Thomas Riede, Sabine Bechtold und Notburga Ott (Hrsg.)

WEITERENTWICKLUNG DER AMTLICHEN HAUSHALTSSTATISTIKEN

SCIVERO

Weiterentwicklung der amtlichen Haushaltsstatistiken

SCIVERO

Weiterentwicklung der amtlichen Haushaltsstatistiken

Thomas Riede, Sabine Bechtold und Notburga Ott (Hrsg.)

SCIVERO

ISBN 978-3-944417-02-8
1. Auflage 2013

© 2013 SCIVERO Verlag, Berlin

SCIVERO ist eine Marke der GWI Verwaltungsgesellschaft für Wissenschaftspolitik und Infrastrukturentwicklung Berlin UG (haftungsbeschränkt).

Dieses Buch dokumentiert die Ergebnisse eines vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) und dem Statistischen Bundesamt veranstalteten Workshops zu Methoden der empirischen Sozialforschung in amtlichen Haushaltsstatistiken im Mai 2013. Dank und Anerkennung gelten ausdrücklich den beteiligten Personen und Institutionen, insbesondere dem Statistischen Bundesamt und dem Bundesministerium für Bildung und Forschung (BMBF), welches die Aktivitäten des RatSWD finanziert und unterstützt.

Lektorat: Thomas Runge
Covergestaltung und Satz: Sören Schumann

Inhaltsverzeichnis

Vorwort 9

Thomas Riede
Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken 13

Dr. Tobias Schmidt, Martin Eisele
Oversampling vermögender Haushalte im Rahmen der Studie „Private Haushalte und ihre Finanzen (PHF)“ 31

Dr. Bernhard Schimpl-Neimanns
Methodische Herausforderungen bei der Erfassung von Bildung und Ausbildung im Mikrozensus 43

Dr. Florian Meinfelder
Datenfusion: Theoretische Implikationen und praktische Umsetzung 83

Prof. Dr. Ralf Münnich, Jan Pablo Burgard, Thomas Zimmermann
Wie genau sind Kreisergebnisse des Mikrozensus - Einsatzmöglichkeiten von Small-Area-Verfahren 101

Vorwort

Auf die amtlichen Haushaltsstatistiken, also die Erhebungen der Statistischen Ämter des Bundes und der Länder bei den privaten Haushalten, kommen aktuell sowohl von europäischer als auch von nationaler Seite neue Anforderungen zu. Um auf diese aktuellen, aber auch auf zukünftige Entwicklungen flexibel und bedarfsgerecht reagieren zu können, hat das Statistische Bundesamt 2012 ein Projekt zur Weiterentwicklung des Systems der Haushaltsstatistiken initiiert. Ziel des gemeinsam mit den Statistischen Landesämtern begonnenen Projektes ist die Etablierung eines kohärenten und zukunftsfähigen Systems der amtlichen Haushaltsbefragungen, welches gleichermaßen den sich ändernden Anforderungen an die Datenproduktion und dem sich verändernden Datenbedarf gerecht wird sowie die Datenqualität sicherstellt.

Im Rahmen des Projektes veranstaltete das Statistische Bundesamt gemeinsam mit dem Rat für Sozial- und Wirtschaftsdaten im Mai 2013 den Workshop „Methoden der empirischen Sozialforschung in amtlichen Haushaltsstatistiken". Mit diesem Workshop war vor allem die Wissenschaft eingeladen, sich mit ihrer Expertise an der Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken zu beteiligen. Nach einer Begrüßung durch den Vizepräsidenten des Statistischen Bundesamts, Dieter Sarreither, griffen die Referenten neben praktischen Fragestellungen bei der Erhebung von Haushaltsdaten auch theoretische Grundlagen auf und lieferten auf diesem Weg wertvolle Beiträge für das Projektvorhaben. Der vorliegende Band sammelt die Vorträge des Workshops und möchte auf diesem Wege eine breite Rezeption und Diskussion der Ergebnisse ermöglichen.

In einem einführenden Beitrag stellt der Gesamtprojektleiter Thomas Riede (Statistisches Bundesamt) das Projekt „Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken" vor. Im Fokus seines Beitrags steht das Grundmodell eines neuen Gesamtsystems, in das die bisher weitgehend unabhängig voneinander geführten Haushaltsstatistiken als Module integriert sind und gemeinsam erhoben werden.

Tobias Schmidt (Deutsche Bundesbank) geht in seinem Beitrag auf den Aspekt des Oversamplings vermögender Haushalte im Rahmen der Studie „Private Haushalte und ihre Finanzen (PHF)" des Forschungszentrums der Deutschen Bundesbank ein. Da Vermögen erheblich stärker konzentriert ist als Einkommen, würde eine proportional angelegte Stichprobe nur wenig Vermögende enthalten. Ein gezieltes Oversampling wohlhabender Haushalte löst dieses Defizit auf

und ermöglicht eine bessere Analyse der Vermögenszusammensetzung und -verteilung.

Bernhard Schimpl-Neimanns (GESIS) vergleicht die Erfassung von Bildung und Ausbildung im Mikrozensus mit der Bildungsstatistik. Insbesondere die Daten zu besuchten Schulen sowie der allgemeinbildenden und beruflichen Abschlüsse unterstreichen dabei, dass noch erhebliche Anstrengungen zur Verbesserung der Datenqualität notwendig sind.

Florian Meinfelder (Otto-Friedrich-Universität Bamberg) erörtert die theoretischen Implikationen und die praktische Umsetzung von Datenfusionen. Diese beschreiben ein spezifisches Datenausfallmuster, das entsteht, wenn (mindestens) zwei unabhängig entstandene Datenquellen ‚übereinandergestapelt' werden, sodass fehlende Variablen einer Datenquelle durch die jeweils andere ergänzt werden können.

Der Beitrag von Ralf Münnich (Universität Trier), den dieser gemeinsam mit Jan Pablo Burgard und Thomas Zimmermann erarbeitet hat, zeigt Möglichkeiten der Verwendung von Small Area-Methoden in deutschen Haushaltsstichproben auf und diskutiert sie analysierend. Dabei geht er insbesondere auf die derzeitige Neustrukturierung der Haushaltsbefragungen ein. Zusätzlich wird die Problematik der Stichprobenziehung behandelt, welche sowohl den Konflikt zwischen Modellbau und Optimierung von Stichproben als auch die Koordination verschiedener Teilstichproben umfasst.

Allen Beteiligen des Workshops, insbesondere den Referenten, sei an dieser Stelle noch einmal ausdrücklich für ihre wertvollen Beiträge gedankt. Ihre wissenschaftliche Expertise wird bei der Weiterentwicklung der Haushaltsstatistiken von großem Nutzen sein.

Sabine Bechtold, Notburga Ott
Wiesbaden, Bochum - November 2013

Die Weiterentwicklung des Systems der amtlichen Haushaltsstatistiken

Thomas Riede

Zusammenfassung

Auf die amtlichen Haushaltsstatistiken – die Erhebungen der Statistischen Ämter des Bundes und der Länder bei den privaten Haushalten – kommen sowohl von europäischer als auch von nationaler Seite kontinuierlich neue Anforderungen zu. Um weiterhin flexibel auf neue Anforderungen reagieren zu können und so eine bedarfsgerechte Datenproduktion auch künftig zu ermöglichen, ist eine Weiterentwicklung der haushaltsstatistischen Erhebungen notwendig.

Hierzu hat das Statistische Bundesamt im vergangenen Jahr das Projekt zur Weiterentwicklung des Systems der Haushaltsstatistiken initiiert und gemeinsam mit den Statistischen Landesämtern begonnen. Ziel des Projektes ist die Etablierung eines kohärenten und zukunftsfähigen Systems der Haushaltsbefragungen der amtlichen Statistik, welches gleichermaßen den sich ändernden Anforderungen an die Datenproduktion und dem sich verändernden Datenbedarf gerecht wird.

Der zentrale Gedanke bei der Neuausrichtung der Haushaltsstatistiken ist die Etablierung eines Gesamtsystems, in dem die bisher weitgehend unabhängig voneinander durchgeführten Haushaltsstatistiken als eine Erhebung realisiert werden. Für das Gesamtsystem wird jährlich eine 1-Prozent-Stichprobe der deutschen Wohnbevölkerung zu einem einheitlichen Kernprogramm befragt. Die Haushaltsstatistiken Mikrozensus (inklusive des Labour Force Surveys der EU), Leben in Europa (EU-SILC) und Informations- und Kommunikationstechnologien in privaten Haushalten (IKT) werden als Module in das Gesamtsystem integriert.

Mit dem Gesamtsystem wird die Basis für ein zukunftsfähiges System der Haushaltsstatistiken ab 2017 geschaffen. Hierbei werden die Aspekte Datenqualität, Möglichkeit der bedarfsgerechten Datenproduktion, Flexibilität, Effizienz und Reduzierung der Befragtenbelastung berücksichtigt.

Schlagworte: Amtliche Haushaltsstatistiken, Weiterentwicklung, modulare Erhebung
JEL: C81 C83

Steigende Anforderungen für haushaltsstatistische Erhebungen

Seit langen Jahren ermöglichen die amtlichen Haushaltsstatistiken – die Erhebungen bei den privaten Haushalten – eine bedarfsgerechte Bereitstellung fachlich und regional tief gegliederter Informationen, die von zahlreichen Nutzergruppen der Gesellschaft genutzt und geschätzt werden. Für diesen Schlüsselbereich der amtlichen Statistik ist in den kommenden Jahren mit einem massiven Anstieg der Anforderungen zu rechnen. So spricht etwa das von der DGINS, der Konferenz der Leiterinnen und Leiter der Statistischen Ämter der Europäischen Union, im September 2012 verabschiedete Wiesbadener Memorandum[1] von einem allgemein wachsenden Bedarf an Sozialdaten sowie der Notwendigkeit, flexibler auf neue Anforderungen reagieren zu können.

Sehr konkret werden diese steigenden Anforderungen bereits für die Erhebung des Labour Force Surveys der Europäischen Union: Aufgrund geplanter Änderungen der entsprechenden EU-Verordnung ist davon auszugehen, dass ab dem Jahr 2017 die ausgewählten Haushalte im Rahmen einer unterjährigen Wiederholungsbefragung nicht mehr wie bisher nur einmal im Jahr, sondern in der Regel zweimal pro Jahr zu befragen sind. Mit einer solchen Veränderung des Erhebungsdesigns ist für den Datenproduzenten ein erheblicher Mehraufwand verbunden. Auch bei anderen Statistiken, wie zum Beispiel der ebenfalls EU-weit vergleichbaren Statistik über Einkommen und Lebensbedingungen der Bevölkerung in der Europäischen Union (EU-SILC; in Deutschland bekannt unter dem Namen LEBEN IN EUROPA) führt der steigende Informationsbedarf zu absehbar steigenden Anforderungen.

Aber auch in anderen Bereichen zeichnet sich eine Veränderung der Datennachfrage bereits ab. Sowohl die Mitteilung der Europäischen Kommission „Das BIP und mehr – die Messung des Fortschritts einer Welt im Wandel“[2] als auch der Abschlussbericht der „Internationalen Kommission zur Messung der wirtschaftlichen Lage und des sozialen Fortschritts“[3] (bekannter als Stiglitz-Sen-Fitoussi Kommission) empfehlen eine Weiterentwicklung der statistischen Berichterstattung. Beide Berichte fordern eine Ausweitung der bisherigen Berichterstattung

1 Director Generals of the National Statistical Institutes (DGINS) and Federal Statistical Office (Destatis) (2011): Wiesbaden Memorandum. https://www.destatis.de/EN/AboutUs/Events/DGINS/Document_Memorandum.pdf?__blob=publicationFile [17.06.2013].

2 Kommission der Europäischen Gemeinschaften (2009): MITTEILUNG DER KOMMISSION AN DEN RAT UND DAS EUROPÄISCHE PARLAMENT. Das BIP und mehr - Die Messung des Fortschritts in einer Welt im Wandel. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0433:FIN:DE:PDF [17.06.2013].

3 Stiglitz, J. E./Sen, A. and Fitoussi, J.-P. (2009): Report by the Commission on the Measurement of Economic Performance and Social Progress. http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf [17.06.2013].

um weitere qualitativ belastbare und aktuelle Indikatoren. So empfehlen beide Berichte eine Erweiterung der Berichterstattung um Aspekte des Wohlbefindens und der Lebensqualität. Auch hier gibt es bereits sehr konkrete Anforderungen an den Bereich der Haushaltsstatistiken, wie sie etwa im Abschlussbericht der Projektgruppe 2 „Entwicklung eines ganzheitlichen Wohlstands- bzw. Fortschrittsindikators" der Enquete-Kommission Wachstum, Wohlstand, Lebensqualität des Deutschen Bundestages deutlich werden.[4]

Durch eine Weiterentwicklung des Systems der Haushaltsstatistiken müssen daher sowohl die Bewältigung der konkret absehbar höheren Anforderungen gewährleistet als auch die notwendige Flexibilität gewonnen werden, um künftig zeitnah auf eine veränderte Datennachfrage reagieren zu können.

Mit dem Projekt für die Weiterentwicklung des Systems der Haushaltsstatistiken sollen aber auch die Qualitätsaspekte amtlicher Haushaltserhebungen weiter gestärkt werden. So kann durch die Etablierung einer unterjährigen Wiederholungsbefragung für den Labour Force Survey ein notwendiger Qualitätsfortschritt im Bereich der kurzfristigen Arbeitsmarktberichterstattung der Statistischen Ämter des Bundes und der Länder erwartet werden. Des Weiteren können durch die Weiterentwicklung des Systems der Haushaltsstatistiken auch Inkohärenzen zwischen verschiedenen Erhebungen reduziert werden.

All diese Anforderungen treffen nicht nur auf zunehmende Budgetrestriktionen der statistischen Ämter, sondern auch auf politische und gesellschaftliche Erwartungen, dass die Bürgerinnen und Bürger von Statistikpflichten entlastet werden.

Vor diesem Hintergrund hat das Statistische Bundesamt Anfang des Jahres 2012 das Projekt „Weiterentwicklung des Systems der Haushaltsstatistiken" initiiert und gemeinsam mit den Statistischen Landesämtern begonnen. Ziel des Projektes ist die Etablierung eines kohärenten und zukunftsfähigen Systems der Haushaltsbefragungen der amtlichen Statistik, das den eingangs skizzierten Mehraufwand für die statistischen Ämter wie auch die Belastung der Befragten durch Nutzung von Synergien soweit möglich begrenzt und damit den sich ändernden Anforderungen an die Datenproduktion ebenso gerecht wird wie dem sich verändernden Datenbedarf.

Im ersten Jahr des bis Ende 2016 angelegten Projektes wurde ein Grobkonzept für die Weiterentwicklung des Systems der Haushaltsstatistiken entwickelt, dem die Amtsleiterkonferenz der Statistischen Ämter des Bundes und der Länder im November 2012 zugestimmt hat. Im folgenden Kapitel wird das Grobkonzept skizziert, welches als „Grundmodell" bezeichnet wird.

---

4 Deutscher Bundestag, Enquete-Kommission „Wachstum, Wohlstand, Lebensqualität" (2013): Schlussbericht – Wege zum nachhaltigen Wirtschaften und gesellschaftlichem Fortschritt in der sozialen Marktwirtschaft, 231 ff. http://www.bundestag.de/bundestag/gremien/enquete/wachstum/Schlussbericht/17-13300.pdf [20.06.2013].

Das Grundmodell des zukünftigen Systems der Haushaltsstatistiken

Der Leitgedanke des Grundmodells besteht darin, die bisher weitgehend separat durchgeführten Erhebungen Mikrozensus (MZ) inklusive Arbeitskräfteerhebung/Labour Force Survey (LFS), Leben in Europa (EU-SILC) und Informations- und Kommunikationstechnologien in privaten Haushalten (IKT) zukünftig als eine Erhebung zu denken und so zu einem System der Haushaltsstatistiken auszubauen, in dem die genannten Erhebungen als einzelne Module eines Gesamtsystems zu verstehen sind. Enge Verbindungen zwischen den einzelnen Statistiken bzw. Modulen bestehen dann sowohl inhaltlich als auch hinsichtlich der Befragungsinstrumente, des Erhebungsmanagements sowie in Bezug auf Auswertungs- und Analyseverfahren.

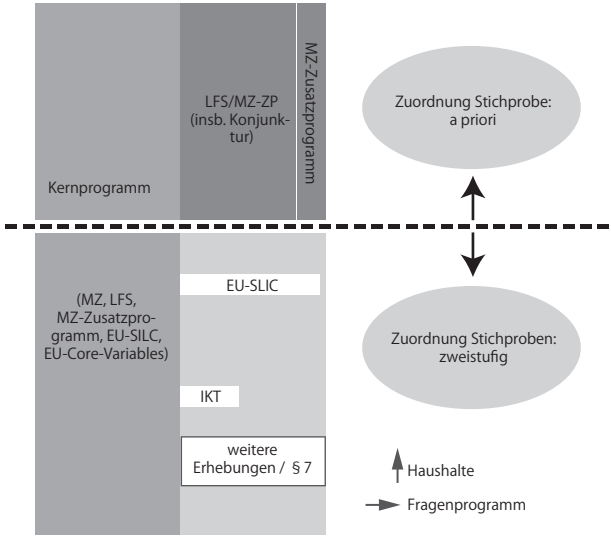


Abbildung 1: Grundmodell[5]

Dieses neue, integrierte System der Haushaltsstatistiken (Abb. 1) kann wie folgt skizziert werden:

Das System der Haushaltsstatistiken basiert auf einer gemeinsamen Zufallsstichprobe (analog zur derzeitigen Mikrozensus-Stichprobe). Die einzelnen Module werden als Substichproben realisiert.

5 Die Größe der Kästchen für die einzelnen Module ist in etwa proportional nach der jeweiligen Zahl der befragten Haushalte (Vertikale) und dem Umfang des jeweiligen Frageprogramms wiedergegeben.

Alle Stichprobenhaushalte bzw. ihre Haushaltsmitglieder werden, wie bisher im MZ, bis zu viermal mit Auskunftspflicht befragt.

Inhaltlich sind die Module über ein Kern-Frageprogramm verknüpft, das bei allen ausgewählten Haushalten erhoben wird. Dieses umfasst neben den EU-Core-Variablen[6] auch den Teil der bisherigen MZ- bzw. LFS-Merkmale, für die zu Auswertungszwecken in fachlicher und regionaler Hinsicht weiterhin ein sehr großer Stichprobenumfang realisiert werden muss. Damit wird auch ein Teil der bisherigen EU-SILC- und IKT-Merkmale abgedeckt.

Zusätzlich zu den Kernmerkmalen werden bei einem Teil der Haushalte die verbleibenden LFS-Merkmale (insbesondere Konjunkturmerkmale) sowie weitere Merkmale des bisherigen MZ-Zusatzprogramms (MZ-ZP) in einem eigenen Modul LFS/MZ-ZP erhoben. Für die verbleibenden LFS-Merkmale und Teile des MZ-ZP besteht Auskunftspflicht[7]. Eine zweite Substichprobe wird zum Modul der EU-SILC-Merkmale befragt, sofern diese Merkmale nicht bereits im Kernprogramm enthalten sind. Mit einem dritten Modul werden IKT-Merkmale bei einer Teilmenge der Haushalte erhoben, wobei jene Merkmale der jetzigen IKT-Erhebung nicht im Modul erfragt werden, die bereits Teil des Kernprogramms sind. Auch für Befragungen, die als Erhebungen nach § 7 BStatG durchgeführt werden, lassen sich zusätzliche Module als Substichproben der Gesamtstichprobe nutzen. Die Module für EU-SILC, IKT sowie Erhebungen nach § 7 BStatG werden auf freiwilliger Basis durchgeführt. Bei Modulen, die ohne Auskunftspflicht erhoben werden, liegen für Haushalte, welche eine freiwillige Befragung verweigern, zumindest die Angaben zum Kernprogramm vor, da für diese Auskunftspflicht besteht. Dies ermöglicht detailliertere Analysen der auftretenden Verzerrungen durch Antwortausfälle und eine bessere Modellierung der Ausfallprozesse für die Hochrechnung als dies gegenwärtig der Fall ist.

Die Gesamtstichprobe wird in Anlehnung an das gegenwärtige Mikrozensus-Design gezogen. Für das Modul LFS/MZ-ZP ist eine a priori-Zuordnung von Stichprobenelementen zur Substichprobe des Moduls möglich. Das heißt bereits bei der Ziehung der Gesamtstichprobe kann eine Teilmenge der Haushalte bestimmt werden, welche für das Modul LFS/MZ-ZP befragt werden. Für die Module EU - SILC und IKT ist aufgrund der wünschenswerten weitergehenden fachlichen, ggf. disproportionalen Schichtung der Substichproben eine zweite Stufe der Zufallsauswahl erforderlich.

---

6 Bei den EU-Core-Variablen handelt es sich um ca. 16 ausgewählte soziodemographische und sozioökonomische Variablen, die auf Empfehlung von Eurostat in allen Haushaltserhebungen in möglichst harmonisierter Form erhoben bzw. bereitgestellt werden sollen. Ziel dieser Harmonisierung ist es, für bestimmte soziale Gruppen und Ergebnisdarstellungen möglichst einheitliche Abgrenzungen über alle Statistiken hinweg zu schaffen.

7 Dies setzt voraus, dass für den LFS durch ein nationales Gesetz eine Auskunftspflicht vorgeschrieben wird. Davon betroffen sind ggf. auch Teile des MZ-ZP.

