Überarbeitete Querschnittshochrechnung der Wellen G-N (1990 bis 1997) des Sozio-ökonomischen Panels (SOEP) unter Einbeziehung der Ergänzungsstichprobe E (Welle O)

von Rainer Pischner

Vorbemerkung:

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

Zusammen mit den üblichen Daten der Welle O werden auch überarbeitete Querschnittshochrechnungen für die Wellen G-N ausgeliefert. Um die Gründe für eine solche Überarbeitung verstehen zu können, ist eine - zumindest kurze - Beschreibung des Prinzips der Querschnittsgewichtungen für das Sozio-oekonomische Panel (SOEP) unerlässlich[1]. Danach wird auf den Anlass für die neue Querschnittsgewichtung eingegangen.

1 Zum Hochrechnungskonzept des SOEP

Eine Panelgewichtung hat sich zwei Zielstellungen zu stellen. Für jede Welle müssen einerseits Querschnittsgewichte bereitgestellt werden, andererseits sind für alle möglichen Kombinationen von Befragungszeiträumen Längsschnittgewichte erforderlich. Weiterhin sollten Quer- und Längsschnittgewichte nach dem selben Prinzip bestimmt werden, damit sie kompatible Ergebnisse liefern.

Gewichte bzw. Hochrechnungsfaktoren[2] werden idealtypisch durch den Kehrwert der Auswahlwahrscheinlichkeit einer Stichprobeneinheit - im SOEP sind das Haushalte und Personen - bestimmt[3]. Somit ist der erste Schritt einer Querschnittsgewichtung die Schätzung der Aus-

[1] Eine ausführliche Beschreibung der Hochrechnung des SOEP findet man in Haisken-DeNew und Frick (1998).

[2] Die Begriffe Hochrechnungsfaktor und Gewicht werden in diesem Aufsatz nahezu synonym verwendet. In der Praxis unterscheiden sich Gewichte und Hochrechnungsfaktoren lediglich durch einen einzigen Faktor. Hochrechnungsfaktoren sind so bestimmt, dass die hochgerechneten Werte einer Stichprobe denen in der Grundgesamtheit gleichen, während Gewichte lediglich die relativen Anteile einzelner Sub-Gruppen verschieben.

[3] Dieses Prinzip wurde erstmals von Horwitz und Thompson (1952) im Rahmen von Querschnittsgewichtungen untersucht.

wahlwahrscheinlichkeit der Stichprobenelemente im Rahmen des angewendeten Stichprobendesigns, also die Schätzung der Wahrscheinlichkeit, mit der ein Stichprobenelement in der Brutto-Stichprobe enthalten ist. Im zweiten Schritt erfolgt die Bestimmung der Wahrscheinlichkeit, dass die Befragung dieses Stichprobenelements realisiert wird. Sind diese Wahrscheinlichkeiten bekannt oder können diese geschätzt werden, kann eine Querschnittsgewichtung erfolgen.
Dieses Prinzip kann auch auf Panel-Befragungen, also für Längsschnitte angewendet werden: Mit jeder neuen Welle kommen weitere zwei-stufige Schätzungen hinzu. Nur tritt an die Stelle der Auswahlwahrscheinlichkeit die Wahrscheinlichkeit der erneuten Kontaktaufnahme[4]. Das Produkt aus Kontaktwahrscheinlichkeit und Antwortwahrscheinlichkeit wird als Bleibewahrscheinlichkeit bezeichnet. Somit ergibt sich das Längsschnittgewicht einer Welle t [5] aus dem Produkt von Längsschnittgewicht der Welle t-1 und der reziproken Bleibewahrscheinlichkeit der Welle t.

Geschätzt werden die Bleibewahrscheinlichkeiten für den SOEP-Datensatz im Rahmen einer umfangreichen Ausfallanalyse, auf die hier nicht näher eingegangen werden soll.[6] Wichtig für das Verständnis der folgenden Ausführungen ist nur, dass die Bleibewahrscheinlichkeiten das Fundament für die nachfolgenden Hochrechnungen bilden. Zu erwähnen ist noch, dass über ein gesondertes Schätzverfahren die Auswahlwahrscheinlichkeiten neuer und wieder in die Stichprobe zurückgekehrter[7] Personen und Haushalte ermittelt werden[8].

