Die Hochrechnung der ersten Welle der Stichprobe F des SOEP

von Rainer Pischner

Vorbemerkung:

Bitte lesen Sie auch den Text „Änderungen am Konzept der Querschnittsgewichtung des Sozio-ökonomischen Panels (SOEP) 1984 – 2001“. Sie finden die Datei im selben Verzeichnis. In diesem Text sind die Änderungen noch nicht eingearbeitet.

Im Jahr 2000 wurde die Stichprobengröße des Sozio-oekonomische Panels (SOEP) mit der Ergänzungsstichprobe F nicht nur ergänzt, sondern auch entscheidend ausgedehnt. Die 6052 neuen befragten Haushalte führten nahezu zu einer Verdoppelung der gesamten Stichprobe (13078 Haushalte). Insgesamt umfasst das SOEP, das 1984 startete, nunmehr sechs Teilstichproben und – für die Teilstichproben A und B - 17 Wellen (1984 – 2000).[1]

- Stichprobe A: Deutsche Haushalte[2] der Bundesrepublik Deutschland (Start 1984)
- Stichprobe B: Ausländische Haushalte[3] der Bundesrepublik Deutschland (Start 1984)
- Stichprobe C: Haushalte der DDR (Start 1990).
- Stichprobe D: Zuwandererhaushalte (Start 1994/95)
- Stichprobe E: Haushalte in Deutschland, Ergänzungsstichprobe (Start 1998)
- Stichprobe F: Haushalte in Deutschland, Ergänzungsstichprobe (Start 2000)

Insbesondere für eher deskriptiv angelegte Analysen ist eine Hochrechnung[4] für den SOEP-Datensatz unbedingt erforderlich. Für die Stichprobe F ist, wenn sie auch Teil eines Paneldatensatzes ist, zunächst eine reine Querschnittshochrechnung erforderlich. Die Beschreibung dieser Hochrechnung erfolgt in diesem Beitrag. Bevor jedoch auf diese Startgewichtung näher eingegangen wird, sind zum besseren Verständnis einige allgemeine Anmerkungen zum Hochrechnungskonzept des SOEP erforderlich.

[1] Eine aktuelle Beschreibung des SOEP findet sich in SOEP Group (2001).
[2] Genauer : Haushalte, deren Haushaltsvorstand nicht türkischer, italienischer, jugoslawischer. griechischer oder spanischer Nationalität ist
[3] Genauer : Haushalte, deren Haushaltsvorstand türkischer, italienischer, jugoslawischer. griechischer oder spanischer Nationalität ist
[4] Da sich Hochrechnungsfaktoren und Gewichte grundsätzlich lediglich um einen einzigen konstanten Faktor unterscheiden, werden im folgendem die Ausdrücke Gewichte und Hochrechnung meist synonym verwendet.

1 Zum Konzept der Hochrechnung des SOEP

1.1 Theoretische Grundlagen

Die Stichprobe des SOEP ist in ihrer Gesamtheit ungewöhnlich komplex. Ihre Gewichtung und Hochrechnung kann nur auf Grundlage eines konsistenten, theoretisch abgesicherten Konzeptes erfolgen. Es ist nicht möglich, an dieser Stelle detailliert hierauf einzugehen. Es erfolgt lediglich eine Beschreibung der wesentlichen Komponenten dieses Konzeptes mit entsprechenden Literaturhinweisen. Außerdem werden beispielhaft die Gewichtungen der bisherigen Teilstichproben dargestellt. [5]

Die Schätzung von Gewichten bzw. Hochrechnungsfaktoren erfolgt nach dem Ansatz von Horwitz und Thompson (1952) prinzipiell über den Kehrwert der Auswahlwahrscheinlichkeit der jeweiligen Stichprobeneinheit. Hochrechnungsfaktoren unterscheiden sich von Gewichten lediglich durch einen Skalarmultiplikator. Während die nach Horwitz und Thompson berechnete Summe der Hochrechnungsfaktoren der Zahl der Einheiten der Grundgesamtheit entspricht, ist die Summe der Gewichte gleich der Fallzahl der Stichprobe.

Horwitz und Thompson stellten ihr Konzept indes nur für reine Querschnittsdaten vor. Stichprobeneinheiten sind beim SOEP primär die Haushalte, sekundär die Personen, die in diesen Haushalten leben. Der Ansatz von Horwitz und Thompson wurde von Galler (1987) für Paneldaten weiterentwickelt, indem die Schätzung von Auswahlwahrscheinlichkeiten für die Startwelle um die Schätzung der Verbleib- bzw. Antwortwahrscheinlichkeit für die folgenden Wellen erweitert wurde. Hierzu muss zunächst die Wahrscheinlichkeit einer erneuten Kontaktaufnahme, darauf folgend die Wahrscheinlichkeit einer erneuten Antwortgewährung bestimmt werden.[6] Eng verknüpft mit dieser Aufgabe ist die Analyse des Ausfallverhaltens der Stichprobeneinheiten von Welle zu Welle.[7] Dieses Konzept ermöglicht es, sowohl Querschnittsgewichte ab Welle 2 jeder Stichprobe sowie Längsschnittgewichte zu ermitteln.