Das neue Kernprogramm wird zwar in Kombination mit unterschiedlichen Modulen erhoben, die darin enthaltenen Variablen werden aber auch unabhängig von diesen ausgewertet. Für die Ergebnisdarstellung werden die erhobenen Daten sowohl für die einzelnen Substichproben als auch für die Gesamtstichprobe (Kernprogramm) hochgerechnet. Aufgrund des großen Gesamtstichprobenumfangs im Kernprogramm können die darin enthaltenen Variablen regional und inhaltlich tief gegliedert ausgewertet werden, vergleichbar mit dem jetzigen Mikrozensus.

Das Modul LFS/MZ-ZP erfüllt die absehbar deutlich erhöhten EU-Anforderungen an den LFS in methodischer und inhaltlicher Hinsicht durch eine Veränderung der Stichprobenrotation. Da die Merkmale des LFS und des MZ-Zusatzprogramms, sofern sie nicht im Kernprogramm enthalten sind, nur bei einer Teilmenge erhoben werden, reduziert sich jedoch die Möglichkeit einer regionalen und fachlich differenzierten Ergebnisdarstellung für diese Merkmale.

Im skizzierten Grundmodell sind die bisherigen Erhebungen der Wirtschaftsrechnungen aufgrund ihrer spezifischen Komplexität sowie der vielfältigen Nutzungen dieser Statistiken nicht enthalten. Bei der weiteren Konzeptualisierung des Systems der Haushaltsstatistiken sowie bei der Entwicklung der eingesetzten Verfahren und Instrumente wird die Prüfung von Alternativen zur derzeitigen Durchführung der Wirtschaftsrechnungen jedoch mit einbezogen. Das Grundmodell soll dabei auch die Flexibilität einer späteren Integration der Erhebungen zu den Wirtschaftsrechnungen bieten. Damit wird nicht nur die Realisierung weiterer Synergieeffekte möglich, sondern auch eine weitere Steigerung des Analysepotenzials.

Stichprobenmethodik

Die Gesamtstichprobe umfasst jährlich ein Prozent der deutschen Wohnbevölkerung. Es wird davon ausgegangen, dass auf Basis der neuen Auswahlgrundlage des Zensus 2011 jährlich etwa 380.000 Haushalte befragt werden können.

In Anlehnung an die Methodik, die gegenwärtig zur Gewinnung der Stichprobe des Mikrozensus Verwendung findet, wird die Gesamtstichprobe als geschichtete Flächenstichprobe realisiert. Hierzu wird die Grundgesamtheit in Auswahlbezirke unterteilt und jährlich 1 Prozent der Auswahlbezirke befragt. Als Schichtungsmerkmale werden – entsprechend der Regelungen des Zensusgesetzes 2011 – die Merkmale verwendet, welche bereits jetzt zur Schichtung der Mikrozensus-Stichprobe verwendet werden (regionale Schichtung, Gebäudegrößenklasse). Die konkrete Ausgestaltung des neuen Auswahlplanes ist noch zu entwickeln.

Die Gesamtstichprobe wird als rotierende Stichprobe umgesetzt. Ein Teil der Haushalte wird hierbei jährlich durch neue Haushalte ersetzt. Durch den Umstand, dass die Haushalte, welche zum Modul LFS/MZ-ZP beitragen, durch die unterjährige Rotation in der Regel nur zwei Jahre befragt werden, erhöht sich der Anteil der jährlich rotierenden Haushalte von gegenwärtig 1/4 auf rund 1/3.

Die Substichprobe des Moduls LFS/MZ-ZP, die etwa 160.000 Haushalte umfasst, wird als geschichtete Stichprobe realisiert, wobei regionale Merkmale und die Gebäudegrößenklasse als Schichtungsmerkmale verwendet werden. Damit lässt sich bereits zum Zeitpunkt der Ziehung der Gesamtstichprobe eine Teilmenge bestimmen, welche zum Modul LFS/MZ-ZP befragt wird.

Für das Modul LFS/MZ-ZP reduziert sich, verglichen mit dem Status quo des MZ, die Möglichkeit, Merkmale fachlich oder regional feingliedrig differenziert darstellen zu können.

Gegenwärtig werden Jahresergebnisse des Mikrozensus veröffentlicht, wenn die hochgerechneten Besetzungszahlen bei mindestens 5.000 liegen. Der Standardfehler liegt bei dieser Häufigkeit bei rund 15 Prozent. Nach ersten vorläufigen Berechnungen bedarf es für Merkmale des Moduls LFS/MZ-ZP einer hochgerechneten Mindestbesetzungszahl von etwa 10.600, um weiterhin unter einem Standardfehler von rund 15 Prozent zu bleiben.

Bezüglich der Module EU-SILC und IKT ist ein weitergehend fachlich geschichtetes, ggf. disproportionales Ziehungsdesign optimal. Da bei diesen Erhebungen besondere Teilgruppen der Bevölkerung im Fokus stehen und besondere Genauigkeitsanforderungen für diese Teilgruppen existieren, ist es wünschenswert, dass die entsprechenden Teilgruppen überproportional in der jeweiligen Substichprobe enthalten sind.

Aus der verbleibenden Teilmenge der Gesamtstichprobe (ohne die Substichprobe des Moduls LFS/MZ-ZP) im Umfang von etwa 220.000 Haushalten – im Folgenden „Teilstichprobe für die freiwilligen Haushaltserhebungen“ genannt – werden die Substichproben für die freiwilligen Haushaltserhebungen gezogen. Die der fachlichen, ggf. disproportionalen Schichtung zugrunde zu liegenden sozioökonomischen Schichtungsmerkmale der Haushalte sind aus dem mit Auskunftspflicht erfragten Kernprogramm verfügbar und können für eine zweite Stufe der Zufallsauswahl genutzt werden. Während mit der ersten Stufe die Teilstichprobe für die freiwilligen Haushaltserhebungen als Teilmenge der Gesamtstichprobe generiert wird, wird mit der zweiten Stufe die jeweils optimale (disproportionale) Substichprobe für das jeweilige Modul gewonnen. In der praktischen Durchführung ergibt sich mit diesem Verfahren für den ausgewählten Haushalt die aus dem bisherigen MZ bereits bekannte Kombination von Fragen mit Auskunftspflicht (hier dem Kernprogramm) und Fragen mit freiwilliger Auskunftserteilung (hier z. B. zum Modul EU-SILC).

Für die Umsetzung der zweiten Stufe der Zufallsauswahl unter Berücksichtigung des Nicht-Vorliegens der Schichtungsinformationen der Haushalte zum Zeitpunkt der Erstbefragung werden durch das Grundmodell zwei Varianten angeboten:

Variante 1: Auswahlgrundlage für die freiwilligen Haushaltserhebungen (AG-HH)

Im Rahmen der Variante 1 ist vorgesehen, die Haushalte, welche in die Teilstichprobe für die freiwilligen Haushaltserhebungen rotieren, im ersten Jahr lediglich zum Kernprogramm zu befragen.

Ab dem zweiten Erhebungsjahr sind für diese Haushalte die Merkmalsausprägungen der Kernmerkmale als Vorjahresinformation bekannt. Auf Basis dieser Vorjahresinformationen können nun disproportional geschichtete Bruttostichproben für Module gezogen werden. Entsprechend kann für jede freiwillige Haushaltserhebung die notwendige Bruttostichprobe aus der AG-HH gewonnen werden.

Ab dem zweiten Jahr – also mit einem „timelag" von einem Jahr gegenüber der ersten Befragung zum Kernprogramm – kann das zweite Interview zum Kernprogramm und das (erste) Interview zu einem freiwilligen Modul zeitgleich für einen Haushalt durchgeführt werden.

Eine Konsequenz des „timelags" ist, dass in einem gegebenen Jahr die in die Teilstichprobe für die freiwilligen Haushaltserhebungen hineinrotierenden Haushalte (erstmalige Befragung zum Kernprogramm) noch nicht für die Befragung zu einem der freiwilligen Module zur Verfügung stehen. Damit reduziert sich die Anzahl der Haushalte, aus denen die Stichproben für freiwillige Module gezogen werden können, von etwa 220.000 Haushalten auf etwa 165.000 Haushalte.

Variante 2: Verwendung einer Live-Stichprobenziehung

Die Konsequenzen des der Variante 1 immanenten „timelags" können mit dem Ziehen einer sogenannten „Live-Stichprobe" vermieden werden. Hierbei werden für die Haushalte aus der Teilstichprobe für die freiwilligen Haushaltserhebungen die Schichtungsinformationen im Rahmen des – computerunterstützten – Interviews zum Kernprogramm erfasst. Die Angaben werden verwendet, um im Anschluss an die Kernbefragung eine Zufallsauswahl durchzuführen. Für die Haushalte bestehen modul- und schichtspezifische Ziehungswahrscheinlichkeiten für die Bruttostichproben. Ein Zufallsverfahren entscheidet nach Erhebung der relevanten Schichtungsmerkmale im Rahmen des Kernprogramms, ob ein Haushalt für die Bruttostichprobe eines freiwilligen Moduls genutzt wird.

Sowohl die AG-HH als auch die Live-Stichprobenziehung erlauben die Ziehung disproportional geschichteter Nettostichproben für die freiwilligen Module mit I Schichten.

Der Netto-Stichprobenumfang der Schicht i ist bei den freiwilligen Modulen jeweils mit *ni* vorgegeben, wobei i=1,⋯,I. Für den Gesamtstichprobenumfang gilt:

$$\sum_{i=1}^{I} n_i = n$$

Für jede Schicht existiert eine geschätzte Responserate[8], die mit $\hat{\theta}_i$ bezeichnet wird. Um die Vorgaben zur Nettostichprobe näherungsweise zu erfüllen, wird aus jeder Schicht eine Bruttostichprobe vom Umfang $m_i = \frac{n_i}{\hat{\theta}_i}$ angestrebt.

a. AG-HH

Aus der AG-HH kann die Ziehung einer Bruttostichprobe erfolgen, wobei die angestrebten m_i genau eingehalten werden können, sofern die AG-HH genügend Haushalte für die einzelnen Schichten aufweist. Hierbei sind die beiden folgenden Aspekte zu beachten:

- Die AG-HH enthält die Schichtzugehörigkeit eines Haushaltes als Vorjahresinformation. Zum Zeitpunkt der Befragung für ein freiwilliges Modul kann sich die Schichtzugehörigkeit verändert haben.
- Haushalte, die umziehen, werden im Grundmodell nicht weiter befragt. In der Regel gehen umziehende Haushalte verloren. Einen Ausnahmefall bilden Haushalte, welche bereits mindestens einmal am Modul EU-SILC teilgenommen haben. Diese Haushalte werden auch am neuen Wohnort weiter befragt.

Die AG-HH erlaubt zwar die Auswahl einer Bruttostichprobe von genau m_i Haushalten für eine Schicht i. Durch den „timelag" von Erhebung der Schichtungsinformation und der Ziehung der Haushalte für die Bruttostichprobe lässt sich der vorgegebene Bruttostichprobenumfang einer Schicht i nur näherungsweise erreichen.

b. Live-Stichprobe

Um die Auswahlwahrscheinlichkeiten der Live-Stichprobe zu ermitteln, bedarf es einer Annahme über die Anzahl der Haushalte der Schicht i in der Substichprobenauswahlbasis. Die Annahme zur Anzahl der Elemente dieser Menge wird mit $\hat{l}_i$ bezeichnet. Unter der Annahme, dass $m_i \leq \hat{l}_i$, ergibt sich die schichtspezifische Ziehungswahrscheinlichkeit in der zweiten Phase mit $p_i = \frac{m_i}{\hat{l}_i}$. Als zweite Phase wird hierbei das Ziehen der Bruttostichprobe aus der Substichprobenauswahlbasis bezeichnet.

Im Rahmen der Live-Stichprobe wird für jeden Haushalt eine unabhängige

8 Gegenwärtig bereitet die amtliche Statistik einen Responseratentest vor, um eine empirische Grundlage für die Beteiligungsbereitschaft der Bevölkerung unter den Randbedingungen des Grundmodells zu erhalten.

Zufallsauswahl durchgeführt. Ein Haushalt der Schicht i wird mit einer Wahrscheinlichkeit p_i für die Bruttostichprobe gezogen. Da es sich um unabhängige Ziehungen handelt, folgt der realisierte Stichprobenumfang der Schicht i einer Binomialverteilung mit dem Erwartungswert m_i, sofern sich die Annahme zu $\hat{l}_i$ als zutreffend erweist.

Kernprogramm und Module des Grundmodells

Kernprogramm

Das Grundmodell geht davon aus, dass die Befragungen der Haushalte zum Kernprogramm gleichmäßig auf alle Kalenderwochen verteilt werden. Jeder Auswahlbezirk und damit jeder Haushalt wird einer bestimmten Berichtswoche zugeordnet, auf die sich die Befragung bezieht (feste Berichtswoche).

Je nachdem, ob die Haushalte zur Substichprobe des Moduls LFS/MZ-ZP oder zur Teilstichprobe für die freiwilligen Haushaltserhebungen gehören, nehmen die Haushalte in der Regel in zwei (LFS/MZ-ZP) oder in vier (Teilstichprobe) aufeinanderfolgenden Jahren an der Befragung zum Kernprogramm teil.

Auskunftspflichtig für das Kernprogramm sind, wie derzeit beim Mikrozensus, alle Haushalte, die in der Berichtswoche im Auswahlbezirk wohnhaft sind. Beim Umzug von Haushalten werden diese nicht weiterverfolgt, stattdessen werden die neu hinzugezogenen Haushalte befragt.

Modul LFS/MZ-ZP

Die Erhebung zum Modul LFS/MZ-ZP erstreckt sich wie das Kernprogramm kontinuierlich mit fester Berichtswoche über das gesamte Kalenderjahr.

Gemäß den zu erwartenden EU-Vorgaben wird der LFS zukünftig auf eine unterjährige Rotation umgestellt. Nach dem aus fachlicher Sicht angestrebten und auch von EUROSTAT favorisierten 2-(2)-2-Rotationsschema werden die Haushalte in zwei aufeinanderfolgenden Quartalen befragt (Bsp.: 1. + 2. Quartal im Jahr t). In den beiden folgenden Quartalen werden die Haushalte nicht befragt (Bsp.: 3. + 4. Quartal im Jahr t). Danach nehmen die Haushalte wiederum in zwei aufeinanderfolgenden Quartalen an der Befragung teil (Bsp.: 1. + 2. Quartal im Jahr t+1). Nur beim jeweils ersten Interview eines Jahres (Bsp.: 1. Quartal im Jahr t und t+1) handelt es sich um eine ausführliche Befragung zu dem vollständigen Kernprogramm, den weiteren LFS-Merkmalen und den weiteren Merkmalen des MZ-Zusatzprogramms. Die jeweils anschließende Befragung im Folgequartal kann durch die Nutzung von Informationen aus der ersten Befragungswelle (dependent interviewing) ggf. abgekürzt werden. Zudem beschränkt sich die Folgebefragung überwiegend auf unterjährig veränderliche Merkmale.

Gemäß diesem Rotationsschema würden die Haushalte i. d. R. nur noch in zwei aufeinanderfolgenden Jahren befragt, wenn an der bisherigen gesetzlichen Vorgabe des Mikrozensusgesetzes von maximal vier Befragungen eines Stichprobenhaushaltes festgehalten wird. Dies führt im Grundmodell zu einer erhöhten Stichprobenrotation für die Substichprobe LFS/MZ-ZP.

Modul EU-SILC

Bei EU-SILC handelt es sich um eine Panelerhebung, bei der die Haushalte auf freiwilliger Basis viermal mit einem Abstand von jeweils einem Jahr befragt werden. Jährlich wird ein Viertel der Stichprobe ausgetauscht.

Für das Modul EU-SILC wird die AG-HH zur Ziehung einer Bruttostichprobe für das neue Rotationsviertel verwendet. Hierbei werden nur solche Haushalte berücksichtigt, deren Befragung zum Kernprogramm im 2. oder 3. Quartal stattfindet. Daneben werden nur solche Haushalte ausgewählt, die bisher nur einmal zum Kernprogramm beigetragen haben.

Die erste Befragung zu EU-SILC findet somit im zweiten Befragungsjahr des Kernprogramms statt. Damit die Haushalte viermal an EU-SILC partizipieren können, müssen diese Haushalte somit fünf Jahre in Folge befragt werden. Im fünften Jahr kann jedoch das Kernprogramm nicht mehr mit Auskunftspflicht erhoben werden. Daneben ist zu beachten, dass für EU-SILC sowohl Haushalte als auch einzelne Personen von Haushalten im Falle eines Umzuges weiterbefragt werden. Auch für diese Haushalte bzw. Personen ist eine Befragung des Kernprogramms mit Auskunftspflicht nicht möglich.

Modul IKT

Der Befragungszeitraum zu IKT beschränkt sich derzeit auf das zweite Quartal eines Jahres, da aus methodischen Gründen das erste Quartal des Kalenderjahres als Referenzzeitraum gesetzlich vorgeschrieben ist. Auch im Grundmodell ist daher eine Beschränkung des Befragungszeitraums für das Modul IKT auf das zweite Quartal vorgesehen. Als Basis für die Substichprobe IKT sollen daher nur Haushalte dienen, die im zweiten Quartal zum Kernprogramm befragt werden. Da es sich bei IKT nicht um eine Panelerhebung handelt, wird jedes Jahr eine neue Substichprobe gezogen. Gegenwärtig ist noch nicht entschieden, ob die Bruttostichprobe für das Modul IKT mit der AG-HH oder der Live-Stichprobe durchgeführt werden wird.

Weitere Erhebungen

Das Grundmodell soll auch Raum für die Aufnahme von neuen bzw. ad hoc-Erhebungen zu aktuellen Fragestellungen bieten und sieht dafür die Möglichkeit

zur Integration weiterer Erhebungen als jeweils eigenständige Module in das System der Haushaltsstatistiken vor. Für diese zusätzlichen Module können im Grundmodell weitere Substichproben aus der Teilstichprobe für die freiwilligen Haushaltserhebungen bereitgestellt werden. Bei weiteren Erhebungen, die ggf. in das System der Haushaltsstatistiken integriert werden sollen, handelt es sich z. B. um Erhebungen nach § 7 BStatG .

Sowohl Erhebungen, die sich über das gesamte Kalenderjahr erstrecken, als auch Erhebungen mit einem beschränkten Befragungszeitraum sind dabei denkbar.

Erhebungsdurchführung

Mit dem neuen System der Haushaltsstatistiken wird ein integriertes System geschaffen, welches die bisher weitgehend separat durchgeführten Erhebungen als Module einer gemeinsamen Erhebung versteht. Ein integriertes System ist mit Synergieeffekten verbunden. Gleichzeitig steigt jedoch auch die Komplexität eines solchen Systems. Entsprechend ist es für eine erfolgreiche Durchführung einer integrierten Erhebung entscheidend, dass hierzu leistungsfähige IT-Systeme zur Verfügung stehen, welche ein effizientes Erhebungsmanagement ermöglichen. Eine umfassende IT-Unterstützung muss das Verwalten von Stichprobendaten, den Einsatz von Interviewern und Interviewerinnen, den Versand von Erinnerungs- und Mahnschreiben und vieles weitere mehr leisten.

Ein weiterer wichtiger Aspekt der IT-Infrastruktur liegt in der Unterstützung verschiedener Erhebungsmodi. Mit der Umsetzung des Grundmodells wird stärker als bisher ein mixed-mode-Ansatz verfolgt. Die Fragen des Kernprogramms und des Moduls LFS/MZ-ZP werden künftig zu einer festen Berichtswoche erhoben, d. h. die Haushalte beantworten die Fragen nicht mehr wie bisher mit Bezug zur Woche vor dem Interview, sondern mit Bezug zu einer festen Berichtswoche. Damit ein befragter Haushalt möglichst zeitnah nach dieser Berichtswoche seine Angaben zum Kernprogramm macht, werden verschiedene Erhebungsmodi flexibel eingesetzt. Außerdem wird erwartet, dass durch das Angebot mehrerer Erhebungsmodi die Teilnahmebereitschaft der Haushalte gesteigert wird. Des Weiteren kann ein mixed-mode-Ansatz auch für Kostenersparnis bei den statistischen Ämtern sorgen.

Für die Erhebungsdurchführung können zum einen Interviewer/innen eingesetzt werden, welche mit Laptopunterstützung im Haushalt die Befragung als face-to-face Interview durchführen (Computer Assisted Personal Interview – CAPI).
Als weiterer Mode kommt ein Telefoninterview in Betracht (Computer Assisted Telephone Interview – CATI). Das Telefoninterview kann hierbei als aktives (der Interviewer/die Interviewerin ruft beim Haushalt an) oder passives (der Haushalt

kontaktiert den Interviewer/die Interviewerin oder das statistische Amt) Interview durchgeführt werden. Daneben kann das Telefoninterview über Telefonstudios oder als „home based interview“ durchgeführt werden.

Mit einer web-basierten Befragung kann ein weiterer computerbasierter Mode eingesetzt werden (Computer Assisted Web Interviewing – CAWI). Hierbei erhält der Haushalt eine Zugangskennung für ein Web-Formular und kann seine Angaben online tätigen.

Zur Ergänzung der computerbasierten Modi kann eine Befragung auch über einen Papierfragebogen realisiert werden. Papierfragebögen sind hierbei mit dem Nachteil verbunden, dass die Filterführung für die Befragten schwieriger ist, keine Plausibilisierungen durchgeführt werden können, der Papierfragebogen an ein statistisches Amt übermittelt werden muss und dort elektronisch erfasst werden muss.

Welche Modi beim Kernprogramm sowie bei den Modulen zum Einsatz kommen können, muss jeweils separat entschieden werden. Hierbei ist zu beachten, dass jeder Mode spezifische Selektions- und Measurementeffekte mit sich bringt. So kann etwa der Einsatz eines Modes zu einem Modul ein selektives Teilnahmemuster erzeugen, in dem die Teilnahmewahrscheinlichkeit von Eigenschaften eines Haushaltes abhängt, die mit den Frageinhalten des Moduls korreliert sind. Bleibt ein solcher Selektionseffekt unkontrolliert, kann durch die Auswahl der Erhebungsmodes gegebenenfalls zwar die Responserate erhöht, jedoch die Ergebnisqualität verschlechtert werden.

