_1^{2*} = \tfrac{1}{3}(3,9^2 + (-3,75)^2 + (-3,17)^2 + 1,17^2 + (-2,15)^2) = 15,10.$$

Für die 3 anderen Wahlkreise ergeben sich die Werte $\hat{\sigma}_2^{2*} = 31,14$, $\hat{\sigma}_3^{2*} = 19,01$ und $\hat{\sigma}_4^{2*} = 5,7$.
Die entsprechend bereinigten Werte y_{it}^{**} und $x_{k,it}^{**}$ finden sich in den Spalten (9), (10) und (11) der Tabelle 3. Führt man mit diesen Werten eine Schätzung der Regressionsfunktion durch, so erhält man für das "***Kmenta***"-Modell folgendes Ergebnis (wiederum muß die Konstante vom Anwender selbst eingelesen werden):

$$SPD_{it}^{**} = 18,07 + 3,75 \cdot ALQ_{it}^{**} + \hat{e}_{it}^{**} \qquad RSS = 11,91$$

Aus dem obigen "***Kmenta***"-Modell läßt sich das Regressionsgewicht β mit einem Wert von 3,75 ablesen, für die Regressionskonstante α ergibt sich ein Wert von 18,07.

Sowohl im "Classical Pooling"- als auch im "***Kmenta***"-Modell haben wir bisher die Gültigkeit einer identischen Regressionsfunktion für die 4 Wahlkreise angenommen. Wie in Abschnitt III.1 argumentiert, könnten wir jedoch durch die Einführung von Dummy-Variablen

Tabelle 3: Beobachtungswerte des "Classical Pooling"-Modells und transformierte Beobachtungswerte des "Kmenta"-Modells

(1) y_{it}	(2) $x_{1,it}$	(3) $x_{2,it}$	(4) $\hat{e}_{it}$	(5) y_{it}^{*}	(6) $x_{1,it}^{*}$	(7) $x_{2,it}^{*}$	(8) $\hat{e}_{it}^{*}$	(9) y_{it}^{**}	(10) $x_{1,it}^{**}$	(11) $x_{2,it}^{**}$
38,46	1	4,32	3,37	37,43	0,97	4,2	3,9	9,62	0,25	1,08
35,32	1	5,86	-7,08	44,17	1,23	6,85	-3,75	11,35	0,32	1,76
30,78	1	3,85	-2,07	38,9	1,23	5,2	-3,17	10	0,32	1,34
35,34	1	4,08	1,4	42,42	1,23	4,97	1,17	10,9	0,32	1,28
30,83	1	3,99	-2,69	38,96	1,23	4,93	-2,15	10,02	0,32	1,27
45,52	1	5,13	6,59	31,11	0,68	3,51	5,6	5,58	0,12	0,63
48,71	1	4,77	11,49	15,48	0,27	1,03	6,64	2,77	0,05	0,18
47,01	1	4,68	10,21	11,45	0,27	1,2	2,01	2,05	0,05	0,22
50,32	1	7,36	0,79	16	0,27	3,94	-3,15	2,87	0,05	0,71
40,05	1	4,28	5,15	3,32	0,27	-1,09	2	0,59	0,05	-0,2
22,86	1	2,39	-3,06	16,33	0,71	1,71	-3,37	3,75	0,16	0,39
18,52	1	1,77	-4,45	2,52	0,3	0,1	-3,6	0,58	0,07	0,02
22,93	1	2,86	-5,22	9,97	0,3	1,62	-1,54	2,29	0,07	0,37
25,02	1	3,15	-4,51	8,97	0,3	1,15	-0,87	2,06	0,07	0,26
35,13	1	4,01	1,52	17,62	0,3	1,81	5,44	4,04	0,07	0,42
41,86	1	6,49	-3,54	37,79	0,9	5,86	-0,28	15,81	0,38	2,45
44,33	1	6,18	0,41	26,33	0,57	3,39	3,36	11,02	0,24	1,42
43,21	1	6,05	-0,09	24,15	0,57	3,39	1,18	10,1	0,24	1,42
46,69	1	7,77	-4,79	28,11	0,57	5,17	-1,17	11,76	0,24	2,16
49,7	1	8,12	-3,44	29,62	0,57	4,78	1,72	12,39	0,24	2

oder aber durch die Bildung von Durchschnitten und anschließender Transformation der Beobachtungswerte unterschiedliche Regressionskonstanten α_i zulassen. Da im allgemeinen die Zahl der Individuen (hier Wahlkreise) sehr groß sein wird, empfiehlt sich aus rechentechnischen Gründen das letztere Verfahren, welches dementsprechend hier zuerst vorge-

stellt wird. Im ersten Schritt müssen für die 4 Wahlkreise jeweils die Durchschnitte des SPD-Anteils und der Arbeitslosenquote über die 5 Wahlperioden bestimmt werden. Es ergeben sich anhand von Tabelle 2 die folgenden Werte:

$\bar{y}_{1.} = (38,46 + 35,32 + 30,78 + 35,34 + 30,83)/5 = 34,15$
$\bar{y}_{2.} = 46,32$
$\bar{y}_{3.} = 24,89$
$\bar{y}_{4.} = 45,16$

$\bar{x}_{2,1.} = (4,32 + 5,86 + 3,85 + 4,08 + 3,99)/5 = 4,42$
$\bar{x}_{2,2.} = 5,24$
$\bar{x}_{2,3.} = 2,84$
$\bar{x}_{2,4.} = 6,92$

Im nächsten Schritt müssen die ursprünglichen Daten als Abweichung der jeweiligen Mittelwerte ausgedrückt werden, für den Wahlkreis 1 ergeben sich die folgenden Werte:

$y_{11}^{lsdv} = 38,46 - 34,15 = 4,31$ $\quad x_{2,11}^{lsdv} = 4,32 - 4,42 = -0,1$
$y_{12}^{lsdv} = 35,32 - 34,15 = 1,17$ $\quad x_{2,12}^{lsdv} = 5,86 - 4,42 = 1,44$
$y_{13}^{lsdv} = 30,78 - 34,15 = -3,37$ $\quad x_{2,13}^{lsdv} = 3,85 - 4,42 = -0,57$
$y_{14}^{lsdv} = 35,34 - 34,15 = 1,19$ $\quad x_{2,14}^{lsdv} = 4,08 - 4,42 = -0,34$
$y_{15}^{lsdv} = 30,83 - 34,15 = -3,32$ $\quad x_{2,15}^{lsdv} = 3,99 - 4,42 = -0,43$

Mit den so bestimmten Werten auch für die anderen 3 Wahlkreise, die in Tabelle 4 in den Spalten (1) und (2) aufgeführt und wie im "Classical-Pooling"-Modell einfach untereinandergeschrieben sind, kann nun die Regressionsfunktion bestimmt werden, wobei allerdings wiederum zu beachten ist, daß kein konstanter Term mitgeschätzt werden darf.

Es ergibt sich folgende Schätzung für das "LSDV"-Modell:

$SPD_{it}^{lsdv} = 3,14 \cdot ALQ_{it}^{lsdv} + (v_{it} - \bar{v}_{i.}) \qquad RSS = 148,79$

Aus der obigen Schätzgleichung kann das für alle Wahlkreise identische Regressionsgewicht β direkt mit einem Wert von 3,14 abgelesen werden. Die individuell verschiedenen Regressionskonstanten der 4 Wahlkreise ergeben sich wie folgt:

$\alpha_1 = 34,15 - 3,14 \cdot 4,42 = 20,27$
$\alpha_2 = 46,32 - 3,14 \cdot 5,24 = 29,87$
$\alpha_3 = 24,89 - 3,14 \cdot 2,84 = 15,97$
$\alpha_4 = 45,16 - 3,14 \cdot 6,92 = 23,43$

Wie erwähnt, kann das "LSDV"-Modell auch durch die Einführung von Dummy-Variablen geschätzt werden, so daß die obige Mittelwertsbereinigung "automatisch" erfolgt. In unserem Fall gehen neben den untransformierten Beobachtungswerten, die aus Gründen der Bequemlichkeit noch einmal in den Spalten (3) und (8) aufgelistet sind, 4 entsprechend kodierte 0-1 Variablen in die Regressionsfunktion ein, die sich in den Spalten (4)-(7) befinden. Der Vorteil dieser Vorgehensweise ist, daß die individuellen Regressionskonstanten nun (bis auf vernachlässigbare Rundungsfehler) direkt aus der Regressionsfunktion ersichtlich sind

$$SPD_{it} = 20,29 \cdot D_{1t} + 29,88 \cdot D_{2t} + 16 \cdot D_{3t} + 23,45 \cdot D_{4t} + 3,14 \cdot ALQ_{it} + v_{it} \qquad RSS = 148,79$$

Da aber bei einer großen Anzahl von Untersuchungseinheiten die Einführung von gleich vielen Dummy-Variablen erforderlich ist, kann dies zu numerischen Problemen bei den Statistik-Programmen führen. Des weiteren, wie weiter unten gezeigt wird, bildet die Mittelwertberechnung den Ausgangspunkt der Schätzung des "EC"-Modells, so daß die Schätzung der Dummy-Variante nur bei einer geringen Anzahl von Individuen sinnvoll ist.

Um zu überprüfen, ob die Einführung von unterschiedlichen Regressionskonstanten für die 4 Wahlkreise sinnvoll war, berechnen wir die F-Statistik, in dem wir einen Vergleich der Abweichungsquadratsumme von "Classical Pooling"- und "LSDV"-Modell, korrigiert durch die Freiheitsgrade, vornehmen. Es ergibt sich eine empirische F-Statistik von

$$F_{emp.} = \frac{(508,26 - 148,79)/3}{148,79/15} = 12,08$$

Demnach, da der Wert von 12,08 deutlich größer ist als die kritische F-Statistik von 5,42 auf einem 1% Signifikanzniveau, können wir die Nullhypothese einer gleichen Regressionskonstanten nicht annehmen, und ziehen das "LSDV"- dem "Classical Pooling"-Modell gegenüber vor.

Als nächster Schritt wäre noch die Möglichkeit in Betracht zu ziehen, das "Error Components"-Modell zu schätzen. Wiederum ist hier eine Transformation der Beobachtungswerte vorzunehmen, für die wir aber schon einige Vorarbeiten geleistet haben. Wie in Abschnitt III.2b ausgeführt, ist für die Schätzung des "EC"-Modells der Faktor θ zu bestimmen, der sich als das Verhältnis zweier Standardabweichungen ergibt. Die "within"-Schätzung des "LSDV"-Modells hat uns schon die erste Varianz geliefert, indem wir die angegebene Abweichungsquadratsumme der Residuen durch die Zahl der Freiheitsgrade dividieren, so daß wir $\hat{\sigma}_v^2 = 148,79/15 = 9,92$ erhalten.

