DIW Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 18

Eine konsistente Haushalts- und Personengewichtung für die DDR-Basisbefragung des SOEP

Rainer Pischner

Deutsches Institut für Wirtschaftsforschung, Berlin
German Institute for Economic Research, Berlin

Diskussionspapier Nr. 18

Eine konsistente Haushalts- und Personengewichtung für die DDR-Basisbefragung des SOEP

Rainer Pischner

Berlin, Januar 1991

Eine konsistente Haushalts- und Personengewichtung für die DDR-Basisbefragung des SOEP

1. Warum ist eine Gewichtung der Stichprobe erforderlich?

Anfang Juli 1990 endete die Feldarbeit für die DDR-Basiserhebung im Rahmen des Sozio-ökonomischen Panels (Ost)[1]. Diese erste Welle des SOEP-Ost konnte von Infratest Sozialforschung, München, trotz der sehr kurzen Vorbereitungszeit dem bundesdeutschen Standard entsprechend als bevölkerungsrepräsentative Stichprobe angelegt werden.

So sorgfältig auch Erhebungen geplant und durchgeführt werden, im Feld kommt es immer zu zufälligen Fehlern und systematischen Verzerrungen, die eine Beeinträchtigung der Stichprobenrepräsentativität zur Folge haben: So standen Infratest zu Beginn der Feldarbeit 3616 Adressen zur Verfügung, die über das für die DDR erstellte Mastersample durch Adressenziehung und Random-Walk ermittelt wurden[2]. Von diesem unbereinigten Adressenbestand wurden 162 Anschriften nicht benötigt, da die Interviewer bereits ihre erforderlichen acht Interviews in den jeweiligen Sample-Points realisieren konnten. Bei weiteren 50 Adressen fanden sie leere Wohnungen vor. Schließlich fielen vollständige Sample-Points mit insgesamt 290 Anschriften aus der Stichprobe, weil sie - bedingt durch die kurze Feldzeit - nicht rechtzeitig bearbeitet werden konnten. Diese 502 Ausfälle gelten zwar als qualitätsneutral, d.h. ohne Einfluß auf die Repräsentativität, doch dürften zumindest die 290 Totalausfälle zu Verzerrungen im Design der Stichprobe, d.h. zu regionalen Ungleichgewichten, geführt

1 Vgl. Schupp, J. und G. Wagner, Die DDR-Basisbefragung des Sozio-ökonomischen Panels.In:Vierteljahrsheft zur Wirtschaftsforschung, 1990 Heft 2-3, S. 152-159.

haben. Jedenfalls verblieben 3114 Adressen, die als bereinigtes Brutto die Ausgangsbasis der Befragung bildeten.

Nach Abschluß der Feldarbeit lagen für 2179 Haushalte und 4453 Personen ab 16 Jahren Interviews vor. Dies entspricht einer - auf das bereinigte Brutto bezogen - Ausschöpfungsquote von 70 vH. Von den ursprünglich 3616 Adressen, die das Gebiet der DDR repräsentativ abbildeten, wurden indes nur 60 vH realisiert. Zwar ist dies ein gutes, verglichen mit dem SOEP-West (57 vH) sogar ein sehr gutes Ergebnis, dennoch ist die Repräsentativität der Stichprobe durch die Ausfälle, wenn sie auch teilweise qualitätsneutral waren, systematisch beeinträchtigt. Hinzu kommen unsystematische Stichprobenfehler, die zufallsbedingt sind.

Obwohl unter bestimmten Bedingungen erwartungstreue Schätzungen auch aus nicht repräsentativen Stichproben gewonnen werden können[3], ist zumindest für die Analyse von Kontingenztabellen, die für eine leicht interpretierbare Darstellung von Sachverhalten unverzichtbar sind, eine Entzerrung der Stichprobenstrukturen, also eine Gewichtung erforderlich. Weiterhin sind zur Gewinnung von Aussagen über die Grundgesamtheit der Haushalte und Personen Hochrechnungsfaktoren erforderlich. Diese unterscheiden sich in der Regel aber nur durch einen einzigen Faktor, mit der die Gewichtungsvektoren noch zu multplizieren sind. Deshalb werden in diesem Aufsatz die Begriffe Gewichte und Hochrechnungsfaktoren synonym verwendet. Sollte eine Unterscheidung notwendig sein, wird explizit im Text darauf hingewiesen.

2. Die Stichprobenanlage

2.1 Die Grundgesamtheit

2 Vgl. Rosenbladt, B.v. und J. Schupp, Die DDR-Stichprobe des SOEP. In: Beiträge zur Arbeitsmarkt- und Berufsforschung Nr. 143.

3 Arminger, G.: Müssen Stichproben repräsentativ sein? In: Unver. Manuskript, Wuppertal.

Die Grundgesamtheit der DDR-Basisbefragung des SOEP bildeten alle Privathaushalte der damaligen DDR mit deutscher Bezugsperson bzw. deutschem Haushaltsvorstand und den in ihnen wohnenden Personen, die zum Befragungszeitpunkt wenigstens 16 Jahre alt waren.

Der Stichprobenplan basiert auf dem von Infratest/München erstellten Mastersample. Dieses ist als einstufige, geschichtete, mehrphasige Stichprobe aufgebaut:

- **Schichtung und Allokation:**

753 Sample-Points wurden proportional zur Einwohnerzahl durch Schichtung von 217 Land- bzw. Stadtkreisen und fünf Gemeindegrößenklassen gezogen.

- **Gemeindeauswahl:**

Die Gemeinden wurden durch eine systematische Auswahl mit Zufallsstart ermittelt.

- **Auswahl der Startadressen:**

Proportional zur Zahl der benötigten Sample-Points je Gemeinde wurden Personenadressen als Startadressen in systematischer Auswahl nach Zufallsstart aus der zentralen Einwohnerdatei gezogen.

- **Auswahl der Haushalte:**

Durch Random-Walk wurden ausgehend von jeder Startadresse zehn Befragungshaushalte und zwei Reservehaushalte ermittelt, von denen schließlich acht Interviews zu realisieren waren.

Eine ausführliche Beschreibung des Stichprobenplans ist zu finden im Methodenbericht von Infratest.[4]

4 Infratest Sozialforschung: Das Sozio-ökonomische Panel - Basiserhebung 1990 in der DDR. Methodenbericht zur Durchführung der Befragung. München 1990.

3. Methodische Grundlagen der Gewichtung

In diesem Aufsatz geht es vornehmlich um die Durchführung der Gewichtung bzw. Hochrechnung für die Basiserhebung des DDR-Panels. Deshalb wird der theoretische Teil, der für das SOEP an anderer Stelle erläutert wird, hier sehr knapp gehalten[5]:

Den Hochrechnungen sämtlicher Stichproben liegt der Grundgedanke von Horvitz/Thompson[6] zu Grunde, daß die inverse Auswahlwahrscheinlichkeit eines Stichprobenelements einen unverzerrten Schätzwert für den Hochrechnungsfaktor jener Elemente darstellt. Unter Auswahlwahrscheinlichkeit soll hier die Wahrscheinlichkeit verstanden werden, daß von einem beliebigen privaten Haushalt in der DDR mit deutscher Bezugsperson ein ausgefüllter Haushaltsfragebogen und Personenfragebogen von allen Personen, die 16 Jahre und älter sind, vorliegt. Sie setzt sich aus vielen (bedingten) Einzelwahrscheinlichkeiten zusammen. Z.B.:

- Auswahlwahrscheinlichkeit der Gemeinden, der Sample-Points
- Wahrscheinlichkeit der Kontaktaufnahme
- Antwortbereitschaft der Haushalte allgemein
- Spezifische Antwortbereitschaft von Haushalten und Personen

Hätte jeder Haushalt und jede Person dieselbe Auswahlwahrscheinlichkeit, dann wären sämtliche Gewichte identisch und gleich eins; die Hochrechnungsfaktoren ergäben sich als reziproker Auswahlsatz, also als Grundgesamtheit dividiert durch die Zahl der Stichprobenelemente. Wie bereits weiter oben angedeutet, sieht die Realität freilich anders aus. Bis tatsächlich die erforderliche Zahl von Interviews vorliegt,

5 Zu den theoretischen Grundlagen der Gewichtung und Hochrechnung für die Panels im Quer- und Längsschnitt sei verwiesen auf: Rendtel: Methodische Konzepte für die Hochrechnung von Panel-Daten, in: Vierteljahrshefte zur Wirtschaftsforschung, 1987, Heft 4, S. 278-290.

6 Horvitz,D. G., D. J. Thompson: A Generalization of Sampling Without Replacement from a Finite Universe, in: Journal of the American Statistical Association, 1952, Volume 47, S.663-685.

können vielerlei Einflüsse die Stichprobe verzerrt haben. Die wesentlichsten Fehlerquellen sind:

1. Die Verteilung der Sample-Points spiegelt regional nicht die wahre Haushalts- und Personenverteilung wider. **Abhilfe**: Anpassung der Gewichte für Sample-Points an die Regionalverteilung.