Im weiteren Projektverlauf ist zu prüfen, wie sich der Modus CATI insbesondere bei der 2. und 4. Befragung zum Modul LFS/MZ-ZPgewinnbringend einsetzen lässt. Bei der 2. und 4. Befragung des Moduls LFS/MZ-ZP handelt es sich um Folgebefragungen, die jeweils drei Monate nach der 1. bzw. 3. Befragung zum Modul LFS/MZ-ZP stattfinden. Der Befragungsaufwand für den befragten Haushalt als auch für die statistischen Ämter kann reduziert werden, wenn für die 2. und 4. Befragung des Moduls LFS/MZ-ZP vorrangig ein ‚dependent‘ Interview geführt werden kann, welches mit dem Modus CATI umgesetzt wird.

Neue Schätzverfahren und Ergebnisdarstellung

Wenn die Weiterentwicklung des Systems der Haushaltsstatistiken die anstehenden zusätzlichen Anforderungen erfüllen und dabei den Ressourceneinsatz wie auch die Befragungsbelastung für die auskunftgebende Bevölkerung möglichst wenig erhöhen soll, ist es unvermeidlich, dass einzelne Module mit weniger Befragten abgedeckt werden, als dies bisher der Fall ist. Im Grundmodell beträfe dies das Modul LFS/MZ-ZP. Eine Umstellung des Moduls LFS/MZ-ZP auf eine unterjährige Rotation nach einem 2-(2)-2-Schema würde unter den genannten

Rahmenbedingungen im Vergleich zum Status quo zu einer Reduktion der für die Erstellung von Quartals- und Jahresergebnissen heranzuziehenden Befragungsbasis führen. Unter Beibehaltung der bisherigen im Mikrozensus angewandten Analyse- und Hochrechnungsverfahren (insbesondere „generalized regression estimation") würde sich damit u. U. eine Einschränkung der Auswertungsmöglichkeit hinsichtlich feingliedrig sachlicher oder regionalisierter Ergebnisdarstellung ergeben. Um dieser veränderten Situation gerecht zu werden, ist eine Erweiterung der verwendeten Methodik in der amtlichen Statistik zu prüfen. Hierbei erscheinen besonders ‚small area estimation'-Verfahren und „statistical matching"-Verfahren gewinnbringend.

In ihrer Intention erlauben „small area estimation"-Verfahren das Schätzen von Ergebnissen, welche auch bei kleinen Stichprobenumfängen hinreichend kleine Stichprobenfehler aufweisen. Die Einsatzmöglichkeiten von „small area estimation"-Verfahren sind im folgenden Projektverlauf daraufhin zu untersuchen, wie die Folgen verminderter Stichprobenumfänge kompensiert werden können. Dabei sind mehrere Aspekte zu berücksichtigen: Zunächst sind „small area"-Schätzungen weitaus komplexer als klassische Schätzverfahren; es sind also nicht nur die Einsatzmöglichkeiten zu prüfen, sondern auch die Anforderungen an die mit dem Einsatz verbundenen notwendigen Personalkapazitäten für die Modellierung und Bewertung der Ergebnisse. Weiterhin sind die Erfahrungen beim Zensus 2011 sowie auch die Erfahrungen anderer nationaler statistischer Ämter einzubeziehen. Insgesamt scheint das System der Haushaltsstatistiken aufgrund der gemeinsamen Stichprobenziehung und der zu erwartenden Menge an Hilfsinformationen zumindest aber eine günstige Grundlage für „small area estimation"-Verfahren zu bieten.

Gleiches gilt auch hinsichtlich der Frage, inwieweit sich neues Analysepotential aus den Möglichkeiten des ‚statistical matching' ergeben könnte. Nach dem Grundmodell liegen Angaben für alle Befragten nur zum Kernprogramm vor. Zu den anderen Modulen werden jeweils nur die Angaben für die befragte Teilpopulation der Gesamtstichprobe erhoben. Beispielsweise liegen für Befragte, welche am Modul LFS/MZ-ZP teilnehmen, somit keine Angaben zu den Variablen von EU-SILC vor. Durch das gemeinsame Stichprobendesign sowie das Vorhandensein des gemeinsamen Kernprogramms eröffnet das skizzierte Modell jedoch auch neue Möglichkeiten, Merkmalsausprägungen von nicht erhobenen Merkmalen zu schätzen. So könnte z. B. mittels eines „statistical matching" ein synthetischer Datenbestand erzeugt werden, in dem die Variablen aus unterschiedlichen Erhebungsmodulen enthalten sind. Dies erlaubt beispielsweise die gemeinsame Analyse von Merkmalen des LFS und EU-SILC. Auch der Einsatz dieser Schätzmethode erfordert ausreichend qualifizierte Personalkapazitäten.

Die Möglichkeiten der Umsetzung dieser Verfahren in der amtlichen Statistik, die dafür erforderlichen Ressourcen sowie die Akzeptanz bei den wichtigsten Datennutzern/-nutzerinnen bedürfen allerdings noch der intensiven Prüfung.

Umfang der Frageprogramme und Anzahl der befragten Haushalte

Umfang der Frageprogramme

Im vorgeschlagenen Grundmodell hängt der Umfang des Frageprogramms, welches von den Haushalten beantwortet wird, im hohen Maße zunächst vom Umfang und Inhalt des Kernprogramms ab. Um die Belastung der befragten Haushalte in Grenzen zu halten wird angestrebt, dass der Umfang des Kernprogramms rund die Hälfte des bisherigen MZ/LFS-Frageprogramms enthält. Ein umfangreiches Kernprogramm verbessert zwar die Analysemöglichkeiten, gleichzeitig bringt es aber auch eine hohe Befragtenbelastung mit sich und verringert dadurch auch die Bereitschaft, danach noch an einer freiwilligen Befragung teilzunehmen.

Ausreichend hohe Responseraten zu den freiwilligen Modulen sind entscheidend für den Erfolg des Grundmodells. Aus der Teilstichprobe der freiwilligen Haushaltserhebungen müssen hinreichend große Bruttostichproben gezogen werden. Die Haushalte der Bruttostichprobe werden um eine Teilnahme an den freiwilligen Modulen gebeten. Haushalte, welche an der freiwilligen Erhebung teilnehmen, bilden die Nettostichprobe. Nur wenn hinreichend große Responseraten erzielt werden können, lassen sich die Nettostichproben mit den notwendigen Stichprobenumfängen gewinnen und qualitativ gute Ergebnisse erzielen. Neben der grundsätzlichen Teilnahmebereitschaft in der Bevölkerung und der mit dem Umfang des Frageprogramms verbundenen Belastung kommen auch dem Auftreten und der Überzeugungskraft der Interviewerinnen und Interviewer sowie einer praxisgerechten Erhebungsunterstützung ein entscheidender Einfluss auf die Bereitschaft der Haushalte zu, zusätzlich zum verpflichtenden Kernprogramm auch zu den freiwilligen Modulen Auskunft zu erteilen. Mit Incentives, v.a. Geldprämien, soll ein zusätzlicher Anreiz für die Teilnahme der Haushalte geschaffen werden. Die Möglichkeiten, positiv auf die Responseraten einzuwirken, sollen im weiteren Projektverlauf empirisch untersucht werden.

Schließlich stellt auch die Harmonisierung/Abstimmung der Erhebungsmerkmale von LFS und EU-SILC auf nationaler und europäischer Ebene einen weitergehenden Arbeitsschwerpunkt dar, weil sich derzeit Definition und Operationalisierung der Erhebungsmerkmale zwischen den Erhebungen teilweise unterscheiden. Nur durch eine umfassende Harmonisierung der Erhebungsmerk-

male lassen sich Mehrfachbefragungen vermeiden und die angestrebten Entlastungseffekte realisieren.

Ein besonderes Augenmerk verdient im Grundmodell der Kreis der Haushalte, der zunächst nicht für eines der Module LFS/MZ-ZP, EU-SILC oder IKT ausgewählt wurde und nur das Kernprogramm bedienen muss. Der Aufwand dieser Haushalte reduziert sich im gewählten Beispiel auf etwa die Hälfte des bisherigen MZ/LFS und so bietet sich die Möglichkeit, diese Haushalte zusätzlich etwa für freiwillige Erhebungen nach § 7 BStatG auszuwählen.

Anzahl der befragten Haushalte

Derzeit werden für die betrachteten Haushaltsstatistiken – ohne Einkommens- und Verbrauchsstichprobe (EVS) und Laufende Wirtschaftsrechnung (LWR) – insgesamt jährlich ca. 377.000 Haushalte befragt. Es ist davon auszugehen, dass sich mit der neuen Auswahlgrundlage aus dem Zensus 2011 die Stichprobengröße für den Mikrozensus um ca. 40.000 auf etwa 380.000 Haushalte erhöhen wird, so dass zukünftig unter Beibehaltung des Status quo für die Haushaltsstatistiken jährlich etwa 417.000 Haushalte befragt werden müssten.

	Anzahl befragter Haushalte
KERN (ausschließlich)	*177.500*
KERN + LFS/MZ-ZP	*160.000*
KERN + EU-SILC	*10.500*
EU-SILC	*3.500*
KERN + IKT	*12.000*
KERN + § 7 Erhebung	*20.000*
Insgesamt	383.500

Tabelle 1: Anzahl der befragten Haushalte pro Jahr

Die Verteilung der Haushalte auf die einzelnen Module, wie sie sich im Grundmodell ergeben wird, ist in Tabelle 1 dargestellt. Hierbei beantworten jährlich 160.000 Haushalte neben den Fragen zum Kernprogramm die Fragen des Moduls LFS/MZ-ZP. Die Nettostichprobe des Moduls EU-SILC umfasst 14.000 Haushalte. Davon werden 10.500 Haushalte neben EU-SILC auch zum verpflichtenden Kernprogramm befragt. Für 3.500 Haushalte kann hingegen kein verpflichtendes Kernprogramm durchgeführt werden, weil diese Haushalte sich bereits im fünften Befragungsjahr befinden. Die Nettostichprobe des Moduls IKT umfasst 12.000 Haushalte. Diese Haushalte beantworten neben dem Modul IKT

ebenfalls verpflichtend das Kernprogramm. Daneben geht Tabelle 1 beispielhaft davon aus, dass eine Nettostichprobe von 20.000 Haushalten für eine zusätzliche § 7 Erhebung genutzt wird. Auch die Haushalte, die zur § 7 Erhebung beitragen, beantworten zusätzlich die Fragen des Kernprogramms. Zusätzlich werden von 177.500 ausschließlich die Kernmerkmale verwendet. Die Gesamtzahl der zu befragenden Haushalte setzt sich zusammen aus der 1-Prozent-Stichprobe von 380.000 Haushalten sowie 3.500 Haushalten, welche sich das vierte Mal an EU-SILC beteiligen.

Sowohl im Hinblick auf die Befragtenbelastung als auch auf den Erhebungsaufwand ist besonders zu beachten, dass die 160.000 Haushalte, die in das Modul LFS/MZ-ZP einbezogen werden, aufgrund der unterjährigen Wiederholungsbefragung in der Regel zweimal im Jahr zu befragen sind.

Fazit

Mit den sich abzeichnenden neuen Anforderungen stehen die Statistischen Ämter des Bundes und der Länder in einem der Schlüsselbereiche der amtlichen Statistiken, den amtlichen Haushaltserhebungen, vor großen Herausforderungen. In dem vorgelegten Grundmodell zur Weiterentwicklung des Systems der Haushaltsstatistiken und dem zugrundeliegenden systemischen Ansatz sehen sie die Basis für ein modular aufgebautes, kohärentes und zukunftsfähiges System der Haushaltsstatistiken ab 2017.

Oversampling vermögender Haushalte im Rahmen der Studie „Private Haushalte und ihre Finanzen (PHF)“

Dr. Tobias Schmidt
Martin Eisele

Zusammenfassung

Die Deutsche Bundesbank hat zwischen September 2010 und Juli 2011 erstmals stichprobenartig private Haushalte in Deutschland über ihre finanzielle Lage und ihr Vermögen befragt. Die Ergebnisse dieser freiwilligen Befragung werden in der Panelstudie „Private Haushalte und ihre Finanzen“ (PHF) zusammengefasst. Im Rahmen der ersten Befragungswelle haben 3.565 Haushalte Auskunft gegeben. Die Studie ist bevölkerungsrepräsentativ angelegt. Jeder private Haushalt in Deutschland (ausgenommen Anstaltshaushalte und Haushalte in Gemeinschaftsunterkünften) mit mindestens einem bei einem deutschen Einwohnermeldeamt gemeldeten Haushaltsmitglied ab 18 Jahre konnte mit einer positiven Wahrscheinlichkeit in die Stichprobe gelangen. Das integrierte Ziehungsdesign der Befragung sah vor, dass wohlhabende Haushalte mit einer höheren Wahrscheinlichkeit in die Stichprobe gelangen. Sie sind absichtlich überrepräsentiert, um die Vermögenszusammensetzung und –verteilung besser analysieren zu können. Die Stichprobenziehung des PHF bestand aus mehreren Stufen, in denen jeweils versucht wurde, ein Oversampling von vermögenden Haushalten zu erreichen. Auf der ersten Stufe wurde nach der Größe der Gemeinde und in Gemeinden mit weniger als 100.000 Einwohnern nach dem Anteil vermögender Haushalte gemäß Einkommenssteuerstatistik geschichtet. Für Gemeinden mit 100.000 und mehr Einwohnern war es möglich, „vermögende“ und „andere“ Straßenabschnitte zu identifizieren und somit zwei weitere Schichten zu bilden. Um die Straßenabschnitte zu klassifizieren, wurden mikrogeografisch-statistische Verteilungen verwendet, welche die Qualität der Wohngegend, den Gebäudetyp sowie die Kaufkraft für jeden Straßenabschnitt darstellen. Das Oversampling wurde durch eine Überrepräsentation von reichen kleinen Gemeinden und vermögenden Straßenabschnitten in Großstädten realisiert. Erste Analysen deuten darauf hin, dass die gewählte Oversampling-Strategie erfolgreich war.

Schlagwörter: Stichprobendesign, Oversampling, Haushaltsbefragung, Finanzen privater Haushalte

JEL: C83, C81

Disclaimer: Der Beitrag stellt die persönliche Auffassung der Autoren dar und gibt nicht notwendigerweise die Position der Bundesbank oder ihrer Beschäftigten wieder.

Einleitung und Überblick

Die Deutsche Bundesbank hat zwischen September 2010 und Juli 2011 stichprobenartig deutsche Haushalte über ihr Vermögen und ihre Finanzen befragt. Die Ergebnisse dieser freiwilligen Befragung werden in der Panelstudie „Private Haushalte und ihre Finanzen" (PHF) zusammengefasst. Die Daten umfassen vor allem die Vermögensbilanzen der Haushalte, ihre Rentenansprüche, die Spartätigkeit, das Einkommen, Daten zur Erwerbstätigkeit, zum Konsum, zu Einstellungen und Erwartungen sowie viele demographische Charakteristika. Die PHF-Studie ist Teil einer neuen, harmonisierten Befragung, die in allen Euroländern durchgeführt wird (HFCS). Die Daten liefern ein umfassendes Bild der Vermögens- und Verschuldungssituation privater Haushalte und ihrer Bestimmungsgründe und ermöglichen so ein besseres Verständnis etwa des Spar- und Konsumverhaltens, der Vermögensverteilung oder der Insolvenzrisiken. Im Rahmen der ersten Befragungswelle haben 3.565 Haushalte Auskunft gegeben[1].

In diesem Aufsatz wird das Stichprobendesign der Studie beschrieben. Der Schwerpunkt liegt dabei auf der Darstellung des Oversamplings vermögender Haushalte. Die PHF-Studie ist grundsätzlich bevölkerungsrepräsentativ angelegt. Jeder private Haushalt in Deutschland (ausgenommen Anstaltshaushalte und Haushalte in Gemeinschaftsunterkünften) mit mindestens einem bei einem deutschen Einwohnermeldeamt gemeldeten Haushaltsmitglied ab 18 Jahre konnte mit einer positiven Wahrscheinlichkeit gezogen werden. Wohlhabende Haushalte sollten in der Stichprobe überrepräsentiert werden, um auch Aussagen über die Zusammensetzung des Vermögens insgesamt und die Ränder der Vermögensverteilung zu ermöglichen. Eine proportional angelegte Stichprobe würde nur wenige Vermögende enthalten, da Vermögen stärker konzentriert ist als Einkommen. Über viele Vermögenskategorien wären dann keine aussagekräftigen Informationen zu gewinnen.

Die Stichprobenziehung des PHF bestand aus mehreren Stufen, in denen jeweils versucht wurde, ein Oversampling von vermögenden Haushalten zu erreichen. Auf der ersten Stufe wurde nach der Größe der Gemeinde und nach dem Anteil vermögender Haushalte gemäß Einkommenssteuerstatistik geschichtet. Unterschieden wurden kleine Gemeinden (<100.000 Einwohner) und große Gemeinden (>=100.000 Einwohner). Die kleinen Gemeinden wurden auf Basis des Anteils von Steuerzahlern mit einem gesamten steuerpflichtigen Einkommen über einem fest definierten Schwellenwert als „vermögende kleine Gemeinden"

1 Weitere Informationen zur PHF Studie und erste Ergebnisse wurden in den Monatsberichten der Deutschen Bundesbank vom Januar 2012 („Das PHF: eine Erhebung zu Vermögen und Finanzen privater Haushalte in Deutschland") und vom Juni 2013 („Vermögen und Finanzen privater Haushalte in Deutschland: Ergebnisse der Bundesbankstudie") publiziert.

bzw. „sonstige kleine Gemeinden“ kategorisiert. Vermögende kleine Gemeinden sind in der Stichprobe überrepräsentiert. Für Gemeinden mit 100.000 und mehr Einwohnern war es möglich, „vermögende“ und „andere“ Straßenabschnitte zu identifizieren und somit zwei weitere Schichten zu bilden. Um die Straßenabschnitte zu klassifizieren, wurden mikrogeografisch-statistische Verteilungen verwendet, welche die Qualität der Wohngegend, den Gebäudetyp sowie die Kaufkraft für jeden Straßenabschnitt darstellen. Entsprechend der Überrepräsentation von reichen kleinen Gemeinden wurden in Großstädten vermögende Straßenabschnitte überrepräsentiert. In den ausgewählten Gemeinden bzw. Straßenabschnitten wurden schließlich Adressen mit einem systematischen Zufallsauswahlverfahren mit fester Schrittlänge und zufälliger Startzahl aus den Einwohnermeldeamtsregistern gezogen.

Das gewählte Design hat Auswirkungen auf die Zusammensetzung der Stichprobe und die Gewichtungsfaktoren. Erste Analysen deuten darauf hin, dass die gewählte Oversampling-Strategie erfolgreich war. So entfallen auf das oberste Quintil der (aus den Daten geschätzten) Vermögensverteilung 37% der realisierten Nettostichprobe. Im obersten Dezil befinden sich 22% der Haushalte in der Stichprobe. Allerdings reduziert sich auch die effektive Fallzahl durch die Designgewichtung um etwa 40%.

Das Ziehungsdesign der PHF-Studie[2]

Das Ziehungsschema der PHF-Studie besteht aus mehreren Stufen, die in Tabelle 1 überblickartig dargestellt sind. Während in Städten mit 100.000 und mehr Einwohnern ein dreistufiges Verfahren zur Anwendung kam, wurde in kleineren Gemeinden ein zweistufiges Verfahren implementiert. Das Oversampling wurde über eine integrierte Ziehung realisiert, d.h. die vermögenderen Haushalte wurden über gezielte Schichtung und entsprechende Auswahlsätze überrepräsentiert und nicht, wie teilweise in anderen Studien, durch eine separate Vermögenden-Stichprobe. In den folgenden Unterkapiteln werden die einzelnen Ziehungsstufen genauer beschrieben und das Oversampling-Konzept näher erläutert.

2 Zusätzliche Tabellen und weitere Informationen zur Methodik der PHF-Studie können dem Methodenband von infas zur ersten Welle entnommen werden. Dieser ist auf der Website der Bundesbank abrufbar: http://www.bundesbank.de/Redaktion/DE/Downloads/Bundesbank/Forschungszentrum/methodenbericht.pdf?__blob=publicationFile

1. Stufe	Explizite Stratifizierung	*Gemeinden nach Größe und "Vermögen"*
	Implizite Stratifizierung	*Ortsgrößenklasse X Kreis in jedem expliziten Stratum*
	Cluster	*Gemeinden (PSU)*
	Ziehungsverfahren	*Probability proportional to size (PPS)* *Basis: Anzahl Haushalte pro Gemeinde*
	Certainty units	*Sehr große Städte immer in der Stichprobe vertreten (Anzahl der Haushalte kann variieren)*
2. Stufe (nur Großstädte mit 100.000 und mehr Einwohnern)	Explizite Stratifizierung	*Straßenabschnitte nach Vermögen*
	Cluster	*Straßenabschnitte (SSU)*
	Ziehungsverfahren	*Probability proportional to size (PPS)* *Basis: Haushalte pro Straßenabschnitt*
3. Stufe	Ziehungsverfahren	*Systematic Sampling von Personen aus Registern der Einwohnermeldeämter*

Tabelle 1: Ziehungsdesign - Überblick

Die erste Ziehungsstufe

Die erste Stufe der Stichprobenziehung stellt auf Gemeinden ab. Die gut 12.200 Gemeinden in Deutschland wurden nach ihrer Größe und nach dem Anteil vermögender Haushalte in der jeweiligen Gemeinde explizit geschichtet.

