Varianzschätzung im faktisch anonymisierten Mikrozensus*

Ulrich Rendtel
Bernhard Schimpl-Neimanns

Vortrag auf der 2. Nutzerkonferenz
„Forschung mit dem Mikrozensus“,
Mannheim, 12.-13. Oktober 2000

Zusammenfassung
Erstmals enthält der faktisch anonymisierte Mikrozensus (FAMZ) 1996 Stichprobeninformationen, die eine Berechnung der Varianz von Populationsschätzern ermöglichen. Nach der Darstellung der Ziehung des Mikrozensus (MZ) und des FAMZ wird ein methodisches Konzept zur Berechnung der Varianz entwickelt und diskutiert, das mit Standard-Software umgesetzt werden kann. Dieses Konzept wird auf die Schätzung von Totals, Verhältniszahlen und Anteilen angewendet. Neu ist die Behandlung von Hochrechnungsergebnissen nach der Anpassung an die Bevölkerungsfortschreibung im Rahmen eines Regressionsansatzes. Schließlich wird überprüft, inwieweit die Design-Zuschlagsfaktoren, das bisher einzige Instrument der FAMZ-Nutzer zur Schätzung der Varianz, brauchbare Ergebnisse auch für den FAMZ liefern. Im empirischen Teil werden für angewählte Merkmale Varianzschätzungen von MZ und FAMZ miteinander verglichen. Es wird gezeigt, daß die Varianz-Komponente bezüglich der zweiten Auswahlstufe der Substichprobenziehung des FAMZ aus dem MZ vernachlässigt werden kann, was die Berechnung der Varianzen erheblich vereinfacht. Ein weiteres Ergebnis besagt, daß die Varianzvergrößerung des FAMZ gegenüber dem MZ geringer ausfällt als dies bei einer einfachen Stichprobenziehung zu erwarten ist. Schließlich wird gezeigt, daß die Approximation der Varianz über die Design-Zuschlagsfaktoren auch für den FAMZ zu brauchbaren Ergebnissen führt, in Einzelfällen aber mit erheblichen Über- bzw. Unterschätzungen der Varianz verbunden ist.

---

* Diese Arbeit entstand als Resultat eines Gastaufenthalt des Erstautors im Oktober 1999 bei ZUMA. Für hilfreiche Diskussionen danken wir Wolf Bihler (Statistisches Bundesamt), Ralf Münnich (Universität Tübingen) sowie Siegfried Gabler, Sabine Häder und Michael Wiedenbeck (ZUMA).

Anschriften:
U. Rendtel, Institut für Statistik und Mathematik, FB Wirtschaftswissenschaften, J.W. Goethe-Universität Frankfurt am Main, Mertonstr. 17, D-60054 Frankfurt, Tel.: ++49 / 69 / 798-237 55, Fax: ++49 / 69 / 798-236 35, Email: Rendtel@em.uni-frankfurt.de
B. Schimpl-Neimanns, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Abt. Mikrodaten, Postfach 12 21 55, D-68072 Mannheim, Tel.: ++49 / 621 / 12 46 – 263, Fax: ++49 / 621 / 12 46 – 100, Email: Schimpl-Neimanns@zuma-mannheim.de

1 Einleitung

Als Folge eines vom Bundesministerium für Bildung und Forschung (BMBF) finanzierten Pilotprojekts haben sich seit Ende 1996 für die Forschung die Nutzungsmöglichkeiten von anonymisierten Mikrodaten der amtlichen Statistik erheblich verbessert (Köhler et al. 2000). Der im Vergleich zu früher vor allem in finanzieller Hinsicht wesentlich erleichterte Datenzugang hat neue Analysemöglichkeiten eröffnet. In kurzer Zeit sind insbesondere die Mikrozensusdaten von der Forschung für die Bearbeitung verschiedenster Fragestellungen herangezogen worden (siehe u.a. die Aufsätze in Lüttinger 1999). Das Pilotprojekt hat auch zu einer verstärkten Diskussion zwischen amtlicher Statistik und Forschung über weitere Möglichkeiten der Zusammenarbeit geführt (vgl. Statistisches Bundesamt 1999).

Mit einem Stichprobenumfang von einem Prozent der Personen und Haushalte ist der Mikrozensus die größte Mehrthemenumfrage der Bundesrepublik Deutschland, in der jährlich vielfältige Informationen über die demographische, soziale und wirtschaftliche Struktur der Bevölkerung erhoben werden. Der Forschung steht eine 70%-Substichprobe mit über 500.000 Personen in rund 230.000 Haushalten zur Verfügung. Dieser große Stichprobenumfang erlaubt die differenzierte Analyse auch kleinerer Bevölkerungsgruppen. Neben der Stichprobengröße ist der praktisch vernachlässigbare Unit-Nonresponse hervorzuheben. Aufgrund der Auskunftspflicht liegt die Teilnahmequote der Haushalte bei rund 97 Prozent, so daß mit Mikrozensusdaten Verteilungs- und Zusammenhangsmaße für die Population verläßlich geschätzt werden können. In diesem Zusammenhang dient der Mikrozensus häufig als Referenzstatistik für die normalen Bevölkerungsumfragen, die vielfach Ausschöpfungsquoten unter 60 Prozent aufweisen.

Wie bei allen Stichproben ist für die Beurteilung der Qualität der Schätzungen die Größe des Stichprobenfehlers von zentraler Bedeutung.[1] Hierbei ist zu berücksichtigen, daß der Mikrozensus keine einfache, uneingeschränkte Zufallsstichprobe, sondern eine geschichtete Flächen- bzw. Klumpenstichprobe ist. Weil die faktisch anonymisierten Mikrozensusdaten bis einschließlich 1995 keine Informationen zum Stichprobenplan enthielten, konnten die Nutzer die Schätzungen nur unter der Annahme einer einfachen Zufallsstichprobe durchführen. Da aber die Klumpung in der Regel mit einer Varianzvergrößerung verbunden ist, wird hierbei

[1] Grundsätzlich sind neben dem Stichprobenfehler aber auch systematische Fehler zu beachten, die z.B. durch falsche Angaben, Kodierfehler und Befragungsausfälle etc. entstehen. In diesem Zusammenhang ist darauf hinzuweisen, daß im Mikrozensus die Zahl geringfügig Beschäftigter systematisch unterschätzt wird (Schupp et al. 1999; vgl. hierzu auch zusammenfassend Kohler et al. (1996).

der Stichprobenfehler unterschätzt. Nur sehr eingeschränkt war mit Hilfe der Fehlerrechnungen des Statistischen Bundesamtes (1998a) eine Korrektur möglich.

Aus Sicht der Nutzer der faktisch anonymisierten Daten war deshalb eine sachgerechte Lösung dieses Problems erforderlich. Um den Forschungszwecken bei der Analyse von Mikrodaten gerecht zu werden, sollten die Nutzer den Stichprobenfehler selbst ermitteln können. Auf Nachfrage von Mikrozensusnutzern stellt nun das Statistische Bundesamt ab dem Mikrozensus 1996 Informationen über den Stichprobenplan in anonymisierter Form zur Verfügung. Damit ist in der Forschungspraxis erstmals eine adäquate Berechnung der Stichprobenfehler möglich geworden. Wie man selbst Varianzschätzungen unter Berücksichtigung des Stichprobendesigns mit den faktisch anonymisierten Mikrozensusdaten vornehmen kann und welche Unterschiede ggf. zu den Fehlerrechnungen der statistischen Ämter bestehen, wird in diesem Beitrag diskutiert. Hierbei sind insbesondere zwei Fragen zu klären, die im wesentlichen damit zusammenhängen, daß sich die faktisch anonymisierten Daten notwendigerweise vom Originalmaterial unterscheiden:[2] Wie wirken sich die aus Datenschutzgründen fehlenden differenzierten Informationen zur regionalen Schichtung aus und wie können Anwender die vom Original-Stichprobenplan abweichende Substichprobenziehung auf Haushaltsebene bei der Varianzschätzung berücksichtigen?

Im folgenden wird zunächst das Erhebungsdesign des Mikrozensus skizziert. Daran schließt sich die Darstellung der Ziehung des faktisch anonymisierten Mikrozensus an. Danach wird gezeigt, wie einfach durchzuführende Schätzungen für die Varianz bestimmter Populationsschätzwerte gemacht werden können. Im einzelnen werden behandelt: Die Schätzung von Populationstotals sowie die Schätzung von Verhältniswerten und Subpopulationen. Weiterhin wird die Anpassung an die Bevölkerungsfortschreibung mit Hilfe des von Särndal et al. (1992) entwickelten Regressionsansatzes behandelt. Schließlich wird die bisher für den Mikrozensus benutzte Methode der linearen Approximation der Design-Effekte für den faktisch anonymisierten Mikrozensus untersucht.

2 Das Erhebungsdesign des Mikrozensus

Der faktisch anonymisierte Mikrozensus (FAMZ) ist eine 70-prozentige Substichprobe des Mikrozensus. Um die Varianz von Populatonsschätzungen auf Basis des FAMZ bestimmen zu können, ist es deshalb notwendig, zunächst das Erhebungsdesign des MZ darzustellen (sie-

[2] Vgl. zu den Anonymisierungsmaßnahmen im Mikrozensus Müller et al. (1991).

he Krug et al. 1994: 239ff.; Meyer 1994).[3] Beim MZ werden zunächst Primäreinheiten (Primary Sampling Units = PSU's) gebildet. Diese werden als Auswahlbezirke bezeichnet. Sie bestehen aus durchschnittlich 9 benachbarten Wohnungen, die in einer Gebäudegruppe oder innerhalb eines größeren Gebäudes liegen. Im Unterschied zu früheren MZ-Erhebungen ist die Anzahl der Wohnungen pro PSU von ca. 23 auf 9 deutlich reduziert worden.