2. Zwar mögen die Sample-Points gleichmäßig über die DDR verteilt gewesen sein; wenn die Ausschöpfung dagegen ungleichmäßig ausfiel, werden bestimmte Gebiete ungleich stark erhoben. Es kommt wenigstens zu regional bedingten Verzerrungen. Das Verhältnis Netto- zu Bruttostichprobe wirkt sich also direkt auf die Gewichtung aus. **Abhilfe**: Anpassung der Gewichte für Sample-points um eben jenes Verhältnis.

3. Die Sample-Points sind stärker geklumpt als erwartet, Haushalte mit speziellen demographischen Eigenschaften werden dadurch über- oder unterrepräsentiert. **Abhilfe**: Berücksichtigung von demographischen Merkmalsverteilungen.

Für die Gewichtung der ersten Welle der Weststichprobe des SOEP sind diese Schritte durchgeführt worden. Dabei zeigte sich, daß der unter 2. genannte Ausgleich zwischen Brutto- und Nettostichprobe allein schon zu einer deutlichen Erhöhung der Varianz in den Hochrechnungsfaktoren führte, wodurch der Stichprobenfehler vergrößert wird. Da durch eine solche Anpassung die Stichprobe vornehmlich nur regional entzerrt wird, regionale Verteilungen aber an anderer Stelle berücksichtigt werden, wurde für die DDR-Basisbefragung zu Gunsten geringerer Stichprobenvarianz auf diesen Schritt verzichtet und folgendes Vorgehen gewählt:

Es werden für die Basisbefragung zwei Gewichtsvektoren benötigt: Haushalts- und Personengewichte. Diese sollen zumindest folgende Bedingungen erfüllen:

1. Da Personen und Haushalte dieselbe Auswahlwahrscheinlichkeit besitzen sollen, müssen die Gewichte

aller Personen eines Haushalts mit dem Haushaltsgewicht übereinstimmen[7].

2. Die Summe aller Hochrechnungsfaktoren für die Haushalte muß die Gesamtsumme aller Haushalte, die der Personen die Wohnbevölkerung in der DDR ergeben.

3. Die demographische Struktur der DDR soll möglichst gut auf die Stichprobe übertragen werden.

4. Die Varianz der Gewichte soll unter den Voraussetzungen 1 bis 3 minimal sein.

Es wurde versucht, dieses durch folgendes Vorgehen zu erreichen: Da das Design der Stichprobe bereits repräsentativ angelegt wurde, wäre im Idealfalle keine Designgewichtung erforderlich. Deshalb bietet es sich an, von einem Gewichtungsvektor für Haushalte auszugehen, der das Ergebnis einer freien Hochrechnung ist. Seine Varianz ist zunächst gleich Null. Anschließend wird der Vektor nach dem Prinzip der Minimierung des Informationsverlusts, also möglichst geringfügig korrigiert, so daß vorgegebene Randbedingungen erfüllt werden. Diese Randbedingungen werden im folgenden Abschnitt näher beschrieben.

4. Der Hochrechnungsrahmen

Jede Hochrechnung, jede Gewichtung kann auch bei bester methodischer Grundlage nur so gut sein wie der ihr zugrundegelegte Hochrechnungsrahmen d.h. so gut wie die Informationen über die anzustrebenden Randverteilungen und Ecksummen.

Seit Mitte 1989 verließen hunderttausende Bürger der DDR ihr Land, um sich in der Bundesrepublik eine neue Existenz aufzubauen. Die letzten zuverlässigen Angaben zur Demographie der Wohnbevölkerung datieren vom Jahresende 1989. Danach hatte die DDR zum 31.12.1989 eine Bevölkerungsbestand von 16,43 Mill. Personen. Die Wanderungen rissen danach nicht ab, auch nach der Befragung im Juni 1990 siedelten noch viele tausend DDR-Bürger in die Bundesrepublik über.[8] Es erhebt sich die

7 Beim SOEP werden alle Haushaltsmitglieder befragt, die 16 Jahre und älter sind. Die Verweigerung nur einer Befragungsperson führt zum Totalausfall des Haushalts. Daher weisen a po-

Frage, warum unter diesen Umständen überhaupt an Eckwerte angepaßt wird, die ohnehin überholt oder gar falsch sind. Im wesentlichen sind zwei Gründe zu nennen:

1. Der Zeitpunkt der DDR-Basiserhebung differiert zeitlich um ungefähr ein halbes Jahr von der aktuellen amtlichen Statistik der DDR zum Jahresende 1989. Im Normalfall, d.h. bei einer in ruhigen Bahnen verlaufenden Bevölkerungsbewegung, wären diese sechs Monate Differenz nahezu unerheblich. Die starken Übersiedlungen im ersten Halbjahr 1990 ergaben indes einen Wanderungsverlust zwischen einem und zwei vH, die sich nicht proportional auf die Bevölkerungsgruppen verteilen. Die Bevölkerungsstruktur änderte sich dementsprechend; freilich liegen keine exakten Informationen darüber vor.. Dennoch ist aus Gründen der Vergleichbarkeit mit anderen Analysen und Untersuchungen der Bezug auf offizielle Zahlen unverzichtbar, auch wenn diese Daten überholt sind. Es bleibt dem Anwender immer noch überlassen, auf Basis der Hochrechnung Aktualisierungen vorzunehmen.

2. Wie weiter unten noch an Beispielen gezeigt werden kann, werden auch in echten Zufallstichproben Stichproben Subpopulationen über- oder unterrepräsentiert. Diese Verzerrungen sind teilweise erheblich stärker ausgefallen als die Veränderungen, die durch die turbulente demographische Entwicklung in den ersten sechs Monaten von 1990 induziert wurden.

Der Hochrechnungsrahmen für die Basiserhebung des SOEP-Ost gliedert sich in drei Teile:

- Verteilung der Wohnbevölkerung nach Geschlecht und Bezirken,
- Verteilung der Wohnbevölkerung nach Geschlecht, Alter und Familienstand,
- Gesamtzahl der Haushalte in der DDR.

steriori Personen und Haushalte dieselbe Auswahlwahrscheinlichkeit auf.

8 Das statistische Amt der DDR geht von ca. 200 Tausend Personen bis Ende Juni 1990 aus, berücksichtigt aber nur ordnungsgemäß abgemeldete Personen. Dagegen spricht das Statistische Bundesamt Wiesbaden von ca. 300 Tausend Bürgern; hier dürften aber Doppelzählungen und Rückwanderungen nicht berücksichtigt sein. Die genannten Zahlen können deshalb als Unter- und Obergrenze der Wanderung im 1. Halbjahr 1990 interpretiert werden.

Die **regionale Gewichtung**, die nach Bezirken vorgenommen wird, weist einige Besonderheiten auf, die erklärungsbedürftig sind:

Im Gegensatz zur Bundesrepublik war die DDR politisch in 15 Bezirke geteilt und nicht nach Ländern gegliedert. Die Bezirke wurden erst mit der Vereinigung im Oktober 1990 durch folgende Zusammenlegung wieder in Länder zurückgeführt.

Länder	**Bezirke**
Mecklenburg-Vorpommern	Rostock, Schwerin, Neubrandenburg
Brandenburg	Potsdam, Frankfurt/Oder, Cottbus
Sachsen-Anhalt	Magdeburg, Halle
Thüringen	Erfurt, Gera, Suhl
Sachsen	Dresden, Leipzig, Chemnitz
nachrichtlich:	
Berlin	Berlin(West) und Berlin(Ost)

Allerdings wurde die Basisbefragung am 30.6.90 zu einem Zeitpunkt abgeschlossen, als die endgültige Zuordnung einiger Kreise zu den neuen Ländern wegen anstehender Volksentscheidungen noch nicht erfolgt war. Die dunklen Flächen in Abbildung 1 zeigen die Regionen, die sich für eine Eingliederung in ein anderes Bundesland entschieden haben und deshalb nicht den oben aufgeführten Zuordnungen entsprechen.

Dies schlug sich auch in der Stichprobe nieder: Von 2179 Netto-Haushalten gibt es in 71 Fällen (3,3 vH) Abweichungen zwischen den Variablen Bezirk und Land.

In Tabelle 1 ist die Verteilung der Haushalte in den Bezirken der der Länder gegenübergestellt worden. In den Blöcken auf der Diagonalen der Tabelle befinden sich die "Normalfälle", die Abweichungen hiervon sind in Klammern gesetzt. Solche Differenzen konnten in der amtlichen Statistik der DDR zum 13.12.89 definitionsgemäß nicht vorkommen. Aus diesem Grunde war es erforderlich, die regionale Anpassung nach Bezirken und nicht nach den neuen Ländern vorzunehmen, zumal die

Abb 1: Länder- und Bezirksgrenzen im Gebiet der DDR 1990

Quelle: Die Wirtschaft, Heft 29, 1990.

Bezirksstruktur auch wesentlich feiner als die Länderstruktur ist.