Für das Oversampling der Haushalte ist insbesondere die Gruppierung der Gemeinden nach dem Anteil der vermögenden Haushalte relevant. Da das Vermögen der Haushalte bzw. Gemeinden nicht aus existierenden Statistiken oder Registern entnommen werden kann, wurde für die Identifikation vermögender Haushalte die Einkommenssteuerstatistik des Statistischen Bundesamts herangezogen. Die Idee dabei ist, dass Einkommen und Vermögen stark miteinander korreliert sind. Das Statistische Bundesamt stellte für jede Gemeinde in Deutschland den Prozentanteil der Steuerzahler mit einem gesamten steuerpflichtigen Einkommen über einen bestimmten Schwellenwert zur Verfügung. Auf Basis dieser

Information wurden anschließend die Gemeinden mit weniger als 100.000 Einwohnern in zwei Schichten, vermögende kleine Gemeinden und sonstige kleine Gemeinden, unterteilt. Der Prozentanteil, ab dem eine kleine Gemeinde als vermögend eingestuft wurde, variierte zwischen Gemeinden mit weniger als 5.000 Einwohnern und Gemeinden mit 5.000 bis 99.999 Einwohnern. Großstädte wurden in der ersten Stufe der Ziehung nicht weiter unterteilt, sondern bilden eine eigene Schicht. Im Ergebnis ergeben sich damit drei Schichten: Schicht 1 („kleine vermögende Gemeinden") mit 1.470 Gemeinden in der Grundgesamtheit, Schicht 2 („kleine sonstige Gemeinden") mit 10.672 Gemeinden und Schicht 3 mit 81 Großstädten.

Die Schichtung der Gemeinden nach dem Anteil vermögender Haushalte ermöglicht das Oversampling. So wurden ebenso viele sog. „Sample Points"[3] in kleinen vermögenden Gemeinden gezogen wie in kleinen sonstigen Gemeinden, obwohl in ersteren nur 8% der Einwohner in der Grundgesamtheit leben, während es in letzteren 61% sind. Auch Großstädte wurden leicht überrepräsentiert. Hier leben 31% der Einwohner in der Grundgesamtheit, es wurden aber 40% der 230 Sample Points aus Großstädten gezogen. Die Gemeindestichprobe bestand am Ende aus jeweils 69 kleinen vermögenden bzw. anderen kleinen Gemeinden und 58 Großstädten.

Um eine regionale Streuung der gezogenen Gemeinden zu erreichen, wurde zusätzlich nach Kreisen und Ortsgrößenklasse implizit stratifiziert. Die tatsächliche Ziehung von Gemeinden innerhalb der drei Schichten wurde mit zur Populationsgröße (Anzahl der Haushalte pro Gemeinde) proportionalen Wahrscheinlichkeiten vorgenommen.

Die zweite Ziehungsstufe

Die zweite Ziehungsstufe betrifft lediglich die in der ersten Stufe gezogenen Großstädte mit 100.000 und mehr Einwohnern. Hier wurde auf eine Untergliederung in vermögende und weniger vermögende Städte verzichtet, da diese als nicht trennscharf genug angesehen wurde.

Das Ziehungsdesign für die Großstädte soll dem Umstand Rechnung tragen, dass sich die Bevölkerung in Großstädten durch eine stärkere Heterogenität auszeichnet als dies in kleineren Gemeinden der Fall ist. Die Schichtung erfolgt daher auf Straßenabschnittsebene[4]. Die für die Schichtung notwendigen kleinräumigen Informationen zu Wohnquartieren und Straßenabschnitten wurden

3 Der Begriff „Sample Point" beschreibt hier keine regionalen Cluster, sonder steht lediglich für eine Zahl von 360 Adressen, die aus den Einwohnermeldeamtsregistern der Gemeinden gezogen wurden.

4 Die Straßenabschnitte enthalten im Mittel etwa 45 Haushalte, wobei die Heterogenität sehr hoch ist.

der Datenbank „infas geodaten“ entnommen. Es handelt sich dabei um drei nach qualitativen Gesichtspunkten ausgewählte mikrogeographische Indikatoren[5].

Da Immobilienbesitz eine große Rolle für das Vermögen eines Haushalts besitzt, wurden zwei Indikatoren gewählt, die die Qualität der Gebäude in der Wohngegend und den Gebäudetyp abbilden. Zusätzlich wurde noch der Kaufkraftindex für die Stratifizierung der Straßenabschnitte verwendet. Ein Straßenabschnitt wurde als vermögend eingestuft, wenn entweder 75% oder mehr der Gebäude in der Wohngegend mindestens von zufriedenstellender Qualität sind oder in 75% oder mehr Fällen der Gebäudetyp als hoch, gediegen oder ausgezeichnet eingestuft werden kann. Zusätzlich muss gelten, dass der Kaufkraftindex für den Abschnitt um mindestens 20 Prozentpunkte über dem durchschnittlichen Kaufkraftindex liegt.

Ein technisches Problem, das sich bei diesem Verfahren ergibt, ist die Tatsache, dass Einwohnermeldeämter typischerweise keine Adressen für Straßenabschnitte, sondern nur für ganze Straßen an die Befragungsinstitute liefern. Bei diesem Vorgehen besteht grundsätzlich die Möglichkeit, dass eine Straße doppelt gezogen wird; nämlich dann, wenn sie mindestens einen „vermögenden“ und einen weniger vermögenden Straßenabschnitt enthält. Die Straßen wurden daher eindeutig einer Schicht zugeordnet, indem sie auf Basis des bevölkerungsreichsten Abschnitts gruppiert wurden.

Das Oversampling wurde hier durch eine Erhöhung des Auswahlsatzes für die vermögenden Straßenabschnitte erreicht. Dabei wurden - unabhängig von der Verteilung der Straßenabschnitte in der jeweiligen Großstadt – jeweils 43% der Adressen aus vermögenden Straßenabschnitten und 57% aus weniger vermögenden Straßenabschnitten gezogen. Ein Sample Point (360 Adressen) setzt sich also in den Großstädten immer aus rund 155 Adressen aus vermögenden Straßenabschnitten und etwa 205 Adressen aus anderen Straßenabschnitten zusammen.

Die Ziehung innerhalb der Schichten in den Großstädten wurde mit zur Populationsgröße (Anzahl der Haushalte pro Straßenabschnitt) proportionalen Wahrscheinlichkeiten vorgenommen. Die Ziehung von Straßenabschnitten aus der jeweiligen Schicht wurde so lange fortgesetzt, bis die Anzahl der Haushalte den oben angegebenen Mengen (155 bzw. 205) entsprach. Die Information über die Anzahl der Haushalte pro Straßenabschnitt wurde der infas geodaten Datenbank entnommen. Im Ergebnis wurden 513 vermögende Straßenabschnitte und 477 sonstige Straßenabschnitte gezogen.

5 Grundsätzlich sind diese Informationen auch für mittelgroße Gemeinden verfügbar, sodass auch für diese eine Schichtung auf Straßenabschnittsebene prinzipiell möglich wäre. Davon wurde aber unter anderem auch deshalb abgesehen, da die kleineren Gemeinden (nach Auskunft des Befragungsinstituts infas) nur selten in der Lage sind, Adressen aus vorgegebenen Straßen zu ziehen.

Die dritte Ziehungsstufe

In der dritten Ziehungsstufe wurden schließlich aus Einwohnermeldeamtsregistern Adressen innerhalb der in Stufe 1 und 2 identifizierten Gemeinden bzw. Straßen(abschnitten) gezogen. Da das Gros der Fragen in der PHF-Studie auf Haushalte abzielt und nicht auf Personen, wäre eine direkte Ziehung von Haushalten wünschenswert gewesen. Die Struktur der Einwohnermeldeamtsregister erlaubt es allerdings nur, Adressen von Personen zu ziehen. Um die Wahrscheinlichkeit zu verringern, dass zwei Personen aus dem selben Haushalt gezogen werden, wurde für die PHF-Studie ein Ansatz gewählt, bei dem aus einer sortierten Liste mit festen Schrittweiten Adressen von erwachsenen Personen gezogen werden. Die Einwohnermeldeämter wurden gebeten, die Register nach dem Nachnamen der Personen ab 18 Jahren zu sortieren und dann nach einem vorgegeben Schema jede x-te Adresse aufzulisten. Da pro kleiner und mittlerer Gemeinde nur 360 Adressen zu ziehen waren, konnte die Schrittweite entsprechend groß gewählt werden. Auch wenn dies noch nicht garantiert, dass kein Haushalt doppelt gezogen wird, ist die Wahrscheinlichkeit vernachlässigbar gering. In den Großstädten wurden, wie oben bereits skizziert, jeweils alle Adressen einer Straße angeliefert. Für diese wurden dann die (eigentlich) zu ziehenden Straßenabschnitte identifiziert und entsprechend den Schichten zugeordnet. Im Feld kamen nicht alle gezogenen Adressen zum Einsatz. Die erste Tranche bestand zum Beispiel nur aus 36 Bruttoadressen pro Sample Point, die zufällig aus den vorhandenen 360 Adressen pro Sample Point gezogen worden sind.

Evaluation des Oversamplings

Die gewählte Oversampling-Strategie hat Auswirkungen auf die Verteilung der Haushalte über die Dezile der Vermögensverteilung und die Designgewichte.

Einen ersten Eindruck, ob die gezwungenermaßen auf Hilfsvariablen (Vermögen ist ex-ante nicht beobachtbar) beruhende Stratifizierung zielführend war, kann ein Vergleich der im PHF gemessenen Vermögen innerhalb der jeweiligen Schichten vermitteln. Wie Tabelle 2 zeigt, verfügen die Haushalte in den als „vermögend" klassifizierten Schichten über ein deutlich größeres Nettovermögen als die Haushalte in den sonstigen Schichten. Die höchsten Werte werden für die vermögenden Straßenabschnitte in den Großstädten erreicht. Die Differenzierung der Straßenabschnitte anhand der mikrogeographischen Informationen scheint also funktioniert zu haben.

	Schicht 1: sonstige kleine Gemeinden	Schicht 2: vermögende kleine Gemeinden	Schicht 3: sonstige Straßenab-schnitte	Schicht 4: vermögende Straßenab-schnitte
Nettovermögen - Mittelwert	251.880 (203.676)	405.746 (298.918)	151.868 (109.325)	782.212 (540.045)
Nettovermö-gen- Median	101.320 (69.284)	236.048 (153.166)	27.220 (17.582)	351.630 (310.940)

Tabelle 2 Nettovermögen der Haushalte nach Schichten in Euro
Hinweise: Gewichtete Werte in Klammern
Quelle: Eigene Berechnungen – PHF Datenstand 2/2013

Die Verteilung der realisierten Nettostichprobe zeigt, dass das Ziel einer Überrepräsentation vermögender Haushalte erreicht werden konnte. So entfallen auf das oberste Quintil der (aus den Daten geschätzten) Vermögensverteilung 37 % der realisierten Nettostichprobe. Im obersten Dezil befinden sich 22 % der Haushalte in der Stichprobe, wie Abbildung 1 zeigt. Auch die unteren sieben Dezile sind mit etwa jeweils rund 8 % der Nettostichprobe noch ausreichend besetzt.

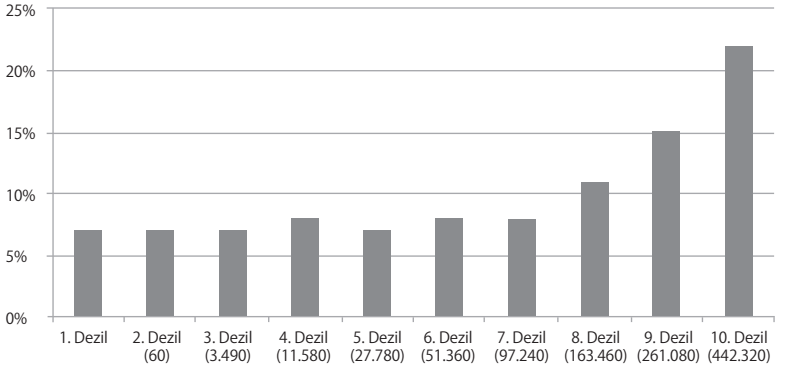


Abbildung 1: Anteil der Haushalte in der Nettostichprobe nach der (aus den Daten geschätzten) Nettovermögensverteilung.
Hinweise: Dezilgrenzen in Euro in Klammern
Quelle: Eigene Berechnungen – PHF Datenstand 2/2013

Die Überrepräsentation der oberen Enden der Verteilung hat natürlich auch Kosten im Sinne eines Effizienzverlusts gegenüber einem Ansatz mit Simple-Random-Sampling. Dies lässt sich am einfachsten an der effektiven Fallzahl erkennen. Während die Bruttostichprobe insgesamt 20.501 Adressen umfasst, ergibt

sich nach Anwendung der Designgewichte nur noch eine effektive Fallzahl von 12.820 Adressen, d.h. es verbleiben 62,5%. Der Effekt für die Nettostichprobe ist vergleichbar, effektiv ergibt sich hier eine Fallzahl von 2.012 Haushalten (bei 3.565 realisierten Haushalten). Auch der Abstand zwischen dem größten und kleinsten Hochrechungsfaktor ist sicherlich ausgeprägter als bei anderen Ziehungsdesigns. Die Varianz der Gewichte ist, wie bei Oversampling-Ansätzen häufig, vergleichsweise hoch. Der feste Auswahlsatz von vermögenden Straßenabschnitten (unabhängig von der Verteilung der Straßenabschnitte in der jeweiligen Stadt) hat die Varianz der Gewichte im PHF zusätzlich erhöht. Für die Designgewichte auf Haushaltsebene ergibt sich eine Standardabweichung von 0,88 bei einem Mittelwert von 1. Das minimale Gewicht liegt bei 0,00045; das maximale bei 3,26.

Fazit

Zusammenfassend lässt sich feststellen, dass das für die PHF-Studie angewandte integrierte Stichprobenkonzept funktioniert hat. Das Ziel, vermögende Haushalte überzurepräsentieren, konnte realisiert werden. Die mikrogeographischen Informationen sowie die Einkommenssteuerstatistik lieferten ausreichend Informationen, um ex-ante die Gemeinden bzw. Straßenabschnitte nach „Vermögen“ zu schichten.

Die nächste Welle der PHF-Studie ist für 2014 geplant. Dann soll zusätzlich zur Panel- eine gesonderte Auffrischungsstichprobe gezogen werden. Auch für letztere ist wieder ein Oversampling vermögender Haushalte vorgesehen. Inwieweit dabei ein vergleichbares Ziehungsdesign zum Einsatz kommen kann, wird sich zeigen. Eine Verbesserung gegenüber der in diesem Aufsatz skizzierten Methode ist bereits angedacht: Das Oversampling von Straßenabschnitten soll nicht mehr nach einem festen Auswahlsatz stattfinden, sondern nach einem Auswahlsatz, der sich an der Verteilung der Straßenabschnitte in der jeweiligen Großstadt orientiert.

Methodische Herausforderungen bei der Erfassung von Bildung und Ausbildung im Mikrozensus

Dr. Bernhard Schimpl-Neimanns

Zusammenfassung

In Bezug auf Bildung und Ausbildung werden in Deutschland durch die Bildungsberichte und auf europäischer Ebene mit der Europa 2020-Strategie der Europäischen Kommission und den Schwerpunktsetzungen des Europäischen Statistischen Programms 2013-2017 einige Anforderungen an das Erhebungsprogramm des Mikrozensus gestellt.
Frühere Vergleiche der Ergebnisse des Mikrozensus und der Bildungsstatistik zum Schulbesuch wiesen auf einige Abweichungen hin. Die Fragen zum Schulbesuch wurden im Mikrozensus 2008 und 2011 geändert. Um die Wirkungen dieser Änderungen zu untersuchen, werden Ergebnisse des Mikrozensus und der Bildungsstatistik für die Zeitpunkte bzw. Schuljahre 2007/2008, 2008/2009 und 2011/2012 verglichen. Für die Klärung von Fragen zur Datenqualität allgemeinbildender und beruflicher Ausbildungsabschlüsse werden Daten des Mikrozensus-Panels 2001-2004 genutzt. Die wiederholten Befragungen ermöglichen die Beurteilung der Antwortkonsistenz im Zeitverlauf. In Kombination mit der jeweiligen Erhebungsart und der Art der Beteiligung können – allerdings eingeschränkt – außerdem Mode- und Proxy-Effekte untersucht werden.
Beim Schulbesuch sind die Differenzen zur Bildungsstatistik teilweise zurückgegangen, insgesamt gibt es aber noch beträchtliche Abweichungen. Diese hängen mit schwierigen, dem Alltagsverständnis nicht direkt entsprechenden Definitionen im Mikrozensus zusammen. Insbesondere berufliche Abschlüsse weisen hohe Antwortinkonsistenzen auf. Insgesamt sind bei schriftlichen Befragungen die Antwortinkonsistenzen höher als bei persönlichen Interviews.

Schlagwörter: Antwortfehler, Datenqualität, Mixed-Mode-Befragungen, Proxy-Interviews, Panel, Mikrozensus, Bildungsstatistik.

Methodological challenges in data collection regarding education and training in the German Microcensus

The German national reports on education, the Europe 2020 strategy of the European Commission as well as the priorities of the European Statistical Programme 2013-2017 put high demands on the German Microcensus regarding the information on individual education and training.
Earlier comparisons of the Microcensus data on school attendance with data from Education Statistics pointed to certain deviations. The questions on school attendance were modified in the 2008 and 2011 Microcensus. To investigate the effects of these changes results of the Microcensus and the Education Statistics for the school years 2007/08, 2008/09 and 2011/12 are compared. The Microcensus Panel 2001-2004 is used to examine data quality issues with regard to general and vocational qualifications. The rotating panel survey allows to assess item stability over time. In combination with the type of interview (interviewer-/self-administered) and the type of reporting (self/proxy) - however limited - mode and proxy effects can be assessed.
In terms of school attendance previous deviations to Education Statistics partially decreased, but overall, there remain substantial differences. Results suggest that official definitions used in the Microcensus questionnaire do not fully match respondent's understanding of the school types and qualification titles. Especially vocational qualifications show high item inconsistencies. Overall, the item stability is lower in self-administered than in interviewer-administered interviews.

Keywords: measurement error; total survey error; mixed modes; self/proxy response; panel; German Microcensus; Education Statistics.
JEL: I21, C81, C83

Einleitung[1]

Bildungsangaben sind in vielen Umfragen als eigenständiges Merkmal oder als Hintergrundmerkmal enthalten (Müller 1979). Aufgrund der Stichprobengröße, die auch Auswertungen für kleine Populationen ermöglicht, zählt der Mikrozensus für die empirische Sozial- und Wirtschaftsforschung zu den wichtigsten amtlichen Datenquellen für bildungsstatistische Analysen. Im Vergleich zur Bildungsstatistik, die als Anstaltserhebung durch Verwaltungshandeln gesammelte Informationen über die Population, wie z. B. die Zahl der Schüler in bestimmten Schularten, aggregiert in Form von Tabellen bereitstellt, lassen sich eine Reihe von Fragestellungen nur mit den Individualdaten des Mikrozensus bearbeiten. Dies betrifft insbesondere alle Sachverhalte, bei denen es um den Haushalts- und Familienkontext geht. Im Mikrozensus stehen Informationen über die sozioökonomische Lage der Eltern oder den Migrationsstatus zur Verfügung, die in der Bildungsstatistik fehlen. Des Weiteren können aufgrund des breit gefächerten Fragenprogramms mit dem Mikrozensus die Verwertung von Bildungsqualifikationen auf dem Arbeitsmarkt und Zusammenhänge zwischen Bildungsstatus und Lebenslage oder Gesundheit untersucht werden. Darüber hinaus dienen Verteilungen des Mikrozensus aufgrund des Stichprobenumfangs und der Auskunftspflicht für viele andere Stichproben in und außerhalb der amtlichen Statistik als Hochrechnungsrahmen oder werden für Prüfungen selektiver Befragungsausfälle herangezogen, die in der Regel eng mit der Bildungsqualifikation verbunden sind.

Weitere und neue Anforderungen an den Mikrozensus stellen sich durch die Integration von Mikrozensusergebnissen in die seit 2006 herausgegebenen Bildungsberichte (Autorengruppe Bildungsberichterstattung 2012, Konsortium Bildungsberichterstattung 2006). Hierfür ist eine hohe Vergleichbarkeit zur Bildungsstatistik notwendig. Auf europäischer Ebene sind mit der Europa 2020-Strategie der Europäischen Kommission (2010) und den entsprechenden Schwerpunktsetzungen des Europäischen Statistischen Programms 2013-2017 (Verordnung (EU) Nr. 99/2013) weitere Zielsetzungen definiert. Für den Einsatz der Daten im Rahmen der Politikevaluation stellen sich gleichfalls hohe Anforderungen an die Datenqualität.

Von 1991 bis 2007 waren Analysen der Bildungsbeteiligung mit dem Mikrozensus stark eingeschränkt, da zum Besuch allgemeinbildender Schulen statt Schularten nur noch Klassenstufen erfragt wurden. Des Weiteren zeigten sich

1 Erweiterte Fassung des Vortrags im Workshop „Methoden der empirischen Sozialforschung in amtlichen Haushaltsstatistiken", 16. Mai 2013, Wiesbaden. Für konstruktive Anmerkungen zu einer früheren Version dieses Aufsatzes danke ich Heike Wirth, Karin Schuller und Robert Herter-Eschweiler. Ganz besonders danke ich Rolf Porst für ergänzende Hinweise.

im Vergleich mit der Bildungsstatistik gravierende Klassifikationsfehler, z. B. waren im Mikrozensus 1996 Schüler der gymnasialen Oberstufe um rund 60 % übererfasst (Schimpl-Neimanns 2006). Erfreulicherweise werden seit 2008 die Schularten wieder erfragt und berufliche Schulen ausführlicher als zuvor erfasst. Weitere Modifikationen des Fragebogens erfolgten 2011. In der Bildungsstatistik liegen Angaben zum Schulbesuch auf Basis von Populationsdaten vor, die den Zeitraum vor und ab den Änderungen der im Mikrozensus verwendeten Fragebögen umfassen. Somit können im Vergleich mit Ergebnissen des Mikrozensus die Gesamtabweichungen (Total Survey Error) ermittelt werden und es lässt sich untersuchen, ob die Fragebogenänderungen zur Reduktion früherer Abweichungen zwischen Mikrozensus und Bildungsstatistik geführt haben.