[5] Die Frage, ob überhaupt eine Stichprobe hochgerechnet werden soll, wird hier nicht diskutiert werden. Siehe zu diesem Thema Rendtel, Pötter (1992)
[6] Siehe hierzu Galler (1987) und Rendtel (1995).
[7] Siehe hierzu z.B. Pannenberg (2000).

Die Gewichtung einer Teilstichprobe des SOEP lässt sich somit in vier Stufen gliedern:

- Ermittlung der Hochrechnungsfaktoren für Welle 1 jeder Stichprobe (Startgewichtungen)
- Schätzung der Querschnittsgewichte ab Welle 2 für alle Stichproben
- Schätzung geeigner Längsschnittgewichte
- Zusammenführung der Stichproben

Einen allgemeine Übersicht zur Hochrechnung des SOEP, die nicht allzu sehr in Detail geht, findet sich in Haisken-De New und Frick (1998)

1.2 Die Startgewichtungen der Stichproben A – E

Querschnittsgewichtungen hängen vom Erhebungsdesign der Stichprobe ab. Keine der Stichproben des SOEP ist freilich identisch angelegt worden. Deshalb weisen auch die Startgewichtungen unterschiedlich Konzepte aus. Dennoch ist das Gerüst jeder Startgewichtung im Prinzip gleich:

1. Zunächst werden sogenannte Designgewichte (inverse Auswahlwahrscheinlichkeit) für die befragten Haushalte erstellt, die weitgehend den durchgeführten Stichprobenplan widerspiegeln sollen. Die Designgewichte berücksichtigen z.B. die Ziehung nach Nationalität. Nicht berücksichtigen können sie Nichtteilnahme ausgewählter Haushalte, z.B. wegen Nicht-Antreffens oder Verweigerung.
2. Um Non-Response auszugeleichen wird auf Basis der Designgewichte eine Randanpassung an wesentliche Eckdaten der amtlichen Statistik oder an andere Merkmale auf Haushaltsbasis vorgenommen. Diese Randanpassung erfolgt nach dem Prinzip des minimalen Informationsverlusts.[8]
3. Ausgehend von den so ermittelten Haushaltsgewichten werden Personengewichte bestimmt. Da grundsätzlich sämtliche Personen, die mindestens 16 Jahre alt sind, an der Befragung teilnehmen sollen, gilt für repräsentative Stichproben prinzipiell die Gleichheit von Haushalts- und Personengewicht. D.h. die geschätzten Haushaltsgewichte können 1:1 auf sämtliche Personen des Haushalts (einschl. der Kinder) übertragen werden. Können nicht alle Personen in einem Haushalt befragt werden, sind Modifikationen notwendig.

[8] Siehe hierzu Merz (1983). Die Randanpassungen wurden durchgeführt mit der MS-DOS-Version des Programms ADJUST von Joachim Merz, Universität Lüneburg.

Stichprobe A und B (1984)

Die Hochrechnungen der Stichproben A und B unterscheiden sich grundsätzlich in der Bestimmung der Designgewichte, da die Ausländerhaushalte (Stichprobe B) über Personenregister gezogen worden sind, was zwangsläufig zu einer Überrepräsentation großer Haushalte führte. Stichprobe B musste deshalb einem sog. Redressment unterzogen werden. Die Designgewichtung berücksichtigte weiterhin die regionale Verteilung der gezogenen Adressen sowie die Ausschöpfung der Sample-Points.[9]

Die Randanpassung beider Stichproben erfolgte an 316 Merkmale. Diese beruhen auf interpolierten Daten der Mikrozensen 1982 und 1985 sowie der EG-Arbeitskräfte Stichprobe des Jahrs 1984. Einerseits wurden die Stichproben A und B sehr detailliert an sozio-demographische Eckzahlen angepasst, andererseits ging diese Genauigkeit zu Lasten einer großen Varianz der Hochrechnungsfaktoren. Diesem Umstand wurde bei den folgenden Startgewichtungen Rechnung getragen.

