Das Sozio-Ökonomische Panel

Seminar
im Sommersemester 2013

Susanne Seidel

20. Juni 2013
Institut für Statistik, LMU

Inhaltsverzeichnis

1 Einführung

2 Panelforschung

Neben Untersuchungen, die auf den momentanen Zustand in der Bevölkerung zielen, interessiert man sich auch häufig für langfristige Entwicklungen und Veränderungen im Zeitverlauf, die durch wiederholte Untersuchung der selben Eigenschaften und dem anschließenden Vergleich dieser ermittelt werden können. Für wiederkehrende Befragungen bei einem gleichbleibenden Kreis von Auskunftsobjekten eignet sich das Forschungsdesign der Panelstudie, welche im Folgenden hinsichtlich ihrer Eigenschaften, Vor- und Nachteile und ihrem Auftreten in der Praxis betrachtet werden soll.

2.1 Definition und Abgrenzung

Gerade in der empirischen Sozialforschung ist ein zentraler Punkt anhand von gesammelten Daten soziale Sachverhalte zu beschreiben und daraus aussagekräftige Schlüsse auf die Meinungs- und Verhaltenszusammenhänge in der Bevölkerung zu ziehen. Um Daten zu erhalten, die das zu untersuchende Merkmal gut beschreiben, wird die Bevölkerung befragt. Anhand von Stichproben (z.B. Zufallsstichproben) werden repräsentative[1] Daten gewonnen, da sie die Gruppenverhältnisse in Grundgesamtheit gut abdecken. Je nach dem Zweck, der verfolgt wird und welcher zeitliche Modus gebraucht wird, kann zwischen drei Erhebungsdesigns gewählt werden: Querschnitt-, Trend und Paneldesign. Bei einem Querschnittdesign handelt es sich um eine einmalige Erhebung der Variablen bei N Untersuchungseinheiten, wohingegen Trend- und Paneldesigns wiederholte Datenerhebungen vorsehen. Die Trenderhebung erfolgt durch eine wiederholte Abfrage derselben Variablen zu unterschiedlichen Zeitpunkten bei jeweils neu gezogenen Auskunftspersonen. Bei einem Panel wiederum handelt es sich um einen gleichbleibenden, repräsentativen Kreis von Untersuchungseinheiten, bei dem wiederholt in regelmäßigen Abständen dieselben Variablen erfragt oder beobachtet werden (siehe Diekmann 2011 S.304-312). Die einzelnen Erhebungen eines Panels werden als Panelwellen bezeichnet. Bei den Erhebungsmethoden unterscheiden man im Wesentlichen zwischen dem Interview, das entweder telefonisch oder bei direkter Anwesenheit (Face-to-Face) durchgeführt werden kann, der postalischen oder Online Befragung und der Beobachtung.

2.2 Vor- und Nachteile eines Panels

Im Vergleich zu den anderen erwähnten Designs hat das Paneldesign den höchsten Informationsgehalt, da sich aus ihm sowohl Querschnittdaten als auch Trenddaten ableiten lassen (wobei der Umkehrschluss nicht gilt!)(Diekmann 2011 S.306). Die Querschnitterhebung liefert nur einen einmaligen Datenbestand zu nur einem einzigen Zeitpunkt.

[1] Repräsentativität: Von Repräsentativität wird gesprochen, wenn sich aus einer Stichprobe zutreffende Rückschlüsse auf eine Grundgesamtheit ziehen lassen. Im engeren Sinne ist eine Stichprobe dann repräsentativ, wenn alle Merkmalsträger der Grundgesamtheit die gleiche Chance besessen haben, Teil dieser Stichprobe zu werden.

Informationsreicher ist zwar die Trenderhebung, die wiederholte Erhebungen bei immer aktuellen Zufallsstichproben produziert. Dennoch ist eine Panelerhebung am produktivsten, da es hier sogar die Möglichkeit gibt einen Menschen „von der Wiege bis zur Bahre“ zu begleiten. Außerdem entsteht eine geringere Gefahr Messungenauigkeiten zu erhalten, wie z.B. Erinnerungsfehler, da sich im Gegensatz zu Querschnitterhebungen die Retrospektiv-Fragen maximal auf das vergangene Jahr beziehen. Auch subjektive Einstellungen werden jeweils für den Erhebungszeitpunkt erhoben und nicht retrospektiv für vergangene Jahre erfragt. Unter anderem dadurch wird die Gefahr von Störvariablen zwischen den Messungen verringert. Bei einer Panelbefragung ist die Trennung von Alters- und Kohorteneffekten[2] möglich, da beispielsweise für jedes Individuum eine Art Lohnprofil im Verlauf des Panels erstellt werden kann. Einer der wichtigsten Vorteile der Panelerhebung ist die ermöglichte Analyse der Veränderungsdynamik auf individueller Ebene, wobei hier Vorsicht geboten ist bei Abschätzung von Trends bei Vorliegen von nur zwei Erhebungswellen. Umso mehr Wellen vorhanden sind, desto sicherer ist die Trendabschätzung (siehe für Details Rendtel 1995).

Allerdings ist zu beachten, dass bei einem Paneldesign nur die erste Welle eine Zufallsstichprobe ist, jedoch nicht jede weitere. Die Bevölkerung kann bei späteren Wellen nur noch durch Gewichtung repräsentativ abgebildet werden, im Gegensatz zur Trenderhebung, bei der jeweils aktuell repräsentative Zufallsstichproben gezogen werden. Problematisch ist bei einer Panelerhebung die Panelmortalität, im Prinzip die Schwundquote. Die Originalstichprobe schrumpft durch Nichterreichbarkeit der Untersuchungseinheiten aufgrund von Tod, Wegzug, Verweigerung etc. von Welle zu Welle. Hier ist die Panelpflege, insbesondere die Adressenpflege besonders wichtig (siehe Kapitel 4.2.2). Da solche Ausfälle durchaus auch systematisch auftreten, da beispielsweise berufstätige Personen schwerer zu erreichen sind und junge Menschen häufiger den Wohnort wechseln, treten systematische Stichprobenfehler aufgrund von Unterrepräsentation solcher Gruppen auf. Zusätzlich bedarf es einiges an Wartezeit bis Generationenanalysen möglich sind, da hierfür eine Panellaufzeit von mindestens 25 Jahren notwendig sein wird. Trotzdem ist die Panelforschung ein wichtiger Teil um Verhaltens- und Zusammenhangsanalysen in der Gesellschaft zu erstellen (siehe für Details Rendtel 1995).

2.3 Anwendungsgebiete und Bedeutung für die Praxis

Der Anwendungsbereich von Panels ist sehr umfangreich. Je nach Themenbereich und zeitlichem Umfang dieser Erhebung sind die unterschiedlichsten Studien möglich: Die Entwicklung von Schulabgängern (z.B. die vom Hochschul-Informations-System HIS durchgeführten Befragungen), das Verhalten von Strafentlassenen, Entwicklung des Konjunktur- und Geschäftsklimas, Konsumentenpanels und viele mehr (siehe Rendtel 1995). Panel-

[2] Eine Kohorte bezeichnet eine Bevölkerungsgruppe, die durch ein gemeinsames, längerfristig prägendes Startereignis definiert ist. Kohorteneffekte bezeichnen Unterschiede, die zwischen verschiedenen Kohorten bestehen und sich somit auf das Vorhandensein unterschiedlicher sozialer und umweltbedingter Einflüsse zurückführen lassen.
Bei Alterseffekten ändert sich durch das Altern ein Merkmal eines Individuums(siehe Diekmann 2011).

studien können experimentell ausgerichtet sein, wie in den Forschungsbereichen der Biometrie, in der die Wirkung unterschiedlicher Behandlungsmethoden auf die Probanden über die Zeit gemessen wird, oder aber auch explorativen Charakter haben, wie bei Haushaltspanels zur Erforschung von individuellen Armutsphasen (siehe Rendtel 1995). Im internationalen Vergleich ist das Sozio-ökonomische Panel mittlerweile eine sehr fortgeschrittene Studie. Die amerikanische Studie „Panel Study of Income Dynamics“ (PSID) startete 1986 mit ca. 18000 Teilnehmern und galt als Vorbild für das SOEP, fragte aber nur nach den ökonomischen Verhältnissen der Amerikaner. Das SOEP als älteste Längsschnittstudie privater Haushalte in Deutschland fragte von Anfang an auch nach dem Verhalten und der Zufriedenheit der Teilnehmer (siehe Götz 2008).