Mess- oder Klassifikationsfehler treten nicht nur beim Schulbesuch, sondern auch beim allgemeinbildenden und beruflichen Abschluss im Mikrozensus sowie auch in anderen Umfragen auf. Auf europäischer Ebene wird deshalb teilweise überlegt, im EU Labour Force Survey Bildungsangaben aus Registern zu übernehmen (Hewat 2011; Thomson und Knight 2010). In Deutschland liegen solche Referenzdaten nicht vor, jedoch können Fragen zur Datenqualität der Angaben zum Ausbildungsabschluss auf Basis des Mikrozensus-Panels untersucht werden. Die wiederholte Befragung ermöglicht die Beurteilung der Antwortkonsistenz bzw. Antwortstabilität. In Kombination mit der jeweiligen Erhebungsart (Interview vs. schriftliche Befragung/Selbstausfüllerbogen; Selbstauskunft vs. Proxy-Interview) können zudem Hinweise zu Mode- und Proxy-Effekten ermittelt werden.

Im Folgenden werden Theorien und Befunde der kognitiven Umfrageforschung skizziert. Anschließend werden die verwendeten Daten und Methoden beschrieben. Der 4. Abschnitt enthält Vergleiche zwischen Mikrozensus und Bildungsstatistik zum Schulbesuch. Deskriptive Analysen zu den Antwortinkonsistenzen bei Bildungsabschlüssen sowie darauf basierende Differenzierungen zu Mode- und Proxy-Effekten sind im 5. Abschnitt zu finden. Der Aufsatz schließt mit einer Zusammenfassung der Ergebnisse und Überlegungen zu den Möglichkeiten der weiteren Verbesserung der Fragebögen und der Feldarbeit.

Forschung zu kognitiven Prozessen bei Umfragen

Die angesprochenen Fragen betreffen nicht-stichprobenbedingte, systematische Fehler, zu denen neben Antwortausfällen insbesondere Antwort-, aber auch Interviewer- und Codierfehler zählen. Für die Untersuchungsziele dieses Aufsatzes sind primär Theorien der kognitiven Umfrageforschung relevant, die erklären, wie Antwortfehler entstehen.

Das kognitive Modell (Groves et al. 2009, Tourangeau/Rips/Rasinski 2000, Strack/Martin 1987) unterscheidet idealtypisch vier Phasen der Beantwortung einer Frage im Fragebogen: Frage- und Begriffsverständnis, Erinnerung relevanter Informationen, (interne) Urteilsbildung und abschließende Antwortformulierung mit Zuordnung zu den vorliegenden Antwortkategorien. In jeder dieser Phasen können Fehler entstehen.

Das Verstehen einer Frage ist für eine sinnvolle Antwort eine elementare Voraussetzung. Befragte können aber beispielsweise eine Frage aufgrund unbekannter Begriffe nicht immer ausreichend verstehen oder korrekt interpretieren. Falls die Frage schwer verständlich ist, versuchen Befragte, den Sinn aus den Antwortkategorien des Fragebogens zu erschließen.

Eingangs wurde die frühere Übererfassung von Schülern der gymnasialen Oberstufe erwähnt, die mit einer Untererfassung von Schülern beruflicher Schulen, die zur Fachhochschulreife oder allgemeinen Hochschulreife führen, einher geht. In Orientierung an das kognitive Modell ist zu vermuten, dass solche Fehlklassifikationen damit zusammenhängen, dass die im Fragebogen verwendeten Begriffe oftmals nicht dem Alltagsverständnis der Befragten entsprechen (Campanelli und Channell 1996). Dies dürfte bei der Unterscheidung zwischen allgemeinbildenden und beruflichen Schulen bzw. Abschlüssen der Fall sein, da einige berufliche Schulen sowohl allgemeinbildende Abschlüsse als auch berufliche Qualifikationen vermitteln (vgl. auch Zühlke 2008).

Schließlich ist in Bezug auf schwierige Begriffe darauf hinzuweisen, dass die Schularten in den Bundesländern unterschiedliche Namen haben und teilweise sogar verschiedene Namen innerhalb eines Landes. Des Weiteren ändern sich diese Begriffe manchmal im Zeitverlauf. Wenn die Vertrautheit mit dem Bildungssystem und den verschiedenen Bildungsabschlüssen wichtig für das Frageverständnis sind, ist zu erwarten, dass Ausländer größere Schwierigkeiten mit den Bildungsfragen haben als Deutsche. In diesem Zusammenhang kann erschwerend hinzukommen, dass es nur Fragebögen in deutscher Sprache gibt und keine Listen zur Vergleichbarkeit von deutschen mit im Ausland erworbenen Abschlüssen eingesetzt werden.

In Bezug auf die zweite Phase des kognitiven Modells, der Erinnerung relevanter Informationen, weisen Campanelli und Channell (1996: 13, 20) darauf hin, dass Befragte Schwierigkeiten haben, wenn der Abschluss längere Zeit zurückliegt oder für die gegenwärtige berufliche Position keine Rolle spielt. Bei Auskünften über den Status anderer Personen sind in der Regel größere Informations- und Erinnerungsprobleme zu erwarten, vor allem, wenn die Auskunft gebende Person nicht das gleiche Bildungssystem wie die zu befragende Person durchlaufen hat (Campanelli und Channell 1996: 65 f.).

Die Befunde zur vermuteten geringeren Datenqualität von Proxy-Interviews im Mikrozensus sind jedoch nicht eindeutig, verweisen auf eher schwache Effekte (Herter-Eschweiler 2009, Köhne-Finster und Lingnau 2009, Zühlke 2008) und sind methodisch nicht einfach zu untersuchen (Blair/Menon/Bickart 1991, Moore 1988, Schwarz und Wellens 1997).

In der vierten Phase der Antwortformulierung können Effekte sozialer Erwünschtheit auftreten. So kann es Befragten unangenehm sein, den Besuch einer Sonder- oder Förderschule oder einen fehlenden Bildungsabschluss anzugeben. In solchen Fällen werden die entsprechenden Antwortkategorien vermieden und andere, sozial erwünschtere Kategorien angegeben (Groves et al. 2009: 207 f.; Tourangeau/Rips/Rasinski 2000: 257 f.).

Krosnick (1991) weist darauf hin, dass nicht immer eine optimale, sondern eine aus Sicht der Befragten zufriedenstellende Antwort („satisficing") gegeben wird. Dies ist abhängig von den kognitiven Eigenschaften und der Motivation der Befragten. Demnach bestehen zwischen der Schwierigkeit von Fragen und den Eigenschaften der Befragten Interaktionsbeziehungen. In dieser Hinsicht lässt sich die oben angesprochene Über- bzw. Untererfassung von Schülern der gymnasialen Oberstufe bzw. beruflichen Gymnasien auch als sogenannter Satisficing-Effekt erklären, wenn man annimmt, dass Befragte die erste einigermaßen passende Antwortkategorie ankreuzen, da sie den Begriff „Gymnasium" enthält.

Die Fragen zum Schul- und Hochschulbesuch sowie zu den Bildungsabschlüssen werden im Mikrozensus erst im letzten Drittel des Fragebogens gestellt. Dies kann mit einer geringeren Aufmerksamkeit und Motivation der Befragten als bei den Eingangsfragen einhergehen und verstärkt zu Satisficing-Effekten führen.

Daten und Methoden

Bildungsstatistik

Die Bildungsstatistik gliedert sich in die Schul- und die Hochschulstatistik, wobei die Schulstatistik in die Statistik der allgemeinbildenden und beruflichen Schulen unterteilt ist (siehe die ausführlichen Datenbeschreibungen in den Qualitätsberichten in: Statistisches Bundesamt 2007a,b). Die Statistiken sind Vollerhebungen und beruhen auf Meldungen der Schulleiter bzw. basieren bei der Hochschulstatistik auf den Verwaltungsdaten der Hochschulen. Ausgenommen von der Auskunftspflicht sind nur Schulen des Gesundheitswesens. Deren Meldungen zum Schulbesuch sind unvollständig und liegen nicht für alle Bundesländer vor. Die Schulstatistik ist in den einzelnen Bundesländern unterschiedlich organisiert. Durch die Orientierung an einem Minimalprogramm können aber Bundesergebnisse zusammengestellt werden. Die Meldungen beziehen sich auf den

Schul- bzw. Hochschulort und umfassen im Fall der Statistik allgemeinbildender Schulen unter anderem die Zahl der Schüler nach Klassenstufe, Geburtsjahr, Geschlecht und Staatsangehörigkeit. Der Berichtszeitraum bezieht sich auf den Beginn eines Schuljahres bzw. des Wintersemesters. Datenbasis für die Vergleiche mit dem Mikrozensus sind die in den jeweiligen Fachserien veröffentlichten Ergebnisse.

Mikrozensus

Im Mikrozensus werden vielfältige Informationen über die demografische, soziale und wirtschaftliche Struktur der Bevölkerung erhoben (siehe die ausführliche Datenbeschreibung im Qualitätsbericht zum Mikrozensus 2011 in: Statistisches Bundesamt 2012a). Die mehrfach geschichtete einstufige Flächenstichprobe umfasst ein Prozent der Personen. Als Auswahlbezirke dienen künstlich abgegrenzte Flächen (Klumpen), die benachbarte Gebäude oder Gebäudeteile umfassen. Aufgrund der Auskunftspflicht liegt die Teilnahmequote der Haushalte bei rund 97 Prozent. Die Haushalte eines Auswahlbezirkes werden vier Jahre lang befragt, wobei jedes Jahr ein Viertel der Auswahlbezirke ausgetauscht wird. Die Rotationsgruppen können zu einem Panel zusammengeführt werden. Dabei ist jedoch zu beachten, dass aufgrund des Prinzips der Flächenstichprobe wegziehende Personen und Haushalte nicht weiter befragt, sondern durch die nachziehenden Personen bzw. Haushalte ersetzt werden. Bis 2004 wurden die Haushalte einmal jährlich befragt. Der Berichtszeitraum (Berichtswoche) bezog sich in der Regel auf die letzte feiertagsfreie Woche im April eines Jahres. Seit 2005 wird die Erhebung unterjährig durchgeführt. Die Berichtswoche ist nun die letzte Kalenderwoche vor der Befragung. Mit der Umstellung auf eine kontinuierliche Befragung sind weitere Modifikationen des Frageprogramms sowie methodische Änderungen der Hochrechnung verbunden.

In Bezug auf die Befragungsmethode dominiert die persönliche Befragung, jedoch steigt der Anteil von Haushalten, die schriftlich Auskunft geben (2001: 16 %; 2011: 21 %). Proxy-Interviews sind möglich, z. B. wenn Eltern die Fragen zum Schulbesuch und Bildungsabschluss ihrer Kinder beantworten. Seit 1999 wird im Rahmen der als Unterstichprobe in den Mikrozensus integrierten EU-Arbeitskräftestichprobe (EU Labour Force Survey) erfasst, ob es sich um Selbst- oder Fremdauskünfte handelt. Insgesamt beruhen 2011 25 % der Angaben von Personen ab 15 Jahren auf Proxy-Interviews. Die meisten Fragen sind auskunftspflichtig. Daneben gibt es Fragen und Themenbereiche, die von der Auskunftspflicht ausgenommen sind. Hierzu zählen Fragen der EU-Arbeitskräftestichprobe.

Für die Auswertungen zum Schul- und Hochschulbesuch werden Daten der Mikrozensus Scientific Use Files 2007 und 2008 verwendet (siehe dazu

die Datendokumentationen auf den Web-Seiten von GESIS und der Forschungsdatenzentren). Dabei handelt es sich um faktisch anonymisierte 70-Prozent-Substichproben des Mikrozensus. Das Ziehungsverfahren berücksichtigt die wesentlichen Designelemente des Mikrozensus und gewährleistet eine sehr hohe Übereinstimmung mit Ergebnissen der Originaldaten. Die Daten des Mikrozensus 2011 basieren auf einer Sonderauswertung des Statistischen Bundesamtes.

Mikrozensus-Panel 2001-2004

In Deutschland gibt es keine Register zu Bildungsabschlüssen, die für Untersuchungen zur Datenqualität verwendet werden könnten. Auch fehlen Test-Retest Studien, die für Vergleiche näherungsweise „wahre" Werte oder einen „gold standard" bieten könnten. Mit Panels ist es jedoch möglich, die Konsistenz der Angaben zu untersuchen.

Für die Analysen der Antwortkonsistenz der Bildungsabschlüsse wird das Mikrozensus-Panel 2001-2004 verwendet. Das Scientific Use File ist eine 70-%-Substichprobe der Auswahlbezirke, in denen im Zeitraum 2001 bis 2004 mindestens ein Interview durchgeführt wurde (siehe zur ausführlichen Datenbeschreibung: Statistisches Bundesamt 2009a). Es enthält Informationen zur Art der Datenerhebung (Interviewer-/Selbstausfüllerbogen) und zur Auskunftsart (Selbst-/Fremdauskunft). Allerdings wurde die Auskunftsart 1999 bis 2004 nur im Rahmen der EU-Arbeitskräftestichprobe ohne Auskunftspflicht erfragt. Bis 2004 wies die Unterstichprobe auf der Ebene der Regierungsbezirke variable Auswahlsätze zwischen 0,4 und 1 Prozent auf. In Verbindung der auf regionaler Ebene variablen Auswahlsätze und der Aufteilung der Stichprobe in Rotationsgruppen liegen beispielsweise für Regierungsbezirke mit einem Auswahlsatz von 0,4 % lediglich für 2 Jahre Daten vor. Deshalb und aus Datenschutzgründen enthält das Panel in den Jahren 2001 und 2004 keine Merkmale der EU-Arbeitskräftestichprobe (Statistisches Bundesamt 2009a: 28, 65).

Methoden

Vergleiche zwischen Mikrozensus und Bildungsstatistik zum Schulbesuch

Um zu klären, ob die Änderungen der Fragen zum Schulbesuch zur Reduktion der Abweichungen des Mikrozensus von den Populationsdaten der Bildungsstatistik geführt haben, sind methodische Anforderungen zu beachten. Für den Vergleich liegen mit den Ergebnissen der Bildungsstatistik zum Schuljahr 2007/2008 und dem Mikrozensus 2007 Daten für den Referenzzeitpunkt vor den Änderungen vor. Im Idealfall sollen die Referenzdaten den „wahren Werten" möglichst nahe

kommen. Davon kann bei der Bildungsstatistik ausgegangen werden. Wichtig für den Vergleich ist außerdem, dass die im Mikrozensus verwendeten Definitionen denen der Bildungsstatistik entsprechen. Dies trifft mit zwei Einschränkungen weitestgehend zu.

Die beiden Ausnahmen betreffen die Schulen des Gesundheitswesens und die Berufsfachschulen. Schulen des Gesundheitswesens werden in der Bildungsstatistik nicht vollständig erfasst und nicht nach 1- vs. 2- bis 3-jährigen Bildungsgängen wie im Mikrozensus ausgewiesen. Berufsfachschulen werden im Mikrozensus bis 2008 nach drei Kategorien differenziert: Schulen, die einen Berufsabschluss vermitteln, zu einem allgemeinbildenden mittleren Abschluss oder zur Fach-/Hochschulreife führen. Im Mikrozensus 2011 werden berufsvorbereitende Bildungsgänge an Berufsfachschulen, die nicht zum Berufsabschluss führen, dem Berufsvorbereitungsjahr gleich gestellt; allerdings wird dies nicht im schriftlichen Fragebogen, sondern nur im Interviewerhandbuch erläutert.

Um die Vergleichbarkeit zu verbessern, können die Schulbesuchsdaten der Bildungsstatistik mit Hilfe von Angaben aus der Datenbank der beruflichen Bildungsgänge (Fest et al. 2010; Statistische Ämter des Bundes und der Länder 2012) gemäß den Kategorien des Mikrozensus typisiert und aufgeteilt werden.[2] Die Datenbank enthält unter anderem Angaben zu den Mindestzugangsvoraussetzungen, den wichtigsten Abschlüssen und der Ausrichtung der Bildungsgänge. Zählt beispielsweise der Erwerb eines mittleren Bildungsabschlusses zu den wichtigsten Abschlüssen eines Bildungsganges und dessen Mindestzugangsvoraussetzung liegt unterhalb des Realschulabschlusses, wird dieser Bildungsgang äquivalent zur Mikrozensus-Kategorie „Berufliche Schule, die einen mittleren Abschluss vermittelt“ zugeordnet. Einschränkend ist anzumerken, dass diese Typisierung nur näherungsweise eine Verteilung der in den Fachserien der Bildungsstatistik berichteten Schüler von Berufsfachschulen und Schulen des Gesundheitswesens wie in der Mikrozensus-Umfrage erlaubt.

Der Berichtszeitraum der Bildungsstatistik zum Schul- und Hochschulbesuch bezieht sich auf den Schuljahresbeginn bzw. Semesteranfang. Für den Vergleich werden im Mikrozensus jedes 4. Quartal Personen am Ort der Hauptwohnung ausgewählt, die in den letzten 4 Wochen vor der Befragung eine Schule oder Hochschule besucht haben. Der Berichtszeitraum des 4. Quartals im Mikrozensus ist zwar nicht identisch mit dem Schuljahres- bzw. Semesterbeginn der Bildungsstatistik. Es kann jedoch angenommen werden, dass die in diesem Zeitraum durch Schulartwechsel oder Ausbildungsabbrüche entstehenden Unterschiede beider Datenquellen gering und für den Vergleich vernachlässigbar sind.

2 Für die Bereitstellung und Erläuterungen zur Datenbank sowie für Aktualisierungen zum Schuljahr 2011/2012 danke ich Herrn Hans-Werner Freitag vom Statistischen Bundesamt, Gruppe H 2.

Methodisch wird im Folgenden die Gesamtabweichung („Total Survey Error") zwischen den Ergebnissen des Mikrozensus (MZ) und den Populationsdaten der Bildungsstatistik (BS) betrachtet. Der Vergleich konzentriert sich auf die relative Differenz der Gesamtwerte beider Datenquellen:

$$\text{Relative Differenz (\%)} = \left(\hat{t}_{MZ} - t_{BS}\right) / t_{BS}$$

Einzelne Gründe für die Abweichungen können mit diesem Verfahren nicht ermittelt werden (siehe Groves und Lyberg (2010) zu Vor- und Nachteilen des Konzeptes „Total Survey Error"). Mit den Scientific Use Files 2007 und 2008 lassen sich aber der Stichprobenfehler und entsprechende 95 %-Konfidenzintervalle der Gesamtwerte schätzen (Schimpl-Neimanns 2011). Da mit der Auswahl eines Berichtsquartals der Stichprobenfehler der geschätzten Gesamtwerte des Mikrozensus steigt, ist es umso wichtiger, ihn beim Vergleich zu berücksichtigen. Schließt das Konfidenzintervall den Gesamtwert der Bildungsstatistik ein, sind die Differenzen zwischen Mikrozensus und Bildungsstatistik durch den Stichprobenfehler zu erklären. Wie die Tabellen im Anhang zeigen, trifft dies bei den meisten Differenzen nicht zu, sodass als Ursache systematische Fehler anzunehmen sind. Für die Angaben zum Schulbesuch im Mikrozensus 2011 liegen keine Schätzungen des Stichprobenfehlers vor. Da die Beschreibung der Abweichungen im vorliegenden Aufsatz im Vordergrund steht, wird auf den Stichprobenfehler nicht weiter eingegangen.

Analysen zu Antwortinkonsistenzen

Vorweg ist anzumerken, dass sich aus der zeitlichen Konstanz von Angaben keine Aussage über deren Validität ableiten lässt. Dennoch wird man im Allgemeinen zeitlich konstanten Angaben mehr Vertrauen schenken. Geringe Antwortstabilitäten weisen auf Verständnisprobleme der Fragen und/oder Antwortkategorien hin. Muster von Übergangswahrscheinlichkeiten deuten darauf hin, dass den Befragten bestimmte Antwortkategorien ähnlich erscheinen. Der Einfachheit halber werden die Übergangswahrscheinlichkeiten auf Basis zusammengefasster Übergangsmatrizen unter der Annahme berechnet, dass die zeitliche Reihenfolge der Antworten keine Rolle spielt.

Bedingt durch Weg- und Zuzüge im Auswahlbezirk liegen im Mikrozensus-Panel nicht bei allen Personen für den 4-Jahreszeitraum Angaben vor. Um besondere Ausfallmodelle zu vermeiden, werden nur räumlich immobile Personen ausgewählt. Außerdem wird auf eine Hochrechnung verzichtet. Bis 2004 waren die Angaben zum Bildungsabschluss nur für Personen bis zum Alter von 50 Jahren auskunftspflichtig. Es werden deshalb nur Personen ausgewählt, die 2001 24 bis 47 Jahre alt waren. Schließlich sollen Veränderungen der Bildungsqualifikationen durch zwischenzeitlich erworbene neue Abschlüsse ausgeschlossen werden. Deshalb erfolgt eine Selektion von Personen, die im Zeitraum 2001 bis 2004 keine Schule oder Hochschule besucht haben.