Stichprobe C (1990)

1990 wurde die DDR-Basisbefragung durchgeführt. Nach Abschluss der Feldarbeiten lagen Daten für 2179 Haushalte vor. Leider fehlten aufgrund der ungewöhnlich rasch durchgeführten Erhebung detaillierte Informationen über die Ausschöpfung der Brutto-Stichprobe (3616 Adressen) aus der Stichprobe C. Eine Designgewichtung, wie sie für die für die Stichproben A und B durchgeführt wurde, war daher nicht möglich. Anstatt eine Designgewichtung allein auf Basis der Schichten der Stichprobe vorzunehmen, wurden regionale Informationen direkt in die Randanpassung einbezogen. Eine weitere Besonderheit war, dass keine Verteilung der Haushalte nach ihrer Größe verfügbar waren. Insgesamt wurden 115 Restriktionen berücksichtigt. Als Startgewichte für die Randanpassung wurden die freien Hochrechnungsfaktoren - d.h. das Verhältnis der befragten Haushalte zur Gesamtzahl der Privathaushalte - verwendet.[10]

[9] Zur Hochrechnung der Stichproben A und B siehe Pischner (1994)
[10] Siehe hierzu Pischner (1991)

Stichprobe D (1994/95)

Der überwiegende Teil der Immigranten, die nach 1984 in die Bundesrepublik einwanderten, konnte vom SOEP auch im Laufe der Zeit nicht erfasst werden. Deshalb wurde eine spezielle Zuwandererstichprobe gezogen. Sie umfasst weniger als 509 Haushalte. Die Designgewichtung ist bei Rendtel et. al. (1997) beschrieben. Die Randanpassung konnte mangels empirischer Grundlagen und wegen der sehr geringen Fallzahl nur an sehr wenige Werte erfolgen und somit lediglich die Übergänge etwas glätten.[11]

Stichprobe E (1998)

Stichprobe E ist die erste von bisher zwei Stichproben, die der Ergänzung des SOEP dienten. Sie wurde 1998 erhoben und umfasste damals 1067 befragte Haushalte. Sie ist nicht nach Befragungsmerkmalen geschichtet; deshalb gab es vom Design her keine systematischen Über- oder Untererfassungen von Haushalten oder Personen. Eine Randanpassung war wie üblich erforderlich.[12] Die Anpassung wurde getrennt nach alten und neuen Ländern durchgeführt. Angesichts der relativ geringen Fallzahl wurden lediglich neun Restriktionen abgesehen von der Unterteilung nach Ost und West berücksichtigt:

- Haushaltsgröße (1- 2 -, 3-, 4 – sowie 5 und mehr Personenhaushalte
- Altersklasse (15-69 Jahre, 70 Jahre und älter)
- Geschlecht (Männlich)
- Nationalität (nicht deutsch)

2 Hochrechnung Stichprobe F

Stichprobe F verdoppelt nahezu den Umfang der „Altstichproben“ A bis E. Sie ist wie Stichprobe E nicht geschichtet, aber es gab – wie üblich – selektive Non-Response.
Angesichts der hohen Fallzahl (6052 neue befragte Haushalte) und dem entsprechenden Gewicht dieser Stichprobe für den gesamten SOEP-Stichprobenbestand kommt einerseits eine sparsame Gewichtung, wie sie für die Ergänzungsstichprobe E durchgeführt wurde, nicht in Frage, andererseits soll Wert darauf gelegt werden, die Varianz der Hochrechnungsfaktoren möglichst klein zu halten.

[11] Siehe hierzu Schupp, Wagner (1995).
[12] Siehe hierzu Pischner (1999).

2.1 Designgewichtung

Der bekannten Untererfassung von Ausländerhaushalten bei nach dem ADM-Mastersample durchgeführten Random-Route-Befragungen[13] konnte teilweise entgegengewirkt werden bei Anlage des Stichprobenplans der Stichprobe F, indem doppelt so viele Adressen wie notwendig erfasst wurden, jedoch von den „überzähligen" Haushalten nur solche befragt wurden, in denen Ausländer leben. Diese Tatsache wurde bei der Generierung der Designgewichte Rechnung getragen.[14] Weiterhin hätten die verfügbaren Informationen es erlaubt - wie bei den Stichproben A und B praktiziert - Ausschöpfungsquoten der Sample-Points zusätzlich zu berücksichtigen. Wie weiter unter ausgeführt, wurde – im Hinblick auf möglichst geringe Varianz der Hochrechnungsfaktoren - ein eleganterer Weg eingeschlagen, auftretende Probleme, die durch eine unterschiedlicher Ausschöpfung von Sample-Points auftreten, zu lösen.

2.2 Haushaltsgewichtung

Die Designgewichte sollten im Idealfall – von einem konstanten Faktor einmal abgesehen – den endgültigen Hochrechnungsfaktoren entsprechen. Dies ist aufgrund selektiven Non-Responses nie der Fall. Deshalb müssen diese Designgewichte so modifiziert werden, dass sie bestimmte Restriktionen erfüllen. Im SOEP sollen sie u.a. die nachstehende Haushaltsstruktur widerspiegeln.