Um zur Varianz $\hat{\sigma}_1^2$ zu gelangen,ist die Durchführung der "between"-Schätzung notwendig. Diese entspricht einer Regression über die schon bestimmten 4 zeitlichen Durchschnitts-

werte der Wahlkreise von SPD-Anteil und Arbeitslosigkeit. Es ergibt sich folgende Regressionsfunktion

$$\overline{SPD}_{i.} = 11,83 + 5,31 \cdot \overline{ALQ}_{i.} + \overline{w}_{i.} \qquad RSS = 61,24$$

wobei die Varianz dieser Schätzung sich wiederum aus dem Verhältnis von Abweichungsquadratsumme und Freiheitsgraden ($N - K - 1 = 4 - 1 - 1 = 2$) ergibt, in diesem Fall $\hat{\sigma}^2_{\overline{w}} = 61,24/2 = 30,62 = \hat{\sigma}^2_1 / T$. Da die Varianz dieser Schätzung mit *T* multipliziert die gesuchte Varianz $\hat{\sigma}^2_1$ ergibt, erhalten wir den Wert 153,1 und für das Verhältnis θ errechnet man $\hat{\theta} = 1 - \sqrt{9,92/153,1} = 0,75$. Im letzten Schritt müssen die Beobachtungswerte der Wahlkreise wie im "LSDV"-Modell von ihren Mittelwerten bereinigt werden, allerdings zieht man im "EC"-Modell nur einen Teil, gegeben durch den Faktor $\hat{\theta} = 0,75$, von den ursprünglichen Werten ab. Für den Wahlkreis 1 ist diese Berechnung exemplarisch wiedergegeben:

$$y^{ec}_{11} = 38,46 - 0,75 \cdot 34,15 = 12,85 \qquad x^{ec}_{2,11} = 4,32 - 0,75 \cdot 4,42 = 1,01$$
$$y^{ec}_{12} = 35,32 - 0,75 \cdot 34,15 = 9,71 \qquad x^{ec}_{2,12} = 5,86 - 0,75 \cdot 4,42 = 2,55$$
$$y^{ec}_{13} = 30,78 - 0,75 \cdot 34,15 = 5,17 \qquad x^{ec}_{2,13} = 3,85 - 0,75 \cdot 4,42 = 0,54$$
$$y^{ec}_{14} = 35,34 - 0,75 \cdot 34,15 = 9,73 \qquad x^{ec}_{2,14} = 4,08 - 0,75 \cdot 4,42 = 0,77$$
$$y^{ec}_{15} = 30,83 - 0,75 \cdot 34,15 = 5,22 \qquad x^{ec}_{2,15} = 3,99 - 0,75 \cdot 4,42 = 0,68$$

Die transformierten Werte auch für die 3 anderen Wahlkreise finden sich in Tabelle 4. Mit Hilfe dieser transformierten Daten kann die Schätzung des "EC"-Modells erfolgen, es ergibt sich die untenstehende Schätzgleichung:

$$SPD^{ec}_{it} = 20,77 + 3,47 \cdot ALQ^{ec}_{it} + w_{it} \qquad RSS = 178,63$$

Das hier präsentierte Rechenbeispiel sollte nur der Veranschaulichung der allgemeinen Vorgehensweise dienen, sowohl *N* als auch *T* sind hier zu klein gewählt, um eine sinnvolle Inferenz oder Entscheidung zwischen den einzelnen Modellen zu ermöglichen. Trotzdem soll exemplarisch die Berechnung der ***Hausman***-Statistik zur Unterscheidung von "LSDV"- und "EC"-Modell vorgestellt werden, da sie nichts weiter als die Schätzung der obigen "EC"-Regressionsfunktion erfordert, die um die transformierte Reihe der Arbeitslosigkeit des "LSDV"-Modells erweitert wird. Es läßt sich die folgende Schätzfunktion ermitteln:

Tabelle 4: Transformierte Beobachtungswerte von "LSDV"- und "EC"-Modell

(1) y_{it}^{lsdv}	(2) $x_{2,it}^{lsdv}$	(3) y_{it}	(4) D_{1t}	(5) D_{2t}	(6) D_{3t}	(7) D_{4t}	(8) $x_{2,it}$	(9) y_{it}^{ec}	(10) $x_{1,it}^{ec}$	(11) $x_{2,it}^{ec}$
4,31	-0,1	38,46	1	0	0	0	4,32	12,85	0,25	1,01
1,17	1,44	35,32	1	0	0	0	5,86	9,71	0,25	2,55
-3,37	-0,57	30,78	1	0	0	0	3,85	5,17	0,25	0,54
1,19	-0,34	35,34	1	0	0	0	4,08	9,73	0,25	0,77
-3,32	-0,43	30,83	1	0	0	0	3,99	5,22	0,25	0,68
-0,8	-0,11	45,52	0	1	0	0	5,13	10,78	0,25	1,2
2,39	-0,47	48,71	0	1	0	0	4,77	13,97	0,25	0,84
0,69	-0,56	47,01	0	1	0	0	4,68	12,27	0,25	0,75
4	2,12	50,32	0	1	0	0	7,36	15,58	0,25	3,43
-6,27	-0,96	40,05	0	1	0	0	4,28	5,31	0,25	0,35
-2,03	-0,45	22,86	0	0	1	0	2,39	4,19	0,25	0,26
-6,37	-1,07	18,52	0	0	1	0	1,77	-0,15	0,25	-0,36
-1,96	0,02	22,93	0	0	1	0	2,86	4,26	0,25	0,73
0,13	0,31	25,02	0	0	1	0	3,15	6,35	0,25	1,02
10,24	1,17	35,13	0	0	1	0	4,01	16,46	0,25	1,88
-3,3	-0,43	41,86	0	0	0	1	6,49	7,99	0,25	1,3
-0,83	-0,74	44,33	0	0	0	1	6,18	10,46	0,25	0,99
-1,95	-0,87	43,21	0	0	0	1	6,05	9,34	0,25	0,86
1,53	0,85	46,69	0	0	0	1	7,77	12,82	0,25	2,58
4,54	1,2	49,7	0	0	0	1	8,12	15,83	0,25	2,93

$$SPD_{it}^{ec} = 11,86 + 5,3 \cdot ALQ_{it}^{ec} - 2,16 \cdot ALQ_{it}^{lsdv} + w_{it}^{*} \qquad RSS = 167,92$$

Dementsprechend ergibt sich eine empirische F-Statistk von

$$F_{emp.} = \frac{178,63 - 167,92 / 1}{167,92 / 17} = 1,08.$$

Da dieser Wert den kritischen F-Wert von 8,40 auf einem 1% Signifikanzniveau nicht übersteigt, kann die Nullhypothese nicht abgelehnt werden. Dies bedeutet, daß man in unserem Beispiel das "EC"- dem "LSDV"-Modell gegenüber vorziehen sollte.

Zusammenfassend läßt sich also festhalten, daß die hier vorgestellten Modelle zur Regressionsanalyse mit Paneldaten im wesentlichen nur die Verwendung von transformierten Beobachtungswerten erfordern. Die Transformationen selbst können recht bequem in Excel durchgeführt werden, da dieses Programm z.B. unter dem Menü-Punkt "Formel" und dem Unterpunkt "Funktion einfügen..." die Möglichkeit bietet, die Berechnung von Mittelwerten für alle Variablen en bloc vorzunehmen. Nachdem die Daten entsprechend umgerechnet und in SPSS eingelesen worden sind, kann anhand einer gewöhnlichen Regressionsanalyse die Schätzung der Parameter erfolgen. Einschränkend soll jedoch angemerkt werden, daß in diesem Beitrag auf wichtige Abweichungen von den Annahmen des klassischen linearen Regressionsmodells nicht eingegangen wurde, so wurde beispielsweise durchgängig von meßfehlerfreien Beobachtungswerten ausgegangen. An dieser Stelle sei allerdings auf die unten angeführte Literatur verwiesen.

Literatur:

Arminger, G. und ***F. Müller*** 1989:
Lineare Modelle zur Analyse von Paneldaten, Opladen: Westdeutscher Verlag.

Baltagi, B.H. 1995:
Econometric Analysis of Panel Data, New York: Wiley.

Dielman, T.E. 1989:
Pooled Cross-Sectional and Time Series Data Analysis, New York/Basel: Marcel Dekker.

Engel, U. und ***J. Reinecke*** 1994:
Panelanalyse: Grundlagen, Techniken, Beispiele, Berlin: de Gruyter

Hsiao, C. 1986:
Analysis of Panel Data, Cambridge: Cambridge University Press.

Johnston, J. 1986:
Econometrics Methods, 3rd ed., New York: McGraw-Hill.

Judge, G. et al. 1985:
The Theory and Practice of Econometrics, 2nd ed., New York: Wiley.

Judge, G. et al. 1989:
Introduction to the Theory and Practice of Econometrics, 2nd ed., New York: Wiley.

Kmenta, J. 1986:
Elements of Econometrics, 2nd ed., New York: Macmillan.

Dörfliche Milieus im vereinigten Deutschland - ein Vergleich qualitativer und quantitativer Daten

von Günter Wolkersdorfer

Zusammenfassung

In Folge der deutschen Vereinigung entstand eine Vielzahl sozialwissenschaftlicher Studien, die sich mit dem gesamtdeutschen Transformationsprozeß beschäftigen. Die ost- und westdeutschen Befindlichkeiten werden darin in mannigfacher sozialstatistischer Weise dargestellt. In der Mehrzahl dieser Arbeiten werden quantitative Methoden genutzt. Demgegenüber stehen wenige qualitative Arbeiten, die versuchen, die regionalisierten Lebenswelten verstehbar zu machen. In der hier vorgestellten Untersuchung wird durch die Nutzung verschiedener methodischer Ansätze die Verbindung zwischen qualitativen und quantitativen Forschungsergebnissen, vor dem Hintergrund eines handlungsleitenden räumlichen Konfliktes, gegeben.

Abstract

The German reunion has given rise to a multitude of studies that deal with the transformation process within the reunited Germany. Conditions and opinions East and West have been described by a variety of social statistics, using for the main part quantitative methods. On the other hand there are a few studies of qualitative approach which aim at the description of regional milieus. The current project means to integrate the findings of both qualitative and quantitative research using various methodological approaches.

1. Einleitung

Die Transformation postsozialistischer Gesellschaften seit der ‘Zeitenwende 1989’ wurde weder von Politikern noch von Sozialwissenschaftlern erwartet. Bei der Suche nach den Ursachen der Revolutionen fand man die Universalerklärung des Transformationsgeschehens in der‘Nachholenden Modernisierung’ der vormodernen sozialistischen Gesellschaften. Diese Erklärung wurde durch eine Vielzahl sozialwissenschaftlicher Studien, die sich mit dem gesamtdeutschen Transformationsprozeß beschäftigen, untermauert. Die westlichen Sozialwissenschaftler versuchten mit Hilfe quantitativer Verfahren nachzuweisen, daß es

sich im Falle Ostdeutschlands um eine rückständige Gesellschaft handelt. Diese An-alysen bestärkten die politischen Akteure in der Handlungsleitlinie, durch eine Übertragung des westlichen Institutionengefüges die rückständige Ostgesellschaft zu modernisieren. Es herrschte die Überzeugung vor, daß durch die Einführung des überlegenen westlichen Systems eine überzeugende Perspektive gegeben werde. Die Zukunft der rückständigen Akteure Ostdeutschlands sollte in den Denkstrukturen Westdeutschlands liegen. Wie man gegenwärtig feststellen kann, handelte es sich dabei um eine problematische Einschätzung. Die in immer stärkerem Maße auftretenden Anomiererscheinungen in Ostdeutschland dokumentieren dies. Dagegen wird in diesem Aufsatz die Feststellung gesetzt, daß die Lebenswelten und Perzeptionsmuster in Ostdeutschland nur scheinbar rückständig sind. Anhand einer empirischen Vergleichsstudie wird der Frage nachgegangen, inwieweit die ostdeutschen Strukturen des Bewußtseins und des Verhaltens Entwicklungspotentiale bergen.