Wie leicht ersichtlich sind Quer- und Längsschnittgewichtung des SOEP Ergebnisse eines komplexen seriellen Prozesses, eine isolierte Überarbeitung der Hochrechnung einzelner Jahre verbietet sich somit.

2 Zum Anlass der überarbeiteten Querschnittsgewichtung

Die Startgewichtung des SOEP, also die erste Querschnittsgewichtung, wurde - wie bei reinen Querschnittsstudien üblich - aus dem Design der Erhebung und anschließender Randanpas-

---

[4] siehe hierzu Rendtel, 1995.

[5] In der ersten Welle einer Panelerhebung sind Querschnittsgewichtung und Längsschnittgewichtung definitionsgemäß identisch.

[6] Siehe hierzu Pannenberg (2000).

[7] Temporäre Ausfälle.

[8] Siehe hierzu Rendtel (1995).

sung an zentrale Eckwerte der amtlichen Statistik vorgenommen. Die Bestimmung der Bleibewahrscheinlichkeiten hingegen erfolgt über eine detaillierte Ausfallanalyse auf Basis der erhobenen Daten. Auch wenn die Annahme, erwartungstreu zu schätzen, erfüllt ist, bleibt immer ein stochastischer Fehler, der sich in Abweichungen von Ecksummen der amtlichen Statistik zeigt. Bei dem SOEP kam hinzu, dass die nach seinem Beginn (Im Jahre 1984) einsetzende starke Zuwanderung nach Westdeutschland nicht im Design berücksichtigt werden konnte (dieses Problem ist weltweit bei allen Haushaltspanels das gleiche). D.h. der überwiegende Teil der Immigranten (einschl. Ausländer), die insbesondere nach 1984 nach Westdeutschland strömten, wurden teilweise vom SOEP nicht erfasst. Aus diesem Grunde wurde in den Jahren 1994/95 eine spezielle Zuwandererstichprobe gezogen. Bis zu diesem Zeitpunkt waren die Zahl und Strukturen der Zuwanderer jedoch unzureichend im SOEP berücksichtigt. Eine Randanpassung soll zumindest die Zeitreihen glätten.[9]

Für die ersten Wellen des SOEP war lediglich eine geringe Unterzeichnung der mit Hilfe des hochgerechneten SOEP Grundgesamtheit zu beobachten. Um diese Unterzeichnung zu korrigieren genügte es, die Querschnittsgewichte jeweils an eine einzige Ecksumme, nämlich der Bevölkerung in Privathaushalte am Hauptwohnsitz in der Bundesrepublik Deutschland anzupassen.[10] Für die Jahre von 1985 bis 1989 wurde so verfahren. In der Folgezeit - nach der deutschen Wiedervereinigung - wurde diese Ecksumme aufgespalten nach neuen und alten Ländern und an zwei Randwerte angepasst.
Mitte der 90er Jahre kam es nicht nur für die Gesamtpopulation, sondern auch für Teilgruppen der Bevölkerung zu größeren Abweichungen von Populationschätzern des SOEP von eben solchen des Mikrozensus. So zeigten Vergleiche vor allem eine signifikante Unterschätzung der Ausländer und der Zahl der Einpersonenhaushalte.

Ein weiteres Problem bei der Konzeption der SOEP-Hochrechnung wurde unterschätzt: Die Erhebung von Haushalten im SOEP unterscheidet sich von der des Mikrozensus, welche auf Grund seiner Stichprobengröße als die zuverlässigste amtliche Stichprobe für Personen- und Haushaltsdaten ist. Der Mikrozensus weißt deutlich mehr Einpersonenhaushalte aus, als die anderen amtlichen Großerhebungen, Volkszählungen (VZ) und Einkommens- und Verbrauchsstichprobe (EVS). Die Ursachen für die Differenzen liegen vermutlich können im

[9] Siehe hierzu Schupp und Wagner (1995).