In Bezug auf die Differenzierung der Antwortinkonsistenz nach Mode- und Proxy-Effekten kann zwischen persönlichen Interviews und schriftlicher Befragung unterschieden werden. Die Angaben zur Auskunftsart liegen jedoch nur für 2002 und 2003 vor. Mode- und Proxy-Effekte lassen sich deshalb nicht systematisch trennen. Die Analysen zur Antwortinkonsistenz in Abhängigkeit von der Erhebungsart (Interview vs. Selbstausfüller) werden für die Jahre 2001-2004 durchgeführt und enthalten auch unbekannte Proxy-Effekte. Um die Heterogenität bei der Analyse von Proxy-Effekten ansatzweise zu reduzieren, konzentrieren sich die Auswertungen auf Antwortinkonsistenzen persönlicher Interviews in den Jahren 2001-2003, wobei in der ersten Befragung die Auskunftsart nicht bekannt ist.

Schulbesuch im Mikrozensus im Vergleich zur Bildungsstatistik

Allgemeinbildende Schulen

Bis 2007 wurden Schüler allgemeinbildender Schulen im Mikrozensus nur nach Klassenstufen erfasst. Ab 2008 werden wieder die Schularten erfragt. Die Frage nach der Klassenstufe wird im Anschluss daran nur gestellt, wenn eine allgemeinbildende Schulart angegeben wurde. Förderschulen sind in der Bildungsstatistik keinem Bereich zugeordnet, sie können erst im Mikrozensus ab 2008 wie in der Bildungsstatistik getrennt ausgewiesen werden. Für das Schuljahr 2007/2008 wurden Verteilungen der Förderschüler unter Verwendung der teilweise vorliegenden Angaben zur Klassenstufe geschätzt. Diese Einschränkung ist bei dem Vergleich zu beachten.

Abbildung 1 zeigt für 2007 insgesamt eine relative Differenz des Mikrozensus (9.520.000 Schüler) im Vergleich zur Bildungsstatistik (9.156.000 Schüler) von 4 % (= (9.520 – 9.156)/9.156). Diese ist 2008 (1 %) zurückgegangen und 2011 (rd. 3 %) wieder leicht angestiegen. Als Grund für die Übererfassung im Mikrozensus kommt infrage, dass hier teilweise auch Schüler beruflicher Schulen erfasst sind (vgl. Tab. A2 und Tab. A3 im Anhang). Die Unterschiede zwischen Mikrozensus und Bildungsstatistik können auch dadurch entstehen, dass die Mikrozensusdaten im Rahmen der Hochrechnung an Ergebnisse der laufenden Bevölkerungsfortschreibung angepasst werden, die allerdings mit zunehmendem Abstand zur Volkszählung 1987 den Bevölkerungsbestand insgesamt überschätzt.

Viel wichtiger als der Vergleich der Schüler allgemeinbildender Schulen insgesamt sind die relativen Differenzen einzelner Bereiche. So zeigt Abbildung 1 für 2007 die aus früheren Untersuchungen bekannte Übererfassung von Schülern der gymnasialen Oberstufe im Mikrozensus, die mit der Untererfassung

beruflicher Gymnasien korrespondiert (siehe Abb. 4). Gemäß der Bildungsstatistik besuchten im Schuljahr 2007/2008 898.900 Schüler die gymnasiale Oberstufe, der Mikrozensus 2007 weist dagegen hochgerechnet 1.434.000 Schüler aus. Die Differenz von über 535.000 Schülern entspricht einer Übererfassung von rund 60 Prozent. Wie im zweiten Kapitel skizziert, ist zu vermuten, dass die Unterscheidung zwischen allgemeinbildenden und beruflichen Schulen in der Befragungspraxis schwierig ist. Die Fehlklassifikation kann auch auf Satisficing-Effekte zurückzuführen sein, wenn Besucher eines beruflichen Gymnasiums (bzw. deren Eltern) die erste passende Antwortkategorie ankreuzen, die den Begriff „Gymnasium" enthält. Diese Übererfassung ist ab 2008 deutlich zurückgegangen.

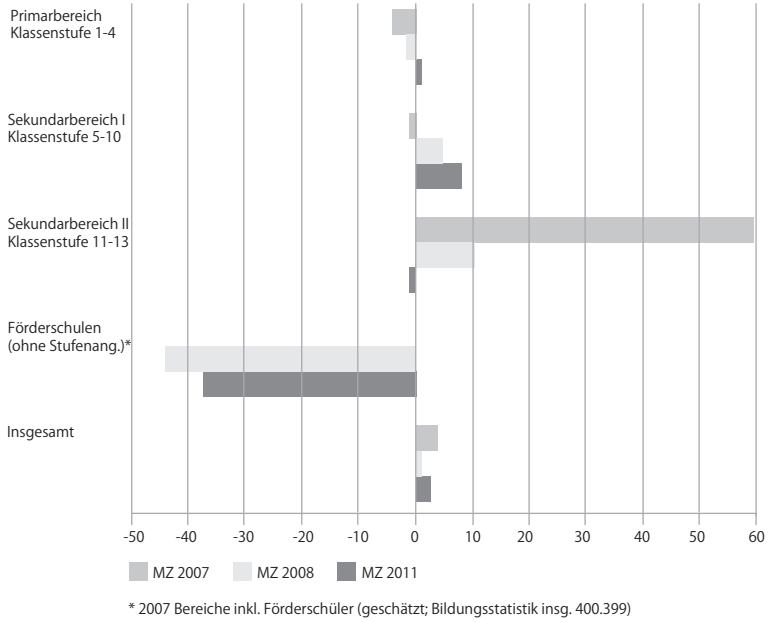


Abbildung 1: Schüler allgemeinbildender Schulen nach Klassenstufen - Relative Differenzen Mikrozensus vs. Bildungsstatistik (in %)

Die starke Untererfassung der Förder- bzw. Sonderschüler im Mikrozensus dürfte mit Effekten sozialer Erwünschtheit zusammenhängen, die häufiger bei persönlichen Interviews als bei schriftlichen Befragungen auftreten. Diese Vermutung wird durch ein Methodenexperiment der Statistischen Ämter der Länder und des Bundes (2010: 24) gestützt, das zeigte, dass die Schulart Förder- bzw.

Sonderschule bei persönlichen Interviews wesentlich seltener genannt wird als bei schriftlichen Befragungen. Nach weiteren Erfahrungen beim Mikrozensus 2008 wurde 2011 die Reihenfolge der Antwortkategorien verändert und stärker in Anlehnung an die Bildungsstatistik differenziert. Die Förderschule wurde von der 8. auf die 3. Stelle verschoben, die Orientierungsstufe von der 4. zur 2., und die Schularten mit mehreren Bildungsgängen von der 5. zur 4. Stelle. Der Besuch der Gesamt- oder Waldorfschule sowie des Gymnasiums oder Abendgymnasiums/Kollegs wurden 2011 in eigenen Kategorien erfragt. Außerdem wurden berufliche Gymnasien 2011 in einer eigenen Kategorie unter der Überschrift „Allgemeinbildende Schulen" genannt. Aus Vergleichbarkeitsgründen werden in Abbildung 2 die zusammengefassten Antwortkategorien des Mikrozensus 2008 verwendet.

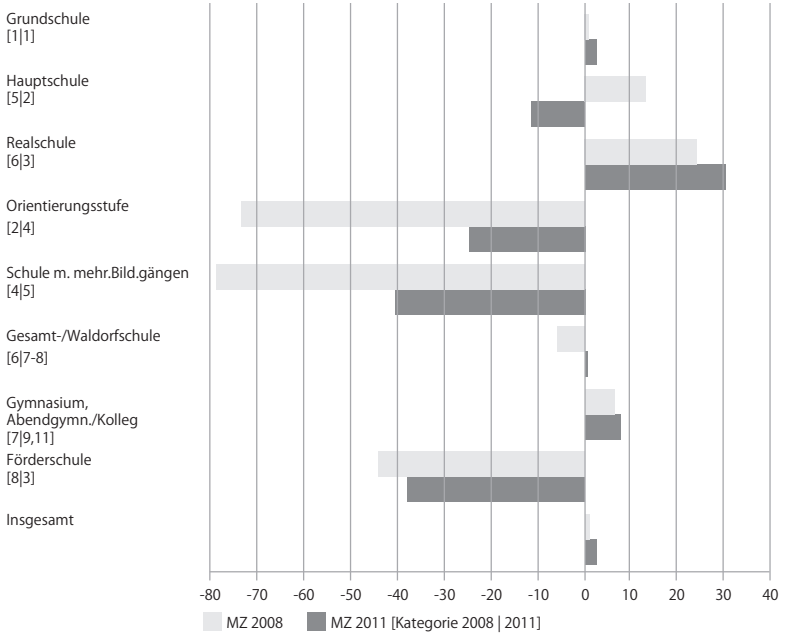


Abbildung 2: Schüler allgemeinbildender Schulen nach Schulart - Relative Differenzen Mikrozensus vs. Bildungsstatistik (in %)

Abbildung 2 dokumentiert, dass sich nach der Veränderung der Reihenfolge der Antwortkategorien die Untererfassung von Förderschülern nur geringfügig von -44 % (2008) auf -37 % (2011) reduziert hat. Dagegen wurde die Untererfassung

von Schülern der Orientierungsstufe (2008: -73 %) um 49 Prozentpunkte und die der Schularten mit mehreren Bildungsgängen (2008: -79 %) um 39 Prozentpunkte reduziert. Diese Veränderungen können sowohl auf eine Reduktion von Satisficing-Effekten zurückzuführen sein als auch damit zusammenhängen, dass im schriftlichen Fragebogen des Mikrozensus 2011 weitere Bezeichnungen der Schularten mit mehreren Bildungsgängen genannt sind. Nach wie vor sind die Differenzen zur Bildungsstatistik gravierend. Aus der Übererfassung von Hauptschülern (2008: 13 %; 2. Antwortkategorie) ist, quasi im Gegenzug, eine Untererfassung (2011: -11 %; 5. Antwortkategorie) geworden. Schüler von Realschulen und Gymnasien sind im Mikrozensus im Vergleich zur Bildungsstatistik übererfasst, wobei die Entwicklung der Differenzen sogar eine steigende Tendenz andeutet. Die Abweichungen bei den Schularten mit mehreren Bildungsgängen werden mit länderspezifischen Analysen auf Basis des Mikrozensus 2008 etwas genauer betrachtet.

Gemäß Bildungsstatistik besuchten im Schuljahr 2008/2009 in Deutschland insgesamt 302.000 Schüler die Schulart mit mehreren Bildungsgängen. Mit rund 228.000 Schülern ist sie in den neuen Bundesländern sehr stark verbreitet. Die Schulart mit mehreren Bildungsgängen führt zum Haupt- oder Realschulabschluss und heißt dort Oberschule, Regionale Schule, Mittelschule, Sekundarschule oder Regelschule. Schüler in der Schulart mit mehreren Bildungsgängen, der Realschule und der Hauptschule sind in Abbildung 3 auf der rechten Seite farblich hervorgehoben. Auf der linken Seite der Grafik sind der Vollständigkeit halber die Zahlen der Schulbesucher anderer allgemeinbildender Schulen abgetragen.

Die Bildungsstatistik weist 31 % (= 228.000/732.000) der 9- bis 18-jährigen Schüler allgemeinbildender Schulen als Besucher der Schulart mit mehreren Bildungsgängen aus. Dagegen besuchen laut Mikrozensus 2008 lediglich 2 % (= 19.000/760.000) der Schüler diese Schulart, aber 27 % eine Haupt- oder Realschule, die es in den neuen Bundesländern bis auf wenige Ausnahmen nicht gibt. Es ist naheliegend zu vermuten, dass sich die Antworten der Befragten an den angestrebten Schulabschlüssen orientierten. Vermutlich treten in Westdeutschland ähnliche Probleme auf. Das lässt sich jedoch nicht sicher prüfen, da es dort Haupt- und Realschulen gibt. Die Untererfassung der Schularten mit mehreren Bildungsgängen kann somit als ein Grund für die oben genannte Übererfassung der Realschüler im Mikrozensus in Erwägung kommen (siehe Abb. 2).

Frühere Analysen weisen darauf hin, dass die Übererfassung der Schüler von Gymnasien mit der Untererfassung bei den beruflichen Schulen, die zur Fachhochschulreife und Hochschulreife führen, (s. u.) verbunden ist. Fasst man die Gesamtwerte beider Schularten zusammen, reduzieren sich die Übererfassungen der Schüler von Gymnasien (siehe Abb. 2; 2008: 6 %; 2011: 8 %) bzw.

Untererfassungen von Schülern beruflicher Gymnasien usw. (siehe Abb. 4; 2008: -30 %; 2011: -32 %) gegenüber den Ergebnissen der Bildungsstatistik auf 1 bis 2 %. Diese Zusammenhänge finden sich jedoch im Mikrozensus 2008 nicht in allen Ländern (z. B. nicht in Nordrhein-Westfalen).

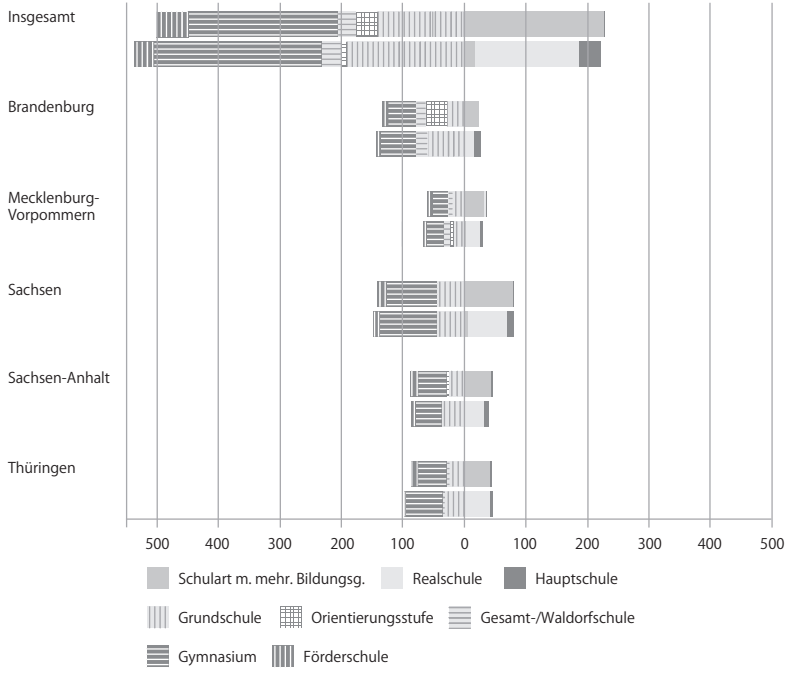


Abbildung 3: 9- bis 18-jährige Schüler allgemeinbildender Schulen in den neuen Bundesländern (ohne Berlin) im Mikrozensus 2008, 4. Quartal, und in der Bildungsstatistik (in 1.000)

Berufliche Schulen

Bis zum Mikrozensus 2007 wurde der Besuch allgemeinbildender und beruflicher Schulen sowie von Hochschulen in drei getrennten Fragen erhoben. Die Schularten beruflicher Schulen wurden nur erfragt, wenn zuvor kein Besuch einer allgemeinbildenden Schule genannt wurde. Während bis 2007 nur vier meist zusammengefasste Gruppen beruflicher Schulen zur Antwortauswahl standen, wurden 2008 neun Kategorien eingesetzt. 2011 erfolgte mit zwölf Kategorien eine weitere Differenzierung in Anlehnung an die Bildungsstatistik. Außerdem wurden 2011 die Untergruppen beruflicher Schulen, die allgemeine Schulabschlüsse vermitteln, stärker gegliedert und beschrieben. Berufliche Gymnasien wurden neu in der Gruppe allgemeinbildender Schulen erfasst. Des Weiteren wurde

2011 die Kategorie Berufsakademie/Duale Hochschule der Gruppe Hochschulen/Fachhochschulen zugeordnet. Um den Vergleich zu vereinfachen, orientiert sich die Darstellung in Abbildung 4 an den Antwortkategorien des Mikrozensus 2007. Ergänzend werden stärker besetzte Untergruppen (dar.: Berufsschule, dar.: Fachschule, dar.: 2-3j. SdG) der Jahre 2008 und 2011 ausgewiesen. Damit kann geprüft werden, ob Veränderungen der relativen Differenzen der Sammelkategorien auch für die einzelnen in dieser Gruppe dominierenden Schularten zutreffen.

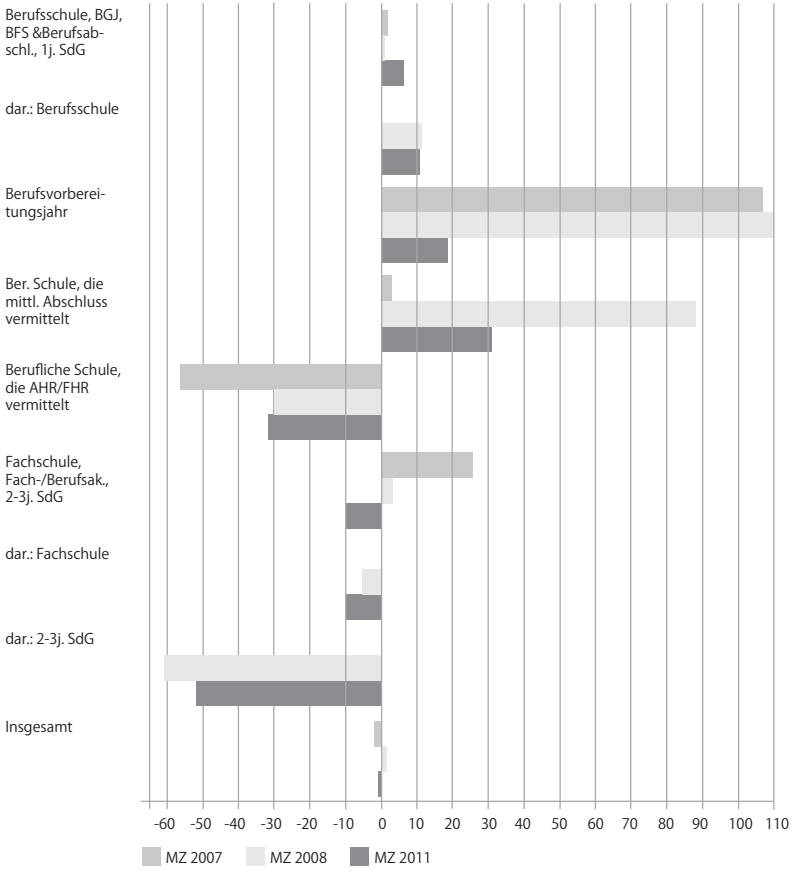


Abbildung 4: Schüler beruflicher Schulen - Relative Differenzen Mikrozensus vs. Bildungsstatistik (in %)

Abbildung 4 zeigt für 2007 große Abweichungen zwischen Mikrozensus und Bildungsstatistik von über 50 % beim Berufsvorbereitungsjahr und beruflichen Schulen, die zur Fachhochschulreife oder Hochschulreife (FHR/AHR) führen. Aufgrund der Anzahl von Schülern der zweiten Gruppe ist ihr das größere Gewicht beizumessen (siehe Tab. A3 im Anhang). Bei beiden Schularten sind die Abweichungen von 2007 bis 2011 deutlich zurückgegangen, sie sind aber nach wie vor beträchtlich.[3]

Die relativen Differenzen der Kategorie „Berufliche Schule, die einen mittleren Abschluss vermittelt (z. B. Berufsfachschule)" haben stark zugenommen. Es ist unklar, worauf die insbesondere 2008 hohe Übererfassung von 88 % zurückzuführen ist.

In der Gruppe „Berufsschule; Berufsgrundbildungsjahr [BGJ]; Berufsfachschule [BFS], die einen beruflichen Abschluss vermittelt; 1-jährige Schule des Gesundheitswesens [SdG]" des Mikrozensus 2007 überwiegen Berufsschüler mit einem Anteil von mehr als 80 %. Die ab 2008 gegebenen Vergleichsmöglichkeiten zeigen eine Übererfassung von rund 11 % gegenüber der Bildungsstatistik.

In der Sammelkategorie „Fachschule …" des Mikrozensus 2007 sind ab 2008 Schüler von Fachschulen mit einem Anteil von rund 50 % und Schüler von 2- oder 3-jährigen Schulen des Gesundheitswesens mit 14 % (2008) bzw. 36 % (2011) vertreten. Von 2007 bis 2011 hat sich die Übererfassung dieser Gruppe in eine Untererfassung gewandelt. Die oben genannte Übererfassung der Berufsschüler dürfte mit der Untererfassung der Schüler von 2- bis 3-jährigen Schulen des Gesundheitswesens in Verbindung stehen und auf die schwierige Zuordnung von Ausbildungen im Bereich der Gesundheits- und Krankenpflege zurückzuführen sein. Diese werden von den Befragten in Umfragen häufig als betriebliche Ausbildung betrachtet (Hall o. J.: 2). Zudem erfolgen diese Ausbildungen nicht nur in Schulen des Gesundheitswesens, sondern auch in Berufsfachschulen oder in Fachschulen.

Hochschulen

Mit den Hochschulen wird der Vergleich zwischen Mikrozensus und Bildungsstatistik abgerundet. Insgesamt sind die Differenzen zwischen Mikrozensus und Hochschulstatistik zurückgegangen (siehe Abb. 5). Studenten in Hochschulen sind über-, in Fachhochschulen[4] und insbesondere in Verwaltungsfachhochschulen untererfasst. Dies könnte damit zusammenhängen, dass Fachhochschulen

3 Im Mikrozensus 2011 ist eine Gegenüberstellung für einzelne berufliche Schulen möglich. Betrachtet man nur die stärker besetzten Schularten, zeigt sich, dass Schüler beruflicher Gymnasien und der Berufsfachschulen mit rund 50 % untererfasst sind, während bei Fachoberschulen nur eine Untererfassung von 3 % gegenüber der Bildungsstatistik vorliegt (hier ohne Nachweis).