2. Der Untersuchungsgegenstand

Empirische Studien zum Transformationsgeschehen verlangen nach einordbaren Beispielen. Auf der strukturellen Ebene bietet sich hier die Analyse eines massiven Ereignisses an, das die Handlungen sämtlicher beteiligter Akteure betrifft. Räumliche Konflikte stellen häufig einen solch massiven Einschnitt dar und sie können den geforderten strukturellen Rahmen bieten.

Für diese Arbeit wurde ein räumlicher Konflikt gewählt, der sowohl in Ost- als auch in Westdeutschland verortet ist und kontrovers diskutiert wird. Den Ansatzpunkt bietet der Zielkonflikt zwischen ökonomischen und ökologischen Interessen, der mehr und mehr überlagert wird von seinen räumlichen Implikationen. Der steigende ökonomische Bedarf der Ballungsräume hat direkte Auswirkungen auf die Peripherräume und führt zu einem Anwachsen der Disparitäten. Der Begriff des ‘ökologischen bzw. sozio-ökonomischen Schattens’ beschreibt sehr gut die Auswirkungen eines Entwicklungsbooms der Kerngebiete auf Kosten der Peripherregionen. In den Untersuchungsregionen beeinflußt der immense Energiebedarf des Verdichtungsraumes Rhein-Ruhr (West) und des Verdichtungsraumes Berlin (Ost) die Entwicklungen in den jeweiligen Peripherregionen. In beiden Regionen wird dieser Bedarf zum großen Teil durch die Verstromung von Braunkohle gedeckt. In dieser Folge manifestiert sich der räumliche Konflikt in Form des Braunkohletagebaus und den in seiner Folge durchgeführten Umsiedlungen (Vertreibungen) von Bewohnern. „Jede Zwangsumsiedlung reißt die Betroffenen aus ihrer bisherigen Lebenswelt und ist, ganz abgesehen von ihren Methoden, ein Eingriff in die Menschenrechte“ (***Brockhaus*** 1967, S. 355). Insofern ist es interessant, zwei Dörfer darzustellen, die historisch wie geographisch in völlig unterschiedlichen Regionen liegen. Die gegenwärtige Entwicklung wird jedoch von der gleichen politischen und sozialen Rahmenhandlung bestimmt. Für die Einwohner beider Dörfer stellt die bevorstehende Umsiedlung das herausragende strukturelle Er-

eignis der Zukunft dar. Weiterhin war wichtig, daß die Dörfer im Hinblick auf die sozialstatistische Situation vergleichbar sind und sich ähnliche Konfliktkonfigurationen zeigen. Deshalb wurden zwei Untersuchungsorte ausgewählt, die beide eine typisch dörfliche Physiognomie aufweisen und sich in der Größe der Einwohnerzahl ähneln. Trotzdem muß klar sein, daß ein einfacher (1:1)-Vergleich beider Dörfer auf Grund der Einzigartigkeit von Individuen und gesellschaftlichen Strukturen weder wünschenswert noch möglich ist. Erst die Analyse des Konfliktablaufes in dem ost- und dem westdeutschen Dorf, vor dem Hintergrund der unterschiedlichen gesamtgesellschaftlichen Entwicklungen, läßt diesen Vergleich sinnvoll werden. Einen wichtigen Beitrag leistet hier die jeweilige individuelle Biographie einer Dorfgemeinschaft. Der historische Hintergrund, gestaltet durch die Handlungen der jeweiligen Individuen, manifestiert sich in der jeweiligen Geographie. Der individuelle Raumausschnitt ist somit Ausgangspunkt der Handlungen und erfährt im gleichen Moment seine Überprägung. Die individuellen Handlungsparameter in den Dörfern werden vorgestellt, um die tatsächlichen individuellen Abwägungen aufzuzeigen. Die strukturellen Einflüsse bilden sich auch auf der individuellen Ebene ab; ihre Darstellung geschieht jedoch mit Rücksichtnahme auf die möglichen Verschleierungen, die diese kollektivistischen Parameter in sich tragen.

3. Überblick über die Untersuchungsmethoden

Es wurde ein möglichst breiter Zugang zu den einzelnen Konfliktparteien und den Dimensionen des Konfliktes angestrebt. Dieser ließ sich am besten durch die Nutzung verschiedener qualitativer und quantitativer Methoden erreichen. Durch die qualitative Analyse werden die erkenntnisleitenden Gesichtspunkte herausgearbeitet. Die quantitative Überprüfung hatte die Aufgabe die Dimension und Reichweite der Ergebnisse zu ermitteln.

3.1. Sondage

Im explorativen Teil wurden qualitative und quantitative Methoden verwendet, um einen Überblick über das Untersuchungsfeld zu bekommen. Es folgte die phänomenologische Darstellung der beiden Untersuchungsräume. Die deskriptive Darstellung wurde mittels Dorf- und Regionalbeschreibungen, sozialstatistischen Erhebungen und Kartendarstellungen der zu untersuchenden Konflikträume gegeben. Es wurden Luftbilder der beiden Dörfer bei den Landesvermessungsämtern ausgewertet, die einen Eindruck von der Physiognomie der Dörfer geben sollten. Es folgten teilnehmende Beobachtungen in den beiden Gemeinden, fotografische Bestandsaufnahmen und 'Experten'-Interviews mit den Ortsvorstehern bzw. Bürgermeistern, den Gemeinderäten, Pfarrern, Kaufleuten, Gastwirten, Vereinsvorständen, den Vertreten des Braunkohletagebauunternehmens etc.

3.2. Qualitativ-interpretativer Teil

Die qualitativen Interviews sollen einen Einblick in die Lebenswelten der Dorfbewohner West und Ost geben. Die Interviews wurden mit biographisch-orientierten Fragen begonnen. Daran schlossen sich themengesteuerte Leitfadeninterviews an, in denen die individuellen Lebenswelten der Befragten in bezug auf die Region und den Konflikt im Mittelpunkt standen. Die Befragten sollten über ihr alltägliches Leben im Dorf berichten. Weiterhin hoben die Interviews darauf ab, die Konfliktorientierung in Abhängigkeit von der Systementwicklung zu erfassen. Der Leitfaden wurde im Laufe der Untersuchung ständig weiterentwickelt.

3.3. Quantitativ-statistischer Teil

Die Erfahrungen aus den Tiefeninterviews gingen in das Design des standardisierten schriftlichen Fragebogens ein. Die Befragten wurden durch ein Zufallsverfahren aus dem Einwohnermelderegister ermittelt. Als Grundgesamtheit waren die Wohnbevölkerung von Horno und Morschenich definiert worden, d. h. alle Personen, die zum Zeitpunkt der Erhebung in den Dörfern wohnten und das 18. Lebensjahr vollendet hatten. In Morschenich waren das 460 und in Horno 305 Personen. In Morschenich wurden 80 und in Horno 70 Fragebögen mit Rückumschlag verteilt. Der Rücklauf wurde bis zum 15.7.96 berücksichtigt. Bis dahin wurden aus Morschenich 29 ausgefüllte Fragebögen zurückgegeben; aus Horno gingen 43 ein. Das entspricht einem Rücklauf von 36 % in Morschenich und 61% in Horno. Im Vergleich mit anderen lokalen schriftlichen Befragungen sind das akzeptable bzw. überdurchschnittliche Werte. Unter strengen Repräsentativitätskriterien muß die Einschränkung gemacht werden, daß die Nettostichprobe, vor allem für das Dorf Morschenich, etwas klein ist, um statistisch signifikante Aussagen treffen zu können. Die erhobenen Daten gewährleisten jedoch Tendenzaussagen.[1]

4. Die Bewertung und Wahrnehmung des gesellschaftlichen und räumlichen Konfliktes in den beiden Dörfern - einige zentrale Ergebnisse

4.1. Sondage

Strukturelle Motivationen

In Morschenich, dem westdeutschen Dorf, finden sich die Ausprägungen eines lange etablierten, kapitalistisch-demokratischen Systems. Dieses System führte zu einer Auflösung der althergebrachten Dorfstrukturen. Die Unterschiede zwischen Dorf und Stadt

1 Um ein Maß für die Qualität der Brutto- und der Nettostichproben zu erhalten, wurden die Prozentverteilungen der Stichprobenvariablen Geschlecht und Alter mit den zur Verfügung stehenden Einwohnerlisten der Gemeindeämter verglichen. Es zeigte sich, daß sich die Merkmale Alter und Geschlecht in der Nettostichprobe ähnlich wie in der Grundgesamtheit verteilten.

verschwanden zunehmend. Sämtliche Indikatoren, seien diese nun Bildung, Einkommensstruktur oder Wertorientierung, belegen eine nivellierende Entwicklung. In dieser Folge wurde die ehemalige solidarische Gemeinschaftsorientierung mehr und mehr auf- und abgelöst durch jene ubiquitäre Orientierung an ökonomischer Macht und individueller Lebensgestaltung. Die zunehmende Spaltung der Gesellschaft trifft die Kernräume ebenso wie die ländlichen Gebiete. Ist diese Entwicklung in den segregierten Großstädten sozialräumlich direkt sichtbar, so funktionieren diese Prozesse in Dörfern subtiler. Ein räumlicher Konflikt mit seinem massiven Veränderungspotential bringt die Dynamik dieser Entwicklungen jedoch direkt an die Oberfläche.

Die strukturellen Hintergründe der Entwicklungen in **Horno**, dem ostdeutschen Dorf, unterscheiden sich hingegen gravierend. Die Transformation in Ostdeutschland verlief rasch als vollständige Übernahme des westlichen Systems. Der Beitritt in das bis dahin funktionierende Staatensystem der Bundesrepublik erschien verlockend und angesichts der gewaltigen Kapitaltransfers wenig problematisch. Zwar brachen direkt nach der Vereinigung weite Teile der Industrie und der alten Handelsbeziehungen weg und es wurden von 1990 bis 1992 rund ein Drittel der Arbeitsplätze abgebaut, dennoch gab dies zu keinen übertriebenen Sorgen Anlaß. Diese Entwicklungen waren, wenn auch nicht in diesem Ausmaß, von den kollektiven Führungseliten einkalkuliert worden und wurden nur als Einbrüche einer kurzen Übergangsphase angesehen. Bald traten jedoch von den politischen und wirtschaftlichen Eliten zum großen Teil nicht vorhergesehene Ereignisse auf den Plan. Der neue Regulationsmodus einer flexiblen Akkumulation bestimmte und bestimmt in immer stärkerem Maße die wirtschaftlichen Abläufe und somit die ostdeutsche Transformationsentwicklung. In diesem neuen globalen Marktsystem hat die ostdeutsche Wirtschaft denkbar ungünstige Ausgangsbedingungen. Die Übertragung des hohen Entwicklungsstands Westdeutschlands, angepaßt an die dortige Entwicklung, stellt für Ostdeutschland eine schwere Hypothek dar. Insofern unterscheidet sich Ostdeutschland gravierend von den osteuropäischen Transformationsstaaten, in denen die Möglichkeit besteht, von einem gemeinsamen niedrigen Niveau auszugehen. Der zunächst als Vorteil angesehene Entwicklungsvorsprung des großen Bruders im Westen verkehrt sich so im Zeitalter eines globalen Entwicklungsmodells zum gewaltigen Nachteil.