[10] Vgl. Pannenberg (2000).

Abrechnungssystem mit den Interviewern. Während für das SOEP Infratest Burke nach Zahl der geführten Interviews abrechnet, entlohnt das Statistische Bundesamt für den Mikrozensus nach der Zahl befragter Haushalte, unabhängig von deren Größe. Auch bei der VZ wird nach der Zahl der Personen entlohnt (während das Anlegen eines „Mantelbogens" für einen Haushalt nicht bezahlt wird). Bei der EVS schätzen sich die Haushalte; die sich alle freiwillig melden, bezüglich ihrer Haushaltsgröße selbst ein. D.h. nur beim Mikrozensus hat der Interviewer einen Anreiz, pro Wohnung, die die Erhebungseinheit ist, möglichst viele Haushalte zu „definieren". So besteht die Möglichkeit, in Mehrgenerationen-Haushalten zwei oder gar drei Haushalte zu „identifizieren". Befinden sich also mehrere Erwachsene in einem Haus oder in einer Wohnung, die nicht miteinander verheiratet sind, ist es oft schwer zu entscheiden, ob diese einen oder mehrere Haushalte bilden. Interviewer des Mikrozensus werden sich verständlicherweise im Zweifel für getrennte Haushalte entscheiden, da sie mit der selben Arbeit mehr verdienen.

Ein Interviewer für das SOEP dagegen hat die Aufgabe, genau einen Haushalt an einer vorgegeben Adresse zu interviewen; er wird, „wenn er schon einmal da ist", ein Interesse daran haben, möglichst viele Personeninterviews zu führen. Abweichungen in der Haushaltsstruktur sind aufgrund der unterschiedlichen Entlohnungsprinzipien unvermeidlich, auch wenn die Interviewer sich an ihre Vorschriften halten, da Ermessensentscheidungen immer wieder verlangt werden.[11]

Wahrscheinlich schätzen weder das SOEP noch der Mikrozensus die Haushaltsstrukturen - gemessen an der Definition einer „gemeinsamen wirtschaftlichen Einheit" unverzerrt. Doch liegt - abgesehen von den sporadisch durchgeführten Volkszählungen - keine genauere Erhebung als der Mikrozensus vor. Deshalb ist es zweckmäßig, dass sich das SOEP an einigen wichtigen Eckgrößen des Mikrozensus orientiert. Deswegen wurde in der ersten Welle (1984) an die Haushaltsstrukturen des Mikrozensus angepasst. Die endogenen Dynamik des SOEP generierte aber tendenziell weniger Einpersonenhaushalte als der Mikrozensus. Die kumulierte Abweichung wurde nun in den 90er Jahren so groß, dass eine Randanpassung sinnvoll erscheint.

[11] Es wäre wünschenswert, wenn sich alle Begragungsinstitutionen - privat oder staatlich - auf prinzipiell gleiche Entlohnungsprinzipien einigen würden, um solche Effekte zu vermeiden.

Bei der Revision der Hochrechnung soll einerseits das Gesamtkonzept der Hochrechnung - nämlich die Fortschreibung über Bleibewahrscheinlichkeiten - erhalten bleiben, andererseits ist kritischen Einwendungen zu begegnen, die sich an auffälligen Abweichungen der SOEP-Querschnittsdaten von den entsprechenden des Mikrozensus stoßen. Deshalb werden den SOEP-Nutzern ab sofort Querschnitts-Hochrechnungsfaktoren bereitgestellt, bei denen - aufbauend auf der traditionellen sequentiellen Modellierung des Ausfallprozesses - wellenspezifisch eine sparsame Randanpassung an Eckdaten des Mikrozensus vorgenommen wird. Diese auf wenige Eckzahlen beruhende Anpassung bezieht sich auf die Erhebungen ab 1990, also seit Einbeziehung der neuen Ländern in das SOEP.

