Methodenberichte

Heft 1
Das Stichprobenverfahren der Einkommens- und Verbrauchs-stichprobe 1998

Carola Kühnen
Gruppe Mathematisch-statistische Methoden

Herausgeber und Vertriebsstelle:

Statistisches Bundesamt, 65180 Wiesbaden

Fachliche Informationen zu dieser Veröffentlichung können Sie direkt beim Statistischen Bundesamt erfragen:
Gruppe IIA, Telefon: 06 11 / 75 35 46, Fax: 06 11 / 75 39 51 oder E-Mail: gruppe-iia@statistik-bund.de

Erscheinungsfolge: unregelmäßig

Erschienen im Mai 2001

Schutzgebühr: DM 8,00 / EUR 4,09 zzgl. Versandkosten

Bestellnummer: 9211010 - 98900

Recyclingpapier aus 100 % Altpapier.

Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:

- im Internet: http://www.statistik-bund.de

oder bei unserem Informationsservice
65180 Wiesbaden
- Telefon: 06 11 / 75 24 05
- Telefax: 06 11 / 75 33 30
- E-Mail: info@statistik-bund.de

Vorwort

Mit dem vorliegenden Heft eröffnet das Statistische Bundesamt eine neue Schriftenreihe, die auf die Darstellung mathematischer Methoden der amtlichen Statistik ausgerichtet ist. Die neue Reihe soll einerseits anhand aktueller empirischer Fragestellungen über Neuentwicklungen von mathematisch-statistischen Methoden im Statistischen Bundesamt informieren. Andererseits soll sie die interessierte Öffentlichkeit, insbesondere die Wissenschaft, zu verstärkter Kommunikation und Kooperation mit dem Statistischen Bundesamt anregen und Anstoß geben für eine stärkere Einbeziehung der Wissenschaft in die Weiterentwicklung des methodischen Instrumentariums der amtlichen Statistik. Denn in vielen statistischen Bereichen sind verstärkt innovative mathematisch-statistische Methoden zu entwickeln und anzuwenden, um die im Zuge der Weiterentwicklung der Bundesstatistik gesetzten Ziele, wie weitere qualitative Verbesserung der Produkte, schnellere Verfügbarkeit wichtiger Ergebnisse und Verringerung der Auskunftslast, zu erreichen. Die Methoden werden schwerpunktmäßig eingesetzt bei der Stichprobenplanung und -analyse, bei der Komponentenzerlegung ökonomischer Zeitreihen für Zwecke der Konjunkturdiagnose und -prognose sowie zur Sicherung der statistischen Geheimhaltung bei Einzeldaten und in Tabellen. In jedem Heft der Schriftenreihe wird ein aktuelles mathematisch-methodisches Thema behandelt. Es ist vorgesehen, jährlich in unregelmäßiger Folge ein bis zwei Hefte zu veröffentlichen.

Im vorliegenden ersten Heft der Schriftenreihe wird das neu entwickelte Stichprobenverfahren der Einkommens- und Verbrauchsstichprobe 1998 vom Auswahlverfahren über das Hochrechnungsverfahren bis zum Verfahren für die Abschätzung der Präzision der Stichprobenergebnisse ausführlich beschrieben und über die bei der Anwendung dieser Verfahren aufgetretenen Probleme berichtet. Das besondere innovative Element des Stichprobendesigns ist das Hochrechnungsverfahren. Das Verfahren ist von Professor Merz (Universität Lüneburg) entwickelt und in der Gruppe Mathematisch-statistische Methoden des Statistischen Bundesamtes in die Stichprobenpraxis umgesetzt worden. Es wird als Hochrechnung nach dem Prinzip des minimalen Informationsverlustes bezeichnet. Das Verfahren macht es möglich, die Stichprobendaten simultan an die Randverteilungen mehrerer Merkmale anzupassen und so die Qualität der Stichprobenergebnisse gegenüber herkömmlichen Verfahren zu verbessern.

Johann Hahlen
Präsident des Statistischen Bundesamtes

Inhalt

Das Stichprobenverfahren der Einkommens- und Verbrauchsstichprobe 1998

1 Einführung

Die Einkommens- und Verbrauchsstichprobe (EVS) der privaten Haushalte ist eine alle 5 Jahre durchgeführte bundesweite Erhebung aus dem Bereich der Wirtschaftsrechnungen privater Haushalte. Das Hauptanliegen der Erhebung besteht darin, die wirtschaftliche und soziale Lage der Haushalte aus dem Blickwinkel der Einkommensverteilung und -verwendung darzustellen. Die Ergebnisse werden u.a. auch in der Volkswirtschaftlichen Gesamtrechnung und zur Anpassung der Gewichte für den Verbraucherpreisindex verwendet. Für die EVS 1998 wurde ein neues Erhebungsdesign entwickelt, durch das in erster Linie die teilnehmenden Haushalte entlastet und die Aktualität der Ergebnisse verbessert werden soll (siehe Chlumsky und Ehling 1997, S. 457 ff.). Außerdem wurden die Erhebungsmerkmale der EVS und der Laufenden Wirtschaftsrechnungen weitgehend aufeinander abgestimmt. In dem vorliegenden Beitrag wird das Stichprobenverfahren für das neue Erhebungsdesign vorgestellt.

2 Das Erhebungsdesign der EVS 1998

Bevor näher auf das Stichprobenverfahren der EVS 1998 eingegangen wird, soll zunächst das neue Erhebungsdesign skizziert werden. Die Erhebung erstreckte sich insgesamt über den Zeitraum eines Jahres. Sie bestand aus den folgenden drei Erhebungsteilen:

- Bei allen Stichprobenhaushalten (Auswahlverfahren siehe 3.1) wurde zu Beginn des Erhebungsjahres ein Einführungsinterview (Stichtag 1. Januar 1998) durchgeführt mit Fragen über die Zusammensetzung des Haushalts, sozioökonomische Merkmale der Haushaltsmitglieder, die Ausstattung mit langlebigen Gebrauchsgütern, Haus- und Grundbesitz und die Wohnverhältnisse. Das Einführungsinterview wurde von geschulten Interviewern durchgeführt, die Haushalte konnten aber auch selbst ausfüllen.
- Die Stichprobengesamtheit der Haushalte wurde in vier Teile zerlegt (Aufteilung siehe 4.5), die jeweils in einem Quartal ein Haushaltsbuch mit Aufzeichnungen über Einnahmen und Ausgaben zu führen hatte. Im einzelnen waren das die Angaben über Einnahmen aller im Haushalt lebenden Personen nach Einkommensarten und -höhe und über alle zum privaten Verbrauch zählenden Ausgaben, z.B. für Nahrungsmittel, Miete, Körperpflege, Verkehrsmittel oder die persönliche Ausstattung. Im Unterschied zur EVS 1993 waren diese Einnahmen und Ausgaben nur über 3 Monate und nicht über das gesamte Jahr anzuschreiben. Im Haushaltsbuch der EVS 1998 wurden außerdem Fragen zur Bildung von Geldvermögen und zu Schulden gestellt. Diese Angaben waren in den vorangegangenen Erhebungen noch gesondert in einem Schlussinterview erhoben worden.
- Die Aufzeichnungen des Haushaltsbuchs wurden ergänzt um ein sogenanntes "Feinaufzeichnungsheft", in dem für einen vorgegebenen Monat die Ausgaben für Nahrungsmittel, Getränke und Tabakwaren detaillierter nach Menge und Preis zu erheben waren. Da die Ausgaben für Nahrungs- und Genussmittel im Vergleich zu anderen Ausgaben geringeren Schwankungen unterliegen, wurden sie nur bei einer Unter-

stichprobe von ca. 20% der Gesamtstichprobe erfragt. Die folgende Übersicht 1 zeigt die Verteilung der Stichprobenhaushalte auf die verschiedenen Erhebungsteile und Monate.

***Übersicht 1**: Erhebungsdesign der EVS 1998*

Zeitraum	**Erhebungsteile**
1.1.98	Einführungsinterview (rd. 74 000 Haushalte**)**
Januar	
Februar	
März	
April	
Mai	
Juni	
Juli	
August	
September	
Oktober	
November	
Dezember	

Führung von Haushaltsbüchern bei rd. 18 500 Haushalten im Durchschnitt je Quartal

Führung von Feinaufzeichnungen bei rd. 1 250 Haushalten im Durchschnitt je Monat

3 Das Auswahlverfahren der EVS 1998

3.1 Zufallsstichprobe oder Quotenverfahren?

Zunächst stellte sich die Frage, ob nicht anstelle der Quotenstichprobe, die in den vorangegangenen Einkommens- und Verbrauchserhebungen eingesetzt wurde, die methodisch vorteilhaftere Zufallsauswahl realisiert werden könnte. Dabei war zu berücksichtigen, dass die Teilnahme an der EVS freiwillig ist, was zwangsläufig zur Problematik der Antwortausfälle führt. Zur Klärung dieser Frage wurde das Ergebnis einer Testerhebung zur Neukonzeption der Laufenden Wirtschaftsrechnungen herangezogen, die 1996 als Zufallsstichprobe in Form eines Random-Route-Verfahrens in 5 Bundesländern durchgeführt worden war (siehe Gertkemper, Kühnen und Wein 1998). Hier zeigte sich, dass sich je nach Land nur zwischen 9% und 15% der angesprochenen Haushalte an der Erhebung beteiligten. Darüber hinaus verteilten sich die Ausfälle nicht zufällig, sondern konzentrierten sich auf bestimmte Bevölkerungsgruppen, wie z.B. Einpersonenhaushalte, Arbeiterhaushalte und Haushalte mit besonders niedrigem und hohem Einkommen. Aufgrund dieser Erfahrungen wurde entschieden, als Auswahlverfahren wieder das Quotenverfahren einzusetzen.

Die Quotenstichprobe hat zwar prinzipielle Nachteile gegenüber einer Zufallsstichprobe, da nicht alle Haushalte eine Auswahlchance bekommen und deshalb Verzerrungen in den Ergebnissen nicht ausgeschlossen werden können. Der stichprobenmethodische Vorteil des Zufallsprinzips bei der Auswahl hat hier aber

wegen der ermittelten geringen und unterschiedlichen Teilnahmebereitschaft erheblich an Bedeutung verloren. Die geringe Teilnahmebereitschaft hätte insbesondere auch zur Folge gehabt, dass ein Vielfaches des stichprobenmethodisch erforderlichen Stichprobenumfangs an Haushalten hätte angesprochen werden müssen, was mit unvertretbar hohen Kosten verbunden gewesen wäre.

Wie bekannt wird bei der Quotenstichprobe – analog zur Schichtung bei Zufallsstichproben – die Grundgesamtheit anhand bestimmter Merkmale (den sogenannten Quotierungsmerkmalen) in Gruppen gegliedert und für jede Gruppe (Quotierungszelle) die Quote der zu befragenden Haushalte vorgegeben. Mit verschiedenen Anwerbeaktionen kann dann versucht werden, die entsprechende Zahl von Haushalten je Gruppe zur Teilnahme zu gewinnen.

Im Unterschied zu vielen konventionellen Quotenstichproben, die in Markt- und Meinungsforschungsumfragen eingesetzt werden, erhalten bei der EVS die Interviewer die tatsächlichen Adressen für eine Kontaktaufnahme und müssen nicht nach Stichprobenhaushalten suchen, die den Quotenanforderungen entsprechen. Damit wird eine subjektive Auswahl der zu befragenden Haushalte seitens der Interviewer ausgeschlossen.

Bei der EVS 1998 wurden mehr als die Hälfte der einbezogenen Hauhalte durch die sogenannte indirekte Werbung (Berichte in den Medien, Verteilung von Informationsmaterial, Mund-zu-Mund-Propaganda) gewonnen. Die übrigen Haushalte konnten mit direkter Werbung, d.h. durch persönliche Anschreiben an Haushalte, die schon an anderen statistischen Erhebungen teilgenommen hatten, gewonnen werden.