Die Bildung dieser PSU's erfolgte separat innerhalb von 4 Gebäudeklassen (sogenannte fachliche Schichtung) auf Basis von Ergebnissen der Volkszählung 1987 bzw. des Zentralen Einwohnerregisters in den neuen Bundesländern ab 1991. Der MZ 1996 enthält über 40.000 PSU's. Alle Haushalte einer ausgewählten PSU werden befragt. Aus diesem Grund wird der MZ auch als Klumpenstichprobe (= Cluster Sample) bezeichnet. Tendenziell vergrößert die Klumpung die Varianz der Populationsschätzer im Vergleich zu einer einfachen Stichprobe.[4]

Das zweite Element des MZ ist die regionale Schichtung, die den umgekehrten Effekt hat, nämlich die Varianz zu verringern. Hierbei bilden Großstädte über 200 Tsd. Einwohner und sonstige Regionen über 250 Tsd. Einwohner eigene Schichten. Innerhalb der Schichten werden jedoch noch weitere Strukturen berücksichtigt. So werden Schichtuntergruppen mit mindestens 100 Tsd. Einwohnern identifiziert, deren regionale Anordnung bei der Auswahl der PSU's berücksichtigt wird.

Für den MZ wurden 20 Vorratsstichproben mit einem Auswahlsatz von 1% gezogen. Hierbei wurden innerhalb der fachlichen und regionalen Schichten die PSU's nach den Merkmalen Kreis, Gemeindegrößenklasse und Gemeinde angeordnet. Innerhalb der Gemeinden wurden die PSU-Nummern, die die regionale Anordnung innerhalb der Schichtuntergruppen wiederspiegeln, zu einer vollständigen Anordnung der PSU's verwendet. Jeweils 100 aufeinander folgende PSU's wurden zu einer Zone zusammengefaßt.

Der Kern des Ziehungsverfahrens besteht in einer zufälligen Zuordnung der jeweils 100 PSU's einer Zone zu den Zahlen 1 bis 100. Die PSU's zu einer Zahl zwischen 1 bis 100 bilden dann jeweils eine MZ-Stichprobe. Aus der Auswahlgesamtheit aller PSU's wurden 20 Vorratsstichproben ausgewählt. Interpretiert man dieses Verfahren innerhalb des Schemas der klassischen Stichprobentheorie, so entsprechen die Zonen Schichten, aus denen jeweils eine Primäreinheit gezogen wird. Für ein derartiges Ziehungsverfahren existieren aber keine Varianzformeln, da hierfür die Varianz der PSU-Totals innerhalb einer Zone benötigt wird. Eine

---

[3] Siehe hierzu auch die Darstellungen der statistischen Landesämter in Frank und Kafurke 1990; Reinders 1993; Schmidt 1990; Werner 1994.

[4] Der Stichprobenfehler ist in der Regel um so größer, je homogener die Klumpen hinsichtlich der interessierenden Merkmale, je größer die Klumpen, und je unterschiedlicher die Klumpengrößen sind.

Möglichkeit, mit dieser Schwierigkeiten umzugehen, besteht darin, die Schichten größer zu definieren und die Zonenstruktur des Ziehungsverfahrens zu ignorieren; d.h. für diese größeren Schichten eine einfache Zufallsstichprobe zu unterstellen.[5] Die Annahme größerer Schichten und das Ignorieren der kleinräumigen Struktur bei dem Ziehungsverfahren führt in der Tendenz zu einer Überschätzung der Varianz. Es gehört zu den vielleicht unerwarteten Ergebnissen, daß schon im Falle des Mikrozensus - so, wie er in den statistischen Ämtern vorliegt - die Stichprobenvarianz nur näherungsweise bestimmt werden kann.

Die Benutzung dieses vereinfachten Berechnungsverfahrens wird auch noch durch einen zweiten Sachverhalt nahegelegt. In den Jahren nach der Ziehung der MZ-Stichproben sind Neubauten entstanden bzw. ganze Flächen neu bebaut worden. Die Neubauten werden bei der Aktualisierung der Stichproben in einer zusätzlichen „Neubauschicht“ berücksichtigt und nach einem gesonderten Ziehungsverfahren ausgewählt. Zunächst werden auf Kreis- bzw. Gemeindeebene Primäreinheiten auf Basis der Bautätigkeitsstatistik gebildet. Diese PSU's werden nach Gebäudegrößenklassen geschichtet und innerhalb jeder Schicht nach der Reihenfolge ihrer Bildung durchnummeriert. Im Unterschied zu dem oben beschriebenen Verfahren erfolgt die Ziehung der PSU's hier aber durch systematisches Ziehen mit festem Intervall und zufälligem Startpunkt. Hierfür wird pro Regionalschicht eine Zufallszahl Z zwischen 1 und 100 gezogen. Die PSU's mit dem Ordnungsnummern $Z, Z+100, Z+200, \ldots$ bilden dann die erste Stichprobe, die PSU's mit den Ordnungsnummern $Z+1, Z+101, Z+201, \ldots$ die 2. Vorratsstichprobe, usw. bis zur 20. Vorratsstichprobe.

Für das systematische Ziehungsverfahren ist bekannt, daß keine erwartungstreuen Varianzschätzer existieren (vgl. Wolter 1985: 248), so daß man sich auch in diesem Fall mit einer Vereinfachung begnügen muß. Erfahrungen aus Simulationsstudien zeigen, daß die Approximation über die Varianz einer in den Schichten einfachen Zufallsstichprobe in der Regel zu einer Überschätzung der Varianz führt (vgl. Wolter 1985: 282).

Nach dem Stichprobendesign des MZ werden 0,25%-Stichproben benötigt, da im Rahmen des Rotationsverfahrens jährlich ein Viertel des Bestandes ausgetauscht wird.[6] Um die 20 1%-Vorratsstichproben in Rotationsviertel zu zerlegen, werden jeweils vier aufeinanderfolgende

---

[5] Diese Strategie wird auch in den Varianzberechnungen durch das Statistische Bundesamt angewendet. Innerhalb jeder Gebäudeklasse bilden die insgesamt über 200 Regionalschichten jeweils eine Schicht bei der Varianzberechnung.

[6] Der MZ ist als rotierende Panelstichprobe angelegt, bei der ein Auswahlbezirk bzw. die darin wohnenden Haushalte vier Jahre lang befragt werden, wobei jährlich ein Viertel der PSU's ausgetauscht wird. Wegziehende Personen bzw. Haushalte werden allerdings nicht weiterbefragt, sondern durch die nachziehenden Haushalte ersetzt (Prinzip der Flächenstichprobe).

Zonen einer Zufallszahl von 1 bis 4 zugeordnet, wobei PSU's mit gleicher Nummer zum gleichen Rotationsviertel zählen.

Da alle Stichproben nach demselben Verfahren ermittelt wurden, ergibt sich die Möglichkeit, die Variation der Schätzergebnisse über diese Vorratsstichproben für die Varianzschätzung zu benutzen. Allerdings ist im FAMZ 96 die Zugehörigkeit der Haushalte zu den Rotationsgruppen nicht dokumentiert. Bei Kenntnis der Rotationsgruppenzugehörigkeit könnte man die Varianz über das Jackknife-Verfahren analysieren (vgl. Wolter 1985, Kapitel 4).

3 Die Ziehung der 70%-Substichprobe

Bei der Ziehung des FAMZ 96 wurden andere Schichtungsmerkmale und Ziehungseinheiten verwendet. Hinzu kommt, daß die Regionalschicht als regionales Schichtungsmerkmal bei der Ziehung des MZ aus Datenschutzgründen nicht in den FAMZ aufgenommen werden konnte. Dies erschwert die Berechnung der Stichprobenvarianz.

Zunächst wurden die Haushalte des MZ 96 in eine Anordnung gebracht, die durch die Variablen Bundesland, Regierungsbezirk, Gemeindegrößenklasse, Zahl der Personen im Privathaushalt, Auswahlbezirksnummer und Haushaltsnummer bestimmt ist. Es wurden alle Haushalte gewählt, deren letzte Platzziffer von 2, 5 und 9 verschieden war. Dieses Vorgehen garantiert eine Auswahl von 70 Prozent aller Haushalte.

Im Anschluß an die Stichprobenziehung wurden die Haushalte umsortiert und erhielten aus Datenschutzgründen eine neue, fortlaufende Nummerierung. Der Verlust der Information über die letzte Platzziffer beim Auswahlverfahren ist insofern bedauerlich, als man aus der Varianz der Schätzergebnisse über die 7 verbliebenden Platzziffern Rückschlüsse auf die zusätzliche Varianzkomponente hätte gewinnen können, die durch die 70%-Auswahl verursacht wird.