Diese Hochrechnungsfaktoren werden anhand zusätzlicher Annahmen generiert, mit denen einzelne Anwender nicht einverstanden sein mögen. Diesen Personenkreis wird die Möglichkeit gegeben, eigenständig Hochrechnungsfaktoren zu bestimmen. Dazu werden ihm **zeitinvariante Designgewichte**[12] zur Verfügung gestellt, auf deren Basis eigenständig der Ausfallprozess über die Zeit modelliert werden kann.

[12] Siehe hierzu Spiess (2000).

3 Prinzip der Neuberechnung der Querschnitts-Hochrechnungsfaktoren seit 1990

Folgende Voraussetzungen soll das erweiterte Konzept der Querschnitts-Hochrechnungsfaktoren für das SOEP erfüllen bzw. beibehalten:

1. Das Hochrechnungsverfahren wird nur auf Privathaushalte bzw. auf Personen in Privathaushalten angewendet. Die endogen errechneten Gewichte der Anstaltshaushalte, für die es keine zuverlässigen Daten in der amtlichen Statistik gibt, bleiben unverändert.
2. Ebenso bleiben die Gewichte, die für die isolierte Stichprobe D (xHHRFD und xPHRFD, mit x = K,...,O) ermittelt wurden, von der Revision ausgeschlossen.
3. Die Anpassung wird getrennt nach alten und neuen Ländern durchgeführt.
4. Angepasst wird an Ergebnisse des Mikrozensus.
5. Das ursprüngliche Hochrechnungskonzept bleibt erhalten. Somit werden die Veränderungen der Querschnittsgewichte über die Zeit weiterhin im wesentlichen auf die Bleibewahrscheinlichkeit zurückzuführen sein.
6. Es wird nur an einfache Ecksummen angepasst, Interaktionen bleiben unberücksichtigt, da diese bereits in der Modellierung des Ausfallprozesses Berücksichtigung finden.

4 Durchführung der sparsamen Anpassung

4.1. Die Eckdaten

Auf Haushaltsebene erfolgte eine Anpassung an die wohnberechtigte Bevölkerung in Privathaushalten auf Basis des Mikrozensus, und zwar einmal an die

Haushaltsgröße

- Einpersonenhaushalte
- Zweipersonenhaushalte
- Dreipersonenhaushalte
- Vierpersonenhaushalte
- Alle größeren Haushalte

zum anderen an
Privathaushalten mit Personen, die

- 15 bis 69 Jahre
- 70 Jahre und älter
- männlich und
- nicht deutscher Nationalität

waren. Zu beachten ist, dass letztere Merkmale ebenfalls als Eigenschaften von Haushalten und als Personenmerkmale zu betrachten sind. Auf diese Weise beeinflussen temporäre Ausfälle von Personen nicht die Hochrechnung für Haushalte. [13]

Auf Personenebene wurde wie bisher nur an die Bevölkerung in Privathaushalten am Hauptwohnsitz angepasst.

---

[13] Angenommen in einem Haushalt leben zwei Personen, Mann und Frau. Die Haushaltsbefragung wird erfolgreich durchgeführt, die Personenbefragung nur für die Frau, da der Mann auf Montage ist. Bei Anpassung auf Personenebene würde das Gewicht des Mannes auf Null gesetzt; die Gewichte der erfolgreich befragten Männer ceteris paribus entsprechend höher gesetzt. Bei Gewichtung auf Haushaltsebene dagegen bleibt der Einfluss des Mannes auf die Haushaltsgewichtung erhalten, da in die Hochrechnung ein Haushalt mit der Eigenschaft: Zweipersonenhaushalt, in dem ein Mann und eine Frau leben, eingeht.

Tabelle 1 zeigt eine Übersicht der verwendeten Eckdaten.

4.2 Die technische Umsetzung der Randanpassung

Es genügt nicht, die neuen Querschnitts-Hochrechnungsfaktoren sparsam an einige zusätzliche Ränder anzupassen. Ziel muss auch sein, möglichst wenige Eingriffe in die bisherige Hochrechnung vorzunehmen. Wie dieses umgesetzt wurde, wird im folgenden dargestellt. Zum leichteren Verständnis werden einige Definitionen eingeführt:

HHRFOLD(h,t) = Alter Hochrechnungsfaktor für Haushalt h in Welle t

HBLEIB(h,t) = Reziproker Wert für die Wahrscheinlichkeit, dass Haushalt h auch in Welle t in der Stichprobe verbleibt.