Tabelle 1

Regionalverteilung der Netto-Haushalte in der Basiserhebung für die DDR im Jahre 1990

Länder Bezirke	Ins-. gesamt	Meckl. Vor-. pommern	Bran-den-burg	Sach-sen-Anhalt	Thue-rin-gen	Sach-sen	Berlin (Ost)
Insgesamt	2179	237	346	401	358	670	167
Rostock	107	107	0	0	0	0	0
Schwerin	80	64	(16)	0	0	0	0
Neubrandenburg	74	66	(8)	0	0	0	0
Potsdam	134	0	134	0	0	0	0
Frankfurt/Oder	85	0	85	0	0	0	0
Cottbus	127	0	103	0	0	(24)	0
Magdeburg	209	0	0	209	0	0	0
Halle	199	0	0	192	(7)	0	0
Erfurt	127	0	0	0	127	0	0
Gera	109	0	0	0	109	0	0
Suhl	99	0	0	0	99	0	0
Dresden	209	0	0	0	0	209	0
Leipzig	190	0	0	0	(16)	174	0
Chemnitz	263	0	0	0	0	263	0
Berlin(Ost)	167	0	0	0	0	0	167

Quelle: Das Sozio-ökonomische Panel (Ost).

Für die 15 Bezirke der DDR weist das Statistische Amt der DDR für Ende 1989 die Wohnbevölkerung nach Geschlecht aus. Hieraus ergeben sich 2*(15-1) = 28 Restriktionen[9].

Tabelle 2 ist die vorgegebene Regionalverteilung zu entnehmen: Für jeden Bezirk - getrennt nach dem Geschlecht - sind folgende Merkmale aufgelistet: Zahl der Personen am 31.12.1989, Anzahl der Personen in der Stichprobe und das Verhältnis aus beiden Größen (= inverse Auswahlwahrscheinlichkeit).

Tabelle 2

Wohnbevölkerung in der DDR und im SOEP -Ost
- nach Bezirken und Geschlecht -

Bezirke	Männer			Frauen		
	Personen DDR	Stichprobe SOEP	DDR/SOEP	Personen DDR	Stichprobe SOEP	DDR/SOEP
Rostock	444731	162	2745	465109	169	2752
Schwerin	285808	123	2324	304363	102	2984
Neubrandenburg	301168	110	2738	314599	109	2886
Potsdam	537263	186	2889	573947	188	3053
Frankfurt/Oder	344944	118	2923	361172	126	2866
Cottbus	426093	181	2354	449488	199	2259
Magdeburg	594518	284	2093	643389	303	2123
Halle	836539	251	3333	911491	281	3244
Erfurt	586458	187	3136	636439	187	3403
Gera	347943	135	2577	380136	144	2640
Suhl	262320	138	1901	282951	155	1825
Dresden	812012	281	2890	901074	313	2879
Leipzig	631097	259	2437	702045	270	2600
Chemnitz	853468	326	2618	964019	338	2852
Berlin(Ost)	611244	208	2928	670274	211	3177
Insgesamt	7873300	2949	2670	8560496	3095	2766

Quellen: Das Sozio-ökonomische Panel (Ost), Statistisches Amt der DDR.

9 Da im folgenden Restriktionsblock die Gesamtzahl der Männer und Frauen in der DDR definiert wird, ist die Verteilung nach Bezirken und Geschlecht schon durch 14 Bezirke hinreichend definiert.

Insgesamt wies die Statistik Ende 1989 eine Wohnbevölkerung von 16433796 Personen aus. Einschließlich der Kinder umfaßt das SOEP-Ost 6044 Personen. Damit ergibt sich eine durchschnittliche Auswahlwahrscheinlichkeit von 0,03678 vH, d.h. eine Person im SOEP-Ost repräsentiert 2719 Personen der Gesamtbevölkerung. Diesert Wert wäre der Faktor für eine freie Hochrechnung. Getrennt nach den Geschlechtern ergeben sich die in der Tabelle ausgewiesenen Werte. Sie weichen nur geringfügig von dem Durchschnittswert ab. Die Männer sind mit einem Faktor von 2670[10] um 1,8 vH überrepräsentiert, die Frauen mit einem Faktor von 2766 um etwa 1,7 vH untererfaßt[11].

Regional gibt es weitaus größere Abweichungen. Zu 20 vh am stärksten unterrepräsentiert sind die Frauen im Bezirk Erfurt (HF=3403), Mit fast 50 vH über dem Soll wurden im Bezirk Suhl die Frauen übererfaßt (HF=1825).

Zu den typischen Randverteilungen, an die Stichproben angepaßt werden, zählt diejenige von Personen nach **Alter, Geschlecht und Familienstand.**

Tabelle 3 zeigt die demographischen Restriktionen, die für die Gewichtung berücksichtigt wurden. In ihr sind nachrichtlich auch die ledigen Personen aufgeführt, obwohl sich diese Daten aus den übrigen Angaben ergeben und deshalb in den Hochrechnungsrahmen nicht eingehen dürfen. Die Ergebnisse vermitteln folgendes Bild: Wie schon erwähnt, sind die Frauen um ca. 1,7 vH untererfaßt[12], eine Abweichung, die angesichts der erwähnten stürmischen Entwicklung nicht sehr ins Gewicht

10 Man könnte diesen Wert auch als "empirischen Hochrechnungsfaktor" bezeichnen, da die Multiplikation der Stichprobenelemente mit diesem Faktor ex definitione die vorgegebene Randsumme ergibt.

11 Generell gilt: Liegen die Werte in den Spalten "DDR/SOEP" über dem Durchschnittswert von 2719, liegt Untererfassung vor, da ein Stichprobenelement mehr Personen zu repräsentieren hat. Werte kleiner als 2719 zeugen von einer überproportionalen Ausschöpfung der jeweiligen Merkmalskombination.

12 Diese und die weiteren prozentualen Abweichungen ergeben sich durch Vergleich der Ist- und Sollzahlen in der Stichprobe: Insgesamt wurden 3095 Frauen erhoben. Aufgrund der oben er-

Tabelle 3

Wohnbevölkerung in der DDR und im SOEP -Ost
- nach Familienstand, Alter und Geschlecht -

Altersklassen	Personen DDR	Stichprobe SOEP	DDR/SOEP	Personen DDR	Stichprobe SOEP	DDR/SOEP
Männer und Frauen insgesamt						
0 -4 J.	545191	229	2381	517254	188	2751
5 -9 J.	577826	272	2124	551517	246	2242
10 -14 J.	518660	264	1965	491694	240	2049
15 -19 J.	502511	195	2577	478014	199	2402
20 -24 J.	633170	194	3264	596900	190	3142
25 -29 J.	719458	247	2913	672170	274	2453
30 -34 J.	649482	270	2405	611288	285	2145
35 -39 J.	631305	247	2556	601112	254	2367
40 -44 J.	414361	197	2103	404597	209	1936
45 -49 J.	579616	203	2855	582704	186	3133
50 -54 J.	598589	216	2771	609967	214	2850
55 -59 J.	456572	158	2890	482116	155	3110
60 -64 J.	353307	100	3533	474354	134	3540
65 -69 J.	250538	68	3684	468452	120	3904
70 -74 J.	126770	34	3729	261042	70	3729
75 -79 J.	162990	33	4939	369108	69	5349
80 -99 J.	152954	22	6952	388207	62	6261
Insgesamt	7873300	2949	2670	8560496	3095	2766
Ledige Männer und Frauen*						
0 -4 J.	545191	229	2381	517254	188	2751
5 -9 J.	577826	272	2124	551517	246	2242
10 -14 J.	518660	264	1965	491694	240	2049
15 -19 J.	500938	194	2582	467691	199	2350
20 -24 J.	510339	161	3170	344060	109	3157
25 -29 J.	283954	79	3594	137160	58	2365
30 -34 J.	117518	33	3561	58982	19	3104
35 -44 J.	100277	27	3713	56829	13	4371
45 -54 J.	59509	11	5410	53287	7	7612
55 -99 J.	29264	6	4877	171088	39	4387
Insgesamt	3243476	1276	2542	2849562	1118	2549
nachrichtlich:						
0 -19 J.	2142615	959	2234	2028156	873	2323
20 -99 J.	1100861	317	3473	821406	245	3353

* Ledige Männer und Frauen wurden nicht direkt in den Hochrechnungsrahmen einbezogen. Sie sind hier nur der Vollständigkeit halber aufgeführt.

wähnten Auswahlwahrscheinlichkeit von 0,03678 vH wären im SOEP-Ost 3149 Frauen zu erwarten gewesen. Somit sind 54 von 3149, also ca. 1,7 vH Frauen, zu wenig erfaßt worden.

noch Tabelle 3

Verheiratete Männer und Frauen

	Männer			Frauen		
Altersklassen	Personen DDR	Stichprobe SOEP	DDR/SOEP	Personen DDR	Stichprobe SOEP	DDR/SOEP
0 -24 J.	116913	34	3439	244999	75	3267
25 -29 J.	392229	156	2514	475388	202	2353
30 -34 J.	467143	218	2143	478827	239	2003
35 -39 J.	490313	215	2281	480438	218	2204
40 -44 J.	336099	173	1943	325393	182	1788
45 -49 J.	481230	184	2615	467694	162	2887
50 -54 J.	513394	198	2593	478856	173	2768
55 -59 J.	402822	145	2778	352096	110	3201
60 -64 J.	315007	87	3621	299208	97	3085
65 -69 J.	219497	61	3598	222942	57	3911
70 -74 J.	105647	28	3773	87641	24	3652
75 -99 J.	201270	36	5591	127947	16	7997
Insgesamt	4041564	1535	2633	4041429	1555	2599
Verwitwete Männer und Frauen						
0 -49 J.	10678	5	2136	39826	13	3064
50 -59 J.	24339	6	4057	98730	25	3949
60 -64 J.	17739	9	1971	99229	17	5837
65 -69 J.	20406	6	3401	166902	43	3881
70 -74 J.	16410	5	3282	136663	38	3596
75 -79 J.	37412	3	5776	238233	53	4495
80 -99 J.	66192	13	6637	297646	50	5953
Insgesamt	193176	47	4110	1077229	239	4507
Geschiedene Männer und Frauen						
0 -29 J.	50307	12	4192	75586	19	3978
30 -34 J.	63895	18	3550	69871	27	2588
35 -39 J.	69267	14	4948	76088	26	2926
40 -44 J.	46025	13	3540	52792	16	3299
45 -49 J.	59551	10	5955	71213	19	3748
50 -54 J.	48678	10	4868	62413	26	2401
55 -99 J.	57361	14	4097	184313	50	3686
Insgesamt	395084	91	4342	592276	183	3236
nachrichtlich nur für Frauen:*						
55 -64 J.	.	.	.	80822	31	2607
65 -99 J.	.	.	.	103491	19	5447

Quellen: Das Sozio-ökonomische Panel (Ost), Statistisches Amt der DDR.