4 Die Berechnung der Varianz im faktisch anonymisierten Mikrozensus

Obwohl der FAMZ somit im strengen Sinne über ein 2-phasiges Ziehungsverfahren[7] auf Haushaltsebene gewonnen wurde, ist das Verfahren in guter Näherung durch ein 2-stufiges Auswahlverfahren beschreibbar. Hierbei entspricht die erste Stufe der geschichteten Auswahl der PSU's im Mikrozensus. Die zweite Stufe ist durch eine einfache 70%-Auswahl von Haushalten aus einer PSU gekennzeichnet.

[7] Phase 1: Auswahl der Haushalte des MZ, Phase 2: Auswahl der Haushalte aus der MZ Stichprobe.

Die Annahme fester Stichprobenumfänge für die zweite Auswahlstufe wird jedoch durch das systematische Ziehungsverfahren nicht immer gewährleistet und bildet die Quelle für eine Variation der Schätzergebnisse, die hier ignoriert wird. Weiterhin sind aufgrund des systematischen Ziehungsverfahrens die Ziehungen der Haushalte zwischen den einzelnen PSU's nicht unabhängig voneinander. Aufgrund des hohen Auswahlsatzes ist aber nicht zu vermuten, daß durch das Ziehungsverfahren hohe Abhängigkeiten in den gemessenen Merkmalen entstehen.[8]

Schließlich enthält der FAMZ nicht alle Schichtungsmerkmale der ersten Auswahlstufe, sondern nur die Merkmale Bundesland, Gemeindegrößenklasse sowie die Gebäudeschicht. Generell vergrößert die Nichtberücksichtigung von Schichtungsmerkmalen die Schätzung der Varianz, so daß man bei der Bildung von Konfidenzintervallen konservativ bleibt.

Die im folgenden benutzte Notation lehnt sich an das Lehrbuch von Särndal et al. (1992) an. Es bezeichne $h \in \{1,...,H\}$ den Schichtindex. Die PSU's werden mit $i, j \in \{1,...,N_h\}$ indiziert, wobei N_h die Anzahl der PSU's in der $h.$ Schicht ist. Haushalte werden mit $k,l \in \{1,...,N_i\}$ indiziert, wobei N_i die Anzahl der Haushalte in der $i.$ PSU ist. $y_{h,i,k}$ bezeichnet den Merkmalswert von Haushalt k in PSU i in Schicht h, y_k verweist auf den Merkmalswert von Haushalt k unabhängig von dessen PSU-Nummer und Schichtzugehörigkeit.

Für ein 2-stufiges geschichtetes Ziehungsverfahren müssen unterschiedliche Grundgesamtheiten berücksichtigt werden. Es sei U_I die Menge aller Primäreinheiten im Erhebungsgebiet (Anzahl = N_I). $U_{I,h}$ sei die Menge der Primäreinheiten in Schicht h (Anzahl = $N_{I,h}$). Die Menge aller Haushalte in der $i.$ PSU sei U_i (Anzahl = N_i). Schließlich bezeichne U die Menge aller Haushalte im Erhebungsgebiet (Anzahl = N).

Den Grundgesamtheiten entsprechen die unterschiedlichen Stichproben: s_I ist die Stichprobe der PSU's von Umfang n_I. Diese Stichprobe verteilt sich auf die H Schichten als $s_{I,h}$ mit dem Umfang $n_{I,h}$. Die Stichprobe der Haushalte aus PSU i sei s_i und habe den Umfang n_i. Schließlich bezeichne s die Stichprobe aller Haushalte mit dem Umfang n.

Weiterhin bezeichne $\pi_{I,i}$ die Ziehungswahrscheinlichkeit von PSU i und $\pi_{k|i}$ die bedingte Ziehungswahrscheinlichkeit von Haushalt k aus PSU i, wenn PSU i gezogen wurde. Im Falle des MZ gilt $\pi_{I,i} = 0{,}01$. Für den FAMZ nehmen wir $\pi_{k|i} = 0{,}7$ an.

[8] Ein solcher Fall würde beispielsweise eintreten, wenn ein Merkmal räumlich so verteilt wäre, daß bei einer bestimmten Realisationen des Ziehungsverfahrens fast alle Merkmalsvertreter in der Stichprobe wären und

5 Schätzung eines Totals

Das Gesamtaufkommen (Total) eines Merkmals y ist gegeben durch:

(1) $t = \sum_{k \in U} y_k = \sum_{i \in U_I} t_i$ wobei $t_i = \sum_{k \in U_i} y_k$

(2) $= \sum_{h=1}^{H} t_h$ wobei $t_h = \sum_{i \in U_{i,h}} t_i$

Hierbei ist (1) die Darstellung über die Totals der Primäreinheiten und (2) die Darstellung über die Schichttotals.

Der π-Schätzer von t basiert auf dem Kehrwert der Ziehungswahrscheinlichkeiten. Man erhält für das PSU-Total die Schätzung:

(3) $\hat{t}_i = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} = \frac{1}{0{,}7} \sum_{k \in s_i} y_k$

Das Schichttotal t_h läßt sich schätzen durch:

(4) $\hat{t}_h = \sum_{i \in s_{i,h}} \frac{\hat{t}_i}{\pi_{I|i}} = 100 \sum_{i \in s_{I,h}} \hat{t}_i$

Man enthält damit als Schätzung für das Gesamttotal:

(5) $\hat{t} = \sum_{h=1}^{H} \hat{t}_h = \frac{100}{0{,}7} \sum_{k \in s} y_k$

Wir wollen im folgenden $V(\hat{t})$, die Varianz von $\hat{t}$, bestimmen sowie eine Schätzung von $V(\hat{t})$ angegeben. Bei Unabhängigkeit der Ziehung zwischen den Schichten erhält man:

(6) $V(\hat{t}) = \sum_{h=1}^{H} V_h(\hat{t}_h)$

(7) $\hat{V}(\hat{t}) = \sum_{h=1}^{H} \hat{V}_h(\hat{t}_h)$

D.h. es genügt, die Varianz des geschätzten Totals $\hat{t}_h$ für die einzelnen Schichten zu bestimmen. Dies soll im folgenden geschehen: Bei einem 2-stufigen Ziehungsverfahren, mit dem auf jeder Stufe eine einfache (simple: SI) Stichprobe gezogen wird (Typ: (SI, SI) in der Notation von Särndal et al. 1992: 142), erhält man:

---

bei anderen Realisationen nur sehr wenige.

(8) $$V_{SI,SI}\left(\hat{t}_h\right) = N_{I,h}^2 \frac{1-f_h}{n_{I,h}} S_{U_{I,h}}^2 + \frac{N_{I,h}}{n_{I,h}} \sum_{i \in U_{I,h}} N_i^2 \frac{1-f_i}{n_i} S_{U_i}^2$$

wobei $f_h = \frac{n_{I,h}}{N_{I,h}} = 0{,}01$, $f_i = \frac{n_i}{N_i} = 0{,}7$ und:

(9) $$S_{U_{I,h}}^2 = \frac{1}{N_{I,h}-1} \sum_{i \in U_{I,h}} (t_i - \bar{t}_{U_{I,h}})^2$$

= Varianz der PSU - Werte in der Schicht *h*

= "Varianz between PSU"

(10) $$S_{U_i}^2 = \frac{1}{N_i - 1} \sum_{k \in U_i} (y_k - \bar{y}_{U_i})^2$$

= Varianz der y - Werte in der *i.* PSU

= "Varianz within PSU"

Für die gegebenen Werte vereinfacht sich $V_{SI,SI}$ zu:

(11) $$V_{SI,SI} = 100^2 \times 0{,}99 \times n_{I,h} \times \text{ "Varianz between PSU"} + 100 \sum_{i \in U_{I,h}} \frac{0{,}3}{0{,}7^2} n_i \times \text{ "Varianz within PSU"}$$

Eine erwartungstreue Schätzung der Varianz $\hat{V}_{SI,SI}$ von $V_{SI,SI}$ erhält man, indem in Gleichung (8) bzw. (11) die Between- und Within-Varianzen durch ihre Stichprobenpendants ersetzt werden. Diese sind gegeben durch:

(12) $$S_{s_{I,h}}^2 = \frac{1}{n_{I,h}-1} \sum_{i \in s_{I,h}} \left(\hat{t}_i - \hat{\bar{t}}_{s_{I,h}}\right)^2$$

= Varianz der geschätzten PSU - Werte in der Stichprobe in der Schicht *h*

(13) $$S_{s_i}^2 = \frac{1}{n_i - 1} \sum_{k \in s_i} \left(y_k - \bar{y}_{s_i}\right)^2$$

= Varianz der y - Werte in der *i.* PSU

Die Berechnung des zweiten Teils von $\hat{V}_{SI,SI}$ verlangt die Ermittlung der Standardabweichung der y-Werte innerhalb von ca. 40.000 PSU's. Dies ist zwar rechenaufwendig, jedoch kein prinzipielles Problem. Särndal et al. (1992: 139ff.) schlagen als Näherung für $\hat{V}_{SI,SI}$ die Berechnung lediglich des ersten Terms über die Between-Varianzen vor. Dies ist gerechtfertigt, wenn die Between-Varianz zwischen den PSU's deutlich größer ist als die Within-PSU Varianz. Wie die numerischen Ergebnisse in Tabelle 1 belegen, gilt dies auch für den FAMZ.