4 Die Duale Hochschule Baden-Württemberg erhielt am Ende des Wintersemesters 2008/2009 den Rang einer Fachhochschule. Sie wird aus Gründen der Vergleichbarkeit bei den beruflichen Schulen berücksichtigt.

immer häufiger den wissenschaftlichen Hochschulen ähnliche Bezeichnungen verwenden. Ebenfalls ist nicht in allen Namen von Verwaltungsfachhochschulen deren Zuordnung direkt erkennbar.

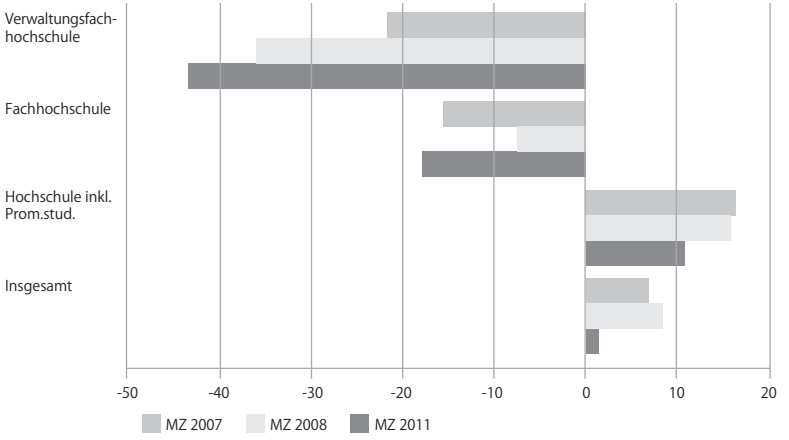


Abbildung 5: Hochschulbesucher - Relative Differenzen Mikrozensus vs. Bildungsstatistik (in %)

Zusammenfassend ist festzuhalten, dass die Änderungen des Fragebogens teilweise auf Verbesserungen hinweisen. Beispielsweise ist sowohl die gravierende Übererfassung von Schülern der gymnasialen Oberstufe als auch die Untererfassung beruflicher Schulen, die zur Fachhochschulreife oder zum Abitur führen, zurückgegangen. Dennoch finden sich weiterhin bedeutende Abweichungen zwischen Mikrozensus und Bildungsstatistik; insbesondere bei Schularten mit mehreren Bildungsgängen, Förderschulen und beruflichen Schulen.

Antwortinkonsistenzen bei Bildungsabschlüssen im Mikrozensus-Panel 2001-2004

Die obigen Vergleiche zwischen Mikrozensus und Bildungsstatistik lassen vermuten, dass die Abweichungen mit den für die Befragten unklaren Begriffen zusammenhängen. Zum Beispiel werden rund 90 % der Abschlüsse der Fachhochschulreife an beruflichen Schulen (v. a. Fachoberschulen und Berufsfachschulen) erworben, die teilweise auch berufsqualifizierende Abschlüsse vermitteln. Hängt die Untererfassung dieser Schularten mit der schwierigen Unterscheidung allgemeinbildender und beruflicher Schulen und/oder den Definitionen und Abgrenzungen dieser beruflichen Schulen voneinander zusammen, wird sich dies auch

in Form geringer Antwortstabilitäten der in diesen Schulen vermittelten Bildungsabschlüsse niederschlagen. Muster von Übergangswahrscheinlichkeiten der Bildungsangaben im Zeitverlauf lassen erkennen, welche Abschlüsse von den Befragten schwer unterscheidbar sind.

Im Folgenden werden solche Muster von Übergangswahrscheinlichkeiten herangezogen, um in einem ersten Schritt und in Orientierung an Befunde der kognitiven Umfrageforschung „schwierige Fragen" bzw. „schwierige Kategorien" zu kennzeichnen. Dabei wird eine Antwortstabilität unter 70 % zugrunde gelegt (Schimpl-Neimanns 2006: 211). Anschließend wird untersucht, wie Antwortinkonsistenzen in Abhängigkeit von der Befragungsart (Interview vs. schriftliche Auskunft) und der Art der Beteiligung (Selbstauskunft vs. Proxy-Auskunft) variieren.

Bildungsangabe (t)	**Bildungsangabe (t+1)**						
(m Beobachtungen)	**0**	**1**	**2**	**3**	**4**	**5**	**9**
0 Kein Abschluss	***54***	*34*	*0*	*3*	*0*	*2*	*7*
1 Volks-/Hauptschule	*2*	***85***	*1*	*7*	*1*	*1*	*3*
2 Polytechnische Oberschule	*0*	*3*	***88***	*4*	*2*	*1*	*2*
3 Realschule	*0*	*9*	*3*	***80***	*3*	*2*	*2*
4 Fachhochschulreife	*0*	*3*	*4*	*14*	***55***	*22*	*2*
5 Allgem./fachgeb. Hochschulreife	*0*	*1*	*1*	*3*	*6*	***86***	*2*
9 Keine Angabe	*5*	*32*	*7*	*17*	*4*	*11*	***25***
Insgesamt (m = 73.782)	2	33	14	24	5	19	3

Tabelle 1: Antwortvariabilitäten des allgemeinen Schulabschlusses im MZ-Panel 2001-2004 - durchschnittliche Übergangswahrscheinlichkeiten - Zeilen-%

In der ausgewählten Teilstichprobe sind Veränderungen der Bildungsqualifikationen durch zwischenzeitlich erworbene neue Abschlüsse ausgeschlossen (siehe Abschnitt 3.2). Für die n = 24.594 Personen des Mikrozensus-Panels 2001-2004 können Übergangsmatrizen für die Übergänge 2001/2002, 2002/2003 und 2003/2004 erstellt werden. Fasst man die drei Übergangsmatrizen zusammen, erhält man durchschnittliche Übergangswahrscheinlichkeiten für m = 73.782 (m = n * 3) Antworten. Tabelle 1 zeigt mit Ausnahme von drei Kategorien Antwortkonsistenzen von über 80 %. Wie erwartet, ist die Konsistenz der Kategorie Fachhochschulreife mit 55 % deutlich geringer. Befragte, die ihren Abschluss nicht genannt haben, wechseln mit einem Anteil von 32 % häufig bei der nächsten Erhebung zum Abschluss der Volks-/Hauptschule. Dies deutet auf Möglichkeiten hin, Panelangaben zur Korrektur von Antwortausfällen (Item-Nonresponse) zu

nutzen. Ein ähnliches Muster trifft für den Übergang von „kein Abschluss“ zu „Volks-/Hauptschulabschluss“ zu. Ob es sich hierbei eventuell um Effekte sozialer Erwünschtheit handelt, ist jedoch ohne weitere Prüfungen nicht zu klären.

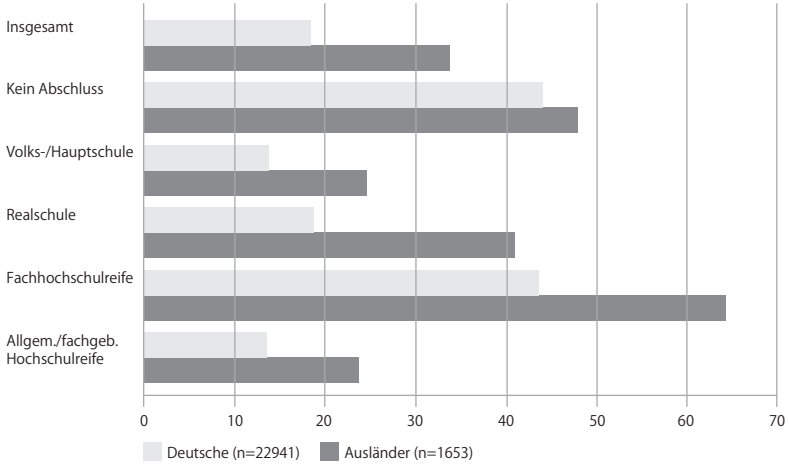
Bildungsangabe (t)	Bildungsangabe (t+1)											
(m Beobachtungen)	0	1	2	3	4	5	6	7	8	9	10	99
0 kein Abschluss vorh.	***70***	*2*	*0*	*18*	*1*	*1*	*0*	*0*	*1*	*1*	*0*	*7*
1 Anlernausb./Praktikum	*24*	***15***	*1*	*43*	*6*	*1*	*1*	*0*	*1*	*0*	*0*	*7*
2 Berufsvorbereitungsjahr	*19*	*8*	***4***	*49*	*7*	*1*	*0*	*1*	*1*	*1*	*0*	*10*
3 Lehre, Vorb. mittl. Dienst	*4*	*1*	*0*	***84***	*3*	*3*	*1*	*0*	*1*	*0*	*0*	*3*
4 Berufsfach-/Kolleg schule	*3*	*1*	*0*	*44*	***29***	*13*	*1*	*0*	*2*	*1*	*0*	*5*
5 Meister/Techniker, BA/ Fachak., Verw.-FH	*1*	*0*	*0*	*19*	*5*	***62***	*2*	*2*	*4*	*1*	*0*	*4*
6 Fachschule DDR	*1*	*0*	*0*	*18*	*3*	*8*	***52***	*1*	*11*	*4*	*0*	*3*
7 Verw.-FH (2002-2004)	*1*	*1*	*0*	*10*	*2*	*6*	*1*	***50***	*20*	*7*	*1*	*2*
8 Fachhochschule	*1*	*0*	*0*	*7*	*2*	*5*	*3*	*5*	***60***	*12*	*2*	*4*
9 Hochschule	*1*	*0*	*0*	*3*	*1*	*1*	*1*	*1*	*9*	***76***	*4*	*3*
10 Promotion	*1*	*0*	*0*	*2*	*0*	*1*	*0*	*1*	*5*	*20*	***66***	*4*
99 keine Angabe	*16*	*2*	*0*	*39*	*3*	*4*	*1*	*1*	*3*	*4*	*1*	***25***
Insgesamt (m = 73.782)	11	1	0	54	4	8	2	1	5	7	1	5

Tabelle 2: Antwortvariabilitäten des beruflichen Abschlusses im MZ-Panel 2001-2004 - Durchschnittliche Übergangswahrscheinlichkeiten - Zeilen-%

Die Antwortinkonsistenzen beruflicher Abschlüsse (siehe Tab. 2) sind erheblich größer als beim allgemeinen Schulabschluss. Lediglich die Kategorien „Abschluss einer Lehrausbildung, Vorbereitungsdienst für den mittleren Dienst in der öffentlichen Verwaltung“ (3) und „Abschluss einer Universität (wissenschaftlichen Hochschule, auch Kunsthochschule)“ (9) weisen Übergangswahrscheinlichkeiten über 70 % auf. Die geringe Antwortkonsistenz bei den folgenden Kategorien korrespondiert mit Über- und Untererfassungen beim Besuch beruflicher Schulen und Hochschulen: Berufsvorbereitungsjahr (2), „Berufsqualifizierender Abschluss an einer Berufsfachschule/Kollegschule, Abschluss einer 1-jährigen Schule des Gesundheitswesens“ (4), „Meister-/Technikerausbildung oder gleichwertiger Fachschulabschluss, Abschluss einer 2- oder 3-jährigen Schule des Gesundheitswesens, Abschluss einer Fachakademie oder einer Berufsakademie,

Abschluss einer Verwaltungsfachhochschule [nur 2001]“ (5), Verwaltungsfachhochschule [extra ab 2002] (7) und Fachhochschule (8).

Abbildung 6: Anteile inkonsistenter Angaben zu ausgewählten Kategorien des allgemeinen



Schulabschlusses im Mikrozensus-Panel 2001-2004 bei Deutschen und Ausländern (in %)

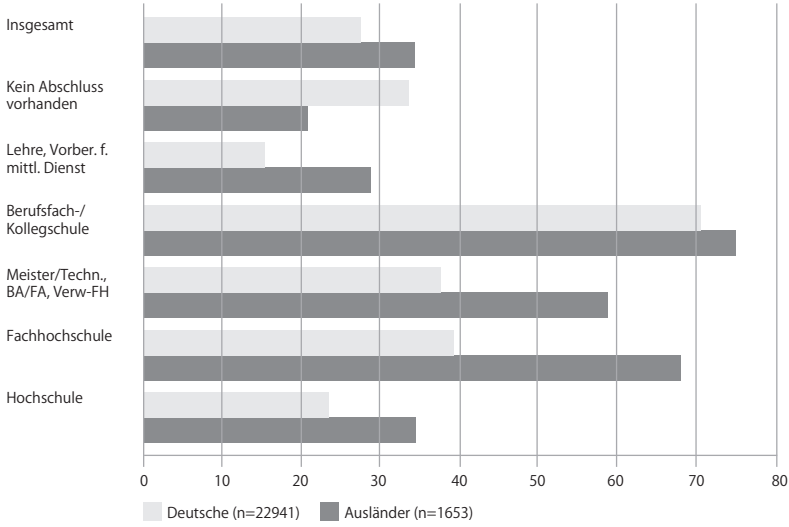


Abbildung 7: Anteile inkonsistenter Angaben zu ausgewählten Kategorien des beruflichen Abschlusses im Mikrozensus-Panel 2001-2004 bei Deutschen und Ausländern (in %)

Gemäß den in Kapitel 2 skizzierten Theorien der kognitiven Umfrageforschung werden für Ausländer größere Antwortinkonsistenzen als für Deutsche erwartet. Als Gründe hierfür kommen eine geringere Vertrautheit mit dem deutschen Bildungssystem und Probleme der Zuordnung von im Ausland erworbenen Abschlüssen infrage. Nicht zuletzt ist zu beachten, dass ausschließlich Fragebögen in deutscher Sprache eingesetzt werden. Die Abbildungen 6 und 7 belegen die Vermutungen sowohl für allgemeine als auch berufliche Abschlüsse. Für die unter Ausländern am stärksten besetzten Abschlüsse Volks-/Hauptschule und Lehrausbildung sind die Inkonsistenzen von Ausländern um mehr als 10 Prozentpunkte höher und im Fall der Lehrausbildung fast doppelt so hoch als bei Deutschen.

Mode-Effekte

Zusammenhänge zwischen Befragungsart und Datenqualität sind methodisch nicht einfach zu untersuchen, da abhängig von der Befragungsart sowohl die Gruppen der Befragten unterschiedlich zusammengesetzt sein können als auch unterschiedliche Antwortverhalten auftreten können (Jäckle/Roberts/Lynn 2010). In einem Literaturüberblick stellt Roberts (2007: 8-16) einige Befunde zusammen: Die Tendenz zur sozialen Erwünschtheit ist in schriftlichen Befragungen geringer als in persönlichen Befragungen durch Interviewer. Missverständnisse aufgrund unklarer Begriffe oder Fragen können sowohl in schriftlichen Befragungen als auch in persönlichen Interviews auftreten. In persönlichen Interviews können eventuelle Verständnisprobleme durch Erläuterungen der Interviewer geklärt werden. Es gibt Hinweise darauf, dass Satisficing-Effekte in schriftlichen Befragungen geringer sind als in Interviews, z. B. weil Befragte weniger unter Zeitdruck stehen. Andererseits sind aber die kognitiven Anforderungen bei schriftlichen Befragungen größer, wodurch Satisficing-Effekte verstärkt werden können.

Im Mikrozensus-Panel 2001-2004 können Mode- und Proxy-Effekte nicht strikt getrennt untersucht werden. Die folgenden Auswertungen konzentrieren sich zunächst auf Wechsel der Befragungsart und können auch Proxy-Effekte enthalten. Außerdem wird die Populationsheterogenität der Gruppen nicht kontrolliert.

Es ist zu vermuten, dass die Antwortinkonsistenzen bei schriftlichen Befragungen größer sind als bei persönlichen Interviews, in denen die Interviewer die Antwortkategorie erläutern können. Allerdings weisen Ergebnisse eines Methodenexperiments zum Mikrozensus auch darauf hin, dass zum beruflichen Abschluss gegebene freie Antworten der Befragten (z. B. nur „Ausbildung“) durch die Interviewer den ihnen plausibel erscheinenden Kategorien („Lehrausbildung im dualen System“) ohne weitere Klärung zugeordnet werden (Statistische Ämter

des Bundes und der Länder 2010: 25-26). Auch deshalb kann es in schriftlichen Befragungen im Vergleich zu persönlichen Interviews tendenziell stärker differenzierte Angaben geben.

Wenn alle Befragungen als persönliche Interviews durchgeführt werden, sind die Antwortinkonsistenzen insgesamt mit 16 % am geringsten (siehe Abb. 8) und steigen auf 30 % bei zwei- und dreimaligem Einsatz des Selbstausfüllerbogens. Wenn in allen Befragungen ein schriftlicher Fragebogen verwendet wurde, sinkt die Antwortinkonsistenz auf 27 %. In ähnlicher Weise schwanken die Anteile beim Abitur, jedoch lediglich zwischen 13 % (nie) und 18 % (3-mal Selbstausfüllerbogen). Dagegen steigen die Antwortinkonsistenzen bei den unteren Abschlüssen mit zunehmender Zahl schriftlicher Befragungen; z. B. beim Hauptschulabschluss von 12 % auf 29 % um mehr als das Doppelte.

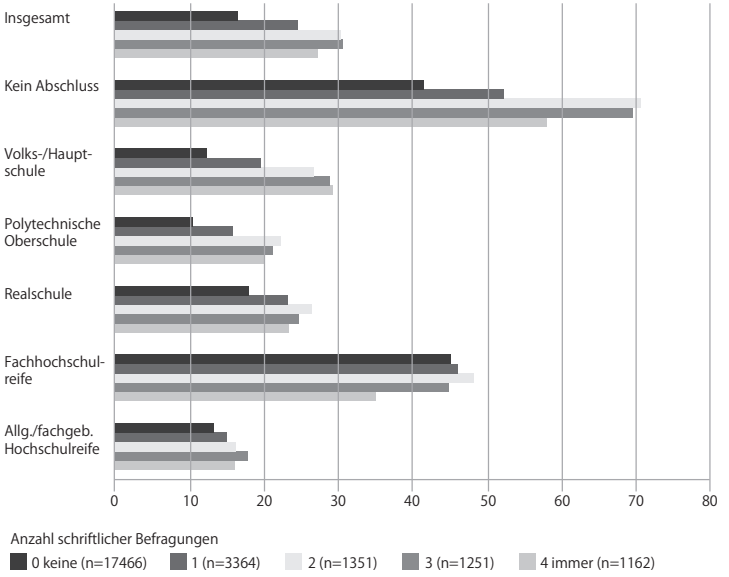


Abbildung 8: Anteil inkonsistenter Angaben zu ausgewählten Kategorien des allgemeinen Schulabschlusses in Abhängigkeit von der Anzahl schriftlicher Befragungen im Mikrozensus-Panel 2001-2004 (in %)

Betrachtet man die Abschlüsse als Indikator kognitiver Kompetenzen der Befragten, lässt dieses Muster auf eine Interaktion mit der Befragungsart schließen. Und zwar in der Weise, dass die Antwortinkonsistenzen bei Personen mit niedriger

Qualifikation bei schriftlichen Befragungen aufgrund mangelnder Fähigkeiten im Umgang mit schriftlichen Dokumenten oder aufgrund fehlender Erläuterungen durch Interviewer größer sind als bei Personen mit hoher Qualifikation.

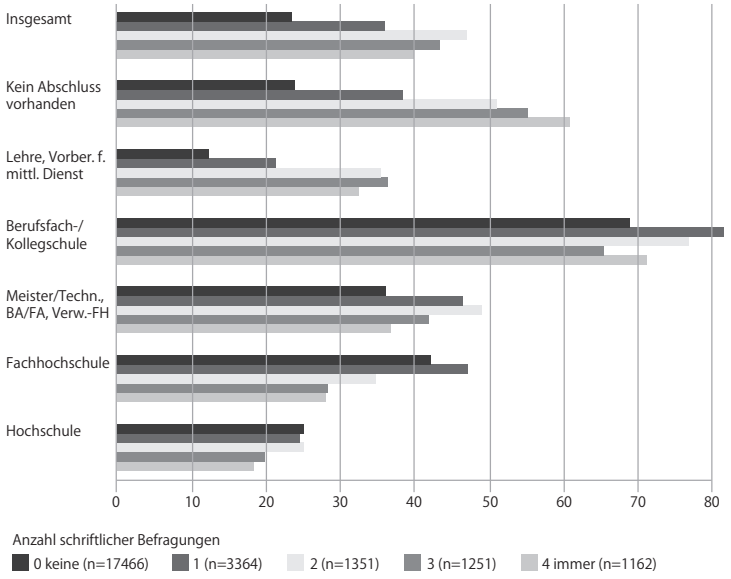


Abbildung 9: Anteil inkonsistenter Angaben zu ausgewählten Kategorien des beruflichen Abschlusses in Abhängigkeit von der Anzahl schriftlicher Befragungen im Mikrozensus-Panel 2001-2004 (in %)

In Bezug auf berufliche Abschlüsse werden in Tabelle 2 nur für die Kategorien Lehre und Universität Antwortkonsistenzen von 70 % und höher festgestellt, sodass alle anderen Abschlüsse als „schwierige Kategorie" gelten können. Abbildung 9 zeigt, wie mit zunehmenden schriftlichen Befragungen insgesamt inkonsistente Antworten von 23 % (nie) auf 47 % (2-mal) steigen, danach aber fallen. Wird immer schriftlich geantwortet, beträgt die Inkonsistenz 40 %. In Bezug auf Antwortinkonsistenzen ist der Zusammenhang zwischen der Anzahl schriftlicher Befragungen und dem Niveau des Abschlusses zwar ähnlich wie beim allgemeinen Schulabschluss, er unterscheidet sich jedoch durch rückläufige Anteile bei den Kategorien Fachhochschule und Universität. Während Inkonsistenzen beim Lehrabschluss zunehmen (nie: 13 %; 4-mal: 32 %), fallen sie beim Hochschulabschluss von 25 % (nie) um 7 Prozentpunkte und beim Fachhochschulabschluss sogar um 14 Prozentpunkte. Für die beiden Kategorien Berufsfachschule usw.

und Meister usw. deuten sich zunächst steigende und dann rückläufige Tendenzen an, die jedoch wegen abnehmender Fallzahlen ab 3 schriftlichen Befragungen nicht weiter interpretiert werden.