* Diese Unterteilung wurde nur für die Frauen vorgenommen, da die Zellenbesetzung bei den Männern in diesen Altersklassen zu gering war.

fällt. Anders ist die Lage zu beurteilen, wenn zur Bewertung der Repräsentativität die Verteilung nach dem Familienstand herangezogen wird. Hier heben sich die Abweichungen von den vorgegebenen andverteilungen deutlicher hervor. So sind Verheiratete um ca. 4 VH überrepräsentiert, für diese Gruppe jedoch kein schlechtes Ergebnis. Auch für die Ledigen errechnet sich insgesamt nur eine relativ geringe Abweichung vom Erwartungswert (ca. +6 vH). Dieses Ergebnis täuscht aber über große strukturelle Abweichungen innerhalb dieser Gruppe hinweg: Die jüngeren Ledigen, die überwiegend noch bei ihren Eltern leben, sind mit 19,4 vH stark überrepräsentiert, die ältere Gruppe mit -20,5 vH entsprechend untererfaßt. Die im Saldo gemessene Übererfassung der Ledigen und Verheirateten hat eine entsprechende Untererfassung der Verwitweten und der Geschiedenen zur Konsequenz. Diese Subpopulation umfaßt mit 2,26 Mill. Personen etwa ein Siebtel der Wohnbevölkerung. Somit errechnet sich eine Minderausschöpfung von nahezu einem Drittel. Statt der zu erwartenden gut 830 verwitweten oder geschiedenen Personen, befinden sich im SOEP-Ost nur 560 Frauen und Männer aus dieser Gruppe. Diese Untererfassung ist nicht - wie vielleicht zu erwarten - allein auf das höhere Durchschnittsalter jenes Bevölkerungskreises zurückzuführen, zieht sich die Minderausschöpfung doch durch fast alle Altersklassen.

Die Alterstruktur der Stichprobe zeigt das erwartete Bild: Deutlich unterrepräsentiert sind - wie bei fast allen Erhebungen - vornehmlich ältere Bürger. Der Bruch zeigt sich etwa ab dem 60. Lebensjahr, ein Alter ab dem viele Menschen in Rente gehen, gesundheitlich nicht mehr auf voller Höhe sind und allmählich weniger aktiv am Leben teilnehmen. Deutlich erkennbar zeigt sich der steile Anstieg des empirischen Hochrechnungsfaktors. Zu niedrige Ausgeschöpfungsquoten finden sich bei den jungen Menschen um Zwanzig, die erfahrungsgemäß überdurchnittlich oft verweigern. Dagegen sind Bürger zwischen 30 und 45 überrepräsentiert.

Insgesamt steuerten diese Randverteilungen mit 86 Eckwerten den größten Teil zur Gesamtzahl der Restriktionen bei.

Die Datenbasis für die **Gesamtzahl der Haushalte** und ihre Struktur für das Jahr 1990 steht auf sehr tönernen Füßen. Vornehmlich der Anteil der Ein-Personenhaushalte ist ungesichert.

Die amtliche Statistik bietet lediglich Ergebnisse der letzten Volkszählung aus dem Jahre 1981 sowie eine "amtliche Fortschreibung" an.[13] Allein Wohnungsstatistiken jüngeren Datums ergänzen dieses Material.[14] Danach schwanken die Anteile der Ein-Personenhaushalte an allen Privathaushalten zwischen 25,1 vH (Volkszählung 1981) und 29,1 vH (amtliche Fortschreibung 1988). Letzlich sind die amtlichen Zahlen aus der DDR schon vor Beginn der Übersiedlerwelle seit Herbst 1989 nicht sehr verläßlich. Die Auswirkungen der starken Wanderungsbewegungen auf die Haushaltsstrukturen sind schon gar nicht quantifizierbar. Gleichwohl ist der Anteil von 15,6 vH der Ein-Personenhaushalte im ungewichteten SOEP mit Sicherheit zu gering.

Für die Hochrechnung besteht allerdings auch das Problem, daß es zuverlässige Zahlen über die Haushaltsstruktur in der DDR nicht gibt. Deshalb soll die Stichprobe nur an die vom DIW geschätzte Ecksumme von 6794 Tausend Haushalten angepaßt werden[15]. Es war Tabelle 3 zu entnehmen, daß die Gruppe der Ledigen, die 20 Jahre und älter sind, stark unterrepräsentiert ist. Dies ist eine Subpopulation, die erfahrungsgemäß häufig alleine wohnt. Damit ist die Wahrscheinlichkeit groß, daß durch eine Hochrechnung, die Geschlecht, Familienstand und Alter berücksichtigt, auch der Anteil der Ein-Personen-

13 Die Einkommensverteilung nach Haushaltsgruppen in der ehemaligen DDR. Bearb.:K.-D. Bedau und H. Vortmann. In: Wochenbericht des DIW, Nr. 47/90, S.656.

14 Privathaushalte und Wohnungsbedarf in Deutschland bis zum Jahr 2000. Bearb.:B. Bartholmai,M. Melzer und E. Schulz. In: Wochenbericht des DIW, Nr. 42/90, S.591-598.

26 Bedau, K.-D. und H. Vortmann,ebenda.

haushalte steigt. Unter Umständen erübrigt sich durch diese Kollinearität eine Korrektur der Haushaltsstruktur.

Mit der Festlegung der Gesamtzahl der Haushalte erhöht sich die Zahl der Restriktionen um eine auf insgesamt 115. Sollte künftig eine bessere Stützung der Haushaltszahl vorliegen, kann diese durch eine einfache skalare Multiplikation berücksichtigt werden. Sollten noch zuverlässige Haushaltsstrukturinformationen bekannt werden, so würden diese ggf. eine neue Hochrechnung erfordern, wenn diese Struktur signifikant abweichen würde[16].

Nach Bestimmung der Haushaltsgewichte werden diese auf sämtliche Personen einschließlich der Kinder unter 16 Jahren der zugehörigen Haushalte übertragen. Auf diese Weise ist die Identität von Haushalts- und Personengewichte immer gewährleistet. Dagegen führt eine von Infratest vorgenommene Hochrechnung, die bei Übergabe der Rohdaten mitgeliefert wird, nicht zu gleichen Gewichten für Personen und Haushalte. Daher werden diese Hochrechnungsfaktoren nicht weiter verwendet[17].

5. Ergebnisse der Gewichtung

Die Auswirkungen der Gewichtung werden an einigen Beispielen demonstriert, indem regionale und demographische Verteilungen vor und nach Gewichtung gezeigt und knapp kommentiert werden. Zunächst jedoch werden einige statistische Eigenschaften der Gewichte betrachtet.