Tabelle 1

Privathaushalte in Deutschland im April 2000

Haushaltsgröße	Deutschland	Früheres Bundesgebiet	Neue Länder und Berlin-Ost
	in 1 000		
Insgesamt	38123	31045	7078
1-Personenhaushalte	13749	11337	2412
2-Personenhaushalte	12720	10269	2451
3-Personenhaushalte	5597	4376	1221
4-Personenhaushalte	4392	3593	799
5 und mehr Pers. HH	1665	1470	195

*) Ergebnisse des Mikrozensus 2000.

[13] Diese Untererfassung beruhte auf der Auswahlgrundlage, den Stimmbezirken der Bundestagswahl, die naturgemäß nur Wahlberechtigte, d.h. keine Ausländer enthalten.

Ein weiteres Merkmal, welches in die in die Randanpassung eingeht, ist die Altersstruktur der Personen die in den Privathaushalten wohnen. Es ist zu beachten, dass dies dennoch keine Personengewichtung darstellt, es handelt sich vielmehr um Haushalte, die eben die Eigenschaft besitzen, Personen dieses oder jenes Alters zu umfassen.

Tabelle 2

Altersstruktur der wohnberechtigten Bevölkerung in Privathaushalten in Deutschland im April 2000

Personen in Privathaushalten im Alter von ... bis unter ... Jahren	Deutschland	Früheres Bundesgebiet	Neue Länder und Berlin-Ost
	in 1 000		
insgesamt	82.473	67.266	15.207
unter 15	12.632	10.656	1.976
15 – 20	4.684	3.615	1.069
20 – 25	4.665	3.712	953
25 – 30	4.882	4.072	811
30 – 35	6.521	5.469	1.052
35 – 40	6.860	5.598	1.262
40 – 45	6.201	5.022	1.179
45 – 50	5.765	4.578	1.187
50 – 55	4.984	4.121	863
55 – 60	5.545	4.454	1.091
60 – 65	5.981	4.810	1.171
65 und älter	13.754	11.159	2.594

*) Ergebnisse des Mikrozensus 2000.

Im Sinne einer insgesamt sparsamen Anpassung werden auch für Geschlecht und Nationalität nur Randsummen berücksichtigt.

[14] Siehe hierzu Spiess (2001).

Tabelle 3

Geschlecht und Nationalität von Personen in Privathaushalten in Deutschland m April 2000			
	Deutschland	Früheres Bundesgebiet	Neue Länder und Berlin-Ost
	in 1 000		
Personen in Privathaushalten insgesamt	82.473	67.266	15.207
darunter:			
Männer	40086	32719	7367
Ausländer	7095	6926	169
*) Ergebnisse des Mikrozensus 2000 .			

Wie oben erwähnt, wurde im Rahmen der Designgewichtung nicht die Ausschöpfung der Sample-Points berücksichtigt, um die Varianz der Hochrechnungsfaktoren gering zu halten. Als Äquivalent für diesen Verzicht, wurden in die Randanpassung Quartier-beschreibungsvariablen einbezogen: Hinter dieser Vorgehensweise steckt folgende Idee: Jeder Sample-Point klumpt in gewisser Weise, spiegelt also nicht die Grundgesamtheit wider: So enthalte Sample-Point X z.B. Haushalte eines Dorfes, Sample Y Haushalte in einem Hochhaus. Man unterstellt wohl zurecht, dass die Bewohner von X und Y

- wahrscheinlich unterschiedlich leben und sich auch anders verhalten und dass
- die Sample-Points X und Y in der Brutto-Stichprobe in dem der Grundgesamtheit entsprechenden Verhältnis vorkommen.

Bei unterschiedlicher Ausschöpfung der Sample-Points werden im allgemeinen Verzerrungen in der Netto-Stichprobe auftreten, da z.B. Sample-Points, in der „Mittelschichthaushalte“ dominieren, weniger Non-Response aufweisen als Sample-Points mit vielen Unter- und Oberschicht-Haushalten. Um dies auszugleichen, kann man die Designgewichte entsprechend über die Ausschöpfungsquote eines jeden Sample-Points auf- oder abwerten. Es ist aber nicht zwingend, dass die Sample-Points in sich tatsächlich so homogen sind, dass sich die Berücksichtigung der Ausschöpfungsquote rechtfertigen ließe. Außerdem wird so die Varianz der Gewichtsfaktoren vergrößert, da jede zufällige Abweichung von der durchschnittlichen Ausschöpfung – auch wenn sie nicht systematisch ist – das Gewicht beeinflusst. Eine Alternative zu diesem Vorgehen bietet die Berücksichtigung von „Quartierbeschreibungsvariablen“, die auch für die Brutto-Stichprobe