Das große Problem jener Übernahme war weiterhin, daß den Beigetretenen keinerlei individueller Handlungsspielraum blieb. Jenes prägende, von den politischen Eliten formulierte, Zitat dieser Epoche:‘Keinem sollte es schlechter und vielen sollte es besser gehen’, führte im Kontext mit den Wendeerfahrungen zu einer sehr hohen Erwartungshaltung.Deshalb wirkten die zunehmend negativen Begleiterscheinungen schockartig auf eine zur Passivität verdammten Bevölkerungsgruppe.Der historisch einmalige Rückgang der Geburtenraten in Ostdeutschland um mehr als die Hälfte, bietet ein eindrucksvolles Zeugnis, ist jed-o-ch nur einer der vielen Indikatoren für eine wachsende Verunsicherung. Neben diesen strukturell

sichtbaren Erscheinungen treten die massiven Unterschiede zwischen ländlichen und städtischen Regionen in den Vordergrund. Auf das traditionelle Leben im ländlichen Raum wurde kaum Reformdruck ausgeübt, während sich das städtische Bürgertum auflöste.

Individuelle Motivationen

Von ganz entscheidender Bedeutung ist, daß sich diese Entwicklungen individuell manifestieren. Die Individuen treiben den Prozeß voran und sie haben dabei einen gewaltigen Handlungsspielraum. Hier zeigt die Untersuchung die unterschiedlichen Reaktionsweisen in den beiden Dörfern. Durch die Durchführung der Interviews und die Auswertung des Fragebogens konnten die jeweiligen Gründe für die verschiedenen Haltungen herausgearbeitet werden. Ziel des qualitativen Teils war es, die unterschiedlichen Lebenswelten und Hintergründe herauszuarbeiten. Bei der Darstellung der in den Interviews gemachten Aussagen wird zwischen den tatsächlichen individuellen Motiven und den strukturellen Antwortmustern unterschieden. Daran anschließend wird die quantitative Verteilung der entdeckten Dimensionen des Alltagshandelns für die Bereiche Wertorientierung, Wohndauer, Konfliktwahrnehmung und Protesthaltung aufgezeigt.

4.2. Qualitativer Untersuchungsteil

Eine Betrachtung der verallgemeinernden Erklärungsmuster zeigt in **Morschenich** folgende generelle Argumentationsmuster an.

Bei denjenigen, die einer Umsiedlung positiv gegenüberstehen, ist eine vorwärts gerichtete Haltung feststellbar. Die zukünftigen persönlichen Entwicklungslinien werden vor allem durch strukturelle Parameter gegeben. *„Wo soll denn all die Arbeit herkommen? Wenn es nach den Grünen geht, gehen hier bald die Lichter aus. Das kann aber nicht sein. Die Sicherung des Standortes Deutschland hängt nun mal daran“*. Viele sehen sich als direkte Interessenvertreter der Tagebaubetreiber und sind als ‘Braunkohlenlobby’ im Dorf zu einer festen Größe geworden. *„Energie ist nun mal wichtig und ich denke es ist nicht gut, wenn wir die nur aus dem Ausland beziehen“*.

Bei denjenigen, die einer Umsiedlung negativ gegenüberstehen, ist eine Orientierung an ehemaligen Werten feststellbar. Es werden Bezüge zu einer Vergangenheit hergestellt, in der solidarische Werte wie Gemeinschaft und Hilfsbereitschaft noch etwas zählten, vieles besser war. *„Die Dorfgemeinschaft war wirklich gut, einer unterstützte den anderen. Wir waren alle in Vereinen, alle zogen da mit“*.

Eine Mischung aus Hoffnungslosigkeit und Gleichgültigkeit bestimmte das individuelle Antwortverhalten. *„Ja man sagt ja oft, Heimat ist das, was einen an eine bestimmte Gegend bindet, aber vielleicht ist das nicht ganz so, dazu gehört auch die Familie. Ich fühle*

mich im ganzen Raum zu Hause, das ist für mich Heimat.... Ich bin ein sehr verstandesorientierter Typ, also wenn ich woanders hingehen muß, gehe ich dahin“.

Die Dominanz des Braunkohletagebaus in den Perzeptionsmustern der Einwohner zeigt sich in der Spaltung des Dorfverbandes. *„Die Leute hier interessiert das Ganze gar nicht mehr so. Am Anfang war da viel mehr los. Bei der ersten Veranstaltung der Bürgerinitiative, wo auch so ein Film über die Umsiedlung gezeigt wurde, waren noch über 60 Leute da. Heute geht da keiner mehr hin. Man möchte sich nicht als Gegner von Rheinbraun zu erkennen geben“* Offensichtlich ist die Aufspaltung der dörflichen Bevölkerung in eine traditionelle und eine neu zugezogene Einwohnerschaft. Während für die traditionelle Gruppe das Dorf noch immer identitätsbildend wirkt und eine starke emotionale Ortsbindung zu finden ist, hat der Ort für die Gruppe der 'Neuzugezogenen' rein ökonomischen Wert. Auch wenn auf die Wichtigkeit identitätsstiftender Einrichtungen, wie etwa Vereine, häufig verwiesen wird, so besteht hier vor allem ein Interesse bezüglich der gemeinsamen Freizeitgestaltung. Die persönlichen Lebenspläne der Einwohner sind sonst jedoch so weit individualisiert, daß für die Entwicklung einer gemeinsamen Zukunft keine Ressourcen zur Verfügung gestellt werden. Die Umsiedlungspläne werden dementsprechend ignoriert oder so in den eigenen Lebensplan eingebaut,daß man für die Umsiedlung schon gerüstet ist.Auch wenn man mit Prognosen vorsichtig sein muß, ist zu erwarten, daß die zukünftige Umsiedlung in Morschenich relativ geräuschlos über die Bühne gehen wird. Insofern wird der räumliche Konflikt um den Braunkohletagebau in Morschenich wohl nicht direkt offenbar werden. In der entsolidarisierten Ortsgemeinschaft werden seine Folgen vor allem individuell wirksam werden.

Auf der strukturellen Ebene der Interviews taucht in **Horno** immer wieder die Erinnerung an die DDR auf. Es würde jedoch zu weit gehen, von einer DDR-Nostalgie zu sprechen. Die Entwicklungen seit der Wende werden von den meisten positiv aufgenommen. Vor allem die politische Freiheit und die Reisefreiheit werden positiv, der Verlust an Sicherheit und Gewohnheit hingegen wird negativ beurteilt. Insgesamt steht man verallgemeinernden Erklärungen kritisch gegenüber *„Für mich ist dies nun schon das dritte System, in dem ich lebe. Da kann ich schon gut unterscheiden“. Früher entschied der Honecker was passierte, heute ist es der, der das Geld hat“.* Auch bezogen auf die Vergangenheit wird die Rolle des Geldes sehr kritisch betrachtet. *„Geld war ja uninteressant. Es gab zwar nichts, aber man half sich immer untereinander. Damit konnte man gut klarkommen. Man mußte nur Beziehungen haben“.* Heute wird das Geld angegeben, wenn nach Gründen für soziale Veränderungen gefragt wird. Prinzipiell wird eine schleichende Aushöhlung der Gemeinschaft konstatiert.

Auf der individuellen Ebene taucht in vielen Interviews die Erfahrung der zweifachen Vertreibung auf. Ein großer Teil der Einwohner Hornos stammt aus den jenseits der Neiße gelegenen Dörfern. Aus diesen wurden sie 1945 vertrieben und fanden in Horno eine neue Heimat. *„Wir mußten plötzlich Strega verlassen, nahezu alles dortlassen. Von der Hoch-*

fläche aus kann man den weißen Kirchturm sehen. Unser Haus steht noch, aber dort war ich seit dem Krieg nicht mehr“. Diese einmalige Vertreibungserfahrung führte bei den Betroffenen zu einer klaren Haltung in der Umsiedlungsdiskussion. Die persönlich erlebten körperlichen und psychischen Folgen einer Umsiedlung sind bei vielen Bewohnern noch so präsent, daß sie sich massiv gegen eine Umsiedlung stellen. *„Ich habe genug verloren. Eine Heimat habe ich in Polen verloren. Noch einmal lasse ich mich nicht vertreiben. Die können es ja versuchen“*. Der Tagebaubetreiber versucht mit großem psychologischem Geschick den Dorfverband auseinander zu dividieren. Die Wirkung dieser Aktivitäten sind jedoch auf Grund des engen, stark durchorganisierten Dorfverbandes eher kontraproduktiv. *„Ich (zentrale Person des Dorfes) höre immer wieder, ich hätte mir schon längst ein Haus in Guben gekauft. Immer wieder werden diese Gerüchte gestreut. Die kann ich nur wegen dem großen Vertrauen und dem ständigen Austausch im Dorf zerstreuen“*.

Ganz explizit wurden mit der gesellschaftlichen Wende verbundene Hoffnungen in die ‘neue Zeit’ genannt. *„1989 dachten wir, wir schaffen jetzt alles“*. *„Wir haben das Grundgesetz wie einen Krimi gelesen“*. Die Einwohner waren schon seit 1977 mit der Umsiedlung konfrontiert worden. Selbst im repressiven DDR-System hatten viele Dorfbewohner gegen die Umsiedlung geklagt, was ihnen eine massive Überwachung durch den Staatssicherheitsdienst eintrug. Dabei war immer klar, daß es keine Möglichkeit gab, der Umsiedlung zu entgehen. Die Ereignisse und Erfahrungen der letzten Jahre habe viele Träume platzen lassen und zu einer neuen resignativeren Beurteilung beigetragen.

Die Betrachtung der individuellen Lebenswelten im ostdeutschen Dorf zeigte, daß in Horno tatsächlich noch eine traditionelle Einwohnerschaft vorherrscht. In dieser homogenen Gruppe tragen einige Schlüsselpersonen die 'Meinung des Dorfes' nach außen. In einer fortwährenden Auseinandersetzung zwischen allen Mitgliedern der Dorfgemeinschaft werden neue Zukunftsstrategien entworfen. In erstaunlicher Einigkeit wurde diese Haltung in allen Erhebungsphasen dargelegt. Der räumliche Konflikt, in den das Dorf involviert ist, eint die Dorfgemeinschaft. Er schweißt die einzelnen Mitglieder der Gemeinschaft so eng zusammen, daß kaum Platz für abweichende Meinungen bleibt. Sämtliche massiven Spaltungsversuche von außerhalb, in den westlichen Braunkohledörfern lange äußerst erfolgreich angewandt, prallen daran ab. Dieser solidarische Dorfverband stellt für viele, vor allem für westdeutsche Entscheidungsträger in Politik und Wirtschaft, ein Phänomen dar. Die üblichen Eingriffsmöglichkeiten, seien sie ökonomischer oder psychologischer Art, verfangen nicht.

4.3. Quantitativer Untersuchungsteil

Nachdem die Einwohner der beiden Dörfer in den Tiefeninterviews ihre persönlichen Lebenswelten offenbarten, sollten diese Einsichten mit einigen quantitativen Ergebnissen verdeutlicht werden. Dabei ging es um die Frage, welche Muster der Identifikation, der

Bindung und der Konfliktorientierung lassen sich für die westdeutschen und ostdeutschen Dorfbewohner auffinden.