Standardprozeduren für die Schätzung der Varianz über die Between-PSU-Varianz innerhalb von Schichten sind mittlerweile in Statistikpaketen wie Stata und SAS implementiert.

Die Tabelle 1 zeigt für einige ausgewählte Merkmale die Schätzung von Totals und vergleicht die Schätzgenauigkeit mit den Angaben des Statistischen Bundesamtes für den MZ. Weiterhin werden die relativen Standardfehler miteinander verglichen sowie die in Abschnitt 9 behandelte Näherungsfunktion auf Basis des Design-Effekt-Faktors. Zunächst einmal fällt auf, daß die geschätzten Totals für den FAMZ stets kleiner sind als für den MZ. Diese systematische Abweichung ist darauf zurückzuführen, daß das Statistische Bundesamt für seine Berechnungen eine Kompensationsgewichtung für die ca. 2,5% Nonresponse-Fälle auf Haushaltsebene benutzt. Die für die einzelnen Haushalte unterschiedlichen Gewichte liegen jedoch für den FAMZ nicht vor und können folglich auch nicht für die Schätzung von Totals benutzt werden.[9]

--- Tabelle 1 ---

Der Vergleich der geschätzten Standardfehler offenbart, daß die Within-Varianz eine zu vernachlässigende Größe im Vergleich zur Between-Varianz ist. Der Vergleich mit den MZ-Werten zeigt, daß die FAMZ-Werte stets über den MZ-Werten liegen. Dies war auch angesichts des um den Faktor 0,7 niedrigeren Stichprobenumfangs zu erwarten gewesen. Allerdings fällt die Vergrößerung des Standardfehlers nur etwa halb so groß aus, als dies bei einer entsprechenden Reduktion des Stichprobenumfangs bei einer einfachen Stichprobe der Fall gewesen wäre.[10] Der Grund für diese Halbierung des Fallzahleffekts liegt in der approximativen Erhaltung der Schichtung durch die Sortierfolge der Haushalte bei der Auswahl des FAMZ aus dem MZ.

Die Betrachtung des relativen Standardfehlers (=Standardfehler/Total; s. Abschnitt 9) belegt die hohe Präzision von Ergebnissen auf Basis des FAMZ, die ihresgleichen bei den üblichen sozialwissenschaftlichen Erhebungen sucht. Selbst bei relativ schwach besetzten Merkmalen, wie z.B. Frauen mit einem Nettoeinkommen unter 600 DM, liegt der relative Standardfehler bei 1 Prozent.[11]

---

[9] Bezogen auf das Standardauswertungsprogramm von über 450 Merkmalen betragen die FAMZ-Totals durchschnittlich 97,7% der entsprechen MZ-Totals.

[10] Statt des Faktors $(1/0{,}7)^{½} = 1{,}195$ beträgt der Faktor für den FAMZ bezogen auf die über 450 Merkmale des Standardprogramms durchschnittlich 1,09.

[11] Nicht zu verwechseln mit einem Prozentpunkt.

6 Schätzung eines Verhältniswerts

Es sei t_y ein Total bezüglich eines Merkmals y und t_z ein Total bezüglich eines Merkmals z. Die zu schätzende Größe ist $R = t_y/t_z$. R wird durch $\hat{R} = \hat{t}_y/\hat{t}_z$ geschätzt. Zur Berechnung der Varianz von $\hat{R}$ wird eine Taylorentwicklung der Funktion $f(t_y, t_z) = t_y/t_z$ benutzt:

$$
\begin{aligned}
f(\hat{t}_y, \hat{t}_z) &= f(t_y, t_z) + \frac{\partial f}{\partial t_y}\Big|_{t_y}(\hat{t}_y - t_y) + \frac{\partial f}{\partial t_z}\Big|_{t_z}(\hat{t}_z - t_z) \\
&= f(t_y, t_z) + \frac{1}{t_z}(\hat{t}_y - t_y) - \frac{t_y}{t_z^2}(\hat{t}_z - t_z) \\
&= \text{Konstante} + \frac{1}{t_z}(\hat{t}_z - R\hat{t}_z) \\
&= \text{Konstante} + \frac{1}{t_z}\sum_{k \in s}\frac{u_k}{\pi_k} \qquad (14)
\end{aligned}
$$

wobei $u_k = (y_k - R z_k)$.

Damit ist bis auf die Konstante $1/t_z$ lediglich die Varianz des π-Schätzers für das U-Total zu bestimmen. Wegen der linearen Approximation der Schätzfunktion wird allerdings nur eine Näherung der Varianz von $\hat{R}$ bestimmt, die für große Beobachtungsumfänge mit $V(\hat{R})$ übereinstimmt (Bezeichnung: Asymptotische Varianz $AV(\hat{R})$):

$$
AV(\hat{R}) = \frac{1}{t_z^2} V(\hat{t}_u) \qquad (15)
$$

Zur Berechnung von $V(\hat{t}_u)$ kann man wieder Gleichung (8)-(11) verwenden. Man hat dabei lediglich y_k durch u_k zu ersetzen. $S_{U_i}^2$ mißt die Varianz der Hilfsgrößen u_k innerhalb einer PSU. Die Hilfsgröße u_k kann als Abweichung der y_k/z_k Werte von dem Populationswert R interpretiert werden.

Eine asymptotische erwartungstreue Schätzung von $AV(\hat{R})$ erhält man, indem die unbekannten Populationswerte t_z und R durch die Schätzwerte $\hat{t}_z$ und $\hat{R}$ ersetzt und Gleichungen (12) und (13) verwendet werden. Hierbei ist $\hat{t}_i$ gegeben durch:

$$
\hat{t}_i = \sum_{k \in s_i}\frac{u_k}{\pi_{k|i}} = \frac{1}{0{,}7}\sum_{k \in s_i}(y_k - \hat{R} z_k) \qquad (16)
$$

$\hat{\bar{t}}_{s_{I,h}}$ ist der Mittelwert der unter (16) geschätzten PSU Werte über alle PSU's der Schicht h.

Die Schätzung der Varianz von $\hat{R}$ bedarf damit keiner gesonderten Programmierung. Man hat lediglich statt des Merkmals y_k das Merkmal $u_k = y_k - \hat{R}z_k$ zu verwenden und darf am Schluß die Division durch $\hat{t}_z^2$ nicht vergessen.

7 Die Varianz von Populationsmitteln

Die meisten Variablen im Mikrozensus sind zwar, wie z.B. Geschlecht oder Beruf, qualitative Merkmale. Um auch für die quantitativen Variablen, wie beispielsweise Alter oder Zahl der Kinder im Haushalt etc., die Varianz von Populationsmitteln bestimmen zu können, kann wie folgt vorgegangen werden.

Der Populationsmittelwert zu einem Merkmal y ist gegeben durch:

$$\bar{y}_U = \frac{1}{N}\sum_{k\in U} Y_k = \frac{t_y}{N}$$

Falls der Populationswert N bekannt ist, läßt sich $\bar{y}_U$ schätzen durch:

$$\hat{\bar{y}}_{U,\pi} = \frac{\hat{t}_y}{N}$$

Eine geringere Varianz hat jedoch häufig das gewichtete Stichprobenmittel $\hat{\bar{y}}_s$ (vgl. Särndal et al. 1992: 182):

$$\hat{\bar{y}}_s = \frac{\sum_{k\in s} Y_k/\pi_k}{\sum_{k\in s} 1/\pi_k}$$

Dieser Schätzer wird von den meisten Programmpaketen benutzt sobald eine Gewichtungsvariable verwendet wird. In dieser Form ist $\hat{\bar{y}}_s$ ein Spezialfall von $\hat{R}$ mit $z_k = 1$, so daß die Varianzschätzung für $\hat{R}$ benutzt werden kann. Man erhält mit $\hat{N} = \sum_{k\in s} 1/\pi_k$ für das Hilfsmerkmal u_k den folgenden Wert:

$$u_k = y_k - \frac{\hat{t}_y}{\hat{N}}\cdot 1 = y_k - \hat{\bar{y}}_s \qquad (17)$$

u_k mißt also die Abweichung der y-Werte vom gewichteten Stichprobenmittel. Der Faktor $1/\hat{t}_z^{\,2}$ ist durch $1/\hat{N}^2$ gegeben.

Die Tabelle 2 zeigt für die Merkmale von Tabelle 1 den geschätzten Anteil (in Prozent) der Merkmalsträger an der Bevölkerung am Hauptwohnsitz bzw. der Privathaushalte insgesamt. Die letzte Spalte in Tabelle 2 zeigt den Wert für den Standardfehler, wenn man die Insgesamt-Werte als bekannte Größe benutzt. Man erhält das auf den ersten Blick verblüffende Phänomen, daß diese Schätzung in den meisten Fällen ungenauer ist als die Schätzung mit variablem Nenner. Dieser Effekt ist umso größer, je häufiger das Merkmal in der Population vertreten ist, d.h. je stärker Zähler- und Nennermerkmal kovariieren. Beispielsweise wird der Anteil der Nichterwerbstätigen um fast 40 Prozent genauer geschätzt, wenn der Nenner ebenfalls geschätzt wird.