3.2 Einteilung der Grundgesamtheit in Quotierungszellen

Zunächst wurde die Grundgesamtheit der Haushalte nach Bundesländern gegliedert, da die Werbung und Befragung der Haushalte von den Statistischen Landesämtern durchzuführen war. Für Berlin erfolgte die Quotierung noch zusätzlich nach West und Ost. Je Bundesland wurde die Haushaltsgesamtheit nach den Quotierungsmerkmalen modifizierter Haushaltstyp (6 Gruppen), soziale Stellung des Haupteinkommensbeziehers/der Haupteinkommensbezieherin[1)] (6 Gruppen) und Haushaltsnettoeinkommen (5 Klassen) gruppiert (vgl. Übersicht 2). Diese Merkmale wurden für die Quotierung ausgewählt, weil sie das Verbrauchsverhalten der Haushalte entscheidend bestimmen und für die Untergliederung der Ergebnisdarstellung von zentraler Bedeutung sind. Die Kombination der Ausprägungen der Quotierungsmerkmale führte theoretisch zu insgesamt 3 060 Quotierungszellen. Quotierungszellen mit weniger als 5 000 Haushalten in der Grundgesamtheit wurden mit benachbarten Zellen zusammengelegt. Dabei wurden in der Regel zunächst die Einkommensklassen, dann die Klassen mit den Ausprägungen der sozialen Stellung und zuletzt die Klassen der Haushaltstypen zusammengelegt. Auf diese Weise wurden insgesamt 1 274 Quotierungszellen gebildet.

4 Das Verfahren der Stichprobenaufteilung

4.1 Gesamtstichprobenumfang

Basis für die Festlegung des Erhebungssolls waren die Ergebnisse des Mikrozensus, der einzigen amtlichen Statistik mit Auskunftspflicht, die auch haushalts- und familienstatistische Angaben liefert. Die Erhebungsgesamtheit umfasste alle im Mikrozensus nachgewiesenen Privathaushalte am Ort der Hauptwohnung, deren

1) Als Haupteinkommensbezieher/-in wird die Person definiert, die den größten Beitrag zum Haushaltsnettoeinkommen leistet und die beim Einführungsinterview von dem befragten Haushalt als solche benannt wird.

monatliches Haushaltsnettoeinkommen weniger als 35 000 DM betrug. Ausgeschlossen wurden also alle Haushalte, deren Bezugsperson am Ort der Nebenwohnung angetroffen wurde und Haushalte, welche die angegebene Einkommensgrenze überschritten. Ebenso waren Personen in Anstalten und Gemeinschaftsunterkünften nicht enthalten[2)]. Zum Zeitpunkt der Werbung der Haushalte waren Ergebnisse des Mikrozensus aus dem Jahr 1995 verfügbar. Für die EVS 1998 wurde – wie auch schon 1993 – ein durchschnittlicher Auswahlsatz von 0,2% der Erhebungsgesamtheit des Mikrozensus 1995 festgelegt, was einem Stichprobenumfang von etwa 73 890 Haushalten entsprach.

Übersicht 2: *Quotierungsmerkmale je Bundesland*

Haushaltstyp
- Einpersonenhaushalte
- Ehepaare/Nichteheliche Lebensgemeinschaften ohne Kinder (ohne weitere Personen)
- Alleinerziehende mit ledigen Kindern unter 27 Jahren und mit mindestens einem Kind unter 18 Jahren (ohne weitere Personen)
- Ehepaare/Nichteheliche Lebensgemeinschaften mit ledigen Kindern unter 27 Jahren und mindestens einem Kind unter 18 Jahren, höchstens ein Partner erwerbstätig (ohne weitere Personen)
- Ehepaare/Nichteheliche Lebensgemeinschaften mit ledigen Kindern unter 27 Jahren und mindestens einem Kind unter 18 Jahren, beide Partner erwerbstätig (ohne weitere Personen)
- Sonstige Haushalte

Soziale Stellung der Bezugsperson
- Selbständige
- Beamte/Beamtinnen
- Angestellte
- Arbeiter(innen)
- Rentner(innen), Pensionäre
- Sonstige Nichterwerbstätige

Haushaltsnettoeinkommen von .. bis unter ... DM
- 0 - 1 400
- 1 400 - 2 500
- 2 500 - 5 000
- 5 000 - 7 000
- 7 000 - 35 000

2) Für die Festlegung des Erhebungssolls musste zunächst die Bezugsperson des Haushalts im Mikrozensus (die beliebig durch ein einzelnes Haushaltsmitglied bestimmt werden kann) nach den Vorgaben der EVS neu definiert werden, d.h. als Bezugsperson wurde die Person mit der höchsten Einkommensangabe festgelegt. Falls von allen Personen eines Mikrozensushaushalts keine Einkommensangabe vorlag, so wurde die erstgenannte Person ausgewählt. In dem Falle, dass mehrere Personen eines Haushalts in dieselbe Einkommensklasse fielen und diese die höchste Einkommensklasse des Haushalts war, so wurde die erstgenannte Person aus dieser Einkommensklasse als Bezugsperson definiert.
Die Zahl der Mikrozensushaushalte ohne Einkommensangabe wurden für jede Gruppe, die aus der Kombination der Quotierungsmerkmale "Haushaltstyp" und "sozialer Stellung der Bezugsperson" gebildet wird, auf alle Einkommensklassen proportional verteilt. Da in der EVS Haushalte mit einem monatlichen Haushaltsnettoeinkommen von 35.000 DM und mehr nicht zur Erhebungsgesamtheit zählen, müssen diese auch aus den Mikrozensuszahlen ausgeschlossen werden. Im Mikrozensus wird diese Einkommensklasse aber nicht gesondert erfasst, die nach oben offene Einkommensklasse liegt bei 12.000 DM und mehr. Die Zahl der Mikrozensushaushalte mit einem Nettoeinkommen von 35.000 DM und mehr wurde aus der Einkommensteuerstatistik mit rund 120.000 Haushalte geschätzt.

4.2 Das Aufteilungsverfahren

Die Aufteilung des Stichprobenumfangs auf die Bundesländer und die Quotierungszellen, die je Bundesland durch Kombination der in 4.1 genannten Quotierungsmerkmale gebildet wurden, erfolgte nach dem "Prinzip der vergleichbaren Präzision für gegliederte Ergebnisse" (Einzelheiten siehe Krug, Nourney und Schmidt 1999, S. 122 ff.). Dieses Verfahren ermöglicht es, die Aufteilung des Stichprobenumfangs auf Schichten so zu steuern, dass eine Abstufung des relativen Standardfehlers des Aufteilungsmerkmals in Abhängigkeit von seinen Schichtwerten erreicht wird gemäß der Beziehung

$$\varepsilon_h = \frac{C}{\hat{X}_h^{\alpha}} \tag{1}$$

mit

ε_h : geschätzter relativer Standardfehler des Aufteilungsmerkmals X in der Schicht h
$\hat{X}_h$: geschätzter Schichtwert des Aufteilungsmerkmals X
α : Exponent der Präzisionsabstufung ($0 \le \alpha \le 0{,}5$)
C : Konstante

Der Exponent α steuert den Grad der Abstufung der Fehler zwischen den Schichten. Der kleinstmögliche Wert 0 des Exponenten bewirkt den gleichen relativen Standardfehler in allen Schichten, der größtmögliche Wert 0,5 hingegen eine besonders starke Abstufung. In diesem Fall wird für Schichtergebnisse eine Präzision angestrebt, die umgekehrt proportional zur Wurzel aus dem Schichtwert ist und für das Gesamtergebnis in der Regel eine höhere Präzision liefert als bei einem kleineren Exponenten. Die Konstante C wird in einem iterativen Verfahren so bestimmt, dass die Formel (1) unter Einhaltung des zu verteilenden Stichprobenumfangs für alle Schichten erfüllt ist.

Für das Quadrat des geschätzten relativen Standardfehlers des Aufteilungsmerkmals X in der Schicht h gilt:

$$\varepsilon_h^2 = N_h \cdot \frac{N_h - n_h}{n_h} \cdot \frac{s_{hx}^2}{\hat{X}_h^2} = \frac{N_h - n_h}{n_h \cdot N_h} \cdot v_{hx}^2 \tag{2}$$

mit

N_h : Anzahl der Einheiten in der Grundgesamtheit der Schicht h
n_h : Stichprobenumfang in der Schicht h
s_{hx}^2 : Varianz des Aufteilungsmerkmals X in der Schicht h
v_{hx} : Variationskoeffizient des Aufteilungsmerkmals X in der Schicht h

Durch Gleichsetzen der Gleichungen (1) und (2) und Auflösen nach dem Stichprobenumfang n_h ergibt sich:

$$n_h = \frac{N_h}{\frac{C^2}{\hat{X}_h^{2\alpha}} \cdot \frac{N_h}{v_{hx}^2} + 1} \tag{3}$$

Im Fall der EVS wird dieses für geschichtete Zufallsstichproben entwickelte Prinzip auf die Bundesländer (siehe 4.3) und innerhalb eines Bundeslandes nochmals auf die Quotierungszahlen (siehe 4.4) angewendet.

4.3 Aufteilung des Stichprobenumfangs auf die Bundesländer

Für die Aufteilung des Gesamtstichprobenumfangs auf die Länder wurde wie schon in der EVS 1993 modellhaft ein Aufteilungsmerkmal mit einheitlichen Mittelwerten ($\overline{X}_L = \overline{X}$) und Variationskoeffizienten ($v_{Lx} = v$) je Land L unterstellt, da von den Ländern keine großen Abweichungen des neuen Stichproben-

umfangs vom bisherigen Stichprobenumfang gewünscht wurden. Das bedeutet, dass die Abstufung der relativen Standardfehler von Land zu Land ausschließlich von der Zahl der Haushalte N_L des jeweiligen Landes abhängig gemacht wurde. Wie bei der Stichprobenaufteilung für die EVS 1993 wurde der Grad der Fehlerabstufung mit $\alpha = 0{,}45$ so gewählt, dass eine starke Abstufung der Präzision von Land zu Land erreicht wird, die in ihrer Wirkung einer proportionalen Aufteilung sehr nahe kommt.

$$\varepsilon_L = \frac{C}{N_L^{0,45}} \tag{4}$$

mit ε_L: geschätzter relativer Standardfehler für das geschätzte Landesergebnis $\hat{X}_L$ eines Merkmals, das in allen Ländern gleichen Mittelwert und Variationskoeffizienten aufweist.

Mit diesem Ansatz sollte erreicht werden, dass die relativen Standardfehler für die Ergebnisse des kleinsten Bundeslandes "Bremen" (Auswahlsatz 0,25%) etwa viermal so groß werden wie die für das größte Bundesland "Nordrhein-Westfalen" (Auswahlsatz 0,18%), sofern die darzustellenden Merkmale einheitliche Mittelwerte und Variationskoeffizienten aufweisen.

Nach den Formeln (3) und (4) und Berücksichtigung einheitlicher Variationskoeffizienten und Mittelwerte ergaben sich die optimalen Stichprobenumfänge n_L der Länder gemäß der Formel

$$n_L = \frac{N_L}{C'^2 \cdot N_L^{0,1}} \qquad \text{mit } C' = \frac{C}{v \cdot \bar{X}^{0,45}} \tag{5}$$

Die Konstante C wurde iterativ so bestimmt, dass die Summe der gerundeten Stichprobenumfänge der Länder mit dem Gesamtstichprobenumfang übereinstimmte. Das Aufteilungsverfahren lieferte folgende Ergebnisse:

Tabelle 1: *Aufteilung des Stichprobenumfangs auf die Bundesländer*

Bundesland	Zahl der Haushalte in 1 000 (Mikrozensus 1995)	Zahl der Stichprobenhaushalte	Auswahlsatz in %
Baden-Württemberg	4 701,7	9 026	0,19
Bayern	5 339,3	10 118	0,19
Berlin-West	1 180,1	2 434	0,21
Berlin-Ost	652,5	1 430	0,22
Brandenburg	1 073,7	2 390	0,22
Bremen	344,6	860	0,25
Hamburg	881,4	2 002	0,23
Hessen	2 707,7	5 496	0,20
Mecklenburg-Vorpommern	760,8	1 750	0,23
Niedersachsen	3 434,5	6 803	0,20
Nordrhein-Westfalen	8 031,8	14 614	0,18
Rheinland-Pfalz	1 757,6	3 719	0,21
Saarland	507,1	1 213	0,24
Sachsen	2 030,2	4 241	0,21
Sachsen-Anhalt	1 200,6	2 644	0,22
Schleswig-Holstein	1 258,5	2 752	0,22
Thüringen	1 075,6	2 398	0,22
Deutschland	36 937,7	73 890	0,20

4.4 Aufteilung des Stichprobenumfangs der Länder auf die Quotierungszellen

Je Bundesland wurden die berechneten Stichprobenumfänge auf die Quotierungszellen *h* aufgeteilt. Die Aufteilung wurde so vorgenommen, dass auch Ergebnisse für nur schwach besetzte Quotierungszellen mit ausreichender Präzision erstellt werden können. Hierfür wurde wieder das in 4.2 beschriebene Verfahren der Präzisionsabstufung eingesetzt. Als Aufteilungsmerkmal wurde das Merkmal "Privater Verbrauch" verwendet, da es ein besonders wichtiges Merkmal für die Ergebnisdarstellung ist. Mit der Wahl des Exponenten $\alpha = 0{,}3$ wurde ein Kompromiss derart geschlossen, dass die Präzision der Teilergebnisse für das Aufteilungsmerkmal "Privater Verbrauch" so abgestuft ist, dass einerseits allzu hohe Auswahlsätze in schwach besetzten Quotierungszellen vermieden werden und andererseits eine gute Präzision des Gesamtergebnisses erreicht werden kann. Mit dieser Vorgabe für die Abstufung des relativen Standardfehlers des Privaten Verbrauchs gemäß

$$\varepsilon_h = \frac{C}{\hat{X}_h^{0,3}} \tag{6}$$

ergibt sich folgende Aufteilungsformel für die Stichprobenumfänge der Quotierungszellen:

$$n_h = \frac{N_h}{\frac{C^2}{\hat{X}_h^{0,6}} \frac{N_h}{v_{hx}^2} + 1} \tag{7}$$

In den Formeln (6) und (7) wurde der Übersicht halber auf einen zusätzlichen Index für das Land verzichtet. Die Konstante *C* wurde wieder so bestimmt, dass je Land die Summe der gerundeten Stichprobenumfänge der Quotierungszellen mit dem Landesstichprobenumfang übereinstimmte. Die für die Stichprobenaufteilung erforderlichen Daten bezüglich des Merkmals Privater Verbrauch wurden aus der EVS 1993 ermittelt.