vollständig vorliegt. Diese werden von den Interviewern erhoben, um Ausfallanalysen[15] und somit eine Hochrechnung zu ermöglichen. Unter der Voraussetzung, dass diese repräsentativ ist, leistet eine Anpassung der Struktur der Quartierbeschreibungsvariablen der Netto-Stichprobe an die der Brutto-Stichprobe ähnliches. Der Vorteil liegt darin, dass – um bei dem Beispiel zu bleiben – tatsächlich und nicht nur vermutlich an die richtige Zahl der „Dorf- bzw. Hochhausbewohner" angepasst wird . Exakt wird dies allerdings auch nicht erfolgen, da die Interviewer möglicherweise ungenaue Angaben machen.
Folgende Quartierbeschreibungsvariable sind für die Stichprobe F erhoben worden:

1. Entfernung des Haushalts zur nächsten Großstadt (100 Tsd. Einwohner und mehr)
2. Gebäudetyp
3. Unmittelbare Umgebung des Wohnhaushalts
4. Soziale Schicht der Mehrzahl der Bewohner im Wohngebiet des zu befragenden Haushalts.

Aus den entsprechenden Anteilen in der Bruttostichprobe wurde die Zahl der Haushalte in Deutschland mit den entsprechenden Eigenschaften geschätzt, an die dann angepasst wurde.

Tabelle 4

Privathaushalte nach Quartiersmerkmalen in Deutschland m April 2000

Quartiersmerkmale	Deutschland	Früheres Bundesgebiet	Neue Länder und Berlin-Ost
	in 1 000		
Lage zur nächst.Großstadt			
Im Geschäftszentrum	*(673)*	*(673)*	*(0)*
<1 km von Großstadt	2079	1866	*(213)*
1-2 km von Großstadt	3263	2629	634
3-5 km von Großstadt	3565	3047	518
> 5 km von Großstadt	5608	4631	976
< 10 km von Großstadt	2274	2051	223
10-25 km von Großstadt	6163	5328	835
25-50 km von Großstadt	7342	5631	1711
> 50 km von Großstadt	7157	5189	1968

[15] Siehe hierzu Däubler (2001).

noch Tabelle 4

Wohnbebauung		0	
Villen	378	312	66
freistehendes Einfam. Haus	14384	11980	2404
freistehendes Mehrf. Haus	7142	6037	1105
freistehendes Gebäude	1709	1421	289
Wohnhochhäuser	821	551	270
Terrassenhäuser	121	121	0
Reihenhäuser	3356	2929	427
Zeilenbauweise	3916	2890	1026
geschlossene Blockrandbebauung	5236	4056	1180
geschl. Bl.randb. mit Höfen	739	485	2542
nichts davon, missing	*(320)*	*(264)*	*(56)*
Umgebung des Hauses			
nur Wohngebäude	22296	18780	3515
vorrangig Wohngebäude	12244	9969	2276
vorrangig Gewerbe	886	576	311
laendl. Bebauung, missings	2696	1720	976
Soziale Schicht			
Unterschicht	1835	1342	493
Untere Mittelschicht	12137	9030	3107
Mittlere Mittelschicht	20549	17341	3208
Obere Mittelschicht	3180	2947	232
Oberschicht, missings	*(422)*	*(385)*	*(38)*

Anmerkung: Kursive Werte in Klammern gingen nur indirekt in die Randanpassung ein.
Ergebnisse des Brutto-Stichprobe F (20001) des SOEP; eigene Berechnungen.

Wie oben erwähnt, erfolgt die Randanpassung nach dem Prinzip des minimalen Informationsverlustes. Dies heißt nichts anderes, als dass die ursprüngliche Struktur der Designgewichte weitestgehend erhalten bleiben soll.

2.3 Personengewichtung

Anders als bei den Startgewichtungen der alten Stichproben konnte das Prinzip Personengewicht = Haushaltsgewicht nicht vollständig beibehalten werden, da diesmal zu viele Haushalte Teilausfälle verzeichneten (8,6% der Haushalte). Insgesamt konnte mit 5,6 % aller Personen, die in zumindest teilweise realisierten Haushalten lebten, keine Interviews geführt werden.

Zur Berechnung der Personengewichte wurde folgendermaßen vorgegangen:

1. Die Trennung zwischen alten und neuen Ländern wurde auch bei der Bestimmung der Personengewichte beibehalten.
2. Im ersten Schritt wurden die Haushaltsgewichte auf die Personen, die in den jeweiligen Haushalten leben (einschl. Kinder), übertragen
3. Für die Bevölkerung in Privathaushalten am Hauptwohnsitz wurde eine eindimensionale Randanpassung an die Altersstruktur vorgenommen, da die temporären Verweigerungen im wesentlichen nur altersabhängig sind. Die Altersstruktur wurde nahezu über die gesamte Altersspanne in 5-Jahreskohorten angepasst (Ausnahme: 90 Jahre und älter). Die Anpassung wurde auf die Altersstruktur beschränkt.