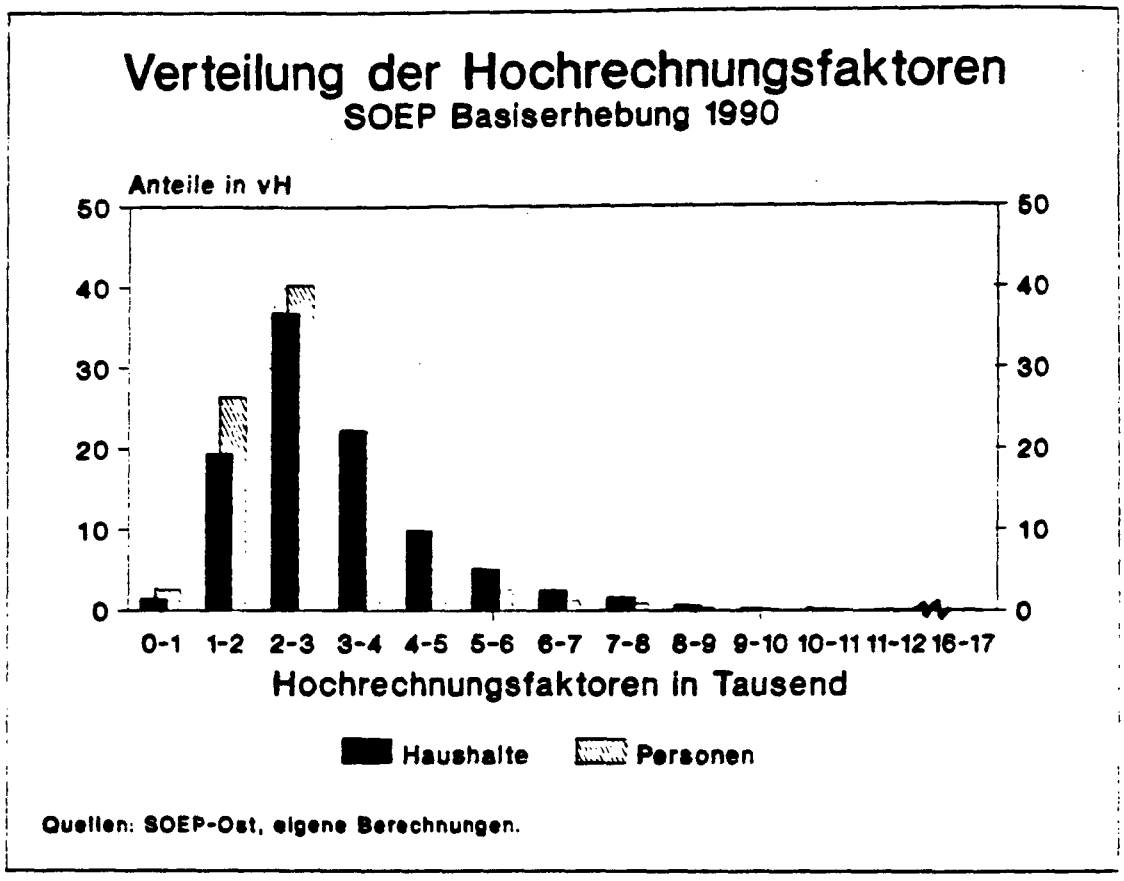
16 Unter Umständen wird man anhand einer Hochrechnung für die 2. oder gar 3. Welle eine Rückrechnung auf die Basiserhebung vornehmen können. Dies setzt allerdings parallele Mikrozensusergebnisse für die neuen Bundesländer und den Ostteil Berlins voraus.

17 Infratest berücksichtigt außerdem nicht den Familienstand für die Gewichtung. Bei rein personenbezogenen Auswertungen, insbesonderere Analysen zur Erwerbsbevölkerung, liefern die Infratest-Gewichte dennoch nahezu gleiche Ergebnisse wie die hier errechneten Hochrechnungsfaktoren.

5.1 Statistische Eigenschaften der Hochrechnungsfaktoren

Definitionsgemäß sind für jeden Haushalt die Personen und Haushaltsgewichte identisch. Unterschiedliche Haushaltsgrößen führen indes zu leicht differierenden Verteilungen. Sie sind in Abbildung 2 dargestellt.

Abb. 2



Quellen: SOEP-Ost, eigene Berechnungen.

Der niedrigste Wert für einen Hochrechnungsfaktor wurde mit 364 ermittelt. Dies heißt, daß die betreffenden Personen und Haushalte nur mit einem Achtel des Durchschnittsgewicht bei Analysen und Auswertungen berücksichtigt werden. Nur 31 Haushalte (1,4 vH) und 157 Personen (2,6 vH) repräsentieren weniger als 1000 Mitbürger und besitzen somit nicht mehr als ein Drittel des Durchschnittsgewichts.

Auswertungen von Teilpopulationen bergen bei allzu großer Varianz der Hochrechnungsfaktoren verstärkt die Gefahr von Verzerrungen. Dies gilt insbesondere für außerordentlich hohe Gewichte. Diese kommen in der Basisbefragung DDR glücklicherweise kaum vor. Nur ein einziger Ein-Personenhaushalt wird mit dem Gewicht von knapp 16800 bewertet und übertrifft den Durchschnittswert um das 5,5-fache. Sechs Haushalte (0,3 vH) mit insgesamt acht Personen (0,1 vH) repräsentieren mehr als 10000 Haushalte bzw. Personen.

So drängt sich erfreulicherweise der weitaus größte Teil der Gewichte um ihren Mittelwert. Fast 60 vH liegen im Bereich von 2000 bis 4000; mehr als 95 vH fallen in das Intervall 1000 bis 7000. Dies wirkt sich auch in den Standardabweichungen aus: Sie beträgt für die Haushalte 1511 und für die Personen 1277.

Die wichtigsten statistischen Kennziffern der Hochrechnungsfaktoren auf einen Blick:

Kennziffer	**Haushalte**	**Personen**
Anzahl	2179	6044
Summe	6793998	16433794
Arithmetisches Mittel	3117,9	2719,0
Standardabweichung	1510,9	1277,1
Minimum	363,9	363,9
Maximum	16798,4	16798,4

5.2 Die Regionalverteilung

In diesem und im folgenden Abschnitt werden die Ergebnisse der Hochrechnung, die mit den vorgegebenen Randverteilungen übereinstimmen, denen einer freien Hochrechnung gegenübergestellt. Freie Hochrechnung heißt für die Personendaten, daß die Zahl der Stichprobenelemente mit dem arithmetischem Mittel der Personengewichte, also mit 2719,0, multipliziert werden. Im wesentlichen werden die im Kapitel

"Hochrechnungsrahmen" bereits diskutierten Eigenschaften graphisch verdeutlicht (Abbildungen 3a und 3b).

Ersichtlich ist die demographische Verteilung durch die freie Hochrechnung (schraffierte Flächen) gut wiedergegeben. Größere Abweichungen von der "wahren" Verteilung (schwarze Flächen) zeigen sich nur in den Bezirken Magdeburg, Suhl, Halle und Cottbus.

5.3 Die demographische Verteilung

Die Abbildungen 4a und 4b zeigen die Verzerrungen in der Altersstruktur, wenn man frei hochrechnet. Auffällig ist der Geburtenausfall nach dem 2. Weltkrieg, der sich 1989/90 in der Klasse der 40-44 jährigen zeigt. Er wird in der Männer-Stichprobe zu schwach, bei den Frauen kaum nachgezeichnet. Besonders hebt die Graphik für die Frauen die unterschiedlichen Ausschöpfungsraten der älteren Bürger gegenüber denen der 30 bis 45-jährigen hervor.

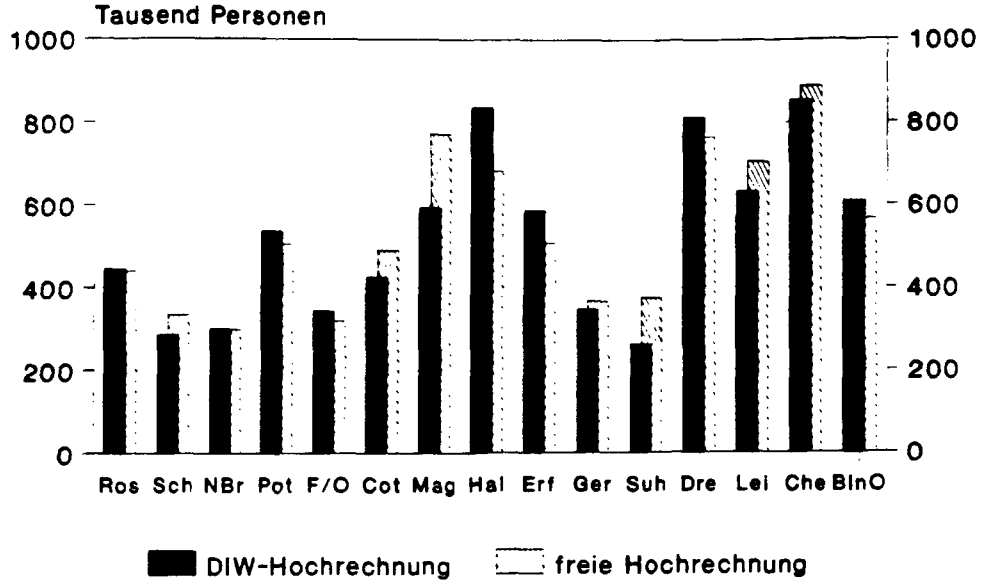
In den weiteren Abbildungen sind Aufgliederungen nach dem Familienstand vorgenommen werden. Auch diese Graphiken erklären sich von selbst. Erwähnt werden soll lediglich noch Abbildung 7b, die die verwitweten Frauen betrifft. Hier sieht man die generelle und die mit dem Alter zunehmende Untererfassung dieser Subpopulation.