--- Tabelle 2 ---

8 Die Varianz von Populationsschätzern nach der Anpassung an die Bevölkerungsfortschreibung

Von den statistischen Ämtern werden die Mikrozensusergebnisse anhand von Eckzahlen der laufenden Bevölkerungsfortschreibung hochgerechnet. Zu diesem Zweck wird ein Hochrechnungsfaktor auf der Basis von regionalen Eckzahlen zu sechs Anpassungsklassen[12] gebildet, der im wesentlichen das Verhältnis der Sollzahlen aus der Bevölkerungsfortschreibung zu den Istzahlen der im Mikrozensus Befragten abbildet (vgl. Heidenreich 1994). Die Nutzer der faktisch anonymisierten Mikrozensusdaten können für diese sogenannte gebundene Hochrechnung auf die im Datensatz enthaltenen Hochrechnungsfaktoren für Personen und Haushalte zurückgreifen.

Die Bevölkerungsfortschreibung ist aber nicht ohne Kritik (Heidenreich 1994, Jäger 1992, Schimpl-Neimanns 1998). Insbesondere bei den Ausländern wird vermutet, daß der Wegzug von Ausländern nur unzureichend erfaßt wird. Die Bevölkerungsfortschreibung führt beim MZ zu einer Hochgewichtung der Ausländer um einen Faktor von ca. 1,5. Eine "Korrektur" in dieser Größe läßt sich nicht aus den Ergebnissen der Feldarbeit herleiten (siehe Heidenreich 1994: 116). Mit der Gewichtung werden die Mikrozensusergebnisse auf Ergebnisse der Bevölkerungsfortschreibung adjustiert, was zur Übertragung von Fehlern der Bevölkerungsfortschreibung auf den Mikrozensus führen kann. Da aber trotz dieser Probleme in der Praxis ein Bedarf an gebundenen Hochrechnungen besteht, wobei sich die FAMZ-Ergebnisse nicht we-

[12] Die Anpassung der MZ-Ergebnisse an die Bevölkerungsfortschreibung nach den Klassen Geschlecht in Kombination mit Staatsangehörigkeit (Deutsche/Ausländer), Soldaten und Wehrpflichtige erfolgt regional auf der Ebene von Anpassungsschichten - das sind regionale Einheiten mit wenigstens 500.000 Einwohnern -

sentlich von den Veröffentlichungen der amtlichen Statistik unterscheiden sollten, ist zu fragen, in welcher Weise die Anpassung an die Bevölkerungsfortschreibung ("Gewichtung") bei der Varianzschätzung berücksichtigt werden kann.

Formal kann die Verwendung von Gewichten, die aus der Anpassung der MZ-Fallzahlen an die Bevölkerungsfortschreibung resultieren, als Regressionsschätzer interpretiert werden. Der hier benutzte Regressionsschätzer basiert auf dem Group Mean Modell, das sich wie folgt darstellen läßt.

Die Grundgesamtheit U läßt sich in G disjunkte Teilmengen $U_g, g \in \{1 \ldots G\}$ zerlegen. Für die Elemente innerhalb jeder Gruppe g gilt:

(18) $$E_\xi(y_k) = \beta_g \quad k \in U_g$$

(19) $$V_\xi(y_k) = \sigma_g^2 \quad k \in U_g$$

Hierbei bezeichnet $E_\xi(\cdot)$ die Erwartungswertbildung bezüglich einer Verteilung ξ über den Merkmalen und $V_\xi(\cdot)$ die entsprechende Varianz. Die Erwartungswertbildung bezieht sich hier also nicht auf den Wert eines Populationsschätzers bezüglich des Erhebungsdesigns, sondern auf die Realisierung der Merkmalswerte bei einer gegebenen Stichprobe.

Die OLS-Schätzung von β_g auf Basis der Daten für die Grundgesamtheit ist:

(20) $$B_g = \frac{1}{N_g} \sum_{k \in U_g} y_k = \bar{y}_{U_g}$$

wobei: N_g = Anzahl Elemente von U_g. Die Gruppendefinition liefert eine Zerlegung der Stichprobe: $s_g = s \cap U_g$. Als Schätzer für B_g benutzt man:

(21) $$\hat{B}_g = \frac{\sum_{k \in s_g} y_k / \pi_k}{\sum_{k \in s_g} 1/\pi_k} = \frac{1}{\hat{N}_g} \sum_{k \in s_g} \frac{y_k}{\pi_k}$$

Hierbei ist $\hat{N}_g$ der geschätzte Wert für den Umfang der Gruppe g. Im Falle des FAMZ ist die Ziehungswahrscheinlichkeiten aller Einheiten gleich. Folglich gilt:

(22) $$\hat{B}_g = \bar{y}_{s_g} \qquad g = (1, \ldots, G)$$

---

bzw. für Soldaten und Wehrpflichtige auf Regierungsbezirksebene.

Es bezeichne $\hat{y}_k$ den durch das Regressionsmodell geschätzten Wert von y_k. Im vorliegenden Fall erhält man:

(23) $\hat{y}_k = \hat{B}_g = \bar{y}_{s_g} \qquad k \in s_g$

Der Regressionsschätzer $\hat{t}_{reg}$ hat für das Group Mean Modell die folgende Gestalt:

(24) $$\hat{t}_{reg} = \sum_{k \in U} \hat{y}_k = \sum_{g=1}^{G} \sum_{k \in U_g} \hat{B}_g = \sum_{g=1}^{G} N_g \hat{B}_g = \sum_{g=1}^{G} \sum_{k \in s_g} \frac{N_g}{\hat{N}_g} \cdot \frac{y_k}{\pi_k} = \sum_{g=1}^{G} \sum_{k \in s_g} w_g \frac{y_k}{\pi_k}$$

wobei $w_g = N / \hat{N}$. Dieser Faktor beschreibt das Verhältnis von N_g = Umfang von Gruppe g in der Grundgesamtheit (= Soll - Vorgabe) zu $\hat{N}_g$ = geschätzter Umfang von Gruppe g (= Ist - Wert).

Bei der Anpassung an die Bevölkerungsfortschreibung wird für N_g jeweils der Wert für bestimmte Gruppen gemäß dieser Fortschreibung gewählt. Im FAMZ ist eine Gewichtungsvariable w_k enthalten, die für jede Person $k \in s_g$ den jeweiligen Wert $N_g / \hat{N}_g$ ("Soll durch Ist") annimmt.[13] Der Regressionsschätzer läßt sich damit als ein "gewichtetes Mittel" darstellen:

(25) $$\hat{t}_{reg} = \sum_{k \in s} w_k \frac{y_k}{\pi_k}$$

Man beachte, daß die w_k von der jeweiligen Stichprobe abhängen.

Eine wesentliche Eigenschaft des Regressionsschätzers liegt darin, daß er für Merkmale, die in die Soll/Ist-Anpassung eingingen, unter jeder Stichprobe wieder die Soll-Werte liefert (vgl. Särndal et al. 1992: 324). Folglich hat $\hat{t}_{reg}$ für diese Merkmale die Varianz Null. Allerdings ist der FAMZ nur eine 70%-Substichprobe aus dem MZ und wurde nicht noch einmal an die Bevölkerungsfortschreibung angepaßt. Daher wird die Varianzschätzung für den Regressionsschätzer auch im Falle der Anpassungsmerkmale positive Werte liefern.

Für die Herleitung von $V(\hat{t}_{reg})$ wird wieder eine Taylorentwicklung von $\hat{t}$ benutzt. Man erhält eine asymptotische Näherung für $V(\hat{t}_{reg})$. Der lineare Teil der Taylorentwicklung ist durch die folgende Hilfsgröße u_k gegeben (vgl. Särndal et al. 1992: 331):

[13] Darüber hinaus liegen Gewichte für die Hochrechnung der Unterstichprobe sowie für Haushalte bzw. Familien vor. Das Haushaltsgewicht ist das arithmetische Mittel der Personenfaktoren im Haushalt.

$$(26)\quad u_k = \frac{N_g}{\hat{N}_g}(y_k - \hat{B}_g) = w_k(y_k - \bar{y}_{s_g}) \qquad k \in s_g$$

Dieser lineare Anteil ist also die mit w_k gewichtete Abweichung des Merkmalswerts y_k von dem jeweiligen Gruppenmittelwert $\bar{y}_{s_k}$ in der Stichprobe. Als asymptotische Varianz wird die Varianz dieses Hilfsmerkmals u verwendet. Bei der praktischen Berechnung hat man also lediglich y_k durch u_k in Gleichung (8) - (13) zu ersetzten.[14]

In Tabelle 3 wird für die Merkmale von Tabelle 1 der Einfluß der Anpassung an die Bevölkerungsfortschreibung auf die Schätzung des Totals und die dazugehörige Varianz dargestellt. Es zeigt sich, daß der relative Standardfehler durch die Anpassung etwas verringert wird. Für Merkmale, die eng mit den Anpassungsklassen zusammenhängen (z.B. „Ausländische Erwerbsperson“, „1-Personenhaushalte, weiblich“) fällt die Reduktion stärker aus. Jedoch besteht ein bemerkenswerter Unterschied zwischen den Schätzwerten in der Größenordnung von hochgerechnet teilweise mehreren Millionen. Dies zeigt die wesentliche Problematik bei der Verwendung der Hochrechnungsergebnisse: Sie führen praktisch kaum zu einer Verringerung der Varianz, sondern verdecken einen Bias. Entweder liefert der MZ und damit auch der FAMZ verzerrte Populationsschätzer oder aber die Bevölkerungsfortschreibung produziert ihrerseits verfälschte Schätzungen.