Der Einstieg in den Fragebogen erfolgte mit einer Vergleichsfrage. Die Frageformulierung entspricht der Anfang 1996 für die Bundesrepublik Deutschland durchgeführten repräsentativen Bevölkerungsumfrage des Instituts für Demoskopie Allensbach zu den Unterschieden zwischen Ost- und Westdeutschen in bezug auf das Freiheits- und das Gleichheitsideal. Das macht eine vergleichende Betrachtung der Ergebnisse für Deutschland Ost und West mit den Ergebnissen der beiden Dörfer möglich. In der folgenden Tabelle werden in Spalte 1 und Spalte 2 die Ergebnisse für Westdeutschland und Morschenich, in Spalte 3 und Spalte 4 die Ergebnisse für Ostdeutschland und Horno präsentiert.[2]

Tabelle 1: Ein Vergleich der Wertorientierungen

Frage: Hier stehen zwei Meinungen. Welcher von beiden würden Sie eher zustimmen, der ersten oder der zweiten? (Angaben in Prozent)

	Deutschland West	Morschenich	Deutschland Ost	Horno
Erste Meinung: „Ich finde Freiheit und Gleichheit eigentlich beide wichtig. Aber wenn ich mich für eines davon entscheiden müßte, wäre mir die persönliche Freiheit am wichtigsten, daß also jeder in Freiheit leben und sich ungehindert entfalten kann“.	56	55	35	12
Zweite Meinung: „Sicher sind Freiheit und Gleichheit wichtig. Aber wenn ich mich für eines davon entscheiden müßte, fände ich Gleichheit am wichtigsten, daß also niemand benachteiligt ist und die sozialen Unterschiede nicht so groß sind“.	28	35	47	51
weder - noch	9	0	9	9
unentschieden	7	10	9	28

Quelle: Werte für Gesamtdeutschland: Institut für Demoskopie Allensbach 1996 in FAZ vom 13.03.1996; Werte für Morschenich und Horno: eigene Erhebung

Die schon beträchtlichen Unterschiede zwischen Freiheits-und Gleichheitszielen,die für die Grundgesamtheiten Deutschland Ost und West ermittelt wurden,verstärken sich in den beiden dörflichen Untersuchungsräumen. Der Wert ‘Gleichheit’ wird in Morschenich und in Horno höher eingeschätzt als in Deutschland West bzw. Ost. Dies erscheint vor dem Hintergrund der Befragung einer dörflichen Einwohnerschaft, in der Gemeinschaftswerte traditionell hoch bewertet werden, nachvollziehbar. Vergleicht man die Untersuchungsräume, so

2 Für die Interpretation der Ergebnisse ist auf die Verwendung unterschiedlicher Forschungsdesigns hinzuweisen. Das Institut für Demoskopie Allensbach wählte die Befragten per Quota-Verfahren aus, während in der aktuellen Untersuchung die Befragten per Random-Verfahren ausgewählt wurden.
Weiterhin sind die Erhebungsarten verschieden: mündliche Befragung per Interviewer vs. schriftliche Befragung.

fällt jedoch auf, daß das Ergebnis in Horno bedeutend stärker vom Ergebnis in Ostdeutschland abweicht. Insgesamt wird dadurch eine völlig unterschiedliche Gewichtung der Ziele Freiheit und Gleichheit in den Untersuchungsdörfern offenbar.

Als nächster Schritt wurde die jeweilige Wohndauer der Einwohner von Morschenich und Horno ermittelt.

Tabelle 2: Die Wohndauer der Bevölkerung

Frage: Wie lange leben Sie schon in Morschenich/Horno ?

Wohndauer in Jahren	Morschenich	Horno
0 - 5	24 %	2 %
5 - 15	10 %	9 %
15- 25	17 %	26 %
über 25	49 %	63 %

Quelle: eigene Erhebung

Im Hinblick auf die Verteilungen in den Stichproben zeigen die Werte in Horno, daß für die meisten Einwohner das Dorf Heimat- und Geburtsort ist. In Morschenich deuten die hohen Zuzugswerte innerhalb der letzten Jahre eine veränderte Entwicklungslinie an. Da die Bevölkerungszahlen in den Dörfern insgesamt sich in den letzten Jahren kaum verändert haben, müssen sich erhebliche demographische Wandlungsprozesse vollzogen haben. Die durch die unterschiedlichen gesellschaftlichen Rahmenbedingungen ausgelösten Entwicklungsprozesse bilden sich in den Wohndauerwerten ab. Neben der unterschiedlichen Entwicklungslinie dörflicher Entwicklung generell, spielen die verschiedenartigen Mobilitätsprozesse eine große Rolle. Beide Dörfer liegen im weiteren Einzugsbereich von Großstädten, so daß sich allein daraus die Unterschiedlichkeiten nicht erklären lassen.

Während in Ostdeutschland der Urbanisierungsprozeß weiter fortschritt und die Dörfer kaum Veränderungsdruck ausgesetzt waren, kam diese Entwicklung in Westdeutschland seit dem Ende der 60er Jahre zum Stillstand. Es entstanden neue Mobilitätsformen, die sich vor allem durch Suburbanisierungsprozesse aus den Kernstädten in die umliegenden Peripherräume auszeichneten. Die Folgen dieser Kern-Rand-Wanderung zeigen sich auch in Morschenich. Die Bevölkerung teilt sich folglich in eine stark mobile, aus den Verdichtungsräumen zugezogene, und in eine immobile, aus dem Ort stammende, Einwohnerschaft auf. Von den Umsiedlungsplanungen sind diese dementsprechend verschieden stark betroffen. Die 'bodenständige' Bevölkerungsgruppe ist besonders von den immateriellen Folgen der Umsiedlungsabsichten betroffen. Es besteht eine enge Beziehung zwischen langer Wohndauer und starker Konfliktbetroffenheit.

Um die zukünftige Haltung zum räumlichen Konflikt zu ermitteln, wurde eine aufgespaltene Befragungsform (Gabelfrage) gewählt. Ziel der ersten Teilfrage war es, zu ermitteln, wie die Einwohner im Fall einer aktuellen Umsiedlungssituation reagieren würden.

Tabelle 3: Zukünftige Protesthaltung

Frage: Falls die Umsiedlung ihres Dorfes tatsächlich anstünde, würden Sie dagegen protestieren?

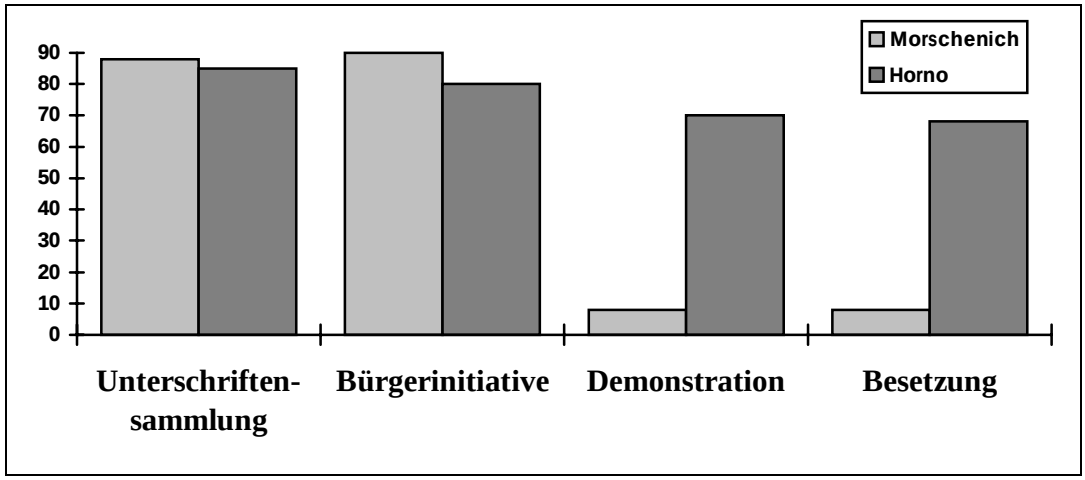
Antwortvorgabe	Morschenich	Horno
ja	45 %	92 %
nein	31 %	4 %
vielleicht	21 %	6 %
weiß nicht	3 %	0 %

Quelle: eigene Erhebung

In Morschenich gab knapp die Hälfte der Einwohner an, im Fall einer konkreten Umsiedlungssituation Widerstand gegen diese Entwicklung in verschiedenen Protestformen zu äußern. In Horno erklärte sich nahezu die gesamte Einwohnerschaft dazu bereit. In der zweiten Teilfrage sollten diejenigen, die die erste Teilfrage positiv beantwortet hatten, angeben, in welcher Form sie ihren Protest äußern würden. Das bedeutete, daß in Morschenich 45 % und in Horno 92 % der Befragten auf diese Frage antworten sollten. Dabei wurden die Antwortvorgaben so gewählt, daß die Schärfe der Form des Widerstands von links nach rechts anstieg. Bei dieser Frage waren Mehrfachantworten zulässig. Durch die Säulenhöhe wird deshalb die Bereitschaft, bezogen auf jede einzelne Protestform, dargestellt.

Tabelle 4: Bereitschaft zu verschiedenen Protestformen

Frage: wenn ja, in welcher Form? (Mehrfachnennung)



Quelle: eigene Erhebung

Die Befragten, die angeben, Widerstand gegen eine zukünftige Umsiedlung äußern zu wollen, sind zu verschiedenen Protestformen bereit. In Morschenich beschränkt sich diese

Bereitschaft im großen und ganzen auf die gemäßigten Protestformen Unterschriftensammlung und Bürgerinitiative. Die Hornoer Einwohner sind zu einem Großteil bereit, sämtliche Formen des zivilen Widerstands, zu nutzen.

5. Fazit und Ausblick

Die Beobachtungen in den beiden Dörfern machten klar, daß man es mit hochkomplexen Gebilden zu tun hat. Das Leben der Dorfbewohner und ihre Reaktion auf den Konflikt ist durch viele Einflüsse geprägt, die möglichst umfassend herausgearbeitet werden müssen. Die beinahe schon inflationär erscheinenden Vergleichsuntersuchungen, die West- und Ostdeutschland implizit vor dem Fokus einer 'Nachholenden Modernisierung' analysieren, erscheinen hier problematisch. Geht man beispielsweise von den ermittelten ost- und westdeutschen Dispositionen aus, so verhielten sich viele Dorfbewohner in Ost und West genau entgegengesetzt den postulierten Klischees, und sie taten dies aus gutem Grund. Jedenfalls deuteten die Ergebnisse der empirischen Studie auf Lebenswelten und Konfliktperzeptionen, die den veröffentlichten sozialen Realitäten entgegenlaufen. Die für Ostdeutschland ja angeblich so bezeichnenden Anomieerscheinungen, aufgezeigt durch Unzufriedenheit und Entfremdung, bestimmen vielmehr den Alltag in dem westdeutschen Dorf. Auf der anderen Seite ist die Lebenswelt in dem ostdeutschen Dorf geprägt von Zuversicht, Mut und Selbstvertrauen.

Literatur

Bertram, H.; Hradil, S.; Kleinherz, S. (Hrsg.) 1996:
Sozialer und demographischer Wandel in den neuen Bundesländern. Opladen.