Tabelle 2 zeigt beispielhaft die Aufteilung des Stichprobenumfangs auf die modifizierten Haushaltstypen, soziale Stellung der Bezugsperson und monatliche Haushaltsnettoeinkommensklassen im Bundesgebiet. Die dargestellten Stichprobenumfänge ergaben sich aus der Summation der Stichprobenumfänge der Quotierungszellen. Relativ hohe Auswahlsätze wiesen beim Haushaltstyp die Gruppen "Alleinerziehende" und "Sonstige Haushalte" auf. Im Hinblick auf die soziale Stellung der Bezugsperson bzw. der Haushaltsnettoeinkommensklasse waren insbesondere Selbständigen- und Beamtenhaushalte bzw. Haushalte mit hohem Einkommen überproportional einzubeziehen. Auf eine Darstellung des vollständigen Quotenplans wird hier wegen des großen Umfangs verzichtet.

In der Praxis zeigte sich, dass manche Quoten trotz großer Anstrengungen nicht erreicht werden konnten. In solchen Fällen wurden dann ersatzweise zusätzliche Haushalte aus ähnlichen Quotierungszellen einbezogen.

Tabelle 2: *Erhebungssoll nach Quotierungsmerkmalen*

Merkmalsausprägung	Zahl der Haushalte in 1000 (Mikrozensus 1995)	Zahl der Stich-probenhaushalte	Auswahlsatz in %
Haushaltstyp			
Einpersonenhaushalt	12 891,5	18 146	0,14
Ehepaar/Lebenspartnerschaft ohne Kinder	10 064,6	19 750	0,20
Alleinerziehende	1 170,9	3 226	0,28
Ehepaar/Lebenspartnerschaft mit Kindern	8 017,3	18 498	0,23
darunter:			
Ehepaar/Lebenspartnerschaft mit Kindern, höchstens ein Partner erwerbstätig	4 006,5	8 608	0,22
Ehepaar/Lebenspartnerschaft mit Kindern, beide Partner erwerbstätig	3 083,1	6 572	0,21
Sonstiger Haushalt	4 793,4	14 270	0,30
Soziale Stellung der Bezugsperson			
Selbständige	2 490,8	11 990	0,48
Beamte/Beamtinnen	1 687,0	5 693	0,34
Angestellte	8 793,8	16 939	0,19
Arbeiter(innen)	7 601,8	13 466	0,18
Rentner(innen), Pensionäre	12 228,4	17 301	0,14
Sonstige Nichterwerbstätige	4 135,9	8 501	0,21
Haushaltsnettoeinkommen von .. bis unter ... DM			
0 -1 400	4 480,1	6 466	0,14
1 400 - 2 500	9 695,7	14 342	0,15
2 500 - 5 000	15 699,2	30 308	0,19
5 000 - 7 000	4 435,1	11 823	0,27
7 000 - 35 000	2 627,6	10 951	0,42
Haushalte insgesamt	36 937,7	73 890	0,20

4.5 Modifikation des Auswahlplans

In der EVS 1993 wurde zusätzlich eine Quotierung für Haushalte mit ausländischer Bezugsperson und für Haushalte von Landwirten durchgeführt mit dem Ziel, auch für diese Haushalte differenzierte Ergebnisse nachzuweisen. Damals konnten allerdings trotz besonderer Anstrengungen der Statistischen Ämter nur rd. 50% des Erhebungssolls für Haushalte von Landwirten und nur 47% für Haushalte mit ausländischer Bezugsperson zur Teilnahme gewonnen werden. Wegen der geringen Teilnahmebereitschaft war es nicht möglich, für diese Gruppen gesonderte Ergebnisse darzustellen. Daher wurden diese Haushalte in der EVS 1998 bei der Quotierung nicht besonders berücksichtigt; die angestrebte Anzahl von Teilnehmern wurde aber auch für sie je Bundesland vorgegeben, um mit Hilfe von besonderen Werbemaßnahmen eine ausreichende Teilnahme anzusteuern. Vorgegeben wurde aber nur je Land die Zahl der anzuwerbenden Haushalte insgesamt und nicht – wie in der EVS 1993 – in Kombination mit weiteren Quotierungsmerkmalen, um eine erfolgreiche Werbung nicht zu gefährden. Für die Festlegung der angestrebten Zahl von Haushalten mit ausländischer Bezugsperson wurden die Auswahlsätze der Länder angewendet. Die Richtwerte für Haushalte von Landwirten ergaben sich aus dem Produkt des Anteils der Landwirte in der Klasse "Selbständige + Landwirte" im Mikrozensus 95 und den für diese Klasse ermittelten Sollvorgaben je Bundesland.

4.6 Aufteilung des Stichprobenumfangs auf die Quartale

Bei der EVS 1993 hatte sich gezeigt, dass die Teilnahmebereitschaft der Haushalte im Laufe des Jahres abnahm. Um bei der EVS 1998 eine gleichmäßige Verteilung der Aufzeichnungen aller Haushalte auf die vier Quartale des Erhebungsjahres zu gewährleisten, wurden – ausgehend von Teilnahmebereitschaftsquoten der EVS 1993 – innerhalb jeder Quotierungszelle die Stichprobenumfänge disproportional auf die Quartale verteilt. Danach ergab sich der Stichprobenumfang $n_{h,q}^{L}$ des Quartals q in der Quotierungszelle h des Landes L aus der Formel

$$n_{h,q}^{L} \bullet \frac{n_h^L}{t_q^L - \sum_{i \bullet 1}^{4} 1/t_i^L} \tag{8}$$

mit

t_q^L : Teilnahmebereitschaftsquote der EVS 1993 des Landes L im Quartal q

n_h^L : Stichprobenumfang des Landes L in der Quotierungszelle h

Diese Aufteilung verfolgte das Ziel, saisonale Schwankungen bei Käufen und Dienstleistungen realistisch abzubilden und die Wahl des Anschreibungsquartals durch die Haushalte zu vermeiden.

Die Stichprobenumfänge für die Quartale wurden zunächst ungerundet mit Nachkommastellen berechnet. Für die Rundung wurde das Niemeyer-Verfahren[3)] verwendet, das die Übereinstimmung der Summen der Stichprobenumfänge der Quartale mit dem Gesamtstichprobenumfang einer Quotierungszelle gewährleistet. Tabelle 3 zeigt die Verteilung der Länderstichprobenumfänge auf die Quartale.

3) Dabei wurden alle Werte nach absteigenden Nachkommastellen sortiert und zunächst abgerundet. Anschließend wurden die abgerundeten Werte nacheinander solange aufgerundet, bis die Summe der Quartalswerte mit dem Stichprobenumfang der Quotierungszelle übereinstimmt. Die Rundung der Stichprobenumfänge führte in einigen Fällen zu Abweichungen mit den Sollwerten für die Randsummen der Quartalswerte der Länder. Um dies zu vermeiden, wurden einzelne Felder der gerundeten Tabelle um eins erhöht oder erniedrigt. Dabei war zu beachten, dass die veränderte Tabelle sich möglichst wenig von der ungerundeten Tabelle unterschied.

Tabelle 3: *Aufteilung des Stichprobenumfangs der Bundesländer auf die Quartale*

Bundesland	Erhebungssoll				
	1. Quartal	2. Quartal	3. Quartal	4. Quartal	Zusammen
Baden-Württemberg	2 197	2 250	2 276	2 303	9 026
Bayern	2 480	2 518	2 550	2 570	10 118
Berlin-West	578	597	621	638	2 434
Berlin-Ost	341	348	365	376	1 430
Brandenburg	570	587	611	622	2 390
Bremen	202	209	220	229	860
Hamburg	477	495	507	523	2 002
Hessen	1 334	1 360	1 386	1 416	5 496
Mecklenburg-Vorpommern	423	433	440	454	1 750
Niedersachsen	1 665	1 689	1 710	1 739	6 803
Nordrhein-Westfalen	3 538	3 661	3 698	3 717	14 614
Rheinland-Pfalz	899	916	940	964	3 719
Saarland	286	297	311	319	1 213
Sachsen	1 037	1 053	1 065	1 086	4 241
Sachsen-Anhalt	636	653	670	685	2 644
Schleswig-Holstein	659	680	696	717	2 752
Thüringen	581	593	605	619	2 398
Deutschland	17 903	18 339	18 671	18 977	73 890

4.7 Auswahl der Unterstichprobe für die Feinaufzeichnungen

Das Erhebungssoll für die Unterstichprobe der Feinaufzeichnungen betrug 15 000 Haushalte insgesamt. Die Aufteilung dieses Stichprobenumfangs auf die Bundesländer und die Quotierungszellen erfolgte wie für die Gesamtstichprobe nach dem unter 4.2 bis 4.6 beschriebenen Verfahren. Die Unterstichprobe wurde auf alle 12 Monate des Jahres verteilt analog zur Aufteilung des Stichprobenumfangs auf die Quartale. Dabei wurde sichergestellt, dass die Quartalswerte der Quotierungszellen eingehalten werden.

5 Das Hochrechnungsverfahren

5.1 Überlegungen zum methodischen Konzept der Hochrechnung

Allgemeines Ziel der Hochrechnung ist es, mit Hilfe geeigneter Schätzfunktionen aus den Stichprobenparametern (Gesamtwert, Mittelwert, Anteilswert, Varianz) auf die Parameter der Grundgesamtheit zu schließen. Einen unverzerrten Schätzwert für den unbekannten Gesamtwert eines interessierenden Merkmals *Y* liefert der sogenannte Horvitz-Thompson-Schätzer

$$\hat{Y}_\pi = \sum_{k=1}^{n} \pi_k^{-1} y_k \tag{9}$$

mit

$\pi_k = P(k \in s)$: Auswahlwahrscheinlichkeit der k-ten Stichprobeneinheit einer Zufallsstichprobe *s*

y_k : Merkmalswert der k-ten Stichprobeneinheit

n : Anzahl der Stichprobeneinheiten

Bei einer Zufallsstichprobe von *n* Stichprobeneinheiten aus *N* Einheiten der Grundgesamtheit haben alle Stichprobeneinheiten die gleiche Auswahlwahrscheinlichkeit $\pi_k = n/N$. Die Hochrechnung wird dann auch als "freie Hochrechnung" bezeichnet, da keine weitere Information für die Hochrechnung verwendet wird.