--- Tabelle 3 ---

9 Design-Effekte

Bisher waren die Nutzer des FAMZ bei der Schätzung der Varianz von Totals auf Verwendung von Ergebnissen der Fehlerrechnung des Statistischen Bundesamts angewiesen. Die Schätzung der Varianz dieser Tabelleneinträge geschah mit Hilfe von Zuschlagsfaktoren bzw. von Design-Effekten. Ein Design-Effekt beschreibt das Verhältnis des Standardfehlers des MZ im Verhältnis zum Standardfehler einer Stichprobeziehung gleichen Umfangs, die jedoch ohne Klumpung und Schichtung durchgeführt wird; also einer einfachen zufälligen Stichprobe. Kern dieses Ansatzes war eine lineare Regression dieses Design-Effekts auf den geschätzten Populationsanteil der durch das Tabellenfeld definierten Merkmalsträger.

Da wegen der Substichprobenziehung die Zahl der Haushalte und Personen pro Auswahlbezirk im FAMZ geringer ist als im MZ, ist davon auszugehen, daß die für den MZ veröffent-

[14] Für Haushaltsauswertungen können die Anpassungsklassen nach den Eigenschaften der Haushaltsbezugsperson gebildet werden.

lichten Zuschlagsfaktoren nicht einfach auf den FAMZ übertragbar sind und bei ihrer Verwendung zu einer Überschätzung der Varianz im FAMZ führen. Wir wollen daher prüfen, ob die Ziehung des FAMZ zu anderen Design-Effekten führt, und wie gut die lineare Approximation der Design-Effekte ist.

Es bezeichne U_d eine Teilmenge der Grundgesamtheit (= Domain), die das Feld einer Tabelle charakterisiert. N_d sei die Anzahl von U_d in der Grundgesamtheit und n_d in der Stichprobe. Mit Hilfe der Indikatorfunktion

$$y_k = \begin{cases} 1 \text{ falls } k \in U_d \\ 0 \text{ sonst} \end{cases}$$

kann man die Schätzung von Domains auf die Schätzung von Totals zurückführen, da in diesem Falle $N_d = \sum_{k \in U} y_k = t_y$ gilt. Für die Schätzung des Populationsanteils $P_d = N_d / N$ durch $\hat{P}_d = \hat{t}_y / N$ erhält man im Fall einer einfachen Stichprobe (vgl. Särndal et al. 1992: 70):

$$(27) \quad \hat{V}_{SI}(\hat{P}_d) = \frac{1-f}{n-1} p_d (1-p_d)$$

wobei $p_d = n_d / n$ das Stichprobenpendant zu P_d ist.

Der relative Standardfehler ist der Variationskoeffizient von $\hat{P}_d$. Für die einfache Stichprobe ist er gegeben durch:

$$(28) \quad cv_{SI} = \frac{\sqrt{\hat{V}_{SI}(\hat{P}_d)}}{\hat{P}_d} = \sqrt{\frac{1-f}{n-1} \frac{1-p_d}{p_d}}$$

Der Design-Effekt $k(\hat{p}_d)$ ist dann gegeben durch:[15]

$$(29) \quad k(\hat{p}_d) = \frac{\sqrt{\hat{V}(\hat{p}_d)}}{\sqrt{\hat{V}_{SI}(\hat{p}_d)}}$$

Dieses Verhältnis kann für jede Subpopulation anders ausfallen. Das Statistische Bundesamt benutzt für 3 Merkmalsgruppen (Bevölkerung und Erwerbstätige (B/E), Ausländer und Erwerbstätige in der Landwirtschaft (A/L) und Haushalte (H)) eine unterschiedliche Einfach-Regression von $k(\hat{p}_d)$ auf $\hat{p}_d$:

[15] Häufig wird der Design-Effekt als das Verhältnis der Varianzen definiert, so z.B. bei Särndal et al. (1992: 54).

(30) $k(\hat{p}_d) \approx a + b\hat{p}_d$

Für gegebene Werte von a und b erhält man unter Verwendung von (29) und (30) die folgende Näherung für den relativen Standardfehler in Abhängigkeit von p_d:

(31) $cv = k(p_d)cv_{SI} = (a + bp_d)\sqrt{\frac{1-f}{n-1}\,\frac{1-p_d}{p_d}}$

Da das Ziehungsverfahren des FAMZ durch die Sortierfolge der Haushalte den Klumpeneffekt bei weitgehender Respektierung der Schichtung reduziert, ist auch eine Verringerung des Design-Effekts zu erwarten. Tabelle 4 vergleicht zusammenfassend die auf Basis des FAMZ jeweils für die einzelnen Gruppen ermittelten Regressionskoeffizienten mit den für den MZ 1990 veröffentlichten Werten und den Berechnungen für den MZ 1996. Insgesamt verlaufen die Geraden für den FAMZ 1996 flacher als für den MZ 1996. Bezogen auf die über 450 Merkmale des Standardauswertungsprogramms beträgt die Reduktion etwa 10 Prozent und ist für die drei Merkmalsgruppen B/E, A/L und H in etwa gleich. Die Spalte "Näherungsfunktion" in Tabelle 1 basiert auf der Anwendung von Gleichung (31). Läßt man einmal außer acht, daß die in der Näherungsfunktion verwendeten Koeffizienten für den MZ 1990 veraltet sind und sich nur auf das frühere Bundesgebiet beziehen, kann festgestellt werden, daß der relative Standardfehler für die hier ausgewählten Merkmale in brauchbarer Näherung wiedergegeben wird.

--- Tabelle 4 ---

Es bleibt noch zu überprüfen, wie gut die lineare Beziehung in Gleichung (30) gilt.[16] Die Abbildungen 1 bis 3 zeigen die Regressionen der Design-Faktoren für die Merkmale des MZ-Standardauswertungsprogramms für die Gruppe B/E, A/L und H. Insgesamt zeigt sich, daß die lineare Approximation zwar ein brauchbares Modell zur Beschreibung des Design-Effekts liefert, aber bei einzelnen Merkmalen doch beträchtliche Abweichungen des jeweiligen Design-Effekts von der Regressionsgeraden vorliegen. In Einzelfällen führt also die Verwendung der Design-Effekt Faktoren zu erheblichen Über- bzw. Unterschätzungen der Varianz, wobei über die Richtung des Fehlers keine Aussage gemacht werden kann. Mit dem Vorliegen der

[16] Für den MZ sind bisher nur für 1990 die entsprechenden Regressionskoeffizienten veröffentlicht worden. Eine Überprüfung der Angemessenheit des einfachen Regressionsmodells ist bis auf den Hinweis, daß die Abweichungen der berechneten von den geschätzten Design-Effekten im Mittel kleiner als 15-20% betragen (Statistisches Bundesamt 1998a: 17), bisher nicht dokumentiert worden.

Information über den Auswahlbezirk und die Gebäudeschicht ist es jedoch nicht mehr nötig, sich auf diese Varianzabschätzung zu verlassen.

--- Abbildungen 1-3 ---

10 Abschließende Bemerkungen

Im vorliegenden Beitrag wurde gezeigt, daß man relativ einfache Varianzschätzungen für den faktisch anonymisierten Mikrozensus erhält, wenn nur die Zugehörigkeit zum Auswahlbezirk sowie einige grobe Regionalmerkmale bekannt sind. Dies eröffnet es den Nutzern die wesentliche Qualität dieses Datensatzes, nämlich die erstaunlich hohe Präzision der Schätzergebnisse bei zugleich sehr geringem Nonresponse, effizient auszuschöpfen.

Die Analyse hat gleichzeitig erbracht, daß es wünschenswert ist, der Forschung weitere feldbedingte Merkmale im Datensatz zur Verfügung zu stellen. Dies betrifft insbesondere die Zugehörigkeit zu den Rotationsgruppen. Bei Kenntnis der Rotationsgruppen kann man die Varianz der Schätzungen durch die Varianz der Schätzergebnisse über die Rotationsgruppen schätzen. Eine Identifikation der Zugehörigkeit zur Rotationsgruppe würde außerdem auch für den Vergleich von Kennwerten auf Basis einzelner Querschnittsangaben (z.B. Veränderungen 1996-1997) verbesserte Varianzschätzung ermöglichen und deshalb von großem Nutzen sein. In diesem Zusammenhang sei angemerkt, daß die ab dem Mikrozensus 1996 mögliche Zusammenführung der einzelnen Mikrozensuserhebungen zu einem rotierenden Panel für die Forschung von allergrößtem Interesse ist.

Weitere Verbesserungsmöglichkeiten des Datensatzes beziehen sich auf die Bereitstellung von Informationen über die Ziehung der 70%-Substichprobe aus dem Mikrozensus. Bei Kenntnis der Haushaltsendziffern, die das Ziehungsverfahren bestimmen, ist es möglich, die Varianzkomponente zu schätzen, die durch diese Stufe der Schätzung bedingt ist. Da die Merkmale Rotationsgruppe und Haushaltsendziffer hinsichtlich des Datenschutzes kein grundsätzliches Problem darstellen dürften, würde die Weitergabe der Variablen neue methodische und inhaltliche Analysen des faktisch anonymisierten Mikrozensus bieten.

Literatur

Emmerling, Dieter, und Thomas Riede, 1997: 40 Jahre Mikrozensus. Wirtschaft und Statistik (3): 160-174.