HHRF*(h,t) = Startwert für ADJUST für Haushalt h in der Welle t

HHRF(h,t) = Neuer Hochrechnungsfaktor für Haushalt h in Welle t

dHHRFOLD(h,t) = HHRFOLD(h,t)/HHRFOLD(h,t-1)
= Veränderungsfaktor des alten Hochrechnungsfaktors für Haushalt h von Welle t-1 nach Welle t

PHRFOLD(p,h,t) = Alter Hochrechnungsfaktor für Person p aus Haushalt h in Welle t

PHRF*(p,h,t) = Vorläufiger neuer Hochrechnungsfaktor für Person p aus Haushalt h in Welle t

PHRF(p,h,t) = Endgültiger neuer Hochrechnungsfaktor für Person p aus Haushalt h in Welle t

dPHRFOLD(p,h,t) = PHRFOLD(p,t)/PHRFOLD(p,t-1)
= Veränderungsfaktor des alten Hochrechnungsfaktors für Person p aus Haushalt h von Welle t-1 nach Welle t

Tabelle 1

Bevölkerung in Privathaushalten Deutschlands
von 1990 bis 1998

Alte Länder

	Haushalte in 1000								
Jahr	1990	1991	1992	1993	1994	1995	1996	1997	1998
Haushalte mit ... Persone(n)									
1	9849	10019	10171	10409	10702	10825	11092	11125	11097
2	8520	8730	8995	9191	9408	9612	9760	9893	10024
3	4712	4680	4715	4710	4618	4571	4500	4470	4402
4	3602	3644	3644	3658	3657	3618	3620	3637	3652
5 und mehr	1493	1510	1498	1528	1522	1518	1499	1491	1461
	Wohnberechtigte Bevölkerung in 1000								
Personen im Alter von									
15 bis unter 70 Jahren	47750	47999	48330	48709	48768	48752	48950	49016	48899
70 Jahren und älter	6161	6398	6630	6793	6981	7091	7163	7290	7397
Ausländer insg.	5085	5511	6031	6534	6724	6797	7019	7056	6981
Männer insg.	30838	31265	31716	32153	32339	32449	32612	32718	32702
	Bevölkerung am Hauptwohnsitz in 1000								
Insgesamt	62380	63207	64000	64770	65400	65401	65703	65950	65930

Quelle: Sonderauszählungen des Mikrozensus 1990-1998.

noch Tabelle 1

Bevölkerung in Privathaushalten Deutschlands von 1990 bis 1998

Neue Länder

	Haushalte in 1000								
Jahr	1990	1991	1992	1993	1994	1995	1996	1997	1998
Haushalte mit ... Persone(n)									
1	1896	1839	1873	1970	2045	2066	2099	2134	2200
2	1985	2132	2161	2198	2216	2246	2279	2328	2365
3	1387	1337	1303	1285	1284	1276	1269	1254	1241
4	1083	1098	1066	1040	1012	979	936	907	875
5 und mehr	301	266	254	240	231	227	227	225	215
	Wohnberechtigte Bevölkerung in 1000								
Personen im Alter von									
15 bis unter 70 Jahren	.	11416	11300	11367	11450	11445	11452	11498	11486
70 Jahren und älter	.	1355	1400	1426	1454	1478	1503	1555	1599
Ausländer insg.	.	112	137	184	216	98	136	146	157
Männer insg.	.	7608	7542	7558	7556	7540	7520	7519	7491
	Bevölkerung am Hauptwohnsitz in 1000								
Insgesamt	16313	15808	15622	15545	15500	15384	15302	15264	15179

Quelle: Sonderauszählungen des Mikrozensus 1990-1998.