Liegt eine geschichtete Zufallsstichprobe vor, so verwendet man die Kehrwerte der Auswahlwahrscheinlichkeiten der Stichprobeneinheiten je Schicht *h* als Hochrechnungsfaktoren:

$$\pi_{hk}^{-1} = \pi_h^{-1} = N_h / n_h \quad \text{für alle Stichprobeneinheiten } k \in h \tag{10}$$

mit

$\pi_{hk} = P(k \in h)$: Auswahlwahrscheinlichkeit der k-ten Stichprobeneinheit in der Schicht *h*

N_h : Anzahl der Einheiten in der Schicht *h* der Grundgesamtheit

n_h : Anzahl der Stichprobeneinheiten in der Schicht *h*

Es ergibt sich folgender erwartungstreuer Schätzwert für den Gesamtwert des Merkmals *Y* bei geschichteter Zufallsauswahl:

$$\hat{Y}_{st} = \sum_{h=1}^{L} N_h / n_h \sum_{k=1}^{n_h} y_{hk} \tag{11}$$

mit

y_{hk} : Merkmalswert der k-ten Stichprobeneinheit in der Schicht h

L: Zahl der Schichten

Zwar können bei der Quotenstichprobe unter strengen stichprobentheoretischen Gesichtspunkten keine Auswahlwahrscheinlichkeiten für die Stichprobeneinheiten berechnet werden, da die Auswahl nicht zufällig erfolgt, sondern von subjektiven Faktoren abhängt. Vernachlässigt man diese methodischen Bedenken, so kann eine Quotenstichprobe mit einer geschichteten Stichprobe verglichen werden, bei denen die Haushalte in den einzelnen Schichten (Quotierungszellen) zufällig gezogen werden. Im Unterschied zu Zufallsstichproben hat man aber i.a. keine zeitlich genau definierte Auswahlgesamtheit zur Verfügung und unterstellt, dass die Stichprobe aus der aktuellen Gesamtheit gezogen wird. Für die Hochrechnung werden daher Informationen aus anderen Quellen über die aktuelle Verteilung der Quotierungsmerkmale benötigt. Sind diese – wie hier aus dem Mikrozensus – vorhanden, so entspricht das Hochrechnungsverfahren formelmäßig der bei Zufallsstichproben üblichen freien Hochrechnung, wobei es sich faktisch um eine Anpassung an die aktuelle gemeinsame Verteilung der Quotierungsmerkmale handelt. Die Anpassung korrigiert die unterschiedlichen

Wahrscheinlichkeiten, die aus den disproportionalen Quotenvorgaben sowie aus der Nichterfüllung der Quotenvorgaben resultieren.

Der Hochrechnungsfaktor wird je Quotierungszelle *h* mit der Formel (10) berechnet. Hierbei bezeichnet N_h die Zahl der Einheiten in der Quotierungszelle *h* der Grundgesamtheit zum Zeitpunkt der Erhebung und n_h die Zahl der verwertbaren Stichprobeneinheiten in der Quotierungszelle *h*. Für jede Quotierungszelle stimmt dann die aus der Stichprobe hochgerechnete Fallzahl mit dem Gesamtwert überein. Um Ergebnisverzerrungen größeren Ausmaßes zu vermeiden, ist darauf zu achten, dass sich N_h und n_h näherungsweise auf den gleichen aktuellen Zeitpunkt beziehen. Außerdem können in der Stichprobe nicht oder sehr schwach besetzte Quotierungszellen auftreten. Diese müssen vor der Hochrechnung mit benachbarten Zellen zusammengelegt werden. Das hier beschriebene Hochrechnungsverfahren wird im Folgenden als "freie Hochrechnung mit aktuellem Hochrechnungsrahmen" bezeichnet. Dieses Verfahren wurde in den vorangegangenen EVS-Erhebungen angewendet.

Die Schätzungen können häufig noch verbessert werden, wenn bei der Hochrechnung zusätzliche Informationen über aktuelle Merkmalsgesamtwerte bekannt sind, die bei der Quotierung nicht genutzt werden konnten und bei denen ein Zusammenhang mit den zu schätzenden Werten wahrscheinlich ist. Die Zusatzinformationen können genutzt werden, um eine Anpassung an die gemeinsame Verteilung mehrerer Merkmale durchzuführen. Dieses Verfahren hat aber den Nachteil, dass viele Anpassungsgruppen, die sich aus der Kombination der Ausprägungen dieser Merkmale ergeben, in der Stichprobe nicht besetzt sind. Um dies zu vermeiden, werden üblicherweise benachbarte Anpassungsgruppen zusammengelegt, was wiederum zur Folge hat, dass die aus der Stichprobe hochgerechneten Fallzahlen für die Anpassungsmerkmale nicht mit den Werten in der Grundgesamtheit übereinstimmen wie nachfolgendes Beispiel zeigt.

Beispiel: *Anpassung an die gemeinsame Verteilung von zwei Merkmalen*

Die Stichprobendaten sollen an die gemeinsame Verteilung der Merkmale Haushaltstyp und Einkommensklasse in der Grundgesamtheit mit jeweils 2 Ausprägungen angepasst werden. Die Kombination der Ausprägungen der Anpassungsmerkmale führt zu Bildung von 4 Anpassungsgruppen.

Haushaltstyp	Einkommens-klasse	Zahl der Stich-probenhaushalte n_h	Zahl der Haus-halte insgesamt N_h	Hochrechnungs-faktor N_h/n_h	Hochgerechnetes Ergebnis
Einpersonen-haushalt	1	2	80	40	80
	2	5	100	20	100
Ehepaare ohne Kinder	1	0 } 4	40 } 120	nicht def. } 30	0 } 120
	2	4	80	20	120
Insgesamt	1	2	120		80
	2	9	180		220

In der Stichprobe ist die Gruppe der Ehepaare ohne Kinder in der Einkommensklasse 1 nicht besetzt und wird daher mit der Einkommensklasse 2 zusammengefasst. Das bedeutet, dass die Ehepaare ohne Kinder mit dem Faktor (40+80)/(4+0)=30 hochgerechnet werden. Hingegen erfolgt bei den Einpersonenhaushalten keine Zusammenlegung. Für das Gesamtergebnis hat dies zur Folge, dass die Summe der hochgerechneten Ergebnisse bezüglich der beiden Einkommensklassen (80 in Einkommensklasse 1 und 220 in Einkommensklasse 2) von den vorgegebenen Daten in der Grundgesamtheit (120 in Einkommensklasse 1 und 180 in Einkommensklasse 2) abweicht.

Aufgrund der Nachteile des beschriebenen Anpassungsverfahrens wurde für die Hochrechnung der EVS 98 ein anderes Anpassungsverfahren eingesetzt, und zwar die sog. "Hochrechnung nach dem Prinzip des minimalen Informationsverlustes" (Merz 1983). Dieses Verfahren hat den Vorteil, dass es eine differenzierte Gliederung der Anpassungsmerkmale erlaubt, ohne dass das Problem der gering oder gar nicht besetzten Gruppen auftritt, da es die Möglichkeit bietet, nur an Eckwert-Gliederungen mehrere Merkmale (Randverteilungen) der Grundgesamtheit anzupassen. Die Ausprägungen der Eckwert-Gliederungen sind in der Regel ausreichend besetzt.

5.2 Hochrechnung nach dem Prinzip des minimalen Informationsverlustes

Bei der Hochrechnung nach dem Prinzip des minimalen Informationsverlustes werden die Stichprobendaten simultan an die Randverteilungen mehrerer Merkmale angepasst. Die Anpassung erfolgt mit Hilfe von Hochrechnungsfaktoren w_k, die so bestimmt werden, dass die geschätzten Gesamtwerte für die Anpassungsmerkmale mit den bekannten Gesamtwerten übereinstimmen:

$$\sum_{k=1}^{n} w_k \mathbf{x}_k = \mathbf{X} \tag{11}$$

Dabei bezeichnet $\mathbf{x}_k = \left(x_{k1},\ldots,x_{kj},\ldots,x_{km}\right)^T$ den Vektor mit den Werten der *m* Anpassungsvariablen des k-ten Haushalts in der Stichprobe s und $\mathbf{X} = \left(X_1,\ldots,X_j,\ldots,X_m\right)^T$ den Vektor der bekannten Gesamtwerte der *m* Anpassungsvariablen.

Unter der Voraussetzung, dass die verwendeten Informationen über die Anpassungsvariablen eine gute Qualität aufweisen, verringert sich durch die Anpassung der systematische Fehler, der auf Antwortausfälle oder andere Gründe zurückzuführen ist und der Stichprobenzufallsfehler für alle korrelierten Merkmale, d.h. die Genauigkeit der Schätzwerte für korrelierte Merkmale wird gegenüber einer freier Hochrechnung verbessert.

Das Prinzip des minimalen Informationsverlustes besteht darin, dass die Hochrechnungsfaktoren nach Anpassung so bestimmt werden, dass sie sich von den ursprünglichen, auf den Auswahlwahrscheinlichkeiten basierenden Hochrechnungsfaktoren so wenig wie möglich unterscheiden. Hierzu wird eine Distanzfunktion definiert, die sich aus der Informationstheorie ableitet, und zwar aus der Entropie einer diskreten Verteilung $\mathbf{p} = (p_1,\ldots,p_n)$ von Wahrscheinlichkeiten p_k ($p_k > 0$ für $k = 1,\ldots,n;\ \sum_{k=1}^{n} p_k = 1$).

Die Entropie ist definiert als

$$H(\mathbf{p}) = \sum_{k=1}^{n} p_k \log\frac{1}{p_k} \tag{12}$$

Die daraus abgeleitete Distanzfunktion

$$I(\mathbf{p},\mathbf{q}) = \sum_{k=1}^{n} p_k \cdot \log\frac{p_k}{q_k} = \sum_{k=1}^{n} p_k\left(\log\, p_k - \log\, q_k\right) \tag{13}$$

beschreibt den Informationsverlust, der bei Ersetzung der ursprünglichen Verteilung $\mathbf{q} = (q_1,\ldots,q_n)$ durch die Verteilung $\mathbf{p} = (p_1,\ldots,p_n)$ entsteht. Der Informationsverlust ist nicht negativ und genau dann Null, wenn es durch die ersetzende Verteilung **p** keine neuen Informationen gibt (p_k=q_k ; siehe Theil 1967, S. 28). In

unserem Fall werden für die Verteilung **q** die ursprünglichen Hochrechnungsfaktoren genommen, die so normiert sind, dass ihre Summe Eins ergibt. p_k sind die normierten Hochrechnungsfaktoren nach der Anpassung. Die Hochrechnungsfaktoren w_k ergeben sich dann als $w_k=p_kN$ (N: Anzahl der Auswahleinheiten in der Grundgesamtheit). Die Distanzfunktion (13) wird minimiert unter Einhaltung der Bedingung (11). Die Lösung des Optimierungsproblems mit Hilfe des Lagrange-Ansatzes führt auf das folgende nichtlineare Gleichungssystem[4]:

$$\sum_{k=1}^{n} q_k e^{\left(\boldsymbol{\lambda}\mathbf{x}_k - 1\right)} \mathbf{x}_k = \mathbf{X}/N \tag{14}$$

wobei $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ den Vektor der Langrangefaktoren bezeichnet. Die gesuchten neuen Hochrechnungsfaktoren ergeben sich mit der Lösung des Gleichungssystems (14) aus der Gleichung:

$$p_k = q_k e^{\left(\boldsymbol{\lambda}\mathbf{x}_k - 1\right)} \quad \text{für alle } k \in s \tag{15}$$

Werden als Anpassungsvariable kategoriale Variablen mit den Ausprägungen $i_s\,(s = 1, \ldots, r_j)$ verwendet, so ist der Vektor $\mathbf{x}_k$ definiert durch:

$$\mathbf{x}_k = \left(\delta_{k1}, \ldots, \delta_{kj}, \ldots, \delta_{km}\right)^T \tag{16}$$

mit

$$\delta_{kj} = \left(\delta_{kj}^{i_1}, \ldots, \delta_{kj}^{i_s}, \ldots, \delta_{kj}^{i_{rj}}\right) \text{ und}$$

$$\delta_{kj}^{i_s} = \begin{cases} 1, & \text{falls die Ausprägung des } k \text{ - ten Haushalts bzgl. der } j \text{ - ten Variablen gleich } i_s \text{ ist} \\ 0, & \text{sonst} \end{cases}$$

und der Vektor der Gesamtwerte durch

$$\mathbf{X} = \left(N_1, \ldots, N_j, \ldots, N_m\right) \tag{17}$$

mit

$$\mathbf{N}_j = \left(N_j^{i_1}, \ldots, N_j^{i_{rj}} \right) \text{ und}$$

$N_j^{i_s}$:Anzahl der Haushalte in der Gesamtheit mit Ausprägung i_s bzgl. der j-ten Variablen.