Frank, Eberhard, und Andrea Kafurke, 1990: Die Mikrozensusstichprobe ab 1990 auf neuer Auswahlgrundlage. Baden-Württemberg in Wort und Zahl (4): 154-164.

Heidenreich, Hans-Joachim, 1994: Hochrechnung des Mikrozensus ab 1990. S. 112-123 in: *Gabler, Siegfried, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs* (Hg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Jäger, Marianne, 1992: Im Westen was Neues? - Im Osten was Besseres? Möglichkeiten der Nutzung von Daten der Einwohnermelderegister für statistische Zwecke. S. 103-124 in: *Statistisches Bundesamt* (Hg.): Volkszählung 2000 - oder was sonst? Band 21 der Schriftenreihe Forum der Bundesstatistik. Stuttgart: Metzler-Poeschel.

Köhler, Sabine, Bernhard Schimpl-Neimanns und Norbert Schwarz, 2000: Pilotprojekt zur Erleichterung der Nutzungsmöglichkeiten von faktisch anonymisierten Mikrodaten, Wirtschaft und Statistik (1): 30-37.

Kohler, Hans, Helmut Rudolph und Eugen Spitznagel, 1996: Umfang, Struktur und Entwicklung der geringfügigen Beschäftigung - Eine Bestandsaufnahme. IAB-Kurzbericht 2/31.1.1996. Nürnberg: IAB.

Krug, Walter, Martin Nourney und Jürgen Schmidt, 1994: Wirtschafts- und Sozialstatistik. Gewinnung von Daten. München: Oldenbourg (3., völlig neubearb. Auflage).

Lüttinger, Paul (Hg.), 1999: Sozialstrukturanalysen mit dem Mikrozensus. ZUMA Nachrichten Spezial, Band 6. Mannheim: ZUMA.

Lüttinger, Paul, und Thomas Riede, 1997: Der Mikrozensus: amtliche Daten für die Sozialforschung, ZUMA-Nachrichten Nr. 41: 19-43.

Meyer, Kurt, 1994: Zum Auswahlplan des Mikrozensus ab 1990. S. 106-111 in: *Gabler, Siegfried, Jürgen H.P. Hoffmeyer-Zlotnik und Dagmar Krebs* (Hg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Müller, Walter, Uwe Blien, Peter Knoche und Heike Wirth u.a., 1991: Die faktische Anonymität von Mikrodaten. Band 19 der Schriftenreihe Forum der Bundesstatistik, herausgegeben vom *Statistischen Bundesamt*. Stuttgart: Metzler-Poeschel.

Reinders, Marlis, 1993: Fehlerrechnung zum Mikrozensus 1990. Statistische Rundschau Nordrhein-Westfalen (8): 398-404.

Särndal, Carl-Erik, Bengt Swensson und Jan Wretman, 1992: Model Assisted Survey Sampling. New York: Springer.

Schimpl-Neimanns, Bernhard, 1998: Analysemöglichkeiten des Mikrozensus. ZUMA-Nachrichten Nr. 42, 22: 91-119.

Schmidt, Gerhard, 1990: Der Auswahlplan des Mikrozensus ab 1990. Bayern in Zahlen (6): 217-221.

Schupp, Jürgen, Joachim Frick, Lutz Kaiser und Gert Wagner (1999): Zur Erhebungsproblematik geringfügiger Beschäftigung. Ein Strukturvergleich des Mikrozensus mit dem Sozio-oekonomischen Panel und dem Europäischen Haushaltspanel. S. 93-118 in: *Lüttinger, Paul* (Hg.): Sozialstrukturanalysen mit dem Mikrozensus. ZUMA Nachrichten Spezial, Band 6. Mannheim: ZUMA.

Statistisches Bundesamt, 1998a: Fachserie 1, Bevölkerung und Erwerbstätigkeit. Reihe 4.1.1, Stand und Entwicklung der Erwerbstätigkeit 1996 (Ergebnisse des Mikrozensus). Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt, *1998b*: Fehlerrechnung Mikrozensus 1996 (nach Kompensation der bekannten Ausfälle). Wiesbaden (unveröffentlichte Tabellen; StBA VIII C; September 1998).

Statistisches Bundesamt (Hg.), 1999: Kooperation zwischen Wissenschaft und amtlicher Statistik - Praxis und Perspektiven. Band 34 der Schriftenreihe Forum der Bundesstatistik. Stuttgart: Metzler-Poeschel.

Werner, Joachim, 1994: Regionalisierung des Mikrozensus. Baden-Württemberg in Wort und Zahl (6): 278-285.

Wolter, Kirk 1985: Introduction to Variance Estimation. New York: Springer.

Tabelle 1: Schätzung von Fallzahlen (Totals) für ausgewählte Merkmale

	Fallzahlen (in 1000)		Standardfehler (in 1000)				relativer Std.fehler (in %)			Design-Effekt	
				FAMZ					Näher.-		
Merkmal	**FAMZ**	**MZ**	**Insges.**	Between	Within	**MZ**	**FAMZ**	**MZ**	**funkt.**	**FAMZ**	**MZ**
Bevölkerung 65 Jahre und älter [1)]	11.965,9	12.264,2	52,6	52,6	2,3	50,1	0,44	0,41	0,44	1,40	1,58
Ausländische Erwerbspersonen [1)]	2.283,7	2.330,6	29,4	29,4	1,2	27,3	1,29	1,17	0,92	1,66	1,83
Sofort verfügbare Erwerbslose [1)]	2.976,7	3.034,2	24,1	24,0	1,2	21,4	0,81	0,71	0,81	1,20	1,26
Überwiegender Lebensunterhalt durch Erwerbstätigkeit [1)]	29.607,9	30.285,0	81,4	81,4	3,2	77,7	0,28	0,26	0,30	1,63	1,85
Nichterwerbstätige [1)]	40.223,4	41.130,9	102,6	102,5	3,8	99,6	0,26	0,24	0,25	2,03	2,36
Weibliche Erwerbstätige mit einem monatlichen Nettoeinkommen unter 600 DM [1)]	1.545,4	1.585,4	15,9	15,8	0,8	13,9	1,03	0,88	1,11	1,08	1,12
Erwerbstätige mit Haupt- (Volks-) schulabschluß [1)]	12.643,9	12.920,9	56,4	56,3	2,4	52,3	0,45	0,41	0,43	1,47	1,61
Erwerbstätige Frauen mit weniger als 36 Stunden normalerw. geleist. Arbeitszeit/Woche, mit ledigen Kindern unter 18 Jahren i.d. Familie [2)]	2.599,4	2.663,9	19,6	19,6	1,0	17,4	0,75	0,65	0,86	1,04	1,09
Privathaushalte mit 1 Person, weiblich [3)]	7.010,1	7.259,6	33,1	33,1	1,5	32,0	0,47	0,44	0,54	1,19	1,35
Haushaltsnettoeinkommen unter 1000 DM [3)]	1.506,1	1.568,2	18,3	18,2	0,7	17,4	1,21	1,11	1,11	1,28	1,43
Nichtehliche Lebensgemeinschaft von verschieden geschlechtlichen Partnern mit ledigen Kindern [3)]	455,6	457,4	8,1	8,1	0,4	6,9	1,78	1,51	1,99	1,02	1,03

FAMZ: faktisch anonymisierte Einzeldaten des Mikrozensus 1996 (ZUMA-File); **MZ**: Frei hochgerechnete Werte nach Kompensation der bekannten Ausfälle (Fehlerrechnungen zum Mikrozensus 1996 (Statistisches Bundesamt 1998b); **Näher.funkt**: Aus den Besetzungszahlen der Tabellenfelder geschätzter relativer Standardfehler unter Verwendung der veröffentlichten Zuschlagsfaktoren zum Mikrozensus 1990 (Statistisches Bundesamt 1998a: 17); Näherungsfunktion; s. Abschnitt 9.
Subpopulationen: **1)** Bevölkerung am Hauptwohnsitz; **2)** Bevölkerung am Familienwohnsitz; **3)** Privathaushalte

Tabelle 2: Schätzung von Anteilswerten für ausgewählte Merkmale

Merkmal Y	Merkmal Z	Anteil R (%)	Standardfehler in % Insges.	Between	Within	N bek.
Bevölkerung 65 Jahre und älter	Bevölkerung	16,7	0,077	0,077	0,004	0,073
Ausländische Erwerbspersonen	Bevölkerung	3,2	0,040	0,040	0,002	0,041
Sofort verfügbare Erwerbslose	Bevolkerung	4,1	0,033	0,033	0,002	0,034
Überwiegender Lebensunterhalt durch Erwerbstätigkeit	Bevölkerung	41,3	0,081	0,081	0,004	0,114
Nichterwerbstätige	Bevölkerung	56,1	0,082	0,082	0,004	0,143
Weibliche Erwerbstätige mit einem monatlichen Nettoeinkommen unter 600 DM	Bevölkerung	2,2	0,021	0,021	0,001	0,022
Erwerbstätige mit Haupt- (Volks-) schulabschluß	Bevölkerung	17,6	0,069	0,069	0,003	0,079
Privathaushalte mit 1 Person, darunter weiblich	Privathaushalte	21,4	0,093	0,093	0,005	0,101
Haushaltsnettoeinkommen unter 1000 DM	Privathaushalte	4,6	0,055	0,055	0,002	0,056
Nichtehliche Lebensgemeinschaft von verschieden geschlechtlichen Partnern mit ledigen Kindern	Privathaushalte	1,4	0,025	0,025	0,001	0,025