3 Das Sozio-Ökonomische Panel (SOEP)

Die Längsschnittstudie Sozio-ökonomisches Panel (SOEP) liefert seit mittlerweile 29 Jahren jährlich Mikrodaten[3] für die Grundlagenforschung und Politikberatung im Bereich der Sozial-, Wirtschafts- und Verhaltenswissenschaften. Dieses Kapitel beschäftigt sich mit den inhaltlichen Schwerpunkten, dem geschichtlichen Hintergrund des SOEP, seine Ziele und Funktion im CNEF, seine Anwendungsmöglichkeiten und dem Service, der rund um das SOEP geboten wird.

3.1 Eigenschaften

Seit 1984 erfasst das Sozio-Ökonomische Panel jährlich ausgewählte private Haushalte und alle darin lebenden Personen ab 17 Jahren zur Beobachtung ihrer Zufriedenheit und Lebenssituation. Der Grobaufbau des Erhebungsprogramms des SOEP ist gegliedert in einen Kernbereich an Variablen, der weitgehend unverändert in jeder Welle abgefragt wird und einem wechselndem Befragungsschwerpunkt, der nur alle paar Jahre wiederholt wird. Der Kern einer Erhebung umfasst Fragen aus diesen Bereichen (siehe Wagner, Göbel, Krause, Pischner, Sieber 2008 S. 305):

- Demographie und Wohnsituation
- Persönlichkeitsmerkmale und Grundorientierung (Präferenzen, Werte, usw.)
- Vorschul- und Schulbildung, berufliche Bildung und Weiterbildung; Qualifikation
- Arbeitsmarkt- und Berufsmobilität
- Einkommen, Vermögen und soziale Sicherheit
- Gesundheit
- Sorgen und Zufriedenheit (allgemeine Lebenszufriedenheit und Bereichszufriedenheiten)

[3]Mikrodaten sind Originaldaten, die sich auf Individuen beziehen (im Gegensatz zu Makrodaten/Aggregierte Daten, die sich auf Gruppen beziehen)

Zum variierenden Bereich gehören biografische Hintergrundinformationen, soziale Herkunft, Nachbarschaft, soziale Sicherung, Vermögen, Weiterbildung, Zeitverwendung, Familie, Arbeitsbedingungen, Zukunftserwartungen und Umwelt (siehe Wagner, Göbel, Krause, Pischner, Sieber 2008 S. 305). Besonderheiten zu den einzelnen Stichproben und Fragebögen folgen in Kapitel 4.1 auf Seite 10f. Das SOEP nutzt verschiedene Zeitbezüge der Fragen um Veränderungen im Lebenswandel auf unterschiedliche Weise erfassen zu können. Neben den zeitpunktbezogenen Fragen, werden zum einen einige Sachverhalte, wie die Anzahl der Arbeitsplätze im letzten Jahr, retrospektiv erfasst und zum anderen auch Fragen bezüglich zukünftiger Erwartungen (beruflich und privat) gestellt. Um einen gewissen Qualitätsstandard bei der Datenerhebung garantieren zu können, müssen bestimmte panelspezifische Kriterien erfüllt werden (siehe 25 Wellen SOEP S.146): Ein Faktor ist die „Vollständigkeitsquote“, die kennzeichnet in wieweit es von Welle zu Welle gelingt alle Mitglieder (ab 17 Jahren) eines teilnehmenden Haushaltes individuell zu befragen. Des Weiteren muss die „Panelstabilität“ im Zeitverlauf gesichert sein. Hier wird wieder von Welle zu Welle berechnet, welche Relation zwischen der realisierten Fallzahl des laufenden Jahres und der des Vorjahres besteht. Es ist sehr wichtig hier eine gute Quote zu erhalten, da bei einer zu hohen Anzahl an Ausfällen, die Startstichprobe viel zu schnell schrumpft und somit irgendwann nicht mehr genug beobachtbare Fälle vorhanden sind um aussagekräftige Analysen zu erhalten. Hier ist die Panelpflege besonders wichtig, die in Kapitel 4.2.2 auf Seite 14 genauer betrachtet wird.

Das SOEP verfolgt zwei Ziele: Zum einen versucht es eine gute Balance zwischen Innovation (z.B. Verbesserung einiger Erhebungsinstrumente und Erweiterungen von Stichproben) und der Sorge um den Bestandserhalt (Risiko, Teilnehmer zu verlieren, aufgrund zu umfangreicher Befragung) zu halten. Zum anderen bemüht sich das SOEP um eine Verzahnung zwischen Service und Forschung. Die Aufbereitung der Daten zu einem bestimmten Bereich, die Mitwirkung am Erhebungsdesign und die Forschung mit den erhobenen Daten werden komplett denselben Personen zugeordet. Dadurch entwickle sich ein forschungsgetriebenes Interesse an hoher Qualität und Aktualität der Daten (siehe 25 Wellen SOEP S.10). All diese Bestandteile haben sich erst nach vielen Jahren der Planung und Verhandlung entwickelt, die letztlich aber dazu führen, dass das SOEP mittlerweile auch internationales Vorbild ist.

3.2 Entstehungsgeschichte

Um so ein großes Projekt wie das Sozio-ökonomische Panel ins Rollen zu bringen, bedarf es einiges an Vorarbeit und Geduld, um Probleme zu erkennen und zu bewältigen. Bei den Überlegungen für die Analyse von Einkommensdaten versuchte man vorerst die Längsschnittanalyse zu umgehen, da kaum Erfahrungen auf dem Gebiet vorhanden waren und die Möglichkeit bestand Paneldaten künstlich durch eine Sequenz von unabhängigen Querschnitten zu erzeugen. Nachdem aber klar wurde, dass zur Prüfung kausaler Hypothesen kein Weg an einer Längsschnittstudie vorbei führte, begann man sich mit der Problematik auseinanderzusetzen. 1971 startete dafür die Zusammenarbeit mit Ökonomen und Soziologen. Infolge der Datenschutzdiskussion hatte sich der wissenschaftliche Zugriff auf amtliche Daten in solchem Ausmaß verschlechtert, dass es

aussichtsreicher war selber Daten zu erheben. Dies konnte mit der Beantragung des Sonderforschungsbereichs 3 "Mikroanalytische Grundlagen der Gesellschaftspolitik"(Sfb 3) und seiner Förderung durch die Deutsche Forschungsgemeinschaft (DFG) umgesetzt werden. Nachdem ein erster Forschungsantrag noch zu unausgereift war und einige Problematiken (z.B. Finanzierung) noch zu überarbeiten waren, begann im Januar 1979 die Vorbereitungsarbeiten für das Haushaltspanel des Sfb 3. Parallel werden in den USA Eindrücke von dort bestehenden Panelprojekten gesammelt und die hohen wissenschaftlichen Maßstäbe übernommen und in Deutschland anschließend umgesetzt. Im August 1982 wurde der überarbeitete Antrag zur Förderung der Forschungsphase vorgelegt.

„Der Entscheidungsprozess war schwierig. Aber wir brauchten alle einen Lernprozess. Wir, aber auch die Gutachter. Denn das was wir machten war ja neu und revolutionär. Wir erhoben Längsschnittdaten - damit hatten wir keine Erfahrung -, wir wollten soziale und ökonomische Faktoren miteinander verbinden und wir wollten Ausländer befragen - das war neu in Deutschland. Und wir wollten viel Geld!“ (Zitat Prof. Dr. Hans Jürgen Krupp, Gründer und Leiter von 1983-88, SOEP-Film: „30 Jahre Leben in Deutschland – die Geschichte des SOEP“)

Nachdem man sich auf eine rein wissenschaftsbezogene (und keine private) Finanzierung geeinigt hatte, erfolgte die Entscheidung am 6. Dezember 1982. Die Bewilligung umfasste die Förderung für vorläufig 5 Umfragewellen, aber auch einige Auflagen und Empfehlungen. Diese beinhalten (siehe 25 Wellen SOEP S.22):

- ein wissenschaftlicher Beirat, der sich grundsätzlich an den Interessen des SFBs, nicht an denen der potentiellen Nutzer der Paneldaten orientieren soll (z.B. Politik),
- eine gute Dokumentation,
- klare Datenschutzregelungen für die Weitergabe an andere Wissenschaftler,
- eine Übergewichtung der Ausländerbevölkerung,
- die Bearbeitung methodischer Probleme, wie die Überprüfung von Retrospektivdaten,
- Forschung zur Gewährleistung der Repräsentativität,
- intensive Zusammenarbeit mit dem Umfrageinstitut.