Dangschat, J. S. 1996:
Raum als Dimension sozialer Ungleichheit und Ort als Bühne der Lebensstilisierung? - Zum Raumbezug sozialer Ungleichheit und von Lebensstilen. In: SCHWENK, O. G. (Hrsg.): Lebensstil zwischen Sozialstrukturanalyse und Kulturwissenschaft. Opladen, S. 99-135.

Gebhardt, W.; Kamphausen, G. 1994:
Zwei Dörfer in Deutschland. Mentalitätsunterschiede nach der Wiedervereinigung. Opladen.

Kollmorgen, R.; Reissig, R. ; Weiss, R. (Hrsg.) 1996:
Sozialer Wandel und Akteure in Ostdeutschland. Opladen.

Opp, K. D. 1996:
Gesellschaftliche Krisen, Gelegenheitsstrukturen oder rationales Handeln? Ein kritischer Theorienvergleich von Erklärungen politischen Protests. In: Zeitschrift für Soziologie 25, H. 3, S. 223-242.

Opp, K. D.; Voss, P.; Gern, C. 1993:
Die volkseigene Revolution. Stuttgart.

Ossenbrügge, J. 1983:
Politische Geographie als räumliche Konfliktforschung. Konzepte zur Analyse der politischen und soziologischen Organisation des Raumes auf der Grundlage anglo-amerikanischer Forschungsansätze. Hamburg.

Werlen, B. 1995:
Sozialgeographie alltäglicher Regionalisierungen. Band 1: Zur Ontologie von Gesellschaft und Raum. Erdkundliches Wissen 116. Stuttgart.

World Congress of Sociology
International Institute of Sociology
University of Cologne, July 7 - 11, 1997

The International Institute of Sociology (IIS)

The IIS was founded in 1893 in Paris by ***René Worms***, thus making it the oldest international association in the social sciences. The 33rd World Congress will be held in July 7-11, 1997 in Cologne. Societies, Corporations, and the Nation-State is the general theme of this Congress, dealing with the problems of the organising of societies into states, and the role of "mediating institutions". Various aspects like social differentiation, a changing world order or the process of globalization are considered.

Plenary Sessions

After the Opening Session (9:30 - 10:30) five PLENARY SESSIONS, meeting during mornings are designed to highlight particular facets of this theme:

1 **Freedom and Control in Contemporary Societies;** (Monday, 11:00 - 13:30)
Elke Koch-Weser Ammassari, Rome, Italy

2 **Culture, Religion, and the Privatisation of Belief;** (Tuesday, 9:00 - 12:30)
Andrew M. Greeley, Chicago, USA / ***Wolfgang Jagodzinski***, Cologne, Germany

3 **Ethnicity, the Nation-State, and Globalization;** (Wednesday, 9:00 - 12:30)
T. K. Oomen, New Delhi, India / ***M. Sasaki***, Tokyo, Japan

4 **Economic Change and the Persistence of Tradition;** (Thursday, 9:00 - 12:30)
Tahar Labib, Tunis, Tunisia

5 **Societal Decay and Institutional Rehabilitation;** (Friday, 9:00 - 12:30)
Alberto Gasparini, Gorizia, Italy

In addition there are about 45 WORKING GROUPS meeting in the afternoon (14:00 - 18:30), e.g:

- Party Development and Social Movements
- Religious Movements, Sects, and Religious Fundamentalism

- Small Scale Wars
- International Migrations as a Source of Social Conflict and Cultural Enrichment
- Comparative Economic Performance: East Asia and the West
- Changing Labour Markets, Unemployment, and Trade Unions
- The Internationalisation of the Mass Media: Forms and Effects in Literate and Non-Literate Societies

The Congress is supported by the High Commission of the European Union; the Federal Ministry of Education, Science, Research, and Technology (BMBF); the Deutsche Forschungsgemeinschaft; the Fritz Thyssen-Foundation; the Ministry of Science and Research of North Rhine-Westphalia; the Robert Bosch-Foundation; the University at Cologne; the City of Cologne; the Cologne Chamber of Commerce; Lengowski & Partner; Lufthansa City Center; and SONY Germany. Lufthansa is the official Carrier.

General Information:

The Congress language is English.
Congress venue: University of Cologne, Albertus-Magnus-Platz, Cologne, Germany.

For further information contact:

Prof. Dr. ***Erwin K. Scheuch***, President, International Institute of Sociology
c/o Kölner Gesellschaft für Sozialforschung,
Liliencronstraße 5, 50931 Köln, Germany
Phone: +49-221-47 69 462-3, Fax: +49-221-47 69 498
email: Scheuch@za.uni-koeln.de
internet: http://www.za.uni-koeln.de/iis-worldcongress/

Buchhinweise

Wolf-Michael Kähler:

Einführung in die statistische Datenanalyse.

Grundlegende Verfahren und deren EDV-gestützter Einsatz,

Vieweg, Braunschweig 1995, DM 49.80, ISBN: 3-528-05526-X.

Ein Buchhinweis von *Steffen M. Kühnel*[1]

Das vorliegende Lehrbuch versteht sich als eine moderne Einführung in die Statistik, bei der die Darstellung von Formeln auf ein Mindestmaß reduziert wird. Anstelle langer Rechenbeispiele wird die Berechnung der Statistiken mit SPSS erläutert. Die Bedeutung und Logik von statistischen Verfahren soll der rezeptartigen Präsentation vorgezogen werden. Neben der Vorstellung von Analysekonzepten enthält das Buch daher auch eine - allerdings knappe - Einführung in die Arbeit mit SPSS für Windows und X-Windows (Unix). Für mich nicht nachvollziehbar ist dabei allerdings die Zweiteilung dieser Einführung in ein kurzes Kapitel und einen zusätzlichen Anhang.

Gelungen erscheint mir die graphische Aufbereitung des Buches. Die Auflockerungsmöglichkeiten moderner Textverarbeitung werden genutzt, ohne dabei zu übertreiben. Die Darstellung der statistischen Inhalte folgt der in der Bundesrepublik - leider - üblichen Aufteilung in beschreibende und schließende Statistik. In einem Einführungs-kapitel werden nach Hinweisen zum Sinn statistischer Datenanalysen die Teile einer Befragung von Gymnasiasten der Oberstufe vorgestellt, deren Daten später weitgehend durchgängig für Anwendungsbeispiele herangezogen werden. Es folgen neben der erwähnten Einführung in SPSS und einem kurzen Kapitel über Meßniveaus zunächst elf Kapitel zur deskriptiven Statistik. Behandelt werden die Darstellung von Verteilungen, Lagemaße, Streuungsmaße einschließlich Schiefe und Steilheit, der Vergleich von Merkmalsausprägungen, nominale und ordinale Zusammenhangsmaße in Kreuztabellen, bivariate lineare Regression und Korrelation und schließlich Rangkorrelation, Konkordanzkoeffizient und der bivariate Zusammen-hang zwischen metrischer und nominalskalierter Variable. Die Darstellung der Zusammenhangsmaße werden durch allgemeine Erläuterungen zum Konzept des statistischen

---

[1] Dr. ***Steffen M. Kühnel*** ist Professor am Institut für Politikwissenschaft, Fachbereich Gesellschaftswissenschaften an der Justus-Liebig Universität in Gießen, Karl-Glöcker-Str. 21 E, 35394 Gießen.

Zusammenhangs eingeleitet und durch zwei Kapitel über Drittvariablenkontrolle und die Beziehungen zwischen Mittelwertunterschieden und Korrelationen abgeschlossen.

In sieben weiteren Kapiteln geht es um Verfahren der schließenden Statistik. Nach einer Vorstellung des Konzepts der Zufallsstichprobe wird zunächst am Beispiel von ***Pearsons*** Chiquadratstatistik zur Prüfung der Unabhängigkeit von zwei kreuztabellierten Variablen in die Logik des Signifikanztests eingeführt. Dabei wird auch die Verwendung dieser Teststatistik als Anpassungstest vorgestellt. Im nachfolgenden Kapitel über Z- und T-Tests zur Prüfung von Mittelwerten werden zusätzliche Hinweise zur Logik des statistischen Testens gegeben. In der Darstellung von Konfidenzintervallen wird deren Logik am Beispiel des Vertrauensbereichs für einen Mittelwert vorgeführt. In den letzten drei Kapiteln werden T-Tests für Mittelwertvergleiche bei abhängigen und unabhängigen Stichproben, nichtparametrischen Tests zum nichtparametrischen Vergleich zweier Verteilungen und schließlich einfache Varianzanalyse bei abhängigen und unabhängigen Stichproben sowie deren nichtparametrischen Äquivalente vorgestellt. Ergänzend enthält die Arbeit 14 Anhänge, die neben zusätzlichen Hinweisen zur Kodierung von Fragebögen, zur axiomatischen Meßtheorie, zur Wahrscheinlichkeitstheorie und zu theoretischen Verteilungen vor allem Tabellen mit kritischen Quantilwerten für die vorgestellten statistischen Tests enthalten. Den Abschluß bilden das Literaturverzeichnis und ein Sachindex.

Die Auswahl des Stoffes orientiert sich eher an den Bedürfnissen von Pädagogen und Psychologen denn an denen von Soziologen und Politologen. So werden unter der Überschrift 'Vergleich von Merkmalsausprägungen' die in der psychologischen Diagnostik verwendeten Berechnungen von Testscores erläutert und der Schwerpunkt der vorgestellten statistischen Tests liegt in varianzanalytischen Fragestellungen. Eindeutig zu kurz kommt dagegen für meinen Geschmack die lineare Regression. Multiple Regression wird ebensowenig thematisiert wie inferenzstatistische Aussagen zur Regressionsanalyse. Auch die Drittvariablenkontrolle wird nach meiner Auffassung viel zu kurz behandelt.

Vom soziologischen Kontext abweichend ist auch die vom Autor verwendete Begrifflichkeit. So bezeichnet der Autor eine Scheinkorrelation (Scheinkausalität) als Beispiel für einen Interaktionseffekt einer Drittvariablen (S.153). Abweichende Wortwahl muß kein Problem sein. Der Leser mag aber in Schwierigkeiten geraten, wenn er den Literaturhinweisen des Autors folgt und dort auf einen anderen Sprachgebrauch trifft. Problematischer erscheint es mir, wenn die Gefahr von Fehlinterpretationen besteht. So schreibt der Autor beispielsweise über Zufallsauswahlen: "Neben einer Zufallsauswahl gibt es weitere Erhebungsformen, mit denen man versuchen kann, sich einen Einblick in eine Grundgesamtheit zu verschaffen. Von besonderer Bedeutung sind dabei z.B. die Techniken "geschichtete Stichproben", "Klumpenstichproben", "zwei- und mehrstufige Stichproben", "Quotenverfahren" und "systematische Stichproben"." (S.167). Sieht man von Quotenverfahren und systematischen Auswahlen ab, sind die aufgezählten Beispiele spezielle Formen von Zufallsauswahlen.