Quelle: faktisch anonymisierte Einzeldaten des Mikrozensus 1996 (ZUMA-File);
N. bek.: Standardfehler bei Verwendung der Fallzahl im Nenner (Z) als bekannte Größe
Bevölkerung = Bevölkerung am Hauptwohnsitz

Tabelle 3: Schätzung von Bevölkerungstotals und relativer Standardfehler für ausgewählte Merkmale mit und ohne Anpassung an die Bevölkerungsfortschreibung

Merkmal Y	Fallzahl (in 1000) mit Anp.	Fallzahl (in 1000) ohne Anp.	relativer Std.fehler mit Anp.	relativer Std.fehler ohne Anp.
Bevölkerung 65 Jahre und älter [1)]	13.388,7	11.965,9	0,45	0,44
Ausländische Erwerbspersonen [1)]	3.609,7	2.283,7	0,72	1,29
Sofort verfügbare Erwerbslose [1)]	3.490,3	2.976,7	0,79	0,81
Überwiegender Lebensunterhalt durch Erwerbstätigkeit [1)]	33.806,1	29.607,9	0,20	0,28
Nichterwerbstätige [1)]	45.919,9	40.223,4	0,15	0,26
Weibliche Erwerbstätige mit einem monatlichen Nettoeinkommen unter 600 DM [1)]	1.740,7	1.545,4	1,00	1,03
Erwerbstätige mit Haupt- (Volks-) schulabschluß [1)]	14.492,7	12.643,9	0,39	0,45
Erwerbstätige Frauen mit weniger als 36 Stunden normalerweise geleisteter Arbeitszeit je Woche, mit ledigen Kindern unter 18 Jahren in der Familie [2)]	2.928,2	2.599,4	0,71	0,75
Privathaushalte mit 1 Person, darunter weiblich [2)]	7.896,6	7.010,1	0,27	0,47
Haushaltsnettoeinkommen unter 1000 DM [3)]	1.760,6	1.506,1	1,20	1,21
Nichtehliche Lebensgemeinschaft von verschieden geschlechtlichen Partnern mit ledigen Kindern [3)]	512,3	455,6	1,78	1,78

Quelle: faktisch anonymisierte Einzeldaten des Mikrozensus 1996 (ZUMA-File);
Subpopulation - verwendeter Hochrechnungsfaktor:

1) Bevölkerung am Hauptwohnsitz - Personen-Hochrechnungsfaktor
2) Bevölkerung am Familienwohnsitz - Haushalts-/Familienhochrechnungsfaktor
3) Privathaushalte - Haushalts-/Familienhochrechnungsfaktor

Tabelle 4: Vergleich der Koeffizienten einer einfachen Regression des Design-Effekts auf den geschätzten Bevölkerungsanteil

Gruppe	Datenbasis	Konstante	Steigung
Bevölkerung und Erwerbstätige (B/E)	FAMZ	1.009	1.84
	MZ96	1.042	2.44
	MZ90	1.136	1.61
Ausländer und Erwerbstätige in der Landwirtschaft (A/L)	FAMZ	1.088	21.69
	MZ96	1.162	25.47
	MZ90	1.169	25.04
Haushalte (H)	FAMZ	0.988	1.01
	MZ96	1.009	1.60
	MZ90	1.119	1.14

Datenbasis
FAMZ: faktisch anonymisierte Mikrozensusdaten 1996 (ZUMA-File)
MZ96: Eigene Berechnungen auf Basis unveröffentlichter Fehlerrechnungen zum Mikrozensus 1996 (Statistisches Bundesamt 1998b)
MZ90: Fehlerrechnung zum Mikrozensus 1990 (Statistisches Bundesamt 1998a: 22)

Abbildung 1: Regression des Design-Effekts (k) für die Merkmalsgruppe Bevölkerung und Erwerbstätige (B/E) auf den Anteilswert für 346 Merkmale des FAMZ 96

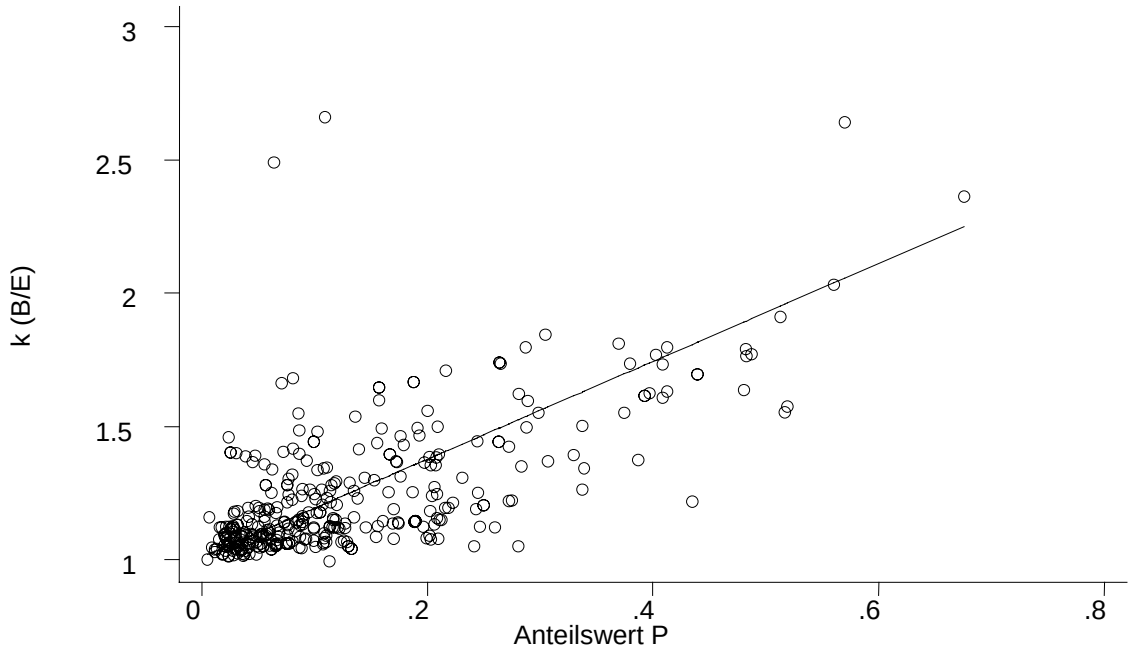

Abbildung 2: Regression des Design-Effekts (k) für die Merkmalsgruppe Ausländer und Erwerbstätige in der Landwirtschaft (A/L) auf den Anteilswert für 18 Merkmale des FAMZ 96

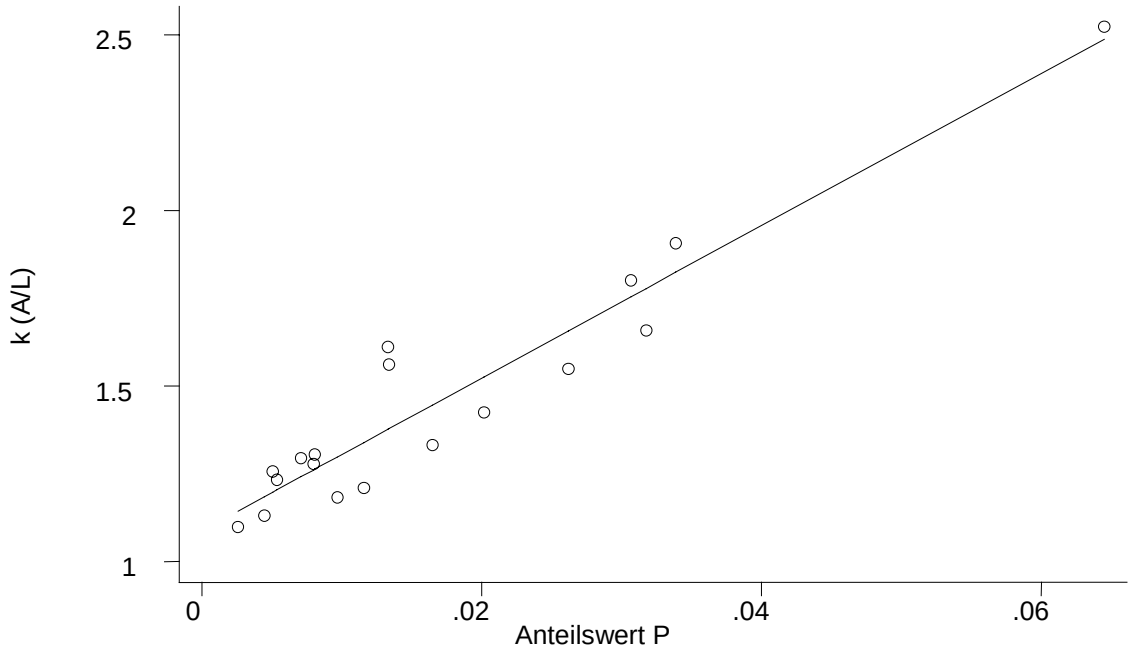

Abbildung 3: Regression des Design-Effekts (k) für die Merkmalsgruppe Haushalte (H) auf den Anteilswert für 94 Merkmale des FAMZ 96