Betrachtet man anstelle der Verteilung der Hochrechnungsfaktoren direkt die Hochrechnungsfaktoren $w_k=p_kN$ und $d_k=q_kN$, so lautet die entsprechende Funktion zu (13)

$$I(\mathbf{w}, \mathbf{d}) = \frac{1}{N} \sum_{k=1}^{n} w_k \left(\log w_k - \log d_k\right) \tag{18}$$

und die entsprechende Gleichung zu (11)

$$\sum_{k=1}^{n} \delta_{kj} w_k = N_j \quad \text{für } j = 1, \ldots, m \tag{19}$$

Aus der Definition des Informationsverlustes geht hervor, dass als Lösungen für das Optimierungsproblem nur nicht-negative Gewichte in Frage kommen, da sonst die Distanzfunktion nicht definiert ist. Die Definition $0 \cdot \log 0 := 0$ bewirkt, dass Hochrechnungsfaktoren mit dem Wert 0 nach der Anpassung wieder den Wert 0 haben. Eine weitere wünschenswerte Eigenschaft dieser Funktion ist ihre strenge Konvexität, womit

4) Aufgrund der Definition der p_k und q_k als relative Häufigkeiten (Wahrscheinlichkeiten) werden die vorgegebenen Gesamtwerte ebenfalls relativiert (X/N).

die Eindeutigkeit der Lösung des Optimierungsproblems sichergestellt ist (siehe Wauschkuhn 1982, S. 42 ff.). Als Bedingung für eine Lösung ist es erforderlich, dass die Verteilung der einzelnen Merkmale in der Stichprobe linear unabhängig ist, das bedeutet, die Ausprägungen eines Merkmals dürfen sich nicht durch Linearkombinationen aus anderen Ausprägungen ermitteln lassen.

Da für nichtlineare Gleichungssysteme i. Allg. keine geschlossene Lösung angegeben werden kann, müssen numerische Verfahren herangezogen werden. Die numerische Lösung des nichtlinearen Gleichungssystems (14) erfolgt iterativ, bis eine vorgegebene Toleranzgrenze für die Differenz der Lösungswerte zu den entsprechenden vorgegebenen Eckwerten unterschritten wird.

Für die numerische Lösung des Optimierungsproblems kann das von Prof. Merz entwickelte Programm ADJUST verwendet werden, das mit Hilfe des modifizierten Newton-Raphson-Verfahrens mit relativ wenig Iterationen die Zielfunktion minimiert[5)].

5.3 Anwendungsbeispiel

Das folgende fiktive Beispiel soll die Berechnung der Hochrechnungsfaktoren gemäß dem beschriebenen Verfahren veranschaulichen.

Schätzung der Einkommensverteilung

Bei einer Stichprobe mit Quotierung nach Geschlecht soll eine Anpassung an die Randverteilungen von Geschlecht und Einkommen erfolgen.

Laut Quotenplan wurden aus einer Gesamtheit von 511 Frauen und 489 Männern 52 Frauen und 48 Männer ausgewählt:

Quotierungszelle	Gesamtheit	Stichprobe
Frauen	511	52
Männer	489	48
Insgesamt	1000	100

Für die Schätzung der Einkommensverteilung von Frauen und Männern stehen ferner folgende Informationen über die Gesamtheit zur Verfügung:

Einkommen	Personen in der Gesamtheit
Niedrig	435
Mittel	296
Hoch	269
Insgesamt	1000

Die ursprünglichen Hochrechnungsfaktoren ergeben sich aus den Auswahlwahrscheinlichkeiten

$$d_k = \begin{cases} 511/52 = 9{,}83 & \text{, falls } k \text{ - te Person Frau} \\ 489/48 = 10{,}19 & \text{, sonst} \end{cases}$$

5) Beim Newton-Raphson-Verfahren werden ausgehend von einem nichtlinearen Gleichungssytem mit m Unbekannten die m Funktionsgleichungen über eine Taylorentwicklung linearisiert und die Lösung des Problems iterativ bestimmt. Das modifizierte Newton-Raphson Verfahren verwendet für die Lösung der linearisierten Gleichungen variable Schrittweiten. Einzelheiten zur numerischen Lösung siehe Merz (1994).

Zu minimieren ist die Funktion

$$I(\mathbf{d},\mathbf{w}) = \frac{1}{1000}\sum_{k=1}^{100} w_k \cdot (\log w_k - \log d_k)$$

$$= \frac{1}{1000}\sum_{k=1}^{52} w_k \cdot (\log w_k - \log 9{,}83) + \frac{1}{1000}\sum_{k=53}^{100} w_k \cdot (\log w_k - \log 10{,}19)$$

(s. Gleichung (18)) unter Einhaltung der Nebenbedingungen:

1) $\sum_{k=1}^{100} \delta_{k1}^{1} w_k = 489$ mit $\delta_{k1}^{1} = \begin{cases} 1, & \text{falls } k\text{ - te Person Mann} \\ 0, & \text{sonst} \end{cases}$

2) $\sum_{k=1}^{100} \delta_{k1}^{2} w_k = 511$ mit $\delta_{k1}^{2} = \begin{cases} 1, & \text{falls } k\text{ - te Person Frau} \\ 0, & \text{sonst} \end{cases}$

3) $\sum_{k=1}^{100} \delta_{k2}^{1} w_k = 435$ mit $\delta_{k2}^{1} = \begin{cases} 1, & \text{falls } k\text{ - te Person niedriges Einkommen hat} \\ 0, & \text{sonst} \end{cases}$

4) $\sum_{k=1}^{100} \delta_{k2}^{2} w_k = 296$ mit $\delta_{k2}^{2} = \begin{cases} 1, & \text{falls } k\text{ - te Person mittleres Einkommen hat} \\ 0, & \text{sonst} \end{cases}$

(s. Gleichungen (19)). Die Restriktionen bezüglich der Zahl der Personen mit hohem Einkommen ergeben sich durch Linearkombinationen aus den 4 aufgeführten Bedingungen.

Die Lösung des Optimierungsproblems wird iterativ mit Hilfe des Programms ADJUST ermittelt. Für die Anwendung des Programms ADJUST sind 2 Dateien zu erstellen. Die erste Datei enthält die Stichprobeninformationen, also die ursprünglichen Hochrechnungsfaktoren und die sogenannte Informationsmatrix mit den Merkmalsausprägungen der Anpassungsmerkmale aller Haushalte. Die Zahl der Datensätze entspricht der Anzahl der Stichprobenhaushalte. Die Zeilenvektoren $\boldsymbol{x}_k^T$ (s. Definition (16)) der Informationsmatrix umfassen die einzelnen Ausprägungen der Merkmale. Wird wie in unserem Beispiel eine Anpassung mit einem Merkmal mit 2 Ausprägungen und einem mit 3 Ausprägungen durchgeführt, so enthält der Zeilenvektor 2+3-1=4 Werte. Die größte Ausprägung des zweiten Merkmals wird weggelassen, um die Lösungsbedingung der linearen Unabhängigkeit der Anpassungsmerkmale zu gewährleisten. Die Informationsmatrix für unser Beispiel enthält demnach 4 Spalten mit den Werten $\delta_{k1}^{1}, \delta_{k1}^{2}, \delta_{k2}^{1}, \delta_{k2}^{2}$ und 100 Zeilen (Anzahl der Personen in der Stichprobe).

Beispiel zum Aufbau der Informationsmatrix

Haushalt	Geschlecht	Einkommen
1	Mann	Hoch
2	Frau	Niedrig
⋮	⋮	⋮
100	Frau	Mittel

$\Rightarrow$

Informationsmatrix

$$\left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 1 \end{array}\right)$$

In der zweiten Datei werden die einzuhaltenden Gesamtwerte (17) aufgeführt.

Das Programm ADJUST liefert folgende Hochrechnungsfaktoren:

Geschlecht	Einkommen	Hochrechnungsfaktor vor Anpassung	Hochrechnungsfaktor nach Anpassung
Frau	niedrig	9,83	11,18
Frau	mittel	9,83	8,62
Frau	hoch	9,83	8,91
u			
Mann	niedrig	10,19	11,85
Mann	mittel	10,19	9,29
Mann	hoch	10,19	9,58

Mit den Hochrechnungsfaktoren der letzten Spalte der vorstehenden Tabelle und der Stichprobenverteilung ergeben sich die Schätzwerte für die Gesamtheit:

Einkommen	Stichprobe			Schätzwerte für die Gesamtheit vor Anpassung			Schätzwerte für die Gesamtheit nach Anpassung		
	Frauen	Männer	Insgesamt	Frauen	Männer	Insgesamt	Frauen	Männer	Insgesamt
Niedrig	23	15	38	226	153	379	257	178	435
Mittel	16	17	33	157	173	330	138	158	296
Hoch	13	16	29	128	163	291	116	153	269
Insgesamt	52	48	100	511	489	1000	511	489	1000

Wie man sieht, stimmen hier nach der Anpassung die Randverteilungen für Einkommen und Geschlecht mit den vorgegebenen Rahmendaten überein. Ohne Anpassung wird die Zahl der Personen mit niedrigem Einkommen deutlich unterschätzt.

5.4 Anwendung des Verfahrens bei der EVS 1998

Die Hochrechnung der EVS 1998 erfolgte getrennt für die drei Erhebungsteile Einführungsinterview, Haushaltsbuch und Feinanschreibungen. Zusätzlich wurden für Deutschland und für die Teilgesamtheiten "Früheres Bundesgebiet" und "Neue Länder einschließlich Berlin-Ost" die EVS-Daten gesondert hochgerechnet, um mit der Hochrechnung strukturtreue Abbildungen der Mikrozensus-Ergebnisse sowohl für den Bund, wie auch für die Teilgebiete zu realisieren. Die für die beiden Teilgebiete ermittelten Hochrechnungsfaktoren wurden auch für die Ergebnisdarstellung der Bundesländer verwendet. Die gesonderte Hochrechnung für Deutschland und die Teilgebiete bzw. Länder hat den Nachteil, dass Bundes- und Länderergebnisse für Merkmale, die nicht bei beiden Anpassungen verwendet worden sind, nicht konsistent sein müssen.

Von dem vorgegebenen Erhebungssoll von 73 890 Haushalten konnten rund 93% zur Teilnahme am Einführungsinterview gewonnen werden. Quartalsanschreibungen mit verwertbaren Angaben lagen nur von rund 84% der Haushalte des Erhebungssolls vor. Dabei war die Teilnahmebereitschaft in den einzelnen Quotierungszellen sehr unterschiedlich, so dass die Quotenvorgaben nicht genau eingehalten werden konnten (zu den Ergebnissen des Einführungsinterviews siehe Münnich und Illgen 1999). Hinsichtlich vieler Merkmale, bei denen ein Zusammenhang mit dem Ausgabeverhalten wahrscheinlich ist, unterschied sich –

bedingt durch die Antwortausfälle und die Nichtzufälligkeit der Auswahl – die Struktur der Stichprobe von der Struktur der Gesamtheit.

Als Anpassungsmerkmale wurden die Quotierungsmerkmale modifizierter Haushaltstyp, soziale Stellung des Haupteinkommensbeziehers und Haushaltsnettoeinkommen gewählt. Die Ausprägungen wurden – entsprechend der Nachweisung der Ergebnisse der EVS 1998 – für die Anpassung tiefer gegliedert als bei der Quotierung. Zum Beispiel wurde für die Anpassung das Haushaltsnettoeinkommen in 10 Klassen statt in 5 und die Nichterwerbstätigenhaushalte weiter nach Studenten/Studentinnen und sonstigen Nichterwerbstätigen gegliedert. Eine Quotierung in einer solch differenzierten Gliederung wäre nicht möglich gewesen, da einerseits viele Quotierungszellen nicht besetzt und andererseits die Quotenvorgaben sehr viel schwerer zu erfüllen gewesen wären. Auch eine Anpassung nach der gemeinsamen Kombination der Quotierungsmerkmale in der gewünschten Untergliederung war nicht sinnvoll, da viele Gruppen in der Stichprobe nicht besetzt gewesen wären. Um dies zu vermeiden, wurde das oben beschriebene Hochrechnungsverfahren "Hochrechnung nach dem Prinzip des minimalen Informationsverlustes" eingesetzt.

Zunächst wurde eine freie Hochrechnung mit aktuellem Hochrechnungsrahmen durchgeführt, d.h. für jede Quotierungszelle wurden Hochrechnungsfaktoren mittels der Quotienten N_h/n_h berechnet, wobei N_h die hochgerechnete Zahl der Mikrozensushaushalte und n_h die Zahl der verwertbaren Haushalte in der Quotierungszelle h bezeichnen (s. Formel (10)). Für die Stichprobendaten des Einführungsinterviews diente der Mikrozensus von 1997 als Hochrechnungsrahmen und für die Quartalsanschreibungen zur Berechnung der Jahresergebnisse der Mikrozensus von 1998.