Tabelle 5

Bevölkerung in Privathaushalten am Hauptwohnsitz im Mai 2000 nach Regionen und Altersgruppen

Alter von ... bis unter ... Jahren	Deutschland	Früheres Bundesgebiet	Neue Länder und Berlin-Ost
	in 1 000		
Insgesamt	81.366	66.303	15.063
unter 5	3.871	3.384	487
5 – 10	4.065	3.568	497
10 – 15	4.651	3.664	986
15 – 20	4.635	3.579	1.057
20 – 25	4.409	3.507	903
25 – 30	4.728	3.935	793
30 – 35	6.416	5.376	1.041
35 – 40	6.790	5.536	1.254
40 – 45	6.141	4.968	1.173
45 – 50	5.707	4.526	1.181
50 – 55	4.926	4.070	856
55 – 60	5.472	4.389	1.083
60 – 65	5.919	4.754	1.165
65 – 70	4.303	3.404	899
70 – 75	3.773	3.062	711
75 – 80	2.933	2.412	521
80 – 85	1.296	1.074	222
85 – 90	992	815	177
90 und mehr	339	281	58

*) Ergebnis des Mikrozensus - Bevölkerung in Privathaushalten am Ort der Hauptwohnung.

2.4 Berücksichtigung des 2. Wohnsitzes

Personen, die zwei Wohnsitze oder mehr haben, besitzen eine höhere Auswahlwahrscheinlichkeit als Personen, die lediglich an einem Wohnsitz zu erreichen sind. Deshalb muss das Personengewicht jener Personen korrigiert werden. Üblich ist die Annahme einer Verdoppelung der Auswahlwahrscheinlichkeit und somit eine Halbierung des persönlichen Hochrechnungsfaktors. Da indes die Informationen über den 2. Wohnsitz erst in der zweiten Welle der Stichprobe F erfasst wird, muss dieser Umstand bei der Bestimmung der Personengewichte zunächst unberücksichtigt bleiben. Diese Korrektur wird im Rahmen der Hochrechnung für die zweite Welle der Stichprobe F vorgenommen werden. Wie die Erfahrung zeigt, dürfte indes diese Berichtigung keinen wesentlichen Einfluss auf die Analysen zeigen.

2.5 Anstaltshaushalte

Eine Gewichtung der Anstaltshaushalte wird nur im Rahmen der Designgewichtung der Haushalte vorgenommen, da es keine zuverlässigen externen Informationen über Zahl und Struktur von Anstaltshaushalten gibt. Ihr Anteil an den Gesamthaushalten wird über alle Gewichtungsschritte hinweg beibehalten. Die im SOEP vorhandenen Anstaltshaushalte dürften systematische Verzerrungen aufweisen, da Anstaltshaushalte nur aus Versehen erhoben werden, stellt doch die Grundgesamtheit des SOEP (in der jeweiligen ersten Welle) nur Privathaushalte dar.

2.6 Ergebnis der Randanpassung

In der folgenden Übersicht sind Strukturen und relativen Differenzen der ungewichteten Stichprobe F und derjenigen nach Randanpassung dargestellt.