Gewünscht hätte ich mir bei diesem Lehrbuch ein noch stärkeres Eingehen auf wichtige statistische Grundkonzepte. Erst vor diesem Hintergrund wird deutlich, daß die Vielzahl der statistischen Modelle Variationen weniger Grundmuster sind. Und erst vor diesem Hintergrund kann ein sozialwissenschaftlicher Anwender meiner Ansicht nach begründet entscheiden, welchen Stellenwert ein spezielles Analysemodell für seine inhaltliche Fragestellung hat. Bei einigen Bereichen ist dies dem Autor durchaus gelungen. Ein Beispiel ist die Behandlung von Konfidenzintervallen. Sehr oft findet man die Auffassung, daß diese die Wahrscheinlichkeit angeben, daß ein gesuchter Grundgesamtheitswert in dem ausgerechneten Intervall liegt. Bei ***Kähler*** wird dagegen explizit festgestellt, daß sich die Wahrscheinlichkeit von Konfidenzintervallen auf die Menge aller Intervalle bezieht und bei geringer Irrtumswahrscheinlichkeit eben nur "die Hoffnung" besteht, daß der gesuchte Wert im Intervall steht (S. 248). Dargestellt wird auch der Zusammenhang zwischen Konfidenzintervallen und Hypothesentests.

Aus meiner Sicht ist diese Einführung in die statistische Datenanalyse vor allem als Lektüre zur Vorbereitung sowie als Ergänzung von Lehrveranstaltungen geeignet. Für das im Klappentext annoncierte Selbststudium würde ich es dagegen weniger empfehlen.

Michael Carlton und *Silvia Schneider* (Hrsg.): Rezeptionsforschung. Theorien und Untersuchungen zum Umgang mit Massenmedien, Opladen, Westdeutscher Verlag, 1997, 289 S., DM 52.00, ISBN 3-531-12825-6.

Ein Buchhinweis von *Eric Mayer*

Das Medienangebot und die Vielfalt des Umgangs mit diesem Angebot wird verschiedenen wissenschaftlichen Disziplinen (Filmwissenschaft, Literaturwissenschaft, Linguistik, Psychologie, Soziologie, Kommunikationswissenschaften, Kognitionswissenschaften, u.v.a.) untersucht. Trotzdem sind viele Fragen hinsichtlich der Mediennutzung verschiedener Medien und der Prozesse, die beim Umgang mit diesen eine Rolle spielen, ungeklärt. Und aufgrund der Pluralität der wissenschaftlichen Disziplinen, die sich spezialisierend mit der Medienforschung beschäftigen, führen - so die Herausgeber - fehlender Erkenntnisaustausch und in Fragestellung von Forschungsfragen und -standards zu kontraproduktiven Ergebnissen. Mit ihrem Band, dessen Mehrzahl der Beiträge auf einer Tagung des SFB 321 *Übergänge und Spannungsfelder zwischen Mündlichkeit und Schriftlichkeit* mit dem Titel *Science Meets Fiction* 1995 vorgestellt wurden, versuchen die Herausgeber, einen Schritt hin zu einer interdisziplinären Medienforschung zu leisten. Die Auswahl der Beiträge erfolgte somit auch unter dem Ziel, Wege zu einer transdisziplinären Medienwissenschaft mit

einem verbindlichen Kanon methodenspezifischer wissenschaftlicher Standards und einem geteilten Problembewußtsein aufzuzeigen.

Eine kritische Auseinandersetzung mit einigen theoretischen Grundlagen der Rezeptionsforschung, dem rezipientenorientierten Ansatz der Medienforschung, findet im ersten Teil des Bandes statt. Die Gegenstandskonzepte der Rezeptionsforschung werden hinsichtlich ihrer Überschneidungen, ihrer Unabhängigkeit und ihrer Ergänzungsfähigkeit untersucht. Dabei werden wirkungs-, kognitions- und handlungstheoretische Konzepte miteinander verglichen. Die gegenstandsbezogenen und wissenschaftstheoretischen Grundlinien einer interdisziplinären Rezeptionsforschung werden in einem *pragmatischen Erklärungsmodell* aufgezeigt und ermöglicht. Der Forschungsansatz der *British Cultural Studies* wird durch eine kritische Präsentation der wichtigsten Jugendstudien vorgestellt und im letzten Beitrag des ersten Teils wird eine interpretative Rezeptionsforschung differenziert betrachtet.

Mit diesem ersten Teil, vor allem mit den ersten beiden Beiträgen, werden (meta-)theoretische Überlegungen für eine mögliche interdisziplinäre Medienwissenschaft angestellt und überzeugend dargelegt. Die Systematik, die sie für die weiteren Beiträge liefern, ist nicht nur eine Hilfestellung für diesen Band, sondern stellt darüber hinaus Orientierungshilfen der zahlreichen Konzepte in der Rezeptionsforschung allgemein bereit. Sie bieten weiterhin eine mögliche Grundlage für eine kritisch-konstruktive Sichtweise, die zur Problematisierung und zum Erkenntniszuwachs in der Rezeptionsforschung beitragen kann.

Der zweite Teil des Bandes stellt einige Formen des Umgangs mit verschieden Medien und Gattungen vor. Das Modell des rationalen Rezipienten, welches die Nachrichtenforschung oft unterstellt, wird kritisch diskutiert. Ein weiterer Beitrag versucht zeitlich überdauernde Nutzungsmuster von verschiedenen Nutzertypen am Beispiel von Action-Filmen voneinander abzugrenzen. Den Abschluß dieses Teils bildet eine psychophysiologische Untersuchung von Schülern, deren Ergebnis unter anderem zeigt, daß Fernsehen im Vergleich zum Schulunterricht eine emotionale Beanspruchungen darstellt, die hinsichtlich Alter und der Viel/Wenig-Seher-Relation differiert.

Der dritte Teil beschäftigt sich mit Wahrnehmungslenkung im Film. Der erste Beitrag dieses Teils untersucht die Wahrnehmungslenkung in Kriegsfilmen, beispielhaft an *Platoon* und *Casualities of War*. Genretypische Merkmale werden herausgearbeitet und in Relation zum Zuschauer gesetzt.

Kinder als Nutzertyp und der Verlust von schützenden, dem Film eingeschriebenen Rahmen durch symbolische und intertextuelle Verweise der markwirtschaftlich ausgerichteten Kul-turindustrie sind Themen des zweiten Beitrags. *Angel Heart* ist der Film (-Text), der im Mittelpunkt des letzten Beitrags dieses Teils steht. Die Autoren arbeiten verschiedene narrative Elemente heraus, die hohe Verstehensleistungen des Rezipienten verlangen und nicht unbedingt koharänte Eindrucksbilder am Ende des Filmes anbieten und damit evtl. zum erneuten Sehen bzw. einer intensiven Folgekommunikation anregen.

Dem vierten Teil liegt als Schwerpunkt ein handlungstheoretischer Ansatz zugrunde. Er nimmt damit im gesamten Band eine dominierende Stellung ein, nicht zuletzt durch die Präferierung dieses Konzeptes der herausgebenden Autoren. Gegenstand der Untersuchungen sind die Anschlußgespräche von Jugendlichen an den Film *Angel Heart,* das Ritual Videoabend, die kognitiven und affektiven Prozesse des Zuschauers bei einem *spannenden Film,* die die Medienrezeption begleitenden emotionalen und motivationalen Prozesse (insbesondere *die Identifikation*) und am Beispiel von HipHoppern, wie Jugendliche kommunikativ über gemeinsame Rezeptionserfahrungen Formen symbolischer Kreativität entwickeln, die ihnen bei ihrer Identitätsbildung hilfreich sind.

Die Teile zwei bis vier zeigen, wie unterschiedlich die Forschungen im Bereich der Medien hinsichtlich ihrer Konzepte,Methoden,Begriffen und Fragestellungen sind.Gleichzeitig wird aber auch deutlich,wie ergänzend und fruchtbar interdisziplinäre Wissenschaft sein kann.

Lorenz Gräf und *Markus Krajewski* (Hrsg.):

Soziologie des Internet. Handeln im elektronischen Web-Werk,

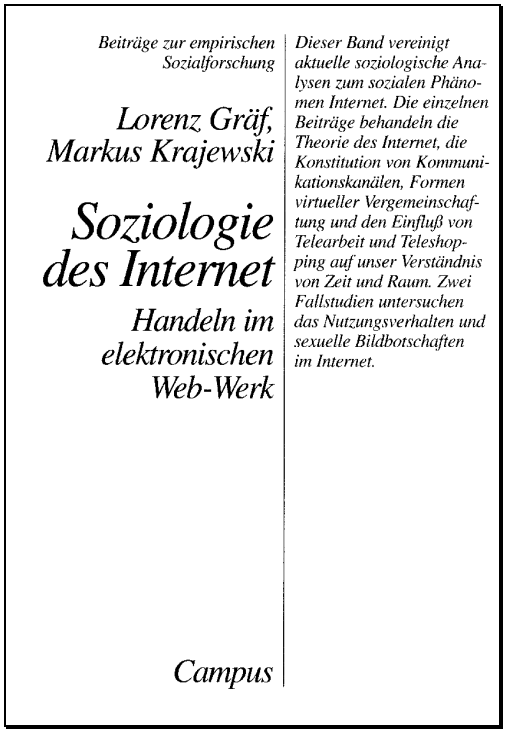
Campus, Frankfurt am Main/New York, 1997.

Preis: 38,- DM, ISBN 3-593-35773-9

Die in diesem Buch aus der Reihe *Beiträge zur empirischen Sozialforschung* versammelten Beiträge versuchen jeder auf seine Weise, eine Bestandsaufnahme dessen, was soziologische und kulturwissenschaftliche Analysen im Augenblick zum Verständnis des Internet beitragen können.

Die theoretische Annäherung an das Phänomen Internet erfolgt in Teil I zunächst über die Techniksoziologie. ***Martin Rost*** untersucht elektronische Netzwerke, indem er die soziale Relevanz von Technik anhand des Begriffs Protokoll hervorhebt. Die virtuelle Gesellschaft mit ihren Ausprägungen wird von ***Achim Bühl*** prognostiziert. ***Markus Krajewski*** spürt der Frage nach, wie ein Hypertext zu lesen ist. Im Anschluß an herkömmliche Lektüren schlägt er zwei differente Lesarten vor.

In Teil II reflektieren ***Matthias Bickenbach*** und ***Harun Maye*** die verschiedenen Metaphern, mit deren Hilfe Diskurse im und über das Netz geführt werden. ***Lorenz Gräf*** entwickelt eine Analogie zwischen dem klassischen Entstehungsprozeß interpersonaler Netzwerke und ihrer elektronischen Variante. Gegen jene Thesen, die vorschnell virtuelle Gemeinschaften im Internet ausmalen, hält ***Josef Wehner*** die Implikationen der nahezu ausschließlichen Kommunikationsart im Internet: der Schriftkultur.



Moderne Kommunikationsmedien verändern das Handeln im beruflichen wie auch im privaten Kontext. Teil III befaßt sich mit den beiden wichtigsten Anwendungsfeldern. ***Karl Kollmann*** beschreibt das soziale und wirtschaftliche Umfeld des privaten Konsums. ***Michael Schack*** fokussiert auf den organisationellen Hintergrund bzw. die soziale Dimension von Telearbeit. Er zeigt, warum soziale Arrangements neben Computernetze treten müssen, wenn durch den Einsatz von Telearbeit die Produktivität erhöht werden soll.

In Teil IV berichten ***Bernad Batinic***, ***Michael Bosnjak*** und ***Andreas Breiter*** von Merkmalen und Verhalten der "Internetler". Die Vorstellungen über Merkmale und Folgen der neuen Medien, die unter den Nutzern des Netzes kursieren, analysiert ***Rupert Schmutzer***.