Anschließend wurden die hochgerechneten Ergebnisse mittels der Hochrechnung nach dem Prinzip des minimalen Informationsverlustes an die Grundgesamtheit angepasst. Im Einzelnen wurde die Anpassung der Daten des Einführungsinterviews an die Randverteilungen der in Übersicht 3 dargestellten Merkmale durchgeführt.

Übersicht 3: *Anpassungsmerkmale für das Einführungsinterview*[6)]

1) *Für Bundesergebnisse*
- Haushaltstyp (6)
- soziale Stellung der Bezugsperson (9)
- Haushaltsnettoeinkommensklasse (10)
- Bundesland (17) ∗ Haushaltstyp (6)
- Bundesland (17) ∗ Erwerbstätigkeit (2)
- Bundesland (17) ∗ Haushaltsnettoeinkommensklasse (5)
- Haushaltstyp (6) ∗ Haushaltsnettoeinkommensklasse (10) ∗ soziale Stellung der Bezugsperson (9)

2) *Für Länderergebnisse und nach früherem Bundesgebiet und neuen Ländern gegliederte Ergebnisse*

a) *Früheres Bundesgebiet*
- Haushaltstyp (6)
- soziale Stellung der Bezugsperson (9)
- Haushaltsnettoeinkommensklasse (10)
- Bundesland (11) ∗ Haushaltstyp (6)
- Bundesland (11) ∗ soziale Stellung der Bezugsperson (4)
- Bundesland(11) ∗ Haushaltsnettoeinkommensklasse (6)
- Haushaltstyp (6) ∗ Haushaltsnettoeinkommensklasse (10) ∗ soziale Stellung der Bezugsperson (7)

b) *Neue Länder einschließlich Berlin-Ost*
- Haushaltstyp (6)
- soziale Stellung der Bezugsperson (9)
- Haushaltsnettoeinkommensklasse (10)
- Bundesland (6) ∗ Haushaltstyp (6)
- Bundesland (6) ∗ soziale Stellung der Bezugsperson (4 bzw. 5)
- Bundesland(6) ∗ Haushaltsnettoeinkommensklasse (7)
- Haushaltstyp (6) ∗ Haushaltsnettoeinkommensklasse (9) ∗ soziale Stellung der Bezugsperson (7)

Die Quartalsanschreibungen für die unter 1) und 2) genannten Gebiete wurden jeweils zusätzlich an folgende Merkmale angepasst:

- Quartal (4) ∗ Haushaltstyp (6)
- Quartal (4) ∗ soziale Stellung der Bezugsperson (4)
- Quartal (4) ∗ Haushaltsnettoeinkommensklasse (4)
- Quartal (4) ∗ Haushaltstyp (6) ∗ Haushaltsnettoeinkommensklasse (8) ∗ soziale Stellung der Bezugsperson (6)

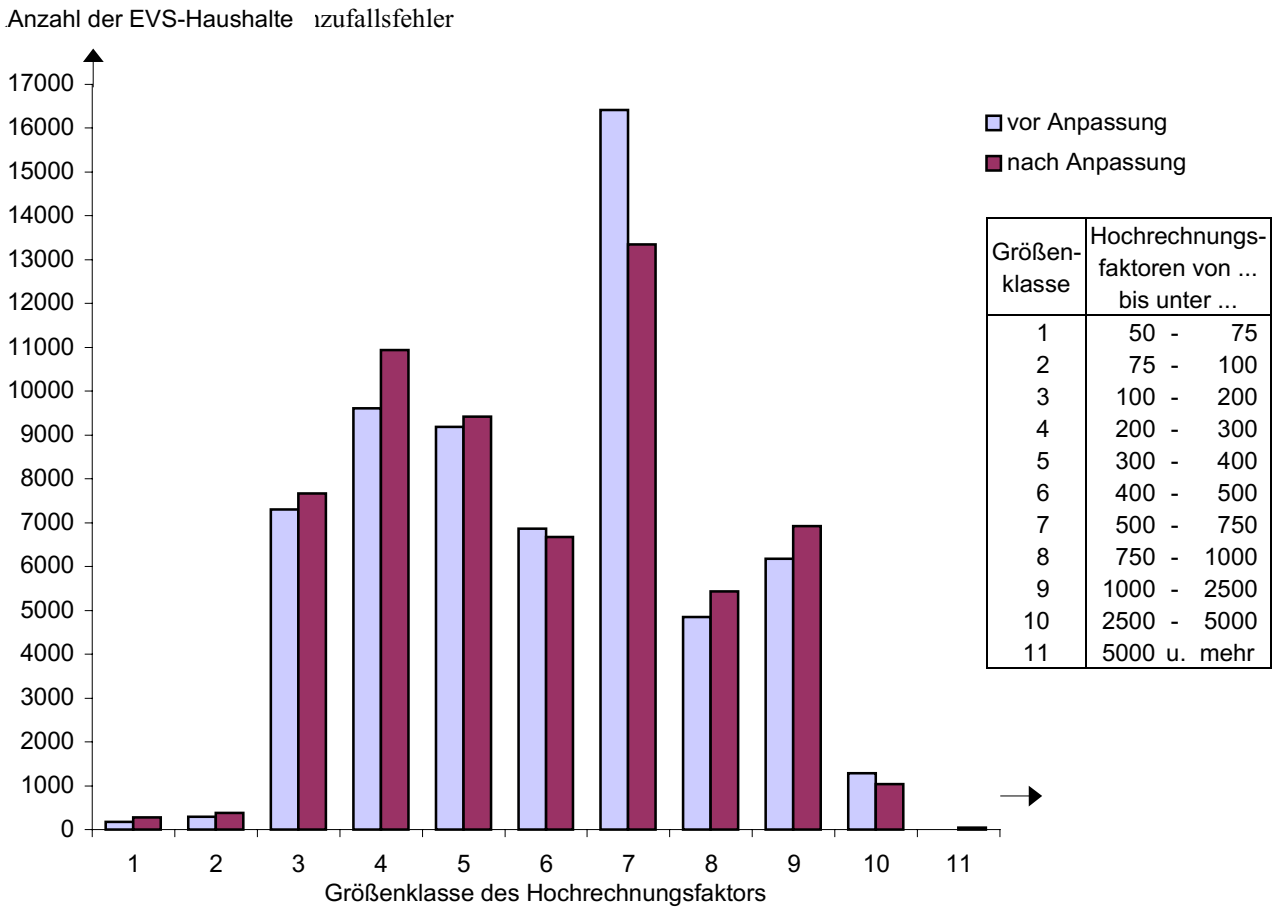
Weitere Anpassungsmerkmale wurden nicht verwendet, denn bei der Anwendung der Hochrechnung nach dem Prinzip des minimalen Informationsverlustes gilt: Je mehr Anpassungsmerkmale verwendet werden, desto größer ist die Spanne zwischen dem kleinsten und größten angepassten Hochrechnungsfaktor zu erwarten. Eine große Spanne ist aber unerwünscht, weil damit höhere Zufallsfehler zu erwarten sind und Ausreißer ein zu hohes Gewicht erhalten können. Bei einigen Merkmalen mussten Ausprägungen

6) Ein Stern zwischen zwei Anpassungsmerkmalen bedeutet, dass an die gemeinsame Verteilung angepasst wird, d.h. für jede Kombination der Ausprägungen stimmt die hochgerechnete Fallzahl mit der Mikrozensuszahl überein. Die Zahl der Ausprägungen der Anpassungsmerkmale ist in Klammern angegeben.

zusammengefasst werden, da selbst die Ausprägungen der Eckwert-Gliederungen in der Stichprobe zu schwach besetzt waren.

Wie schon oben erwähnt, hat das Minimierungsproblem keine Lösung, wenn sich die Ausprägungen eines Merkmals durch Linearkombination aus anderen Ausprägungen errechnen lassen. Unter Beachtung dieser Lösungsbedingung wurden bei der Hochrechnung des Einführungsinterviews bzw. der Quartalsanschreibungen für das Bundesgebiet insgesamt 385 bzw. 671, für das Frühere Bundesgebiet 323 bzw. 624 und für die Neuen Länder 214 bzw. 473 Mikrozensus-Eckwerte vorgegeben, an die die Stichprobendaten anzupassen waren. Die hochgerechneten Werte stimmten bis auf eine Differenz von jeweils weniger als 100 mit den vorgegebenen Werten überein. Die ursprünglichen Hochrechnungsfaktoren wurden mit Faktoren überwiegend im Bereich von 0,4 bis 2,5 modifiziert. Schaubild 1 zeigt die Verteilung der Hochrechnungsfaktoren des Einführungsinterviews für das Bundesgebiet vor und nach der Anpassung.

*Schaubild 1**: Verteilung der Hochrechnungsfaktoren vor und nach der Anpassung mit "ADJUST"***



Größen-klasse	Hochrechnungs-faktoren von ... bis unter ...
1	50 - 75
2	75 - 100
3	100 - 200
4	200 - 300
5	300 - 400
6	400 - 500
7	500 - 750
8	750 - 1000
9	1000 - 2500
10	2500 - 5000
11	5000 u. mehr

6 Abschätzung des Stichprobenzufallsfehlers

6.1 Das Schätzverfahren

Um Aussagen über die Qualität der EVS-Ergebnisse treffen zu können, ist insbesondere eine Abschätzung der Präzision der Ergebnisse erforderlich. Die Präzision der Ergebnisse wird anhand der Stichprobenzufallsfehler beurteilt, deren Größenordnung mit Hilfe der relativen Standardfehler zuverlässig abgeschätzt werden kann. Hierbei ist zu beachten, dass eine Abschätzung der Stichprobenzufallsfehler streng genommen nur für Zufallsstichproben zulässig ist. Um dennoch Aussagen über die Präzision der Ergebnisse machen zu können, wird hilfsweise unterstellt, dass die Fehlerwerte der Quotenstichprobe näherungsweise den Fehlerwerten einer geschichteten Zufallsauswahl entsprechen, wobei die Schichtungsmerkmale die Quotierungsmerkmale sind. Die berechneten Standardfehlergrößen können also nur als Anhaltswerte dienen.
Neben dem Stichprobenzufallsfehler als wichtigster Komponente trägt auch eine Verzerrung des Schätzverfahrens zum gesamten Stichprobenfehler bei. Diese Verzerrung des Schätzverfahrens wurde nicht analysiert, da sie bei großen Stichprobenumfängen gegenüber dem Zufallsfehler in der Regel vernachlässigt werden kann. Neben den Stichprobenfehlern treten bei einer Erhebung auch Nichtstichprobenfehler auf. Sie werden im wesentlichen durch Antwortausfälle, unzutreffende und fehlende Angaben sowie Fehler bei der Datenaufbereitung verursacht. Um die Genauigkeit der Stichprobenergebnisse zu beschreiben, müssen auch diese Fehlerkomponenten betrachtet werden. Nichtstichprobenfehler sind aber nicht aus der Stichprobe abschätzbar, sondern können nur durch aufwendige Kontrollerhebungen nachgewiesen werden, was im Rahmen dieser Erhebung nicht realisierbar war.

Der relative Standardfehler $v_{\hat{Y}}$ (in %) für den hochgerechneten Totalwert $\hat{Y}$ des Merkmals Y wird geschätzt gemäß

$$v_{\hat{Y}} = \frac{s_{\hat{Y}}}{\hat{Y}} \cdot 100 \tag{20}$$

mit

$s_{\hat{Y}}$: Schätzwert für den Standardfehler von $\hat{Y}$.

Für die Berechnung des Standardfehlers kann bei einer Hochrechnung nach dem Prinzip des minimalen Informationsverlustes keine geschlossene Formel angegeben werden, da die Hochrechnungsfaktoren mit numerischen Verfahren bestimmt werden. Deville und Särndal (1992) zeigten, dass die Varianzen von Schätzfunktionen für eine große Klasse von Distanzmaßen, die bei einer Anpassung an vorgebende Eckwerte verwendet werden, bei nicht zu kleinen Stichprobenumfängen näherungsweise der Varianz folgender Regressionsschätzfunktion entspricht:

$$\hat{Y}_{reg} = \sum_{k=1}^{n_h} w_k y_k = \hat{Y}_\pi + (\mathbf{X} - \hat{\mathbf{X}}_\pi)^T \hat{\mathbf{B}}_s \tag{21}$$

mit:

$\mathbf{X} = \sum_{k=1}^{N} \mathbf{x}_k$ Vektor der Totalwerte der *m* Anpassungsmerkmale

$\hat{\mathbf{B}}_s = \left(\sum_{k=1}^{n} \pi_k^{-1} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum_{k=1}^{n} \pi_k^{-1} \mathbf{x}_k y_k$ Gewichteter Schätzer der Regressionskoeffizienten

$\hat{Y}_\pi$ bezeichnet den Horvitz-Thompson-Schätzer des Erhebungsmerkmals Y und $\hat{\mathbf{X}}_\pi$ den Horvitz-Thompson-Schätzer der Anpassungsmerkmale (analog zu Formel (9) in Vektorschreibweise).