Tabelle 6

Strukturen von Privathaushalten der Stichprobe F des SOEP Gewichte und ungewichtete Verteilungen und deren relativen Abweichungen						
Merkmale, die in die Randanpassung eingingen	Ge-wichtet	Unge-wichtet	Abwei-chung	Ge-wichtet	Unge-wichtet	Abwei-chung
	Anteile %		in %	Anteile %		in %
Haushaltsstruktur						
1-Personenhaushalte	36,5	26,7	-26,9	34,1	28,1	-17,7
2-Personenhaushalte	33,1	35,2	6,4	34,6	36,3	4,9
3-Personenhaushalte	14,1	15,6	10,9	17,3	17,3	0,1
4-Personenhaushalte	11,6	15,2	31,5	11,3	14,6	29,6
5 und mehr Pers. HH	4,7	7,2	52,7	2,8	3,7	35,8
Altersstruktur						
unter 15	15,8	8,5	-46,2	13,0	7,7	-40,4
15 – 20	5,4	7,4	37,8	7,0	9,0	27,8
20 – 25	5,5	5,4	-2,3	6,3	6,7	6,7
25 – 30	6,1	5,4	-10,0	5,3	4,4	-18,4
30 – 35	8,1	8,4	3,7	6,9	6,6	-3,9
35 – 40	8,3	10,0	19,9	8,3	8,5	2,5
40 – 45	7,5	9,6	28,0	7,8	10,1	30,7
45 – 50	6,8	8,0	17,0	7,8	7,8	-0,6
50 – 55	6,1	7,2	16,9	5,7	6,2	9,5
55 – 60	6,6	6,7	1,6	7,2	6,4	-10,6
60 – 65	7,2	7,9	10,4	7,7	9,9	29,0
65 und älter	16,6	15,5	-6,5	17,1	16,6	-2,5
Nationalität u. Geschlecht						
Ausländer	10,3	9,5	-8,2	1,1	1,1	-0,3
Männer	48,6	54,8	12,7	48,4	51,9	7,1
Lage zur nächst.Großstadt						
Im Geschäftszentrum	2,2	1,7	-23,6	-	-	.
<1 km von Großstadt	6,0	5,7	-4,5	3,0	3,0	-1,3
1-2 km von Großstadt	8,5	6,8	-20,0	9,0	8,1	-9,8
3-5 km von Großstadt	9,8	8,6	-12,4	7,3	7,1	-3,5
> 5 km von Großstadt	14,9	14,6	-2,3	13,8	12,1	-12,4
< 10 km von Großstadt	6,6	7,0	6,6	3,1	3,6	13,4
10-25 km von Großstadt	17,2	17,5	1,9	11,8	11,9	0,9
25-50 km von Großstadt	18,1	19,1	5,1	24,2	24,9	3,1
> 50 km von Großstadt	16,7	19,1	14,0	27,8	29,4	5,8
Wohnbebauung						
Villen	1,0	1,0	-1,1	0,9	1,1	18,7
freistehendes Einfam. Haus	38,6	43,4	12,3	34,0	36,5	7,4
freistehendes Mehrf. Haus	19,4	19,1	-2,0	15,6	15,3	-1,9
freistehendes Gebäude	4,6	3,9	-15,4	4,1	3,3	-18,7
Wohnhochhäuser	1,8	1,0	-45,2	3,8	2,8	-26,4
Terrassenhäuser	0,4	0,3	-25,7	-	-	.
Reihenhäuser	9,4	10,3	9,5	6,0	5,8	-4,1
Zeilenbauweise	9,3	7,9	-14,8	14,5	15,4	6,1
geschlossene Blockrandbebauung	13,1	10,9	-16,6	16,7	15,1	-9,7
geschl. Bl.randb. mit Höfen	1,6	1,1	-28,4	3,6	4,1	13,6
nichts davon, missing	0,8	1,2	33,3	0,8	0,7	-15,0

noch Tabelle 6

Umgebung des Hauses						
nur Wohngebäude	60,5	63,0	4,1	49,7	49,0	-1,4
vorrangig Wohngebäude	32,1	29,5	-8,1	32,2	34,1	6,1
vorrangig Gewerbe	1,9	1,2	-34,1	4,4	3,8	-12,8
ländl. Bebauung, missings	5,5	6.3	13,4	13,8	13,1	-5,0
Soziale Schicht						
Unterschicht	4,3	3,3	-24,3	7,0	6,5	-7,2
Untere Mittelschicht	29,1	28,0	-3,6	43,9	39,7	-9,5
Mittlere Mittelschicht	55,9	58,0	3,8	45,3	49,6	9,4
Obere Mittelschicht	9,5	9,7	1,9	3,3	3,8	16,6
Oberschicht, missing	1,2	1,1	-14,0	0,5	0,4	-18,9
Mikrozensus 2000; SOEP-Stichprobe F; eigene Berechnungen.						

3 Zusammenführung der Stichproben

Wie bereits bei der Ziehung der Teilstichprobe E wird die neue Teilstichprobe F nunmehr in die Stichproben A-E „integriert". Zur Integration musste, wie bereits für Teilstichprobe E, ein optimales Konvexgewicht zur Minimierung der Varianz der Gewichte geschätzt werden. Danach sollen die Teilstichproben A-E mit einem Gewicht von 0,55 und die Teilstichprobe F mit dem Gewicht 0,45 in die Hochrechnungsfaktoren eingehen.

Die im Datensatz enthaltenen Querschnitts-Hochrechnungsfaktoren QHHRF und QPHRF sind Hochrechnungsfaktoren, die für die (Querschnitts-) Verwendung aller Teilstichproben (A-F) bestimmt sind; die Summe der Hochrechnungsfaktoren entspricht also der Anzahl der Personen bzw. der Haushalte in Deutschland. Im Gegensatz zur Zusammenführung der Stichproben A, B und C , wo die Menge der zu repräsentierenden Haushalte disjunkt war, ist es nicht trivial, die Zuwandererstichprobe[16] bzw. die Ergänzungsstichproben in die Altstichproben einzubeziehen, d.h. optimale Faktoren zur Zusammenführung der alten und der neu hinzugekommenen Stichproben E und F zu ermitteln. [17]
Für ein separate Auswertung der Teilstichproben A-E sind die Hochrechnungsfaktoren QHHRFAE, für eine separate Auswertung der Teilstichprobe F sind die Hochrechnungsfaktoren QHHRFF zu verwenden.