1984 konnte dann die erste Welle erhoben werden. Damals wurden 5921 Haushalte ausgewählt, in denen 12245 persönlich befragte Erwachsene und 3928 Kinder lebten. In den darauf folgenden Jahren wurden praktische Erfahrungen gesammelt und um die Weiterfinanzierung und somit um das Überleben des Panels gekämpft. Dennoch kann dieses Jahr, trotz der damaligen Planung von nur 5 Wellen, die 30. Erhebungswelle realisiert werden.

3.3 Einbettung in das CNEF

Für die Realisation einer international vergleichenden Auswertung sozial- und wirtschaftswissenschaftlicher Daten wurde an der Cornell University in den USA das „Cross National Equivalent File“ (CNEF) ins Leben gerufen. Das CNEF stellt die Mikrodaten mit einheitlichen und benutzerfreundlich definierten Variablen aus den verschiedenen Studien zusammen. Zu den mittlerweile acht teilnehmenden Ländern[4] mit ihren jeweiligen Längsschnittstudien gehören die US-amerikanische Panel Study of Income Dynamics (PSID), das Sozio-oekonomische Panel (SOEP) aus Deutschland, der British Household Panel Survey (BHPS) aus Großbritannien, der Survey of Labour and Income Dynamics (SLID) aus Kanada, der Household Income and Labour Dynamics Survey (HILDA) aus Australien, das Schweizer Haushalt-Panel (SHP), die Korea Labor Income Panel Study (KILPS) sowie der Russia Longitudinal Monitoring Survey (RLMS-HSE) aus Russland. In den jeweiligen Datensätzen sind unter anderem Informationen über die Haushaltsstruktur, das Einkommen, die Erwerbstätigkeit und die demografischen Merkmale enthalten. Ziel des CNEF ist es einen vereinfachten Zugang zu international vergleichenden Daten zu bieten und weiterhin Richtlinien für die Definition vereinheitlichender Variablen zu liefern. Dadurch, dass das SOEP (nach dem PSID) eine Art Vorbildfunktion für die folgenden Panelstudien war, ist es durch die Teilnahme der hinzugekommenen Länder mit ihren Sudien ein weiterer Fortschritt für die Wissenschaft nun auch länderübergeifend forschen zu können (siehe für Details 25 Wellen SOEP S.110-125).

3.4 Relevanz des Panels

Das SOEP mit seinen Mikrodaten ist in vielerlei Hinsicht unabdingbar für die heutige Forschung und Politikberatung geworden. Neben der eben genannten Funktion im CNEF, ermöglicht es empirische Analysen für die Sozial- und Wirtschaftswissenschaften sowie die Verhaltenswissenschaft und ist zudem auch für die Geografie, Umwelt- und Energiewissenschaften nützlich. Die hier betrachteten Haushalte, Lebensläufe und das daraus resultierende Verhalten und die Zufriedenheit bieten nicht nur ökonomische Aspekte, die betrachtet werden wollen, sondern werden auch aus psychologischer und soziologischer Sicht betrachtet. Dadurch hat das SOEP einen stärkeren multidisziplinären Charakter als die ursprüngliche Vorbildstudie aus den USA (PSID). Viele Publikationen (siehe Wagner, Göbel, Krause, Pischner, Sieber 2008 S. 303), die auf Basis des SOEP-Datensatzes entstanden sind, beschäftigen sich mit Familien-Dynamik, Intergenerationaler Mobilität, Übergang in den Ruhestand, Anpassung von Zufriedenheit, Migration und noch vielen anderen Themen. Hierbei wird auch deutlich das das SOEP nicht nur ein Haushaltspanel ist sondern auch eine Kohortenstudie (siehe Wagner, Göbel, Krause, Pischner, Sieber 2008 S. 303).

[4] http://www.human.cornell.edu/pam/research/centers-programs/german-panel/cnef.cfm)

3.5 SOEPservice

Da der ursprüngliche Gedanke war, die Datengrundlage für die Sozialwissenschaften zu verbessern wurde mit der Zeit die Dienstleistung des SOEPservice immer weiter ausgebaut, um die Daten unter der Einhaltung der Datenschutzregeln an andere Wissenschaftler weitergeben zu können. Der SOEPservice ist somit ein nicht-wissenschaftlicher Service für Nutzerinnen und Nutzer der SOEP-Mikrodaten. Er ist zuständig für die allgemeinen Formalitäten und Dokumentationen. Im Speziellen bedeutet dies (siehe 25 Wellen SOEP S.131):

- Organisation der SOEPhotline als zentrale Anlaufstelle für alle Fragen rund um die Nutzung der SOEP-Daten, insbesondere zum Vertrags- und Datenmanagement, sowie die Vermittlung von AnsprechpartnerInnen bei speziellen Anwendungsfragen
- Bereitstellung möglichst aller Informationen, von Variablenlabels bis zu Dokumentationen und Webseiten, auch oder ausschließlich in englischer Sprache
- Abschluss und Verwaltung der Datenweitergabeverträge
- Bereitstellung der anonymisierten Mikrodaten auf DVD, Bestellwesen und Versand
- Suche und Erfassung von Belegstücken für SOEPlit
- Pflege und permanente Weiterentwicklung der Homepage
- Herausgabe der SOEPpapers
- Organisation von Nutzerschulungen
- Organisation der Internationalen Nutzerkonferenz im 2-Jahres-Rhythmus
- Öffentlichkeitsarbeit durch Pressemitteilungen und Journalistenbetreuung

Die hier bereits erwähnte Dienstleistung SOEPpapers ist eine Plattform für die zentralen und weltweiten Veröffentlichungen von Forschungsergebnissen, die auf den SOEP-Daten basieren. Auch das SOEPlit ist eine Ansammlung von Berichten, die sich allerdings auf die Struktur und Methodik des Surveys beziehen (siehe 25 Wellen SOEP S.131). Zusätzlich zum SOEPservice wurde das SOEPinfo ins Leben gerufen, dass im Allgemeinen eine Übersicht über die gesammelten Informationen, die Itemkorrespondezlisten, die die korrespondierenden Variablen über die Zeit abbilden sowie interaktiv erstellbare Syntax-Codes für die Statistik-Pakete SAS, SPSS und Stata zur Verfügung stellt (siehe 25 Wellen SOEP S.103). Erwähnenswert ist auch SOEPcampus, das bereits für die Statistik-Ausbildung gedacht ist. Hierbei soll eine Schulung und Förderung des Umgangs mit den SOEP-Daten ermöglicht werden. Dies sind alles Leistungen, die nicht selbstverständlich sind aber dennoch auf internationaler Ebene erwartet werden. All diese Bereiche sollen auch noch in Zukunft weiter ausgebaut werden, um eine möglichst gute Nutzung mit den Daten zu ermöglichen.

4 Das Panel

Im folgenden Kapitel soll ein Überblick über die Struktur der Daten des Panels verschafft werden. Nach der Übersicht über die verschiedenen Stichproben, Fragebögen und der allgemeinen Stabilität des Datenbestandes, betrachten wir den Ablauf der Feldphase, durchgeführt durch TNS Infratest Sozialforschung in München und die notwendige Panelpflege. Anschließend erfolgt die Prüfung und Aufbereitung der Daten im DIW Berlin. Zum Ende des Kapitels soll noch auf die Datenstruktur und auftretende Schwierigkeiten eingegangen werden.

4.1 Datengrundlage

4.1.1 Stichproben

Das Sozio-ökonomische Panel hatte nicht von Anfang an den vollen Umfang, den es heute besitzt. Gestartet mit nur einer Stichprobe, umfasst das Panel nunmehr 9 Stichproben die seit Beginn 1984 ins Leben gerufen wurden. Dem Datennutzer wird aber nur eine Gesamtstichprobe der jeweiligen Befragungswelle zur Verfügung gestellt, obwohl es sich eher um ein Komplex aus Teilstichproben handelt. Eine Übersicht über die Stichproben A-H erfolgt in Tabelle 1.