In der EVS werden durchschnittliche Ausgaben und Einnahmen von Haushalten dargestellt. Für die geschätzte Varianz des Schätzwertes des Mittelwertes

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{k=1}^{n} w_k y_k \tag{22}$$

gilt wegen (21):

$$s^2_{\hat{\bar{Y}}} \doteq s^2_{\hat{\bar{Y}}_{Reg}} = \frac{1}{N^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\frac{\pi_{ki}-\pi_k\pi_i}{\pi_{ki}} w_k e_{y_k} w_i e_{y_i} \tag{23}$$

mit

$$\pi_{ki} = P(k \wedge i \in s) = \begin{cases} \dfrac{n(n-1)}{N(N-1)} & , \textit{falls } k \neq i \\ \pi_k = \dfrac{n}{N} & , \textit{sonst} \end{cases}$$

Wahrscheinlichkeit, dass die Einheiten k und i in die Stichprobe gelangen

$$e_{y_k} = y_k - \mathbf{x}_k^T \cdot \hat{\boldsymbol{\beta}}_y$$

Schätzwert der Residuen

$$\hat{\boldsymbol{\beta}}_y = \left(\sum_{k=1}^{n} w_k \mathbf{x}_k \mathbf{x}_k^T\right)^{-1}\left(\sum_{k=1}^{n} w_k \mathbf{x}_k y_k\right)$$

m-dimensionaler Vektor der Schätzwerte der Regressionskoeffizienten

Für geschichtete Zufallsstichproben gilt:

$$\pi_{ki} = \begin{cases} \dfrac{n_h(n_h-1)}{N_h(N_h-1)} & , \textit{falls } k \neq i \\ \pi_k = \dfrac{n_h}{N_h} & , \textit{sonst} \end{cases} \tag{24}$$

n_h bezeichnet die Zahl der EVS-Haushalte in der Schicht *h* und N_h die Zahl der hochgerechneten Mikrozensushaushalte in der Schicht *h*. Damit erhält man aus (23) die Formel:

$$s^2_{\hat{\bar{Y}}} \doteq \frac{1}{N^2}\sum_{h=1}^{L}\left(\sum_{k=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)(w_k e_{y_k})^2 - \sum_{\substack{k=1 \\ k\neq i}}^{n_h}\sum_{i=1}^{n_h}\frac{1}{n_h-1}\left(1-\frac{n_h}{N_h}\right) w_k e_{y_k} w_i e_{y_i}\right) \tag{25}$$

Die Ergebnisse der EVS werden nicht nur für die Haushalte insgesamt nachgewiesen, sondern auch gegliedert nach verschiedenen Merkmalen. Der Schätzwert für den Mittelwert einer Untergruppe *g* lautet:

$$\hat{\bar{Y}}_g = \frac{\hat{Y}_g}{\hat{N}_g} = \frac{\sum_{k=1}^{n_g} w_k y_k}{\sum_{k=1}^{n_g} w_k} \tag{26}$$

Die Varianz für den Mittelwert entspricht der Varianz eines kombinierten Verhältnisschätzers (siehe Krug, Nourney und Schmidt 1999, S. 116 ff.):

$$s^2_{\hat{\bar{Y}}_g} = \frac{1}{\hat{N}_g^2}\sum_{h=1}^{L}\left(s^2_{\hat{Y}_{gh}} - 2\hat{\bar{Y}}_g Cov\left(\hat{Y}_{gh}, \hat{N}_{gh}\right) + \hat{\bar{Y}}_g^2 s^2_{\hat{N}_{gh}}\right) \tag{27}$$

Zusammen mit (25) ergibt sich die Varianz für den Mittelwert einer Untergruppe g approximativ aus der Gleichung:

$$
\begin{aligned}
s^2_{\hat{\bar{Y}}_g} \doteq & \frac{1}{\hat{N}_g^2}\sum_{h=1}^{L}\sum_{k=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\left(\left(w_k e_{y_{gk}}\right)^2 - 2\hat{\bar{Y}}_g w_k{}^2 e_{y_{gk}} e_{N_{gk}} + \hat{\bar{Y}}_g^2\left(w_k e_{N_{gk}}\right)^2\right) \\
& - \frac{1}{\hat{N}_g^2}\sum_{h=1}^{L}\sum_{\substack{k=1\\k\neq i}}^{n_h}\sum_{i=1}^{n_h}\frac{1}{n_h-1}\left(1-\frac{n_h}{N_h}\right)\left(w_k e_{y_{gk}} w_i e_{y_{gi}} - 2\hat{\bar{Y}}_g w_k e_{y_{gk}} w_i e_{N_{gi}} + \hat{\bar{Y}}_g^2 w_k e_{N_{gk}} w_i e_{N_{gi}}\right)
\end{aligned}
\tag{28}
$$

mit

$$
e_{y_{gk}} = \begin{cases} y_k - \mathbf{x}_k^T \cdot \hat{\boldsymbol{\beta}}_{y_g}, & \text{falls } k \in g \\ -\mathbf{x}_k^T \cdot \hat{\boldsymbol{\beta}}_{y_g}, & \text{sonst} \end{cases}
$$

$$
e_{N_{gk}} = \begin{cases} 1 - \mathbf{x}_k^T \cdot \hat{\boldsymbol{\beta}}_{N_g}, & \text{falls } k \in g \\ -\mathbf{x}_k^T \cdot \hat{\boldsymbol{\beta}}_{N_g}, & \text{sonst} \end{cases}
$$

$$
\hat{\boldsymbol{\beta}}_{y_g} = \left(\sum_{k=1}^{n} w_k \mathbf{x}_k \mathbf{x}_k^T\right)^{-1}\left(\sum_{k=1}^{n_g} w_k \mathbf{x}_k y_k\right)
$$

$$
\hat{\boldsymbol{\beta}}_{N_g} = \left(\sum_{k=1}^{n} w_k \mathbf{x}_k \mathbf{x}_k^T\right)^{-1}\left(\sum_{k=1}^{n_g} w_k \mathbf{x}_k\right)
$$

6.2 Analyse der Stichprobenzufallsfehler

Die Stichprobenzufallsfehler wurden für verschiedene Kategorien der Ausgaben, Einkommen und Einnahmen der Haushalte, gegliedert nach Bundesländern und sozialer Stellung des Haupteinkommensbeziehers bzw. Haushaltsgröße, abgeschätzt. Schaubild 2 zeigt beispielhaft relative Standardfehler der Ausgaben für die Hauptgruppen des Privaten Verbrauchs der Haushalte des Früheren Bundesgebiets. Wie man dem Schaubild entnehmen kann, liegen die relativen Standardfehler für Ausgaben gegliedert nach der sozialen Stellung des Haupteinkommensbeziehers überwiegend unter 5%. Deutlich höhere Fehlerwerte ergeben sich nur für die relativ heterogenen Ausgabengruppen "Gesundheitswesen", "Verkehr" sowie "Bildungswesen" bei Arbeitslosenhaushalten. Aber auch die Ergebnisse für diese Positionen sind von der Präzision her noch uneingeschränkt aussagefähig. Ein ähnliches Bild ergibt sich für die relativen Standardfehler der Einkommen und Einnahmen (Schaubild 3). Lediglich für die Bruttoeinkommen aus selbständiger Arbeit und Einkommen aus nichtöffentlichen Transferzahlungen ergaben sich deutlich höhere Fehlerwerte als 5%.
Hier nicht dargestellte, tiefer als nach Hauptgruppen gegliederte Ergebnisse der Ausgaben für Waren und Dienstleistungen nach der Ausgabenart hatten zwar generell etwas höhere relative Standardfehler, es wurden jedoch nur bei einem geringen Anteil aller Fälle Werte über 5% erreicht. Deutlich höhere Fehler zeigten sich bei Ausgabenkategorien mit geringen Fallzahlen und bei Ausgabenkategorien, die besonders selten anfallen, wie zum Beispiel Ausgaben für Kraftfahrzeuge oder Pauschalreisen.

In Schaubild 4 werden Stichprobenschätzwerte des oben beschriebenen Hochrechnungsverfahrens mit entsprechenden Schätzwerten einer freien Hochrechnung in Verbindung mit aktuellem Hochrechnungsrahmen verglichen, um den Effekt des Anpassungsverfahrens auf Ergebnisse und Präzision dazustellen. Als Vergleichskriterium dienten Mittelwerte und relative Standardfehler der EVS im Verhältnis zu Mittelwerten und relativen Standardfehlern, die sich bei freier Hochrechnung mit aktuellem Hochrechnungsrahmen ergeben hätten.
Das Schaubild zeigt, dass die Verhältniszahlen der Mittelwerte für alle Ausgabenpositionen und Berufsgruppen nahe bei dem Wert Eins liegen, d.h. die Schätzwerte für die durchschnittlichen Ausgaben bei Anwendung des gewählten Hochrechnungsverfahrens unterscheiden sich nur wenig von denen der freien Hochrechnung mit aktuellem Hochrechnungsrahmen. Daraus kann allerdings nicht der Schluss gezogen werden, dass die Auswirkungen der unterschiedlichen Hochrechnungsverfahren auf die Gesamtwerte gering waren, da Zähler und Nenner der Mittelwerte in ähnlicher Weise betroffen sein können.

Die Verhältniszahlen der relativen Standardfehler zeigen fast durchweg leichte Verbesserungen bei der Präzision der Ergebnisse durch die Verwendung des gewählten Hochrechnungsverfahrens. Größere Präzisionsunterschiede konnten nicht erwartet werden, da für die Quotierung und Anpassung dieselben Merkmale verwendet wurden. Es kann aber davon ausgegangen werden, dass Verzerrungen durch Antwortausfälle auf diese Weise erheblich verringert werden konnten. In wenigen Fällen ist das Verhältnis der relativen Standardfehler größer als 1. In diesen Fällen dürfte die Variabilität der Hochrechnungsfaktoren nach Anpassung, die für die Schätzung des Ergebnisses innerhalb der Nachweisungsgruppe verwendet wurden, relativ groß sein oder ein geringer statistischer Zusammenhang zwischen dem Erhebungsmerkmal und den Anpassungsmerkmalen bestehen.
Betrachtet wurden hier nur die Ausgaben der Haushalte aus dem Früheren Bundesgebiet, gegliedert nach der sozialen Stellung des Haupteinkommensbeziehers. Vergleichbare Resultate wurden bei der Untersuchung der Ausgaben der Neuen Länder oder bei einer Gliederung der Haushalte nach der Haushaltsgröße erzielt.