5.4 Die Haushaltstruktur

Als "Restriktion" ist in die Hochrechnung für die Anpassung der Haushalte nur die Ecksumme eingegangen. Die Gründe hierfür sind im Abschnitt "Hochrechnungsrahmen" genannten worden. Die Verteilung nach Haushaltsgröße mußte mangels geeigneter Rahmendaten allein dem Stichprobenergebnis sowie den

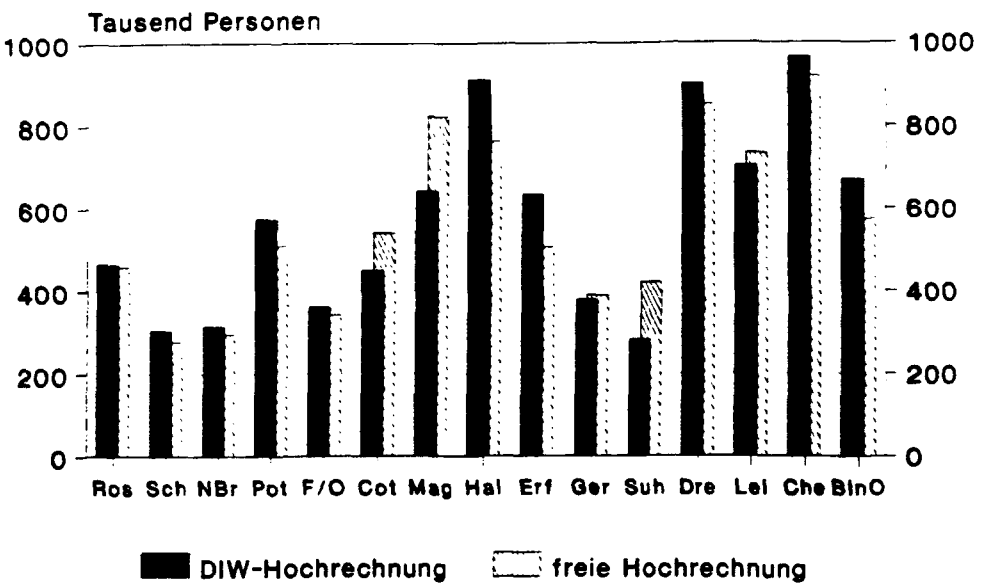
Abb 3a und 3b

Männer in der DDR nach Bezirken
Hochrechnung: SOEP Basiserhebung '90



Quellen: SOEP-Ost, eigene Berechnungen,
Statistisches Amt der DDR.

Frauen in der DDR nach Bezirken
Hochrechnung: SOEP Basiserhebung '90

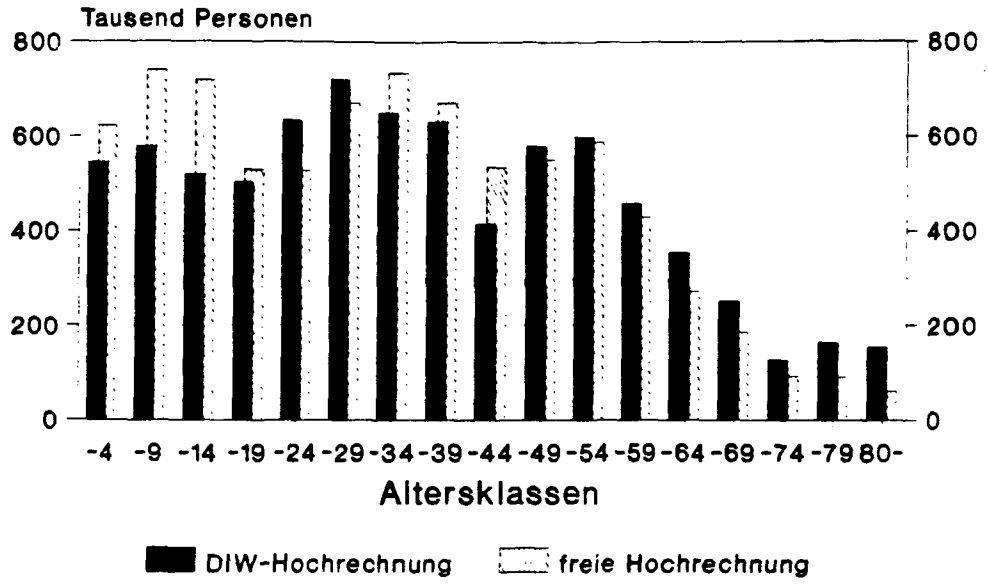


Quellen: SOEP-Ost, eigene Berechnungen,
Statistisches Amt der DDR.

Abb. 4a und 4b

Männer in der DDR nach Alter

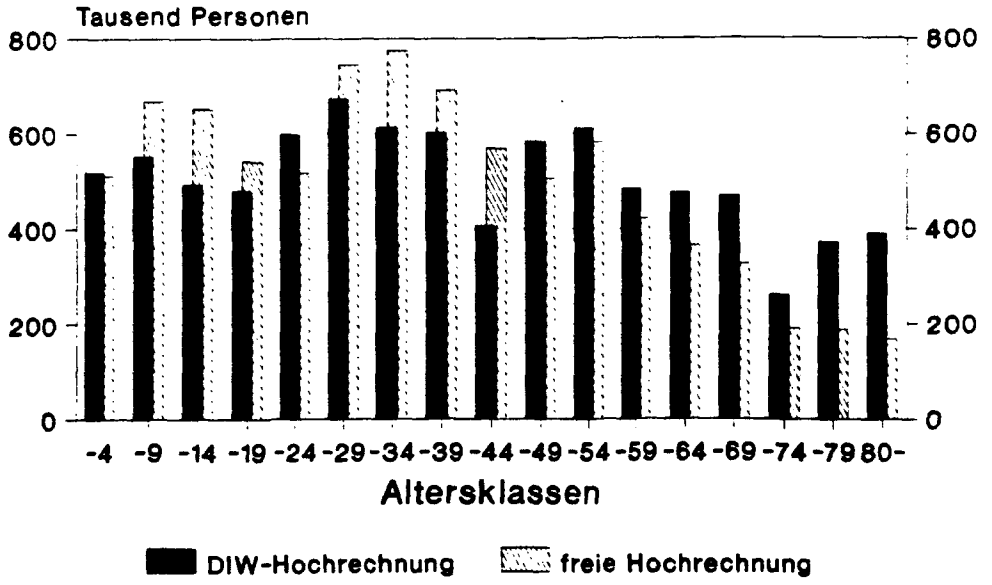
Hochrechnung: SOEP Basiserhebung '90



Quellen: SOEP-Ost, eigene Berechnungen,
Statistisches Amt der DDR.

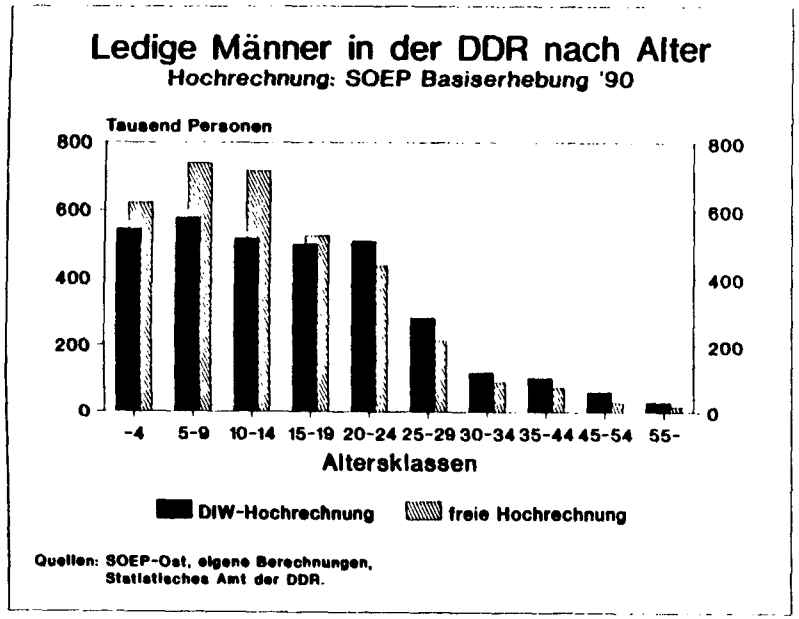
Frauen in der DDR nach Alter

Hochrechnung: SOEP Basiserhebung '90



Quellen: SOEP-Ost, eigene Berechnungen,
Statistisches Amt der DDR.