Michael Schetsche schließlich berichtet von einer explorativen Studie über sexuelle Kommunikation im Internet. Bildmaterial, hier insbesondere mit verbotenen pornographischen Motiven, so das Ergebnis seiner Studie, ist sehr viel schwerer zu bekommen, als häufig in den Medien behauptet wird.

Muß ein Buch, das elektronische Netzwerke zum Thema hat, nicht auch im Internet selbst Präsenz zeigen? Nicht zuletzt um elektronische Kommunikation mit herkömmlicher zu vermengen, besitzt dieses konventionelle Druckwerk eine eigene Internet-Adresse: http://infosoc.uni-koeln.de/sozdesnet/. Hier befindet sich das Online-Forum, das Wider-, Zu- oder Einsprüche ermöglicht, um Diskussionen, neue Daten und Thesen beizutragen.

Thomas A. Wetzstein, Hermann Dahm, Linda Steinmetz, Anja Lentes, Stephan Schampaul und *Roland Eckert*: Datenreisende. Die Kultur der Computernetze. Opladen: Westdeutscher Verlag 1995, Preis: 54,00, ISBN: 3-531-12796-9

Ein Buchhinweis von *Frank Perschmann*

Von 1991 bis 1994 haben Mitarbeiter der„Arbeitsgemeinschaft sozialwissenschaftlicher Forschung und Weiterbildung an der Universität Trier e.V“zu Kommunikationsformen und Kommunikationsthemen in nichtkommerziellen Computernetzen(Mailboxen) gearbeitet

und Daten zu den Nutzern dieser Netze zusammengetragen. Die Autoren berichten von den Aktivitäten der Nutzer, analysieren Formen und Inhalte der Kommunikation in den Mailboxen, beschreiben unterschiedliche Aneignungsprozessen, benennen Formen der elektronischen Demokratie und liefern eine soziodemographische Beschreibung der 'Netzbewohner'. Für das Verständnis der Computernetze zum Ende dieses Jahrzehnts sind die vorgelegten empirischen Befunde nur von eingeschränktem Wert. Zu sehr unterscheidet sich das Internet 1997 von der Mailboxszene und dem Internet zu Beginn der 90er Jahre. Anderes gilt jedoch für einige Theorieaspekte. Die Aufarbeitung der metakommunikativen Aspekte des Netzaustauschs und die Befunde zu Aneignunsprozesse und den Voraussetzungen des Selbstlernens sind für das Verständnis des Internet in seiner heutigen Form gewinnbringend. Eine Anschaffung des Buches kann daher trotz teilweise veralteter Deskription empfohlen werden.

Rainer G. Haselier und Klaus Fahnenstich:
Word 7 für Windows 95. Textverarbeitung mit Windows 95. Addison-Wesley Publishing Company, Bonn Paris u.a. (1.Aufl.,), inkl. CD-ROM, 1995, 998 S., DM 79,90, ISBN 3-89319-978-0,
Ein Buchhinweis von *Bruno Hopp*

Das Autorenteam ***Haselier*** /***Fahnenstich*** hat sich zum wiederholten Male der Aufgabe gewidmet, ein komplexes Standardprogramm zu erklären. Für die mit Windows noch nicht vertrauten Einsteiger ist hier ebenso gesorgt wie für Experten, die auf zahlreiche Erfahrungen mit diesem oder einer anderen Software zurückblicken können. Auf den ersten Blick läßt ein Buch von rund eintausend Seiten zurückschrecken - wer soll das alles lesen? Dank guter Gliederungsfunktionen ist die Masse an Information übersichtlich: so wird man in einer Marginalienleiste am Rand per Symbol hingewiesen, wenn Funktionen der Programmversionen 2 oder 6 sich zur besprochenen Version 7 unterscheiden - wichtig für alle Umsteiger. Auch die unter Windows allgegenwärtigen bunten Icons zum Anklicken tauchen in der Marginalienleiste auf und machen gesuchte Tips schnell zugänglich.

Das Buch ist aufgeteilt in neun Abschnitte mit insgesamt 51 Kapiteln, zusätzlich bietet es einen guten Index. Neben den grundlegenden Auswahlmöglichkeiten während des Installationsvorganges und den wichtigsten Neuerungen für Umsteiger von Vorgängerversionen werden im einzelnen die Elemente des Word-Bildschirmes erklärt (Kap. 4), für Anfänger auch so „normale Dinge" wie das Erstellen, Speichern und Laden von Dokumenten (Kap. 6), Texteingabe inkl. der Funktionen der Zwischenablage und der Sammlung (Kap. 7). Grundlegende Formen der Textgestaltung (Kap. 8) durch Zeichenformatierungen, Tabellen,

Grafiken einfügen usw. dürften geübteren Textverarbeitern nicht neu sein (beliebtes Rätsel: was unterscheidet einen Absatz vom Abschnitt in Word?). Mit dem Kapitel über Formatvorlagen, ihrer Verwendung und Erstellung wird es recht interessant. Für wissenschaftliche Texte ist sicher die Beschreibung von „Numerierungen, Aufzählungen und Listen“ (Kap. 17) interessant. Ebenso sind dem Seitenlayout (inkl. Abschnitten und Spaltensatz), dem Umgang mit Kopf- und Fußzeilen, Seitennumerierung und Positionsrahmen eigene Kapitel gewidmet. Jedem Anwender sei empfohlen, das Kapitel über Dokumentvorlagen (Kap. 21) zu lesen - es dürfte informativer sein als das vergleichbare Kapitel im Originalhandbuch des Herstellers (bes. zu globalen und lokalen Vorlagen). Der Anpassung der Programmoberfläche an individuelle Benutzerbedürfnisse via Symbolleisten, Tastatur-Shortcuts und benutzerdefinierten Einstellungen im Menüpunkt „Extras - Optionen“ sind eigene Kapitel reserviert. Automatische Funktionen, etwa die unter Windows-95 Software zahlreicher gewordenen Assistenten, Autotexteinträge (in älteren Winword-Versionen „Textbausteine“ genannt), AutoFormatfunktionen und Autokorrekturmöglichkeiten werden ausführlich dargestellt. Neben den mit Word 7 mitgelieferten Zusatzprogrammen (Formel Editor, MS Graph2, Orgchart, MS Draw, Word Art) sind die Glanzlichter die Kapitel zu den Gliederungen und Zentraldokumenten sowie den Überarbeitungs- und Anmerkungsfunktionen (Kap. 48). Zentraldokumente bzw. das Anlegen von Filialdokumenten sind regelmäßig bei sehr umfangreichen Publikationen ratsam und die Überarbeitungsfunktion könnte problemlos auch die Überschrift „workgroup computing“ (d.h. im Netzwerk bzw. LAN) tragen. Die Darstellung dieser Kapitel ist gewohnt gründlich, selbst langjährige Praktiker können hier hilfreiche Tips finden.

Das Schlußkapitel zur Makroprogrammierung fällt eher mager bis enttäuschend aus - was Kenner der Materie kaum verwundern wird - dies ist zu komplex, als daß es erschöpfend auf knappen dreißig Seiten abgehandelt werden kann. Darum wird der Dialog-Editor ebenfalls nicht weiter erklärt, da er für „Normalanwender“ scheinbar ohne Funktion ist. Allein zu diesem Thema ließe sich leicht ein mehrere hundert Seiten umfassendes Buch zusammenstellen. Daran ändert dann auch die lexikonartige Kurzfassung gängiger Makrobefehle nichts, denn hier ist Fachwissen um sinnvolle und effiziente Programmierung nicht durch angelesenes Handbuchwissen ersetzbar.

Zusammenfassend kann das Buch von ***Haselier*** und ***Fahnenstich*** allen Anwendern empfohlen werden, die tägliche Schreibarbeit effizient mit Word 7 erledigen wollen. Die Autoren zielen auf eine sehr breite Leserschaft, die sich vom Neueinsteiger bis zum an älterer Software gestählten Anwender erstreckt. Durch diesen Anspruch wird viel Erklärungsbedarf (für Selbstverständliches, würden Profis einwenden) erzeugt, durch üppige Bebilderung (Screenshots) und Beispiele der Umfang des Buches aufbläht. Andererseits wollen die Autoren auch ambitionierten Anwendern noch Feinheiten aufzeigen, was ihnen auch gelingt. Der umgangssprachliche Stil der pädagogisch aufeinander aufbauenden Kapitel läßt keine Ängste vor „Fachchinesisch“ aufkommen. Leider ist die dem Buch mitgegebene CD-ROM

nur bedingt empfehlenswert: außer den im Buch erwähnten Beispielen enthält sie working models (also Testversionen) der Grafikprogramme „Corel Draw 6 für Windows 95“ und der „MicrografX ABC Graphics Suite für Windows 95“ (Laufzeit: 90 Tage). Eine kleine Auswahl recht guter Zusatzprogramme für Winword aus dem Sharewaremarkt, etwa einige Adreßverwaltungen oder spezialisierte Makro- und Vorlagensammlungen hätte den Nutzen der CD für die Leserschaft noch weiter erhöhen können. Alles in allem: lesenswert, auch in Auszügen.

Herbert Schubert:

Anforderungen von Migranten an Wohnungen und Gewerbestandorte: Marktstudie für das Projekt Internationales Wohnen und Gewerbe am Kronsberg.

In der Nähe des Messegeländes und späteren EXPO-Geländes von Hannover entsteht der neue Stadtteil Kronsberg. Dort wird das Projekt "Internationales Wohnen und Gewerbe am Kronsberg" geplant. Neben einer hohen Versorgungsqualität für den neuen Stadtteil soll es zugleich die räumliche Integration von Migranten zum Ziel haben. Das Modell ist innovativ, weil es mit einem Konzept des "Internationalen Wohnens und Arbeitens" einen grundsätzlich neuen Beitrag zum Städtebau leisten will.

Die Vorstellung des Internationalen Wohnens geht von der Neudefinition des Integrationsbegriffs aus. Die alte Vorstellung, daß Integration über Mischungsproportionen erreicht wird, nach denen ein bestimmter Anteil von Migrantenhaushalten als Schwelle nicht überschritten werden darf, ist mit Blick auf Differenzierung der Migrantenbevölkerung nicht mehr hinreichend. Statt dessen ist ein pluralistischer Integrationsbegriff in praktische Wohnmodelle umzusetzen. Kennzeichen des pluralistischen Verständnisses ist es, daß eine bunte Mischung verschiedener Nationalitäten und Ethnien einen anerkannten Teil der ebenso bunt gemischten deutschen Lebensstile darstellt, ohne daß einzelne Bewohnergruppen ihre Einstimmigkeit aufgeben müssen.

Zur Realisierung des innovativen Ansatzes sind Marktstudien erforderlich, die die Anforderungen von Migranten an das Wohnen und an Räumlichkeiten für Gewerbe untersuchen. Das Institut für Entwicklungsplanung und Strukturforschung an der Universität Hannover hat im Frühjahr des letzten Jahres eine Umfrage bei Migrantenhaushalten nach ihren Wohnbedürfnissen und Wohnanforderungen durchgeführt. Dazu wurde jetzt der oben zitierte Beitrag von ***Herbert Schubert*** im IES-Bericht 203.96 vorgelegt.

Institut für Entwicklungsplanung u. Strukturforschung GmbH an der Universität Hannover
Lister Straße 15, 30163 Hannover, Tel. 0511 39970