Tabelle 1: Die Stichproben des Sozio-ökonomischen Panels

Stichprobe A	(Deutsche)[1]Haushalte[2]in Westdeutschland	(Start:1984)
Stichprobe B	Ausländische Haushalte[3]in Westdeutschland	(Start:1984)
Stichprobe C	Privathaushalte in der DDR	(Start:1990)
Stichprobe D	Zuwanderer-Privathaushalte in Deutschland	(Start:1994/95)
Stichprobe E	Ergänzungsstichprobe	(Start:1998)
Stichprobe F	Ergänzungsstichprobe	(Start:2000)
Stichprobe G	Hocheinkommens-Privathaushalte in Deutschland	(Start:2002)
Stichprobe H	Ergänzungsstichprobe	(Start:2006)
Stichprobe I	Innovationsstichprobe	(Start:2009)

[1] Genauer: Haushalte, deren Haushaltsvorstand zum Zeitpunkt der Ziehung nicht türkischer, italienischer, jugoslawischer, griechischer oder spanischer Nationalität war.(war zu 99% der Fall).

[2] Anstaltshaushalte sind bei der Stichprobenziehung nicht eingeschlossen; sie werden zwar auch nicht ausgeschlossen, wenn sie beim Random-Walk gelistet werden und sind insofern im Bruttobestand enthalten, werden aber bei Durchführung der Befragung in der Regel von den Interviewern bei neuen Samples nicht berücksichtigt. Anstaltshaushalte werden in der Regel erst bei der weiteren Befragung durch Weiterverfolgung per Interviewer erfasst (Umzug ins Altersheim etc.); die in den Folgewellen einbezogene Anstaltspopulation ist aber nicht repräsentativ für die Grundgesamtheit.

[3] Genauer: Haushalte deren Haushaltsvorstand zum Zeitpunkt der Ziehung türkischer, italienischer, jugoslawischer, griechischer oder spanischer Nationalität war.

Quelle: 25 Wellen SOEP S.86

Hierbei ist anzumerken, dass die jeweiligen Stichproben jeweils auf Haushalts- und Per-

sonenebene realisiert werden. Als Panelteilnehmer auf Haushaltsebene gelten Haushalte, für die der HAUSHALTS-Fragebogen und mindestens ein PERSONEN-Fragebogen auswertbar vorliegen. Dies ist die Mindestbedingung (siehe Methodenbericht 2010 S.57). Allerdings sollen darüber hinaus möglichst alle Teilnehmer des Haushaltes ab 17 Jahren befragt werden, um fehlerhafte Angaben, die bei Proxy-Interviews[5] entstehen würden, zu vermeiden.

Gesondert zu betrachten ist die jüngste Stichprobe, die Innovationsstichprobe I (gegründet: 2009), die einige Besonderheiten aufweist (siehe Methodenbericht 2010 S.50 ff.): Neben der ursprünglichen Aufgabe, der Panelmortalität entgegenzuwirken, wird diese Stichprobe genutzt um innovative Survey-Methoden zu testen. Zum einen wurden verschiedene Incentivierungsmaßnahmen[6](siehe Kapitel 4.2.2) in Bezug auf die Wirkung auf die Teilnahmebereitschaft erprobt. Genau wie in Stichprobe F kommt ein „Oversampling“[7] von nicht-deutschen Haushalten zum Einsatz, für das ein onomastisches Verfahren[8] zur Repräsentation ausländischer Haushalte genutzt wurde. Letztlich erhält man auch durch eine zusätzliche „Non-Response-Nacherhebung“[9] und einer verbesserten Wohnumfelddokumentation weitere Informationen über die Gruppe der nicht teilnehmenden Haushalte.

4.1.2 Fragebögen

Für die Befragung der Teilnehmer ist ein Mindestalter von 17 Jahren festgelegt. Um jedoch Informationen über die Zeit davor zu erhalten, werden Fragebogen konstruiert, die sich an die Eltern des Betroffen richten. Das ganze Fragebogenkonstrukt hat sich mittlerweile soweit ausgedehnt, dass es nun möglich ist, Informationen „von (der Zeit vor) der Wiege bis zur (Zeit nach der) Bahre “ zu gewinnen. Von dem Vorstand eines Haushaltes, meist der Mann, werden bei dem HAUSHALTS-Fragebogen Merkmale über den Haushalt erhoben. Zusätzlich wird, wie schon erwähnt, allen zusätzlichen und befragbaren Haushaltsmitglieder der PERSONEN-Fragebogen vorgelegt. Zusätzlich werden bei Bedarf spezielle Fragebögen je nach Altersgruppe und betreffender Situation durchgearbeitet.

Wie bereits in Kapitel 3.1 erwähnt, behandelt der HAUSHALTS-Fragebogen Themen wie Wohnen und Wohnkosten, Hilfe- und Pflegebedürftige im Haushalt, Putz- und Haushaltshilfen und als wichtigstes Merkmal Einnahmen und Ausgaben. Der PERSONEN-Fragebogen, ist nicht jedes Jahr gleich, sondern behandelt bestimmte Themen nur alle paar Jahre. Zur Auswahl stehen grundsätzlich Zufriedenheit in verschiedenen Lebensbereichen, Erwerbstätigkeit, Einkommen, gesundheit und Krankheit, Spenden und Geldanlagen, emotionale Stabilität, politische Beteiligung, Herkunft und familiäre Situation und Leistungen privater Unterstützung. Zu den ZUSATZ-Fragebögen, die nur für spezi-

[5]Interview bei dem Dritte über die eigentliche Zielperson befragt werden. Hier: Auskunft von einem Haushaltsmitglied über ein anderes.
[6]Verschiedenen Varianten der Sondervergütung
[7]Oversampling = Überproportionale Vertretung
[8]Onomastik-Verfahren: Verfahren zur sprachlichen Analyse von Personennamen
[9]Non-Response-Nacherhebung = Nacherhebung bei Nichterreichbarkeit im Befragungszeitraum

elle Altersgruppen geeignet sind (siehe Methodenbericht 2010 S.25), gehören:

- LEBENSLAUF-Fragebogen
- JUGEND-Fragebogen
- DJ-Fragebogen („Lust auf DJ“)
- MuKi A („Mutter und Kind“)
- MuKi B („Ihr Kind im Alter von 2 bis 3 Jahren“)
- MuKi C („Ihr Kind im Alter von 5 bis 6 Jahren“)
- ELTERN-Fragebogen („Ihr Kind im Alter von 7 bis 8 Jahren“)
- LÜCKE-Fragebogen („Personenfragebogen 2009 - Nachbefragung Kurzfassung“)
- Zusatzfragebogen „Die verstorbene Person“

Bei dem Fragebogen „Lust auf DJ“ (Denksport und Jugend) handelt es sich um einen kognitiven Test, der den Entwicklungsstand des Jugendlichen erfassen soll. Der LÜCKE-Fragebogen soll den Längsschnitt vervollständigen, falls eine Befragungsperson in einer Welle nicht teilnehmen konnte und somit fehlende Daten entstanden sind. Hier werden hauptsächlich Eckdaten zum Erwerbsverlauf nachgetragen. Die folgende Tabelle 2 soll einen Überblick über die Erhebungsinstrumente und über die Häufigkeit der Anwendung liefern.

Auch hier erkennt man wieder die Komplexität des Panels und seine Relevanz nicht nur im ökonomischen sondern auch im sozialen bzw. verhaltensbezogenen Bereich.

4.1.3 Bestandsentwicklung

Die Personen aus einem für die erste Welle gezogenen Haushalts bilden die sogenannten Stammpersonen oder „Original Sample Members “ (OSM). Natürlicherweise kommt es vor, dass einzelne Personen aus diesem Haushalt ausziehen (z.B. Auszug aus dem elterlichen Haushalt bei jungen Personen, Auszug eines Partners bei Scheidungsfällen) und eventuell einen neuen Haushalt bilden oder in einen bestehenden Haushalt einziehen. Diese Tatsache wird genutzt um neue Teilnehmer für das Panel zu gewinnen, indem die Ausgezogenen nachverfolgt werden und die neuen Haushaltsmitglieder dieser Person mit in das Panel einbezogen und als „Non Original Sample Members “ (NOSM) bezeichnet werden. Durch diese Methode entsteht eine Art Schneeballstichprobe, die dazu dient die OSM besser im Kontext betrachten zu können. Dieser Schneeball-Effekt wird durch eine Gewichtung der Daten einberechnet (siehe 25 Wellen SOEP S.86).