Abb. 5a und 5b



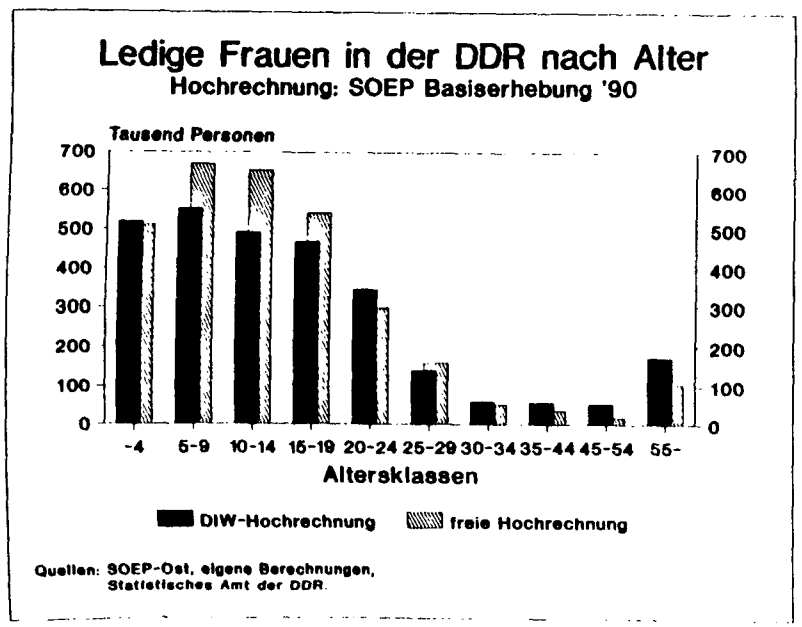
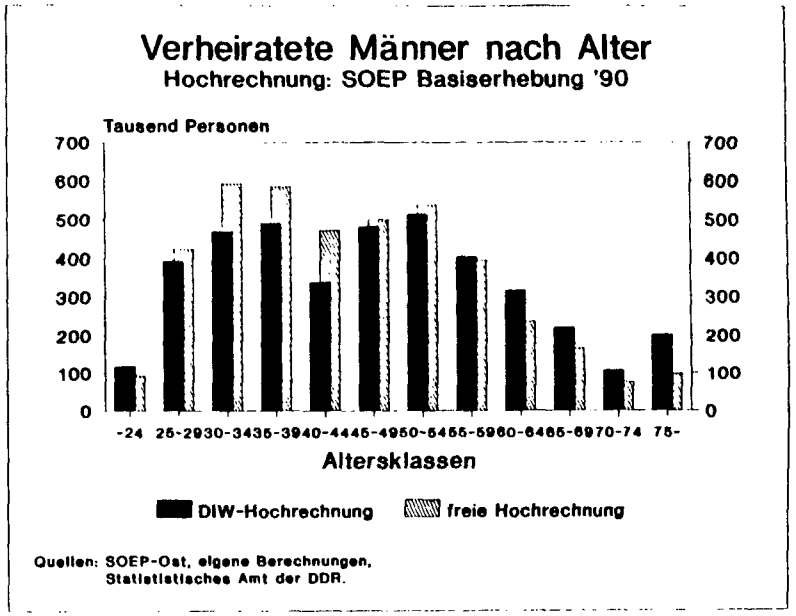


Abb. 6a und 6b



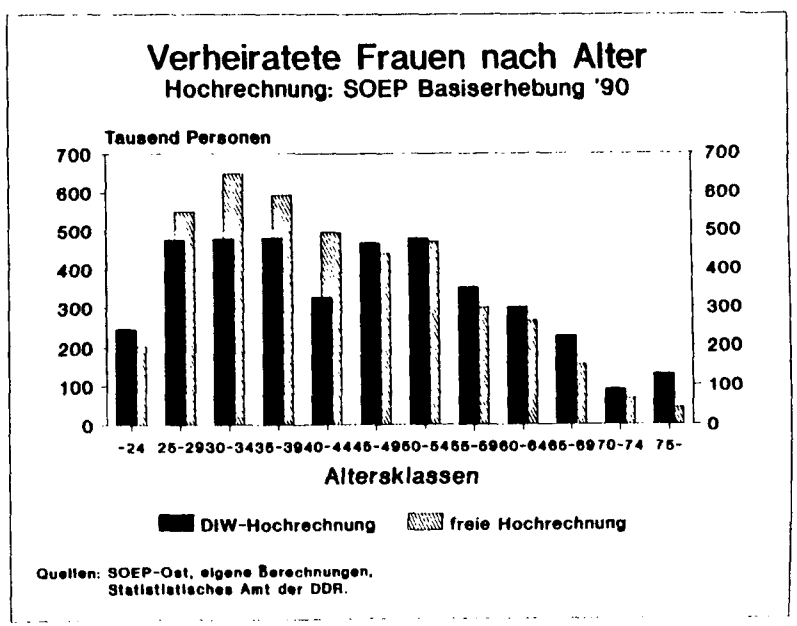
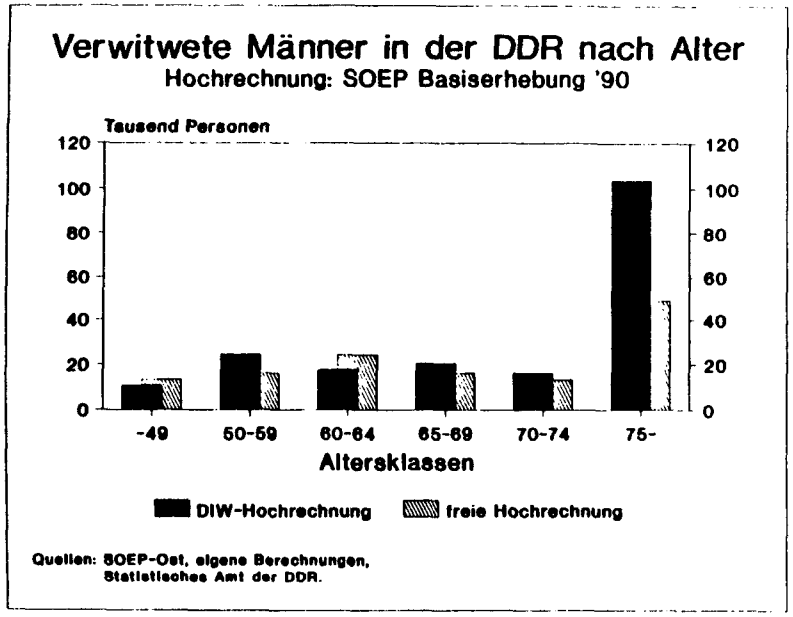


Abb. 7a und 7b



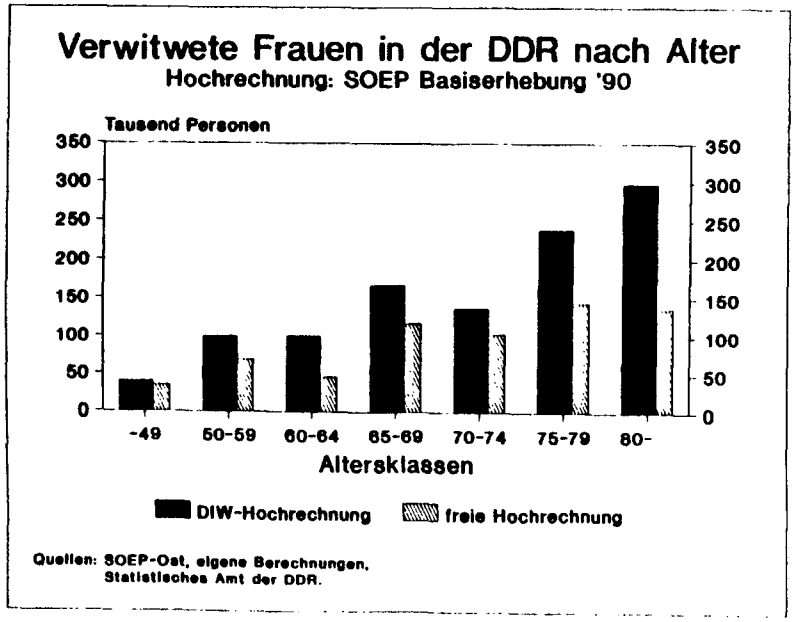
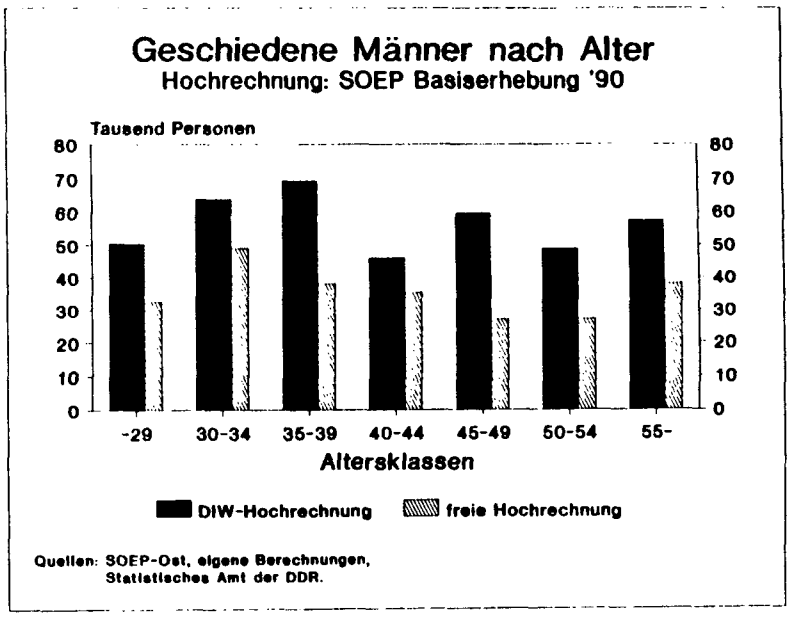
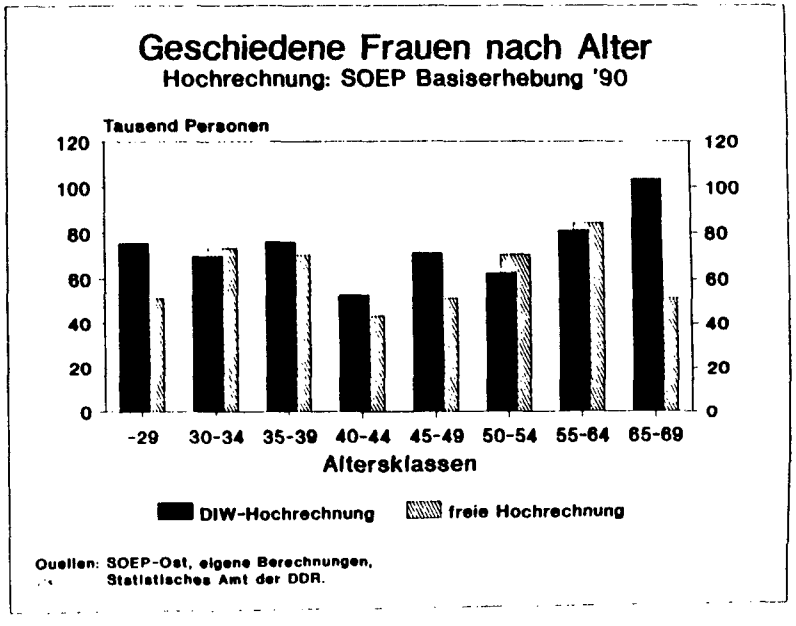