Der Algorithmus besteht darin, dass für jede Welle im ersten Schritt neue Haushaltsgewichte bestimmt werden. Dazu werden getrennt nach alten und neuen Länder die alten Hochrechnungsfaktoren HHRFOLD(h,1990) als Startwerte für die Randanpassung, vorgegeben:

(1) HHRF*(h,1990) = HHRFOLD(h,1990)

Angepasst wird an die vom Mikrozensus vorgegebenen Eckdaten aus Tabelle 1 für das Jahr 1990 . Die Randanpassung erfolgt nach dem Prinzip des minimalen Informationsverlustes[14] und führt zu den neuen Hochrechnungsfaktoren HHRF(h,1990).

Im nächsten Schritt ist die Anpassung auf Personenebene vorzunehmen. Schritt (1) galt nur für das Startjahr 1990. Die folgenden Ausführungen gelten für jede nachfolgende Welle t. Prinzipiell sind die Hochrechnungsfaktoren für Haushalte und Personen identisch, solange nicht eine oder mehrere Personen eines Haushalts über einen Zweitwohnsitz verfügen und alle Befragungspersonen Interviews gewähren[15]. Aus diesem Grunde dürfen die alten Hochrechnungsfaktoren für Personen PHRFOLD(p,h,t) nicht unabhängig von der Haushaltsgewichtung an ihre Ecksumme angepasst werden. Daher wird in zunächst das Verhältnis der neuen Haushaltsgewichte zu den alten Haushaltsgewichte auf die Personengewichte übertragen:

(2) PHRF*(p,h,t) = PHRFOLD(p,h,t) * HHRF(h,t) / HHRFOLD(h,t) für jede Person p des Haushalts h.

Dann erst erfolgt die Randanpassung an die Bevölkerung in Privathaushalten am Hauptwohnsitz. Diese reduziert sich in diesem Fall auf die Berücksichtigung eines einfachen Korrekturfaktors, der sich aus dem Verhältnis der Bevölkerung am Hauptwohnsitz und der Summe aller vorläufigen Gewichte PHRF*(p,h,t) für jede Welle t ergibt.

(3) PHRF(p,h,t) = (PHRF*(p,h,t) * Korrekturfaktor

[14] Siehe hierzu Merz (1983).

[15] Diese Personen haben ungefähr die doppelte Auswahlwahrscheinlichkeit; ihr Gewicht wird deshalb halbiert.

Damit sind die neuen Querschnitts-Hochrechnungsfaktoren für Welle t erstellt.
Für 1991 und die folgenden Jahre müssen die neuen Startgewichte nach einem anderen Verfahren erzeugt werden. Sie müssen zwei Eigenschaften mitbringen: Zum einen hat die Bleibewahrscheinlichkeit BLEIB(h,t) als wesentliche Komponente Eingang in die Hochrechnung zu finden, zum anderen müssen die Änderungen in den neuen Hochrechnungsfaktoren aus der Vorperiode t-1 berücksichtigt werden.
Wie weiter oben erwähnt, ist bei der Fortschreibung der Querschnittsgewichte zwischen alten, neuen und vorübergehend nur in der Vorperiode ausgefallenen Haushalten zu unterscheiden. Die Hochrechnungsfaktoren für die alten Haushalte werden über die Bleibewahrscheinlichkeit, die übrigen über eine gesonderte Schätzung ihrer Auswahlwahrscheinlichkeit ermittelt. Für die erste Gruppe ergaben sich bisher die Gewichte der Welle t aus dem Produkt

(4) HHRFOLD(h,t) = const*HRFOLD(h,t-1)*BLEIB(h,t) mit const als Randanpassung.

Der Veränderungsfaktor ist somit

(5) dHHRFOLD(h,t) = const*BLEIB(h,t)

und enthält als wesentliche Komponente die Bleibewahrscheinlichkeit.