Zusätzlich dazu kommt es auch zur Mobilität kompletter Haushalte, z.B. bei dem Umzug aller Haushaltsmitglieder. Diese werden im Rahmen der Panelpflege verfolgt und im Panel weiterhin aufgeführt. Letztlich kann es aber auch zu Haushaltsauflösungen kommen, z.B. im Falle eines Todes oder bei einem Wegzug ins Ausland. In solchen Fällen scheidet der Teilnehmer vorerst, aber meist endgültig, aus dem Panel aus.

Tabelle 2: Erhebungsinstrumente und Datenanreicherung im SOEP

Wiederholt pro Lebenslauf	**Einmalig pro Lebenslauf**
Adressprotokoll	
Fragebögen	
Personenfragebogen für älte"Personen (grün bis 1993)	Personenfragebogen für "neue"Personen (blau bis 1993)
Personenfragebogen für alle Personen	
Nacherhebungsfragebogen für Lückefälle	Lebenslauffragebogen (1984 und seit 1987 für alle neuen Personen)
	Jugendfragebogen (seit 2000)
	Muki A (Kinder im Alter von 0-1 Jahren)
	Muki B (Kinder im Alter von 2-3 Jahren)
	Muki C (Kinder im Alter von 5-6 Jahren)
Andere Erhebungsformen und Tests	
Greifkrafttest	Kognitive Leistungsfähigkeit (17-jährige)
Kognitionstests	
Weitere Datenanreicherungen und Recherchen	
Experimente	Verbleibstudie bei Ausfällen
Kleinräumige Regionalinformationen	Wegzug ins Ausland

Quelle: 25 Wellen SOEP S.84

Im Zuge der Panelmortalität, also dem Ausscheiden einer Person oder eines Haushaltes aus dem Panel, ist zu erwähnen, dass es im Grunde nur zwei Ausfallgründe gibt: entweder existiert ein Haushalt nicht mehr oder er ist zur weiteren Teilnahme am Panel nicht mehr bereit (siehe Methodenbericht 2010 S.48). Eine Nicht-Erreichbarkeit tritt in den seltensten Fällen auf, da die Feldphase über eine lange Zeitspanne (mehrere Monate) vollzogen wird. Als Rückkehrer werden Personen bzw. Haushalte bezeichnet die im Vorjahr vorläufig als „Ausfall“ deklariert wurden, aber in der laufenden Welle erfolgreich wieder in das Panels aufgenommen werden konnte (siehe Methodenbericht 2010 S.49).

Als Stabilität der Stichprobe wird das Saldo aus der negativen Größe Panelmortalität und den positiven Größen „Rückkehrer“ und „neue Haushalte“ bezeichnet (siehe Methodenbericht 2010 S.48).

4.2 Erhebung durch Infratest

4.2.1 Feldphase

Von der Konstruktion der Fragebögen bis hin zur Bereitstellung der Daten für die Nutzer vergehen rund 2 Jahre. Gerade mal 8 Monate macht dabei die hauptsächliche Feldarbeit aus. Im Februar des Befragungsjahres, nachdem die Teilnehmenden Haushalte im Rahmen der Panelpflege informiert wurden, gehen die Interviewer los um die Haushalte zu befragen. Bereits im April sind knapp 80% der Interviews durchgeführt worden. Danach werden hauptsächlich noch die „schweren Fälle“ bearbeitet, dass sind solche Haushalte,

die entweder umgezogen sind oder neu entstanden sind (hier ist Recherchearbeit notwendig), nicht erreichbare Personen oder solche die sich während der Befragung plötzlich nicht mehr befragungswillig zeigen, wodurch der Fall zur weiteren Abklärung an die zentrale Bearbeitungsgruppe geleitet wird (siehe Methodenbericht 2010 S.34f). Am Ende des jeweiligen Befragungsjahres werden die Daten nach einer ersten großen Prüfung an das SOEP-Team im DIW Berlin weitergeleitet.

Bis vor einigen Jahren wurden die Erhebungen im Januar gestartet. Man hat sich aber seit 2005 für den Feldstart am 1. Februar entschieden, da so ein einmonatiger Abstand zum vorherigen Kalenderjahr besteht und Informationen deutlich neutraler abgegeben werden können (z.B. in Bezug auf die Zufriedenheit im vergangenen Jahr). Die Feldphase des Innovationssamples I fand im Vergleich zu den Haupt-Samples A-H bisher im Zeitraum September bis Anfang Februar statt.

Im Laufe der Jahre konnte man feststellen, dass durch Interviewer, die intensiv geschult werden und im Idealfall immer dieselben Haushalte befragen, die Teilnahmebereitschaft sehr hoch gehalten werden kann. Bei den Interviewmethoden des SOEP handelt es sich um ein Mixed-Mode-Design. Je nach Situation kann dann entschieden werden ob die Befragung standardmäßig, also im mündlich-persönlichen Interview (Face-to-Face) mit Paper and Pencil (PAPI) oder mit dem Computer Assisted Interviewing (CAPI) verlaufen kann, oder ob die Befragung durch Selbstausfüllen des Fragebogens durch die Befragungsperson erfolgt, beispielsweise bei viel beschäftigten Berufstätigen, die zu möglichen Interview-Zeiten nicht verfügbar sein können. Diese Auswahl an Erhebungsmethoden erhöht die Teilnahmebereitschaft, da situativ entschieden werden kann, welche Variante am ehesten zum Haushalt passt, was gerade bei den schweren Fällen von Nutzen ist(siehe Methodenbericht 2010 S.42).

Der Zeitliche Rahmen, in dem das Interview stattfindet, ist durch das SOEP im Vorhinein festgelegt worden. Der Aufwand für den HAUSHALTS-Fragebogen sollte um die 15 Minuten betragen und der für den PERSONEN-Fragebogen jeweils 30 Minuten. Bei einem 2-Personen-Haushalt bedeutet das eine Belastung von 75 Minuten. Allerdings beträgt der wahre Wert eine weitaus höhere zeitliche Belastung. Für das Jahr 2010 betrug die Überlänge laut TNS Infratest fast 30 Minuten. Da das eine enorme zusätzliche Belastung bedeutet, sollte man über Kürzungs- und Entlastungsmöglichkeiten nachdenken. Der jeweilige Befragungsmodus und Interviewer ist pro Beobachtungseinheit codiert und im Datensatz mit erfasst. Somit lassen sich auch mögliche Interviewereffekte analysieren und gegebenenfalls durch explizite Schulung vermeiden.

4.2.2 Panelpflege

Wie bereits erwähnt, erhöht eine gewisse Routine in den Interviews die Teilnahmebereitschaft. Sei es der immer gleiche Interviewer, der die Teilnehmer durch die Befragungen führt oder eine wachsende Vertrautheit zu den Fragebögen, die sich bis auf die jeweiligen Schwerpunktthemen immer sehr ähnlich sind. Die Panelpflege hat sehr viele Facetten zu bieten, die über die Jahre ausgereift wurden und sich auch immer weiterentwickeln werden. Das Mixed-Mode-Design geht auf die individuellen Wünsche des Befragten ein und vermindert so das Risiko, einen Teilnehmer zu verlieren. Viele Interviewer bekommen

spezielle Schulungen um die Erfolgsquote der Interviews zu erhöhen. Zusätzlich gibt es unterschiedliche Incentivierungsmaßnahmen. Zum einen erhält jeder Teilnehmer eines im Panel enthaltenen Haushaltes ein 5-Euro-Los der ARD-Fernsehlotterie „Ein Platz an der Sonne“, das kurz vor dem Feldstart mit einem Ankündigungsschreiben verschickt wird. Zusätzlich werden auf Haushaltsebene eine Deutschlandkarte, ein FSC-zertifizierter Holzkugelschreiber sowie ein Post-it-Notizblock verschickt. Wie bereits erwähnt wird im Innovationssample I eine neue Incentivierungsmaßnahme bzw. -höhe getestet. Verschiedene Cash-Varianten werden hierbei im Vorfeld einzelnen Personen angekündigt und im Nachhinein wird beobachtet ob sich die variierenden Höhen der Geldbeträge unterschiedlich auf die Teilnahmebereitschaft auswirken. Die „Low-Cash“ Variante, also 5 Euro pro HAUSHALTS-Interview und 5 Euro pro PERSONEN-Interview lieferte die höchste Teilnahmebereitschaft(für weitere Details siehe Methodenbericht 2010 S.50ff). Weitere Zusatzgeschenke, wie das „Uhrenmännchen“ für erstmals befragte Jugendliche und ein Fotoleporello für Mütter neugeborener Kinder die den Fragebogen „Mutter und Kind“ ausgefüllt haben, gehören auch zur Panelpflege.