Schaubild 2: *Relative Standardfehler* $v_{\hat{\bar{Y}}_g}$ *der Ausgaben für ausgewählte Positionen des Privaten Verbrauchs (Früheres Bundesgebiet)*

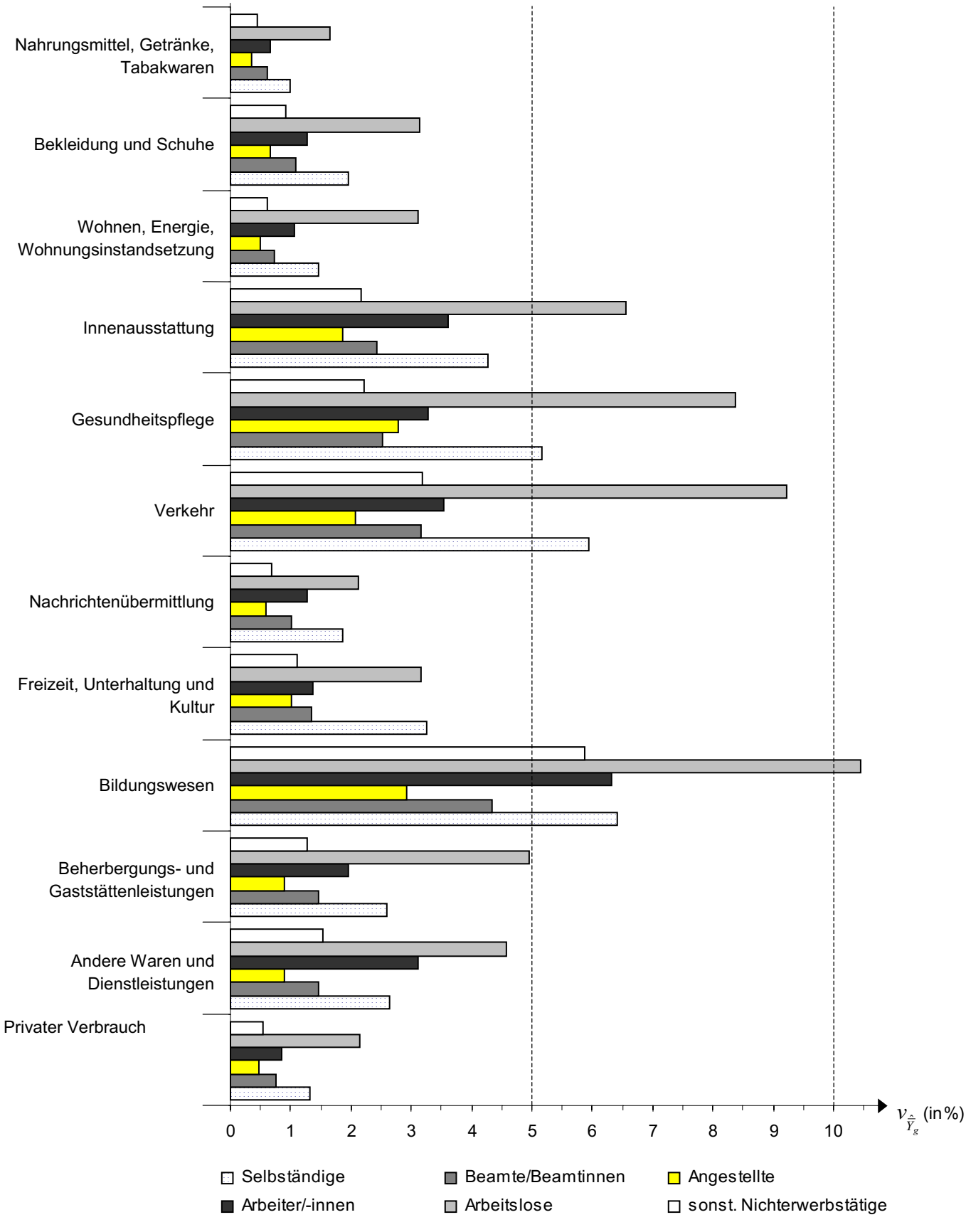


Schaubild 3: *Relative Standardfehler* $v_{\hat{\bar{Y}}_g}$ *der Einkommen und Einnahmen*

(Früheres Bundesgebiet)

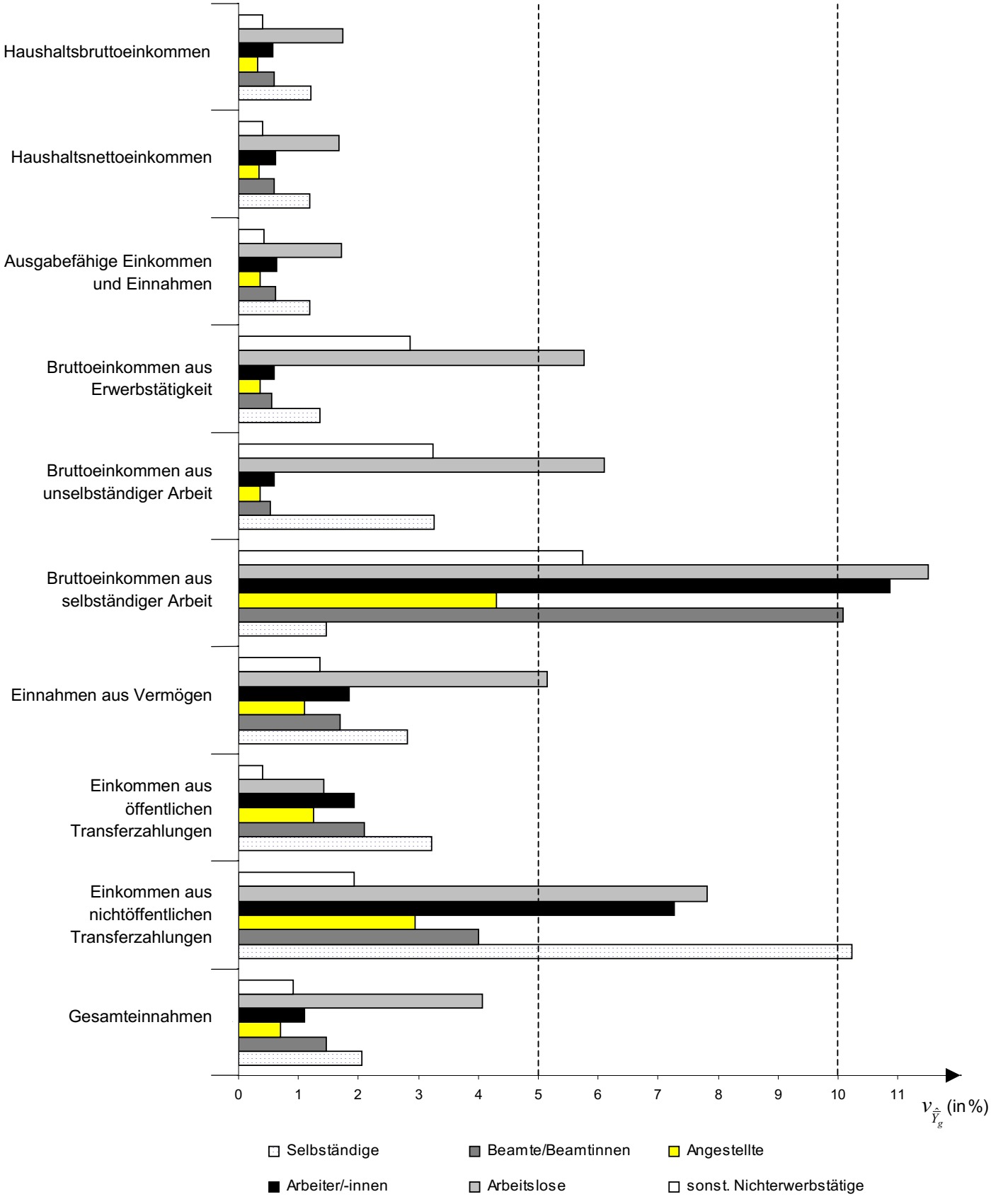
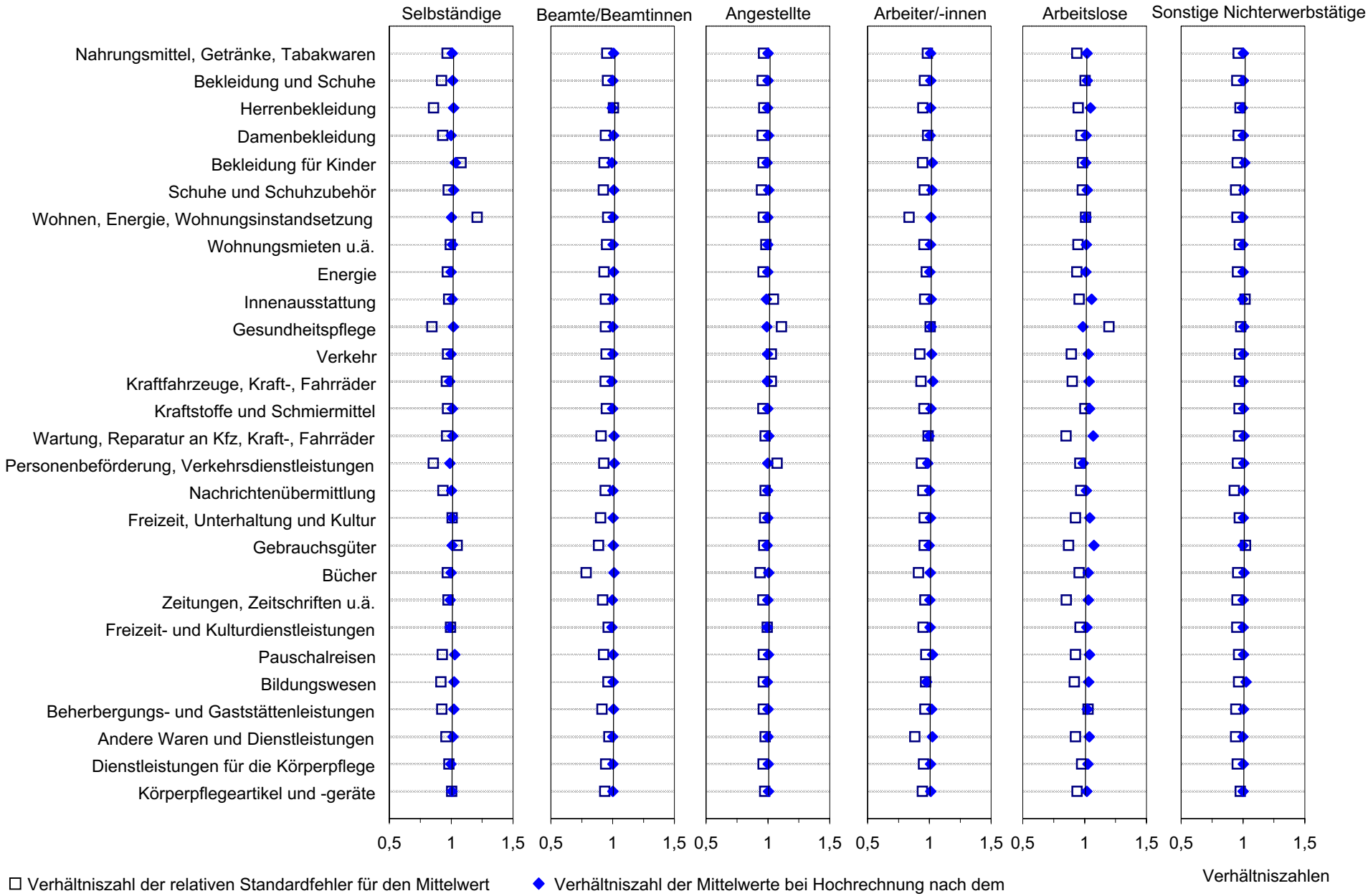


Schaubild 4: *Effekt des Hochrechnungsverfahrens mit minimalem Informationsverlust auf die Ergebnisse für ausgewählte Ausgabenpositionen des Privaten Verbrauchs und deren Präzision (Früheres Bundesgebiet)*



☐ Verhältniszahl der relativen Standardfehler für den Mittelwert bei Hochrechnung nach dem Prinzip des minimalen Informationsverlusts und bei freier Hochrechnung

◆ Verhältniszahl der Mittelwerte bei Hochrechnung nach dem Prinzip des minimalen Informationsverlusts und freier Hochrechnung

7 Literatur

Chlumsky, J./Ehling, M. (1997): *Grundzüge des künftigen Konzepts der Wirtschaftsrechnungen der privaten Haushalte,* in: Wirtschaft und Statistik 7/1997, S. 455 - 461.

Deville, J.-C./Särndal, C.-E. (1992*): Calibration Estimators in Survey Sampling,* in: Journal of the American Statistical Association, 87, S.376 - 382.

Gertkemper, F./Kühnen, C./Wein, E. (1998): *Ergebnisbericht der Testerhebung zur Neukonzeption der Laufenden Wirtschaftsrechnungen,* Wiesbaden.

Krug, W./Nourney, M./Schmidt, J. (1999): *Wirtschafts- und Sozialstatistik – Gewinnung von Daten,* Oldenbourg-Verlag, München, Wien, 5. Auflage.

Merz, J. (1983): *Die konsistente Hochrechnung von Mikrodaten nach dem Prinzip des minimalen Informationsverlustes,* in: Allgemeines Statistisches Archiv, Bd. 67, S. 342 - 366.

Merz, J. (1994): *Microdata adjustment by the minimum information loss principle,* FFB-Discussion Paper No. 10, Department of Economics and Social Sciences, University of Lüneburg.

Münnich, M./Illgen, M. (1999): *Ausstattung privater Haushalte mit langlebigen Gebrauchsgütern – Erste Ergebnisse der Einkommens- und Verbrauchsstichprobe (EVS) 1998,* in: Wirtschaft und Statistik 1/1999, S. 46 - 54.

Theil, H. (1967): *Economics and Information Theory,* Amsterdam.

Wauschkuhn (1982): *Anpassung von Stichproben und n-dimensionalen Tabellen an Randbedingungen,* Bericht Nr. 138 der GMD, München, Wien.