Es ist deshalb plausibel, die Startwerte für die Wellen ab 1991 wie folgt zu definieren:

(6a) HHRF*(h,t) = HHRF(h,t-1) * dHHRFOLD(h,t) / const, wenn HHRF(h,t-1) existiert

(6b) HHRF*(h,t) = HHRFOLD(h,t) / const ,wenn HHRF(h,t-1) nicht existiert

Damit liegen die Startgewichte für die Randanpassung vor. Sind die neuen Hochrechnungsfaktoren HHRF(h,t) ermittelt, wird mit der Ermittlung der Personengewichte ab Formel (2) fortgefahren. Das serielle Verfahren wird wiederholt, bis sämtliche Querschnitts-Hochrechnungsfaktoren bestimmt worden sind.

5 Stichprobe E

Im Jahr 1998 wurde die erste Welle der Ergänzungsstichprobe E erhoben. Sie umfasst 1067 befragte Haushalte, in denen 2064 Personen leben.[16] Die Stichprobe wurde für Privathaushalte repräsentativ angelegt, deshalb gab es vom Design her keine systematischen Über- oder Untererfassungen von Haushalten oder Personen. Theoretisch hatte jeder Haushalt, hatte jede Person die selbe Auswahlwahrscheinlichkeit.[17] Wie bereits weiter oben erwähnt, gibt es Unterschiede in der Feldarbeit zwischen dem mit der SOEP-Befragung betrautem Institut „Infratest Burke“ und dem Statistischen Bundesamt, dass den Mikrozensus erstellt. Dies führt in der Praxis zu Unterschätzungen der ausländischen Bevölkerung und der Einpersonenhaushalte. Somit ist eine Randanpassung auch für die Stichprobe E unerlässlich. Angesichts ihrer relativ geringen Fallzahl bietet es sich an, die selbe sparsame Anpassung an die Eckzahlen des Mikrozensus vorzunehmen, wie es für die Altstichproben beschrieben worden ist.

Als Startwerte für die Randanpassung wird der freie Hochrechnungsfaktor vorgegeben, der sich als Quotient aus der Zahl Privathaushalte insgesamt und der Zahl der Privathaushalte der Stichprobe E ergibt.

(1a) $HHRF^*(h,1998)_E = 37532000/1064 = 35208$

Als Ergebnis der Randanpassung erhält man die endgültigen Haushaltsgewichte $HHRF(h,1998)_E$. Diese bestimmen die vorläufigen Personengewichte:

(2a) $PHRF^*(p,h,1998)_E = HHRF(h,1998)_E$, wenn Person p keinen 2. Wohnsitz hat

$= HHRF(h,1998)_E / 2$ sonst.

Alle übrigen Schritte stimmen mit den oben genannten überein.

[16] Ein Haushalt wurde als Anstaltshaushalt ermittelt.

[17] Dies war bei den früheren Stichproben ebenso nur für die DDR-Stichprobe der Fall.

6 Zusammenführung der Stichproben A bis E

Die Hochrechnungen der Ergänzungsstichprobe E ist vollständig in den Hochrechnungsrahmen des SOEP integriert worden. Hierzu wurde eine „konvexe Gewichtung“ gewählt.[18] Die Konvexgewichte wurden so gewählt, dass die Varianz der gemeinsamen Schätzer möglichst klein wird. Mit der angebotenen Lösung erhält die Stichprobe E ein gegenüber ihrer Fallzahl überproportionales Gewicht, da sich aufgrund der selektiven Panelmortalität die Varianz der Schätzer in den Altstichproben im Laufe der Zeit vergrößert hat.

Für die korrekte Gewichtung und Zusammenführung der Altstichproben A bis D mit der Ergänzungsstichprobe E werden die Hochrechnungsfaktoren der Altstichproben ab 1998 mit dem Faktor 0,8, die der Stichprobe E mit 0,2 multipliziert[19]. Für getrennte deskriptive Analysen ist die Umrechnung leicht rückgängig zu machen.

[18] Siehe hierzu Spiess u. Rendtel (2000).

[19] Die Fallzahlen würden Faktoren von ca. 0,86 für die Altstichproben und 0,14 für die Stichprobe E implizieren.

Literatur

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*

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

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 (2000) „Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP), *DIW Diskussionspapier No. 197*

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