Nicht realisierbare Interviews werden von den Interviewern im Anschluss an die zentrale Bearbeitungsgruppe weitergegeben. Hier wird noch einmal Kontakt zu den betreffenden Befragten aufgenommen und ihnen weitere Möglichkeiten dargeboten um dem Panel doch noch erhalten zu bleiben. Auch diese Telefonate werden durch ein extra geschultes Team durchgeführt. Auf Nachfrage werden den Teilnehmern zusätzliche Informationsmaterialien zugesendet, wie zum Beispiel Berichte, die auf den SOEP-Daten basieren. Nach der Befragung erhält jeder Haushalt eine Dankes-Postkarte zugeschickt. Zusätzlich wurde eine Internet-Seite eingerichtet (www.leben-in-deutschland.info) auf der sich die Teilnehmer über das SOEP erkundigen können oder z.B. Adressänderungen eingeben können. Ein wichtiger Teil der Panelpflege ist, neben der Motivation der Teilnehmer zur Weiterarbeit, die Adressenpflege. Bei Ausgezogenen oder Umgezogenen ist es wichtig die neuen Adressen zu recherchieren um diese Untersuchungseinheiten nicht zu verlieren. Erste Hinweise auf Adressänderungen ergeben sich durch Zuschicken von Info-Material im Rahmen der Panelpflege, während des Feldphase durch die Interviewer oder durch die Haushalte selbst, die den Umzug melden. Falls auf diesem Wege keine Adressermittlung möglich ist, wendet man sich an das Einwohnermeldeamt oder die Post (siehe Methodenbericht 2010 S.35). Generell geht die Teilnahmebereitschaft in den letzten Jahren deutlich zurück. Deshalb sollte man in Zukunft über Aufwandsentschädigungen nachdenken oder den Bekanntheitsgrad der Studie beispielsweise durch Werbung erhöhen, um möglichen zukünftigen Panelteilnehmern die Scheu vor einer Ihnen noch unbekannten Sache zu nehmen.

4.3 Prüfung und Aufbereitung der Daten

Nach der Erhebung durch TNS Infratest, müssen die Daten verarbeitet und aufbereitet werden, bevor sie für den Nutzer zur Verfügung gestellt werden können. Alle Schritte unterliegen 3 Qualitätsaspekten, die in Tabelle 3 erklärt werden:

Tabelle 3: Qualitätsaspekte des TNS Infratest

Produktqualität	Aussagekraft, Generalisierbarkeit, Vergleichbarkeit und Repräsentativität der durch TNS Infratest erhobenen SOEP-Daten
Prozessqualität	Nachvollziehbarkeit, Wiederholbarkeit und Transparenz der bei TNS Infratest geleisteten Arbeitsschritte
Servicequalität	Anschlussfähigkeit der von TNS Infratest erbrachten Leistungen an die Prozesse und Bedarfe des DIW im Rahmen der Datenweiterverarbeitung, z.B. für den so genannten Scientific Use File des SOEP

Quelle: Methodenbericht 2010 S.64

Der Ablauf der Datenverarbeitung lässt sich in 4 Blöcke unterteilen (siehe Methodenbericht TNS Infratest S.67), die aber ineinander übergehen.

- Datenerfassung: Bereitstellung erhobener Daten für die weitere Verarbeitung (Scanprogramme einrichten etc.)
- Datenprüfung: Prüfung mit Hilfe festgelegter Prüfkriterien und gegebenenfalls Editierung der Rohdaten
- Datenbereinigung: direkter Eingriff in die Daten auf Einzelfallebene, wie die Korrektur von Datenfehlern zur Herstellung der Längs- und Querschnittkonsistenz
- Datenanreicherung: Generierung von neuen Informationen als neue Variablen, Vercodung offener Angaben und Hinzufügen regionale Kennzeichen

Eine detaillierte Beschreibung der Prozessschritte befindet sich im Anhang. Die Prüfung der Daten lässt sich wiederum in 3 Teilbereiche gliedern. Zuerst erfolgt die Bruttobezogene Basisprüfung, die sich um Inkonsistenzen oder Unvollständigkeiten in den Daten kümmert. Darauf folgen die Nettobezogenen Prüfprozesse, die für die Prüfung und Korrektur der Filterführung und unzulässiger Mehrfachnennungen, Plausibilitäts- und Summenprüfungen, sowie für die Überprüfung von Wertebereichen zuständig ist. Zusätzlich werden offenen Textangaben bearbeitet. Bei der Abschlussprüfung erfolgt die Vollständigkeitsprüfung und Registration eventueller Auffälligkeiten der Daten(siehe Methodenbericht 2010 S.70f). Sobald all diese Schritte durchlaufen wurden, sind die Daten bereit zur Weitergabe.

4.4 Auftretende Schwierigkeiten

Besonders bei Face-to-Face Befragungen ist das Risiko des Auftretens von Interviewereffekten besonders groß. Durch die Schulung der Interwiever und durch die Interviewer-Befragung sollen solche Effekte erkannt und möglichst gut beseitigt werden. Dies gelingt jedoch nicht in vollem Ausmaß. Selten treten auch von Interviewern gefälschte Interviews auf, die zwar im Querschnitt nicht sichtbar waren, aber aufgrund der Längsschnittstruktur

der Daten im Nachhinein identifiziert werden können. Dies geschieht zwar in sehr geringem Maß, ist aber dennoch erwähnenswert. Des Weiteren ist das Thema Datenschutz ein Problem, das viel Arbeit beansprucht um einen Datenverlust zu vermeiden. Beispielsweise müssen Klartextangaben vercodiert werden um Rückschlüsse auf die jeweilige Erhebungseinheit zu verhindern. Ebenso müssen sensible Daten, wie detaillierte Regionalinformationen (z.B. Postleitzahl), Wohnumfeldsmerkmale bzw. Nachbarschaftsdaten und Berufsbezeichnungen durch aufwendige Verfahren aufbereitet, vercodet und wieder integriert werden. Solche Daten sind nur von registrierten Nutzern innerhalb des DIW-Berlin auswertbar. Auf ganz sensible Daten haben sogar nur wenige Personen Zugriff, wodurch ein Missbrauch der Daten nahezu ausgeschlossen ist (siehe 25 Wellen SOEP S.82). Ein weiterer Aspekt betrifft die Gewichtung der Daten und der richtige Umgang mit fehlenden Werten anhand von statistischen Methoden, die im folgenden Kapitel genauer betrachtet werden sollen.

5 Statistische Methoden im Umgang mit den Daten

Wie bereits erwähnt, besteht, aufgrund der unterschiedlichen Ziehungsdesigns der Stichproben, ein großes Problem bei der richtigen Gewichtung der Daten. Auch der Umgang mit ersten Wellen ist ein besonderer Fall bei der Bestimmung der Gewichte. Ebenso kritisch ist der Umgang mit fehlenden Werten. Welches Verfahren ermöglicht eine gute Approximation um die Stabilität der Daten zu erhalten? Im diesem Kapitel sollen diese Problematiken genauer betrachtet werden.