4 Ausblick

Mit Erhebung der Stichprobe F hat sich das SOEP nahezu verdoppelt. Konzeptbedingt sind die Altstichproben A-E und die neue Stichprobe F unterschiedlich hochgerechnet worden. Die

[16] Siehe hierzu Daschke, Rendtel (1996).

Fallzahlen erlauben es, für die Welle Q getrennte Analysen mit den Stichproben A bis E sowie F durchzuführen und zu vergleichen. Da die Hochrechnungsfaktoren wegen bisher fehlender Berücksichtigung zweiter Wohnsitze ohnehin vorläufiger Natur sind, bleibt Zeit, die Analysen auf Basis beider Stichproben zu erstellen und kritisch zu beleuchten. Auffälligkeiten sollten der SOEP-Gruppe mitgeteilt werden.

Literatur

Däubler, Thomas (2001) Ausfallanalysen für die erste Welle der Stichprobe F des SOEP. *DIW-Materialien (In Vorbereitung)*

Galler, Heinz P. (1987) Zur Längsschnittgewichtung des Sozio-oekonomischen Panels. *In: Krupp/Hanefeld: Lebenslagen im Wandel: Analysen 1987, Band 2 der Reihe: Sozio-oekonomische Daten und Analysen für die Bundesrepublik Deutschland, Frankfurt, S. 295-317.*

Haisken-De New, John P. and Joachim R. Frick (1998) Eds. "DTC - Desktop Companion to the German Socio-Economic Panel Study „ mimeo.

Horwitz, D. and D. Thompson (1952) „A Generalisation of Sampling without Replacement From a Finite Universe", *Journal of the American Statistical Association,47,* S. 663-685.

Merz, Joachim (1983) „Die konsistente Hochrechnung von Mikrodaten nach dem Prinzip des minimalen Informationsverlusts", *Allgemeines Statistisches Archiv, 67,* S.342-366.

Pannenberg, Markus (2000) „Documentation of the Sample Sizes and Panel Attrition in the German Socio-Economic Panel (GSOEP)", *DIW Diskussionspapier No. 196*

Pischner, Rainer (1991) „Eine konsistente Haushalts- und Personengewichtung für die DDR-Basisbefragung des SOEP und für die Ost-Pilotstudie des Wohlfahrtssurveys. *In: Vierteljahrshefte zur Wirtschaftsforschung. Heft 1 – 2,Berlin, S.50-64.*

Pischner, Rainer (1994) „Quer- und Längsschnittgewichtung des Sozio-oekonomischen Panels. *In: S.Gabler, J. H. P. Hoffmeyer-Zlotnik, D. Krebs (Hrsg.): Gewichtung in der Umfragepraxis, Opladen 1994, S. 166-187.*

Pischner, Rainer (1999) „Überarbeitete Querschnittshochrechnung der Wellen G-N (1990 bis 1997) des Sozio-oekonomischen Panels (SOEP) unter Einbeziehung der Ergänzungsstichprobe E. *Siehe: Querschnittshochrechnung 90-97 auf der Website: http://www.diw.de/deutsch/sop/service/doku/index.html*

Rendtel, Ulrich und Ulrich Pötter(1992) „Über Sinn und Unsinn von Repräsentationsstudien. *DIW-Diskussionspapier 61, Berlin*.

[17] Siehe hierzu Spiess , Rendtel (2000)

Rendtel, Ulrich (1995) „Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität“, Campus, Frankfurt - New York.

Rendtel, Ulrich (1997), Markus Pannenberg und S. Daschke: Die Gewichtung der Zuwandererstichprobe des Sozio-oekonomischen Panels (SOEP). *In: Viertelljahrshefte zur Wirtschaftsforschung,66, S. 271-285.*

SOEP Group (2001): The German Socio-Economic Panel (GSOEP) after more than 15 years - Overview. In: Elke Holst, Dean R. Lillard und Thomas A. DiPrete (Hg.*): Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP2000), Vierteljahrshefte zur Wirtschaftsforschung, Jg. 70, Nr. 1, S. 7-14.*

Schupp, Jürgen und Gert G. Wagner (1995) „The German Socio-Economic Panel: a Data-base for Longitudinal International Comparisons, *Innovations, 8(1), S.95-108.*

Spiess, Martin (2001) „Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP), *DIW-Materialien / Research notes No.5*

Spiess, Martin und U. Rendtel (2000) „Combining an ongoing panel with a new cross-sectional sample“, *DIW-Diskussionspapier No. 198*