Abb. 8a und 8b



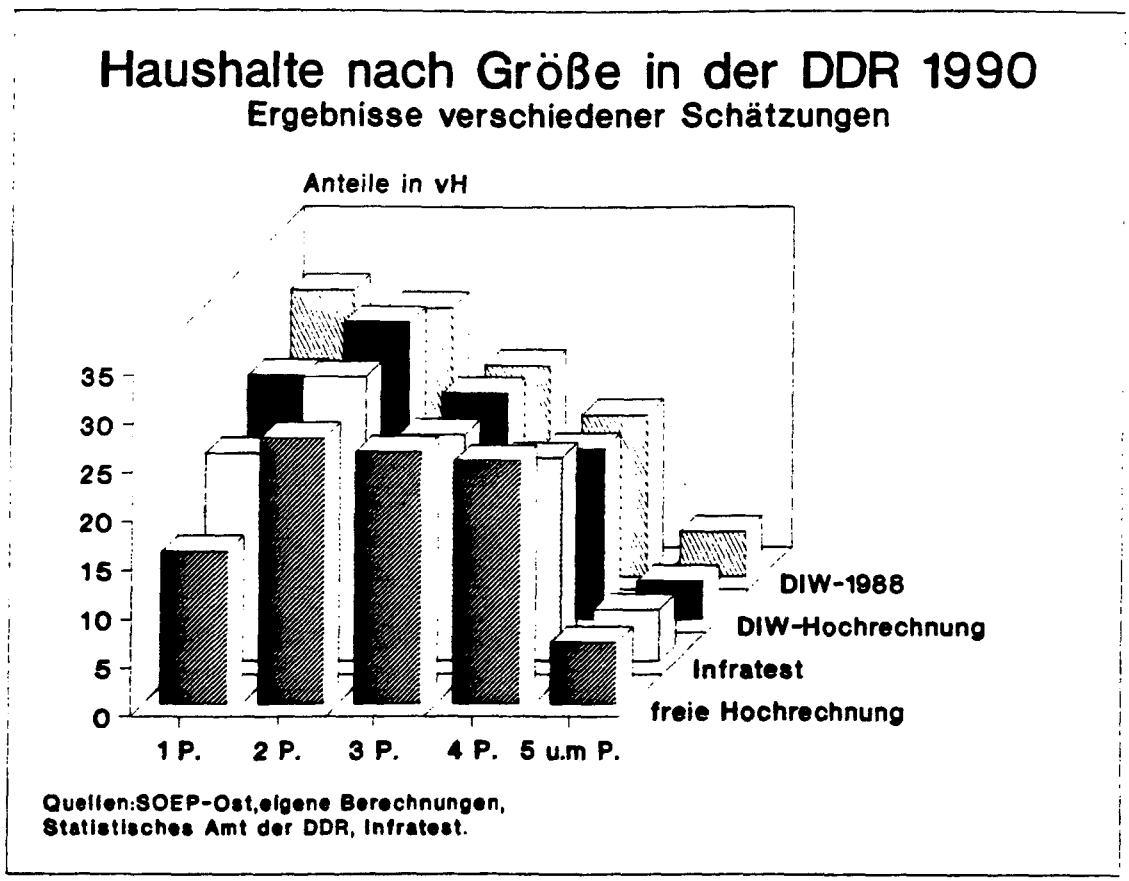


Auswirkungen der übrigen Anpassung an Personenmerkmale überlassen werden:

In Abbildung 9 sind Haushaltsstrukturen dargestellt worden, die sich aus vier verschiedenen Verfahren ergaben:

DIW-Hochrechnung	Vorliegende Gewichtung des DIW mit 115 Restriktionen.
freie Hochrechnung	Ungewichtete Hochrechnung der Basiserhebung DDR.
DIW-1988	Schätzung des DIW für das Jahr 1988 [18]
Infratest	Gewichtung von Infratest[19]

Abb. 9



18 Bedau, K.-D. und H. Vortmann,ebenda.

19 Infratest Sozialforschung,ebenda.

Während sich die Anteile für die 2, 3 und 5 u.m.-Personenhaushalte relativ wenig voneinander unterscheiden, führen die Verfahren in den beiden übrigen Klassen doch zu erheblichen Abweichungen. Vor allem in den Ein-Personenhaushalten sind die Differenzen signifikant. Zwei Hauptgründe sind dafür verantwortlich, daß EinPersonenhaushalte in nahezu jeder Stichprobe unterrepräsentiert sind: Zum einen ist die Antreffwahrscheinlichkeit und somit die Antwortwahrscheinlichkeit geringer, weil nur eine Person im Haushalt wohnt, zum anderen leben in ihnen häufig sehr junge oder relativ alte Personen, deren Verweigerungsquote - wenn auch meist aus unterschiedlichen Gründen - überdurchschnittlich hoch ist. Der erzielte Anteil von 15,6 vH, der mit dem Wert der freien Hochrechnung identisch ist, erscheint auch unter diesen Umständen als ausgesprochen niedrig[20]. Auf der anderen Seite erscheint der Schätzwert des DIW für 1988 mit 29,4 vH eher eine Obergrenze zu sein[21], zumal die Infratest-Gewichtung mit 21,3 vH einen erheblich niedrigeren Wert erbrachte. So gesehen ist das erzielte Ergebnis der Hochrechnung mit 25,2 vH als plausibel anzusehen, das sich vornehmlich aus der Anhebung der Gewichte für Ledige über 19 Jahre und ältere Personen ergeben hat.

6. Zusammenfassung und abschließende Bemerkungen

In diesem Beitrag wurde die erste Hochrechnung und Gewichtung des SOEP-Ost vorgestellt. Grundgedanke für die Anlage des Gewichtungsverfahren war, die von Infratest Sozialforschung repräsentativ angelegte Stichprobe weitgehend zu erhalten, d.h. von einer Gleichverteilung der Gewichte ausgehend, diese

20 Zum Vergleich: In der Basiserhebung des SOEP-West lag der Anteil der Ein-Personenhaushalte bei 22,6 vH.

21 Die DDR-Volkszählung aus dem Jahr 1981 ergab einen Anteil der Einpersonenhaushalte an allen Haushalten von rund 25 vH. Zwar ist es aufgrund von demographischen Verschiebungen und Verhaltensänderungen wahrscheinlich, daß dieser Anteil gestiegen ist. Ein Anteil von fast 30 vH ist angesichts der Wohnungsprobleme in der ehemaligen DDR indes eher unwahrscheinlich.

nach dem Prinzip des minimalen Informationsverlustes nur so weit zu verändern, daß eine vorgegebene Anzahl von Randbedingungen erfüllt sind. Diese Randbedingungen waren 114 Restriktionen zur Personenverteilung nach Region, Geschlecht, Familienstand und Alter. Nur eine Restriktion - die Zahl der Haushalte in der DDR insgesamt - wurde aus empirischen Gründen für die Haushaltsebene vorgesehen, so daß insgesamt 115 Restriktionen in die Gewichtung eingingen. Sicher bestand die Möglichkeit, weitere Restriktionen z.B. Personen nach Bezirken, Geschlecht und "arbeitsfähigem Alter" oder nach Gemeindegrößenklassen aufzunehmen. Indes ist es nicht unbedingt von Vorteil, die Zahl der Restriktionen nach Belieben auszudehnen. Mit jeder zusätzlichen Restriktion nimmt die Varianz der Hochrechnungsfaktoren zu. Diese überträgt sich grundsätzlich auch auf die Stichprobenvarianz. Dies wiederum führt zu größeren Konfidenzintervallen der zu schätzenden Parameter und somit zu Signifikanzverlusten.

Wenn die Ergebnisse der ersten Mikrozensen für die fünf neuen Ländern vorliegen, sollten Stichprobenergebnisse und die Hochrechnung mit diesen verglichen werden. Sollte sich erweisen, daß die hier vorgelegten Gewichte zu erheblich abweichenden Strukturen führen, wäre eine Überarbeitung der ersten Gewichtung angezeigt.