5.1 Gewichtung

5.1.1 Gewichtung und Hochrechnung

Aufgrund der unterschiedlichen Auswahlwahrscheinlichkeiten bei der 1. Welle ist es notwendig Gewichte einzusetzen um eine Überrepräsentation von Teilgruppen der Grundgesamtheit auszugleichen. Zum anderen müssen Non-Response-Fälle, also die fehlende Teilnahmebereitschaft in der Start und/oder in einigen Folgewellen gesondert gewichtet werden. Je nach interessierenden Sachverhalt ist zwischen zwei Ansätzen zu wählen. Die erste Variante ist der Design-basierte Ansatz, bei dem ausschließlich die durch das Erhebungsdesign implizierten Ziehungswahrscheinlichkeiten Aussagen über die Häufigkeitsverteilung eines Merkmals treffen können. Hier werden die Merkmale von Einheiten in der endlichen Grundgesamtheit als unbekannte Parameter betrachtet. Bei dem zweiten Ansatz handelt es sich um einen Modell-basierten Schätzer, bei dem Aussagen über die Verteilung der Merkmale gemacht werden, indem sie als zufällige Realisierungen gemäß eines statistischen Modells betrachtet werden (siehe Rendtel S.47f,169). Hier werden neben der Informationen über die Ziehungswahrscheinlichkeiten noch zusätzliche Informationen über Zusammenhänge zwischen zwei Merkmalen genutzt. In der Amtlichen Statistik und in der Sozialberichterstattung ist die design-basierte Schätzung jedoch geläufiger.

Nach dem Ansatz von Horvitz und Thompson (1952), der allerdings für Querschnittsdaten ausgelegt ist, erhält man eine unverzerrte Schätzung der Hochrechnungsfaktoren durch die inverse Auswahlwahrscheinlichkeit der jeweiligen Stichprobeneinheit. Eine beispielsweise größere Chancen einer Einheit gezogen zu werden, wird somit durch eine Herunter-Gewichtung ausgeglichen. Da im SOEP aber Paneldaten vorhanden sind, erfolgte eine Weiterentwicklung dieses Ansatzes durch Galler (1987) für Längsschnittanalysen, indem anstelle von Stichprobeneinheiten zu einem Zeitpunkt zeitliche Verläufe als Erhebungseinheit betrachtet werden. Die Grundgesamtheit im Sinne der Stichprobentheorie ist nicht mehr die Population zu einem bestimmten Zeitpunkt t, sondern die Menge der zeitlichen Verläufe im betrachteten Zeitraum. Hier geht es also im Speziellen um die Wahrscheinlichkeit, dass ein bestimmter zeitlicher Verlauf in der Stichprobe erhoben wird. Das ist die Grundlage für die Ermittlung der Gewichte für die Längsschnittanalyse.

Laut Galler (1987) ist die Beobachtung in der Stichprobe vom Erhebungsdesign und vom Antwortverhalten abhängig, was dazu führt, dass sich die totale Auswahlwahrscheinlichkeit ausdrücken lässt als die Kombination aus designbestimmter und verhaltensbestimmter Komponente. Die designbestimmte Komponente bezieht sich auf die direkte Auswahl bei der Stichprobenziehung, wohingegen die verhaltensbestimmte Komponente die bedingte Wahrscheinlichkeit ausdrückt, dass eine ausgewählte Einheit auch befragungswillig ist, sprich man das Interview auch tatsächlich realisieren kann. Die Auswahlwahrscheinlichkeit zu einem bestimmten Zeitpunkt t entspricht der Wahrscheinlichkeit ausgewählt zu werden und bis zum betrachteten Zeitpunkt in der Stichprobe zu bleiben. Anzumerken ist hier noch, dass auch diese Bleibewahrscheinlichkeit durch die beiden genannten Komponenten bestimmt ist. Voraussetzung für die folgenden Schritte ist, dass das Antwortverhalten von der Auswahlentscheidung unabhängig ist.

Sei $d(i,t)$ eine Indikatorvariable, die den Wert $d(i,t) = 1$ annimmt, wenn die Einheit i zum Zeitpunkt t nach dem Erhebungsdesign ausgewählt wird, und sonst Null ist, und ist r(i,t) eine entsprechende Indikatorvariable für die Antwortbereitschaft. Man erhält die Wahrscheinlichkeit, dass eine Einheit vom Zeitpunkt t_1 bis zum Zeitpunkt t_k in der Stichprobe beobachtet werden kann, folgendermaßen:

$$P(d(i,t_k), r(i,t_k), ..., d(i,t_1), r(i,t_1)) =$$

$$= P(d(i,t_k), ..., d(i,t_1) | r(i,t_{k-1}), ..., r(i,t_1)) * P(r(i,t_k), ..., r(i,t_1))$$

$$= P(d(i,t_1), r(i,t_1)) \quad *$$

$$* \quad \prod_{j=2}^{k} P(d(i,t_j) | d(i,t_{j-1}), r(i,t_{j-1})) \; * \; P(r(i,t_j) | r(i,t_{j-1}), ..., r(i,t_1))$$

$$= P(d(i,t_1), r(i,t_1)) \quad * \\ * \quad \prod_{j=2}^{k} P(d(i,t_j)|d(i,t_{j-1}), r(i,t_{j-1})) \quad * \quad \prod_{j=2}^{k} P(r(i,t_j)|r(i,t_{j-1}), ..., r(i,t_1)) \tag{1}$$

Anhand der Rechnung kann man erkennen, dass sich wie vorher schon beschrieben, die Wahrscheinlichkeit von Beginn an in der Stichprobe zu sein und bis zu einem Bestimmten Zeitpunkt t in der Stichprobe zu bleiben, aufteilen lässt in die Wahrscheinlichkeit der Aufnahme in das Panel, der design-bestimmten Komponente (in der jeweiligen Welle gezogen zu werden) und der verhaltens-bestimmten Komponente (befragungswillig zu sein). Gezogen heißt hier nicht, dass eine neue Stichprobe gezogen wird, sondern dass aus der vorhandenen Menge an Probanden im Panel die einzelnen Einheiten im Rahmen der Panelpflege auffindbar sind. Anhand der Auswahlwahrscheinlichkeit einen bestimmten Ablauf zu ziehen, kann nun die Gewichtung durch die Inverse vorgenommen werden. Diese Grundlagen werden nun verwendet um die unterschiedlichen Gewichte im Rahmen des Sozio-ökonomischen Panels im Folgenden betrachten zu können.

Startwellengewichte

Um die Startwellengewichte bestimmen zu können, muss, wie schon angedeutet, zuerst einmal die Ziehungswahrscheinlichkeit aus dem jeweiligen Sampleverfahren gefolgert werden. Anschließend erfolgt die Ausfallkorrektur, da nicht bei allen gezogenen Einheiten erfolgreiche Interviews durchgeführt werden können.

$$\frac{1}{P(\text{In Welle 1 gezogen} \cap \text{antwortbereit})} \quad (Startwellengewicht) \tag{2}$$

Zum Schluss erfolgt noch die Anpassung an Ränder des Mikrozensus mit den Variablen Region, Alter, Geschlecht, Haushaltsgröße und Nationalität und die Korrektur bei Zweitwohnsitzen, da Personen mit einem zweiten Wohnsitz eine doppelte Auswahlchance haben. Ihr Gewicht wird demnach halbiert (siehe 25 Wellen SOEP S.90).

Längsschnittgewichte

Bei den Längsschnittgewichten ist jetzt interessant, ob beispielsweise erfolgreiche Realisationen der ersten Welle in der zweiten Welle wieder beobachtbar sind oder nicht. Dies kommt in zwei Schritten zu Stande: Zuerst erfolgt die Kontaktaufnahme zu Personen, die in der ersten Welle beobachtbar waren (design-bestimmte Komponente). Ist diese Maßnahme erfolgreich (Adressermittlung bei Umzug etc.), folgt im zweiten Schritt die Entscheidung, ob eine weitere Teilnahmebereitschaft vorhanden ist oder nicht, also ob ein erfolgreiches Interview stattfindet oder nicht (verhaltens-bestimmte Komponente). Das Produkt beider Wahrscheinlichkeiten ergibt die Bleibewahrscheinlichkeiten und der

Kehrwert davon die Bleibefaktoren (siehe 25 Wellen SOEP S.91). Besonderheiten entstehen hierbei durch Haushalte die abgespalten oder fusioniert wurden. Hier sind die Modelle entsprechend anzupassen. Für detaillierte Informationen siehe Spieß, Kroh, Pischner und Wagner(2008). Das Längsschnittgewicht der Einheit über zwei Wellen erhält man also aus dem Produkt des Gewichts der Vorwelle und des Bleibefaktors der aktuellen Welle. Anschaulich bedeutet das im Speziellen Fall von $k = 2$ Wellen (k-te Welle $= W_k$): Produkt aus Startwellengewicht und Bleibefaktor der 2. Welle.

Fall $k = 2$:

$$\underbrace{\underbrace{\frac{1}{P(W_1 \text{ gezogen} \cap \text{Antwort})}}_{\text{Startwellengewicht}} * \underbrace{\frac{1}{P(W_2 \text{ ermittelt} \mid \text{Startwelle}) * P(\text{Antwort } W_2 \mid \text{Antwort } W_1)}}_{\text{Bleibefaktor}}}_{\text{Längsschnittgewicht}} \tag{3}$$

Abschließend erfolgt wieder die Anpassung an die Ränder des Mikrozensus und die Korrektur für den Zweitwohnsitz.

Querschnittsgewichte ab Welle 2

Um Aussagen über die Struktur der Grundgesamtheit zu bestimmten Zeitpunkten ableiten zu können, also gewichtete Querschnittsanalysen zu ermöglichen, werden sogenannte „Rohgewichte "gebildet. Die Rohgewichte für die Welle k ergeben sich analog wie oben aus dem Kehrwert des Produkts aus Beobachtungswahrscheinlichkeit in Welle $k - 1$ und Bleibewahrscheinlichkeit in Welle k. Anschließend erfolgt wieder die Anpassung an die Ränder des Mikrozensus und die Korrektur für den Zweitwohnsitz (siehe 25 Wellen SOEP S.92)

5.1.2 Varianzschätzung

Die oben erklärten Gewichte für die Daten erfüllen den Zweck Verzerrungen der Schätzungen durch die Ziehung und die Ausfälle zu verhindern.

Durch die Gewichtungen sollen Verlaufsanalysen ermöglicht werden, sodass ein Zusammenhang zwischen den beobachteten Verläufen und den Veränderungen in aufeinander folgenden Querschnittbeobachtungen hergestellt werden kann. Diese Gewichtungen sollen unverzerrte Schätzer liefern. Neue Beobachtungen erhöhen die Effizienz, obwohl sie „Non Original Sample Members“ (siehe Kapitel 4.1.3) sind, da insgesamt mehr Informationen vorhanden sind. Die Schätzer müssen effizient sein, d.h. bei gegebener Stichprobe müssen varianzminimale Schätzer gewählt werden und die vorhandene Information muss weitestgehend ausgeschöpft werden (siehe Galler 1987)

Generell ist keine allgemeingültige Varianzschätzung möglich, da je nach Ziehungsdesign und Teilstichprobe eine andere Schätzung geeigneter erscheint. Zur Auswahl stehen Resamplingmethoden wie Jackknife und Bootstrapping oder die vom SOEP unterstütze

Methode über „Random Groups“. Die Idee dahinter besteht darin, eine Unterteilung der gezogenen Stichprobe in R (möglichst unabhängige) Unterstichproben (Random Groups) zu finden, so dass jede der Unterstichproben als eine Realisation des ursprünglichen Ziehungsexperiments mit verringertem Stichprobenumfang angesehen werden kann (siehe Rendtel 1995 S.94 ff.).

5.2 Fehlende Daten

Das Problem der fehlenden Werte in Umfragen lässt sich auf mehrere Arten lösen. Der fallweise Ausschluss ist im SOEP keine Option, da hierbei zu viele Informationen verloren gehen. Es gibt jedoch die Möglichkeit der Imputation, also das Füllen von Lücken durch geschätzte Werte. Die Entscheidung für die jeweilige Variante muss individuell getroffen werden. Zum einen ist das die Single-Imputation, die für jeden fehlenden Wert einen neuen Wert generiert. Bei der zweiten Variante, der sogenannten multiplen Imputation, werden für jeden fehlenden Wert mehrere Werte generiert und anschließend kombiniert. Für den so großen Umfang des SOEP ist die Multiple-Imputation jedoch noch nicht umsetzbar (siehe 25 Wellen SOEP S.96).

Bei der Single-Imputation gibt es verschiedene Methoden, die sorgfältig zu formulieren und schätzen sind, denn von der Qualität der Schätzwerte hängt die Qualität der letztlich interessierenden Schlussfolgerung ab (siehe 25 Wellen SOEP S.96). Besonderes Augenmerk sollte in diesem Sinne auch auf dem Standardfehler der Schätzungen liegen.

Die folgenden Techniken basieren auf Vorschläge in Engel/Reinecke (1994). Die wohl einfachste Methode ist die Mittelwertsersetzung um die zugrunde liegenden Daten zu vervollständigen. Nachteil ist hier jedoch, dass eine systematische Unterschätzung der Varianz auftritt, da die ersetzten Werte keine Streuung aufweisen. Abhilfe kann hier das Einsetzen von Klassenmittelwerten schaffen, indem Merkmalsklassen gebildet werden, anstatt die gesamte Stichprobe zu verwenden. Für den Fall des Regressionsverfahrens gilt ähnliches. Hier wird eine Regression zwischen zwei Variablen x und y auf Basis vollständiger Fälle durchgeführt. Hierbei werden anhand der Regressionsgerade passende Werte gefunden. Die entstehenden Mittelwerte sind zwar unter Normalverteilungsannahme gute Schätzer, weisen aber ebenso eine Unterschätzung der Varianz und Kovarianz auf. Stützt man sich auf die in den einzelnen Panelwellen vorhandene Information, ohne Imputationen vorzunehmen und schätzt die Mittelwerte, Varianzen und Kovarianzen aus allen Panelwellen unter Berücksichtigung des unterschiedlichen Stichprobenumfangs erhält man ein Maximum-Likelihood-Schätzverfahren, das für kontinuierliche Paneldaten geeignet ist (siehe Engel/Reinecke S.262f). Weitere Techniken wie die Hot-Deck-Imputation (Ersetzen durch einen beobachteten Wert eines möglichst ähnlichen Falls) und die Zeilen-und-Spalten-Imputation (Ersetzen durch einen zu einem anderen Zeitpunkt beobachteten Wert) werden hier nicht genauer behandelt.

6 Ergebnisse aktueller Analysen

7 Ausblick

Tabellenverzeichnis

Literatur

[1] Diekmann, Andreas. *Empirische Sozialforschung.* Reinbek bei Hamburg, Rowohlt-Taschenbuch-Verlag, 2011.

[2] DIW-Berlin, Vierteljahreshefte zur Wirtschaftsforschung 2008. *25 Wellen Sozio-oekonomisches Panel.* Berlin, Duncker & Humblot Berlin, 2008.

[3] Engel, U., Reinecke, J. *Panelanalyse: Grundlagen, Techniken, Beispiele.* Berlin, New York, de Gruyter, 1994.

[4] Galler, H.P.: Zur Längsschnittgewichtung des Sozio-oekonomischen Panels. In: Krupp, H.-J., Hanefeld, U. (Hrsg.):. *Lebenslagen im Wandel: Analysen 1987, Band 2 der Reihe: Sozio-oekonomische Daten und Analysen für die Bundesrepublik Deutschland.* Frankfurt/Main, New York, Campus-Verlag, 1987. S. 295-317.

[5] Götz, Eva-Maria (Deutschlandradio). Eine Goldgrube für die Wissenschaft (vom 17.07.2008). `http://www.dradio.de/dlf/sendungen/studiozeit-ks/816713/`. (letzter Aufruf: 21:18, 16.06.2013).

[6] Rendtel, Ulrich. *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität.* Frankfurt/Main, New York, Campus-Verlag, 1995.

[7] SOEP-Gruppe, Huber, S., Jänsch, A., Siegel, N.A.,. SOEP 2010 Methodenbericht zum Befragungsjahr 2010 (Welle 27) des Sozio-oekonomischen Panels. München, TNS Infratest Sozialforschung, 2011.

[8] Wagner, G., Göbel, J., Krause, P., Pischner, R., Sieber, I. Das Sozio-oekonomische Panel (SOEP): Multidisziplinäres Haushaltspanel und Kohortenstudie für Deutschland - Eine Einführung (für neue Datennutzer) mit einem Ausblick (für erfahrene Anwender). Berlin, Springer, 2008. S. 301-328.

[9] Website Cornell University. CNEF. `http://www.human.cornell.edu/pam/research/centers-programs/german-panel/cnef.cfm`. (letzter Aufruf: 21:18, 16.06.2013).

[10] Website Statista Lexikon. `http://de.statista.com/statistik/lexikonListe`.