Proxy-Effekte

Antworten Befragte für andere Personen im Haushalt, weil diese beim Interview nicht anwesend sind, kommen zu den oben angesprochenen Schwierigkeiten beim Beantworten der Bildungsfragen zusätzlich Informations- und Erinnerungsprobleme hinzu (Campanelli und Channell 1996: 18 f.). Wenn Auskünfte für Ehegatten gegeben werden, können diese Probleme gering sein, weil deren Abschlüsse bekannt sind und i. d. R. in der gleichen Lebensphase bzw. im historisch gleichen Bildungssystem erworben wurden. In anderen Fällen, wenn beispielsweise Familienfremde, Kinder für ihre Eltern oder Eltern für ihre Kinder antworten, ist mit größeren Informationsproblemen zu rechnen. Ergebnisse zur Pisa-Studie weisen darauf hin, dass Proxy-Angaben von Jugendlichen zur beruflichen Bildung ihrer Eltern stärkere Inkonsistenzen zu den von den Eltern genannten Abschlüssen aufweisen als Angaben zu den allgemeinen Schulabschlüssen der Eltern (Kreuter et al. 2010; siehe in diesem Kontext auch den Literaturüberblick in: Lipski 2000). Die bisherigen Analysen zum Mikrozensus ergaben eher schwache Proxy-Effekte (vgl. zu Bildungsfragen Herter-Eschweiler 2009 und Zühlke 2008, zur Erwerbstätigkeit Köhne-Finster und Lingnau 2009).

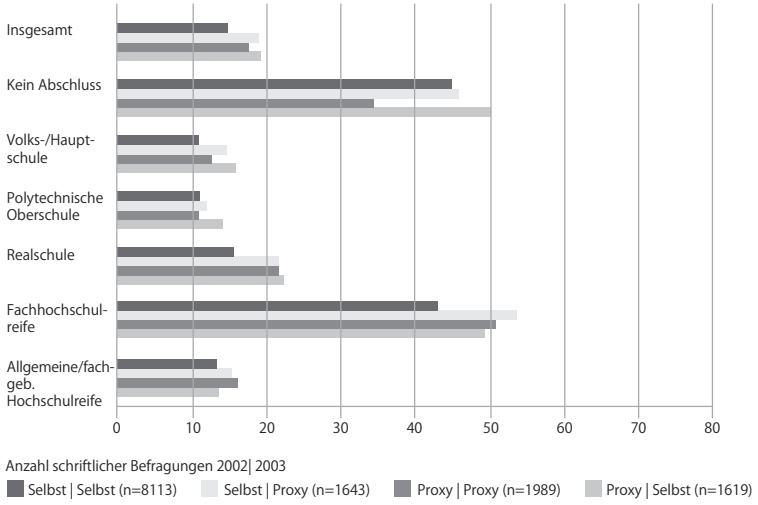


Abbildung 10: Anteil inkonsistenter Angaben ausgewählter Kategorien des allgemeinen Schulabschlusses nach Art der Beteiligung (in %)

Bei den Auswertungen des Mikrozensus-Panels werden Informationen über die Auskunft gebende Personen nicht berücksichtigt. Die Auswertungen beschränken sich auf persönliche Interviews von 2001 bis 2003, in denen nur für 2002 und 2003 Angaben zur Auskunftsart vorliegen. In den Abbildungen werden nur Kategorien mit wenigstens m = 100 Übergängen gezeigt.

Beim allgemeinen Schulabschluss beträgt der Anteil inkonsistenter Angaben insgesamt 15 %, wenn 2002 und 2003 jeweils Selbstauskünfte vorliegen (siehe Abb. 10). Falls 2002 und 2003 die Angaben durch Proxy-Interviews erfolgten, steigt die Quote geringfügig um 3 Prozentpunkte. Größere Veränderungen von 6 und 8 Prozentpunkten sind nur bei den Abschlüssen Realschule und Fachhochschulreife zu erkennen.

Für berufliche Abschlüsse variiert der Anteil inkonsistenter Angaben insgesamt ähnlich gering wie für allgemeine Abschlüsse. Ausgenommen davon ist die Kategorie „Berufsqualifizierender Abschluss an einer Berufsfachschule/ Kollegschule, Abschluss einer 1-jährigen Schule des Gesundheitswesens". Hier variieren die Anteile systematisch mit der Anzahl von Proxy-Interviews. Wurden 2002 und 2003 Proxy-Interviews durchgeführt, liegt der Anteil inkonsistenter Angaben um rund 10 Prozentpunkte höher, als wenn in beiden Jahren Selbstauskünfte gegeben wurden.

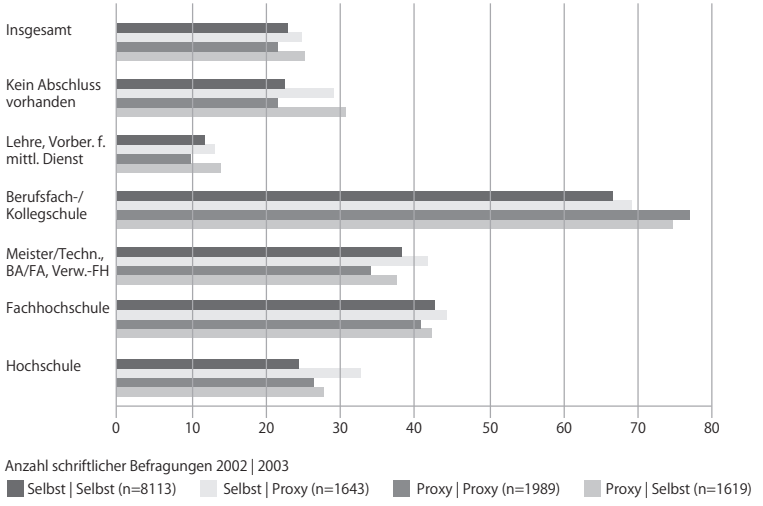


Abbildung 11: Anteil inkonsistenter Angaben ausgewählter Kategorien des beruflichen Abschlusses nach Art der Beteiligung (in %)

Schluss

Mit der 2008 wieder aufgenommenen Erfassung der Schularten allgemeinbildender Schulen bietet sich für die Bildungs- und Sozialforschung die Möglichkeit, frühere Analysen zur sozialen Ungleichheit der Bildungsbeteiligung wieder durchzuführen und fortzuschreiben. Die Änderungen der Fragebögen zeigen bei Vergleichen mit der Bildungsstatistik und im Vergleich zur Ausgangslage 2007 teilweise Verbesserungen, d. h. geringere Abweichungen zwischen Mikrozensus und Bildungsstatistik.

So ist zum Beispiel die gravierende Übererfassung von Schülern der gymnasialen Oberstufe deutlich zurückgegangen. Zugleich sind Schüler beruflicher Schulen, die zur Fachhochschulreife oder zum Abitur führen, weniger stark untererfasst als zuvor. Daneben gibt es aber eine ganze Reihe von Abweichungen. Dies betrifft unter anderem Schularten mit mehreren Bildungsgängen und Förderschulen sowie (ab 2008) die Übererfassung von Schülern weiterführender Schulen (Realschule und Gymnasium) bei den allgemeinbildenden Schulen. Insgesamt führen Mikrozensus und Bildungsstatistik zu unterschiedlichen Ergebnissen.[5]

Die Analyse von Gesamtabweichungen („Total Survey Error") erlaubt nur sehr begrenzt Schlussfolgerungen auf die Abweichungsgründe. In Verbindung mit theoretischen Ansätzen der kognitiven Umfrageforschung lässt sich vermuten, dass die Untererfassung der Schüler von Förder- oder Sonderschulen auf Effekten sozialer Erwünschtheit beruht. Am Beispiel der Schularten mit mehreren Bildungsgängen in den neuen Bundesländern zeigt sich sehr deutlich, dass dieser Begriff in der Befragungspraxis zumindest teilweise nicht dem Alltagsverständnis von Schülern und Eltern entspricht. Insbesondere treten bei den beruflichen Schulen größere Differenzen zwischen Mikrozensus und Bildungsstatistik auf. Auch hier ist als Grundproblem zu vermuten, dass die aus der Bildungsstatistik übernommenen Fachbegriffe für viele Befragte nicht klar genug sind.

Vor allem die beruflichen Bildungsgänge haben sich in den letzten Jahren stark ausdifferenziert und bestimmte allgemeinbildende Abschlüsse sind in verschiedenen Schularten zu erreichen. Nicht zuletzt gibt es in den Ländern unterschiedliche Bildungsangebote mit zum Teil unterschiedlichen Bezeichnungen. Den sich daraus ergebenden Anforderungen für die Formulierung von praxistauglichen Fragen in Umfragen wird sich der Mikrozensus auch künftig nicht entziehen können. Für die Feldarbeit können länderspezifische Listen der Schularten sowie eine besondere Schulung der Interviewer hilfreich sein. In Anbe-

5 Davon ist auch die deutsche EU-Arbeitskräftestichprobe betroffen, da die darin vorliegenden Codierungen der Bildungsangaben nach der Internationalen Standardklassifikation des Bildungswesens (ISCED) auf den Angaben im Mikrozensus beruhen.

tracht des steigenden Anteils schriftlicher Antworten wäre die Aufnahme von Erläuterungen im schriftlichen Fragebogen zu überlegen. Dies dürfte aber nicht den Kern des Problems treffen.

Im Mikrozensus 2008 und noch stärker 2011 sind in Anlehnung an die in der Bildungsstatistik berichteten Schularten frühere Zusammenfassungen aufgelöst worden bzw. neue Kategorien hinzugekommen. Der Vergleich zwischen Ergebnissen des Mikrozensus 1989 und der Bildungsstatistik zum Besuch der gymnasialen Oberstufe (Schimpl-Neimanns 2006) lässt dagegen vermuten, dass die damals gestellten Fragen zum Schulbesuch mit gröberen Antwortkategorien der Befragungspraxis angemessener waren. Vor diesem Hintergrund ist deshalb zu fragen, ob eine Eins-zu-eins-Umsetzung der Schularten nach den Definitionen der Bildungsstatistik im Mikrozensus notwendig und sinnvoll ist. Vielmehr wäre, mit den Formulierungen von Campanelli und Channell (1996: 34), die Schlüsselfrage anzugehen: „To avoid problems, surveys should mirror closely the terminology currently being encountered by respondents." Hierzu bedarf es Pretests, um zu klären, wie Befragte Fragen und Begriffe zum Schulbesuch und zu den Bildungsabschlüssen verstehen.

Die Analysen zur Antwortinkonsistenz der Bildungsabschlüsse ergaben nur schwache Proxy-Effekte. Jedoch stammen mehr als 60 % aller Angaben der Schulbesucher im Alter ab 15 Jahren von den Eltern. Zudem sind Interaktionen zwischen den Erfassungsproblemen der von den Kindern besuchten Schularten und den Bildungsabschlüssen der Eltern zu vermuten. Es ist deshalb sinnvoll, den Einfluss von Proxy-Interviews in diesem Bereich näher zu untersuchen

In den deskriptiven Analysen der Antwortinkonsistenzen zu Bildungsangaben des Mikrozensus-Panels 2001-2004 konnten Mode- und Proxy-Effekte nicht getrennt untersucht werden.[6] Aufgrund weiterer methodischer Einschränkungen lassen sich deshalb nur vorläufige und grobe Hinweise geben. Es zeigt sich, dass die Befragungsarten persönliches Interview und schriftliche Befragung mit erheblichen Unterschieden hinsichtlich der Antwortkonsistenz verbunden sind. Insbesondere scheinen kognitive Anforderungen bei schriftlichen Befragungen für Personen mit formal geringerer Qualifikation mit einem höheren Anteil inkonsistenter Angaben einherzugehen. Das Ergebnis steht im Einklang mit Theorien der kognitiven Umfrageforschung. In einigen Fällen korrespondieren hohe Antwortinkonsistenzen der Bildungsabschlüsse mit Über- und Untererfassungen beim Schulbesuch entsprechender Schularten. Dies spricht ebenfalls für den bereits genannten Bedarf nach Pretests.

Die Angaben des Mikrozensus beruhen zurzeit im Wesentlichen auf persönlichen Interviews. Der Anteil von Haushalten mit schriftlicher Auskunft

6 Dieses Problem wird bei Paneldaten des Mikrozensus ab 2005 nicht mehr auftreten. Die Bereitstellung aktueller Paneldaten für die Forschung durch die statistischen Ämter wäre daher sehr zu begrüßen.

steigt. Künftig kommen auch weitere Befragungsformen, wie z. B. per Telefon oder Internet, infrage. Im EU Labour Force Survey variieren die von den nationalen statistischen Ämtern eingesetzten Befragungsarten erheblich (siehe Eurostat 2013: 9). Zu berücksichtigen ist auch, dass Ergebnisunterschiede zwischen Gruppen (hier: Bildungsniveaus) und Ländern teilweise mit Mode-Effekten behaftet sind (Statistische Ämter des Bundes und der Länder 2010: 4). In dieser Hinsicht bemerkt Roberts (2007: 19): „(...) any continuous, longitudinal or repeated cross-sectional surveys seeking to incorporate new modes into their data collection designs should do so only after careful assessment of the likelihood of any mode effects and their impact on the data" (Roberts 2007: 19). Jäckle/Roberts/Lynn (2010) beschreiben die zu beachtenden methodischen Probleme. Weiteren Methodenstudien zu Mode-Effekten beim Mikrozensus ist daher eine zunehmende Bedeutung beizumessen.

Literatur

Autorengruppe Bildungsberichterstattung (Hrsg.) (2012): Bildung in Deutschland 2012. Ein indikatorengestützter Bericht mit einer Analyse zur kulturellen Bildung im Lebenslauf. Bielefeld: Bertelsmann.

Biemer, P. P. and Lyberg, L. E. (2003): Introduction to Survey Quality. Hoboken (NJ): Wiley.

Blair, J./Menon, G. and Bickart, B. (1991): Measurement effects in self vs. proxy responses to survey questions: an information-processing perspective. In: Biemer, P.P./Groves, R.M./Lyberg, L.E./Mathiowetz, N.A. and Sudman, S. (eds.): Measurement Errors in Surveys. New York: John Wiley and Sons, 145-166.

Campanelli, P., and Channell, J. (1996): The conceptualisation of qualifications by individuals. London: HMSO.

Europäische Kommission (2010): Europa 2020 - Eine Strategie für intelligentes, nachhaltiges und integratives Wachstum. Brüssel. www.bmwi.de/BMWi/Redaktion/PDF/E/eu-kommission-mitteilung-europa-2020.pdf [22.04.2013].

Eurostat (2013): Quality report of the European Union Labour Force Survey 2011. Luxembourg: Publications Office of the European Union. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-008/EN/KS-RA-13-008-EN.PDF [22.04.2013].

Fest, M./Freitag, H-W./Fritzsch, B. und Skripski, B. (2010): Zuordnung der beruflichen Bildungsgänge nach der ISCED auf Länderebene – Anschlussfähigkeit an die nationale und internationale Bildungsberichterstattung. In: Bundesministerium für Bildung und Forschung (Hrsg.): Indikatorenentwicklung für die Bildungsberichterstattung in Deutschland – Grundlagen, Ergebnisse, Perspektiven. Bildungsforschung Band 33. Bonn/Berlin: BMBF, 95-105.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (o. J.): Informationen zu Mikrozensus Scientific Use Files. http://www.forschungsdatenzentrum.de/bestand/mikrozensus/index.asp [22.04.2013].

GESIS (o. J.): Informationen zu Mikrozensus Scientific Use Files. http://www.gesis.org/unser-angebot/daten-analysieren/amtliche-mikrodaten/mikrozensus/grundfile/ [22.04.2013].

Groves, R.M./Fowler, F.J.Jr./Couper, M.P./Lepkowski, J.M./Singer, E. and Tourangeau, R. (2009): Survey methodology. Hoboken, NJ: Wiley.

Groves, R.M., and Lyberg, L. (2010): Total Suvey Error: Past, Present, and Future. Public Opinion Quarterly 74(5), 849-879.

Hall, A., (o. J.): Arbeitsmarkterfolg von Absolventen des dualen Systems und von Berufsfachschulabsolventen im Vergleich – Analysen auf der BIBB/IAB-Erhebung 1998/99. Bonn. www.bibb.de/dokumente/pdf/DS-BFS_BIBBIAB.pdf [29.04.2013].

Herter-Eschweiler, R. (2009): Datenqualität beruflicher Ausbildungsabschlüsse im Mikrozensus & Regionalkennung im Scientific Use File. Präsentation zur 6. Nutzerkonferenz: „Forschung mit dem Mikrozensus. Analysen zur Sozialstruktur und zum sozialen Wandel", Mannheim. http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/gml/Veranstaltungen/6.NK_2009/Praes/Herter.pdf [22.04.2013].

Hewat, C. (2011): Using education databases to improve the quality of the LFS. Paper presented at the 6th Workshop on LFS Methodology, Wiesbaden, Germany on the 12th May 2011. https://www.destatis.de/EN/AboutUs/Events/LFS/PapersP/E1_UsingEducationDatabases_Hewat.pdf?__blob=publicationFile [22.04.2013].

Jäckle, A./Roberts, C. and Lynn, P. (2010): Assessing the Effect of Data Collection Mode on Measurement. International Statistical Review 78(1), 3-20.

Köhne-Finster, S., und Lingnau, A. (2009): Untersuchung der Datenqualität erwerbsstatistischer Angaben im Mikrozensus – Ergebnisse des Projekts „Nachbefragung im Mikrozensus/LFS". Wirtschaft und Statistik 12/2008, 1067-1088.

Konsortium Bildungsberichterstattung (2006): Zur langfristigen Sicherstellung der Datenbasis für die Bildungsberichterstattung. Interner Bericht. Frankfurt/M. http://www.bildungsbericht.de/daten/datenstrategie.pdf [22.04.2013].

Kreuter, F./Eckman, S./Maaz, K. and Watermann, R. (2010): Children's Reports of Parents' Education Level: Does it Matter Whom You Ask and What You Ask About? Survey Research Methods 4(3), 127-138.

Krosnick, J.A. (1991): Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology 5, 213-236.

Lipski, J. (2000): Zur Verläßlichkeit der Angaben von Kindern bei standardisierten Befragungen. In: Heinzel, F. (Hrsg.): Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive. Weinheim: Juventa, 77-86.

Moore, J.C. (1988): Self/Proxy Response Status and Survey Response Quality. Journal of Official Statistics 4 (2), 155-172.

Müller, W. (1979): Schulbildung und Weiterbildung als soziologische Hintergrundsvariablen. In: Pappi, F. (Hrsg.): Sozialstrukturanalysen

mit Umfragedaten – Probleme der standardisierten Erfassung von Hintergrundsmerkmalen in allgemeinen Bevölkerungsumfragen. Königstein/Ts.: Athenäum Verlag, 217-240.

Roberts, C. (2007): Mixing modes of data collection in surveys: A methodological review. NCRM Methods Review Papers NCRM/008. London: Centre for Comparative Social Surveys, City University. http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf [22.04.2013].

Schimpl-Neimanns, B. (2006): Zur Datenqualität der Bildungsangaben im Mikrozensus am Beispiel des Besuchs der gymnasialen Oberstufe und des allgemeinen Schulabschlusses. In: Faulbaum, F. und Wolf, C. (Hrsg.): Stichprobenqualität in Bevölkerungsumfragen – Tagungsberichte 12 Bonn: Informationszentrum Sozialwissenschaften, 197-218.

Schimpl-Neimanns, B. (2011): Schätzung des Stichprobenfehlers in Mikrozensus Scientific Use Files ab 2005. AStA Wirtschafts- und Sozialstatistisches Archiv 5(1), 19-38.

Schwarz, N. and Wellens, T. (1997): Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Actors and Observers. Journal of Official Statistics 13(2), 159-179.

Statistische Ämter des Bundes und der Länder (2012): Datenbank der beruflichen Bildungsgänge der Länder. Aktualisierung für das Schuljahr 2010/2011. [Unveröffentlichte Tabellen bereitgestellt durch Statistisches Bundesamt, Gruppe H 2].

Statistische Ämter des Bundes und der Länder (2010): Ergebnisse des Projekts Q-MED/LFS - Quantifizierung von Methodeneffekten unterschiedlicher Erhebungsinstrumente auf die Datenqualität im Labour Force Survey. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2012a): Qualitätsbericht Mikrozensus 2011. Wiesbaden. https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/Bevoelkerung/Mikrozensus2011.pdf?__blob=publicationFile [24.04.2013].

Statistisches Bundesamt (2012b): Fachserie 11, Reihe 1, Bildung und Kultur, Allgemeinbildende Schulen, Schuljahr 2011/2012. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2012c): Fachserie 11, Reihe 2, Bildung und Kultur, Berufliche Schulen, Schuljahr 2011/2012. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2012d): Fachserie 11, Reihe 4.1, Bildung und Kultur, Studierende an Hochschulen, Wintersemester 2011/2012. Wiesbaden: Statistisches Bundesamt.

Statistisches Bundesamt (2010a): Fachserie 11, Reihe 1, Bildung und Kultur, Allgemeinbildende Schulen, Schuljahr 2007/2008. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (2010b): Fachserie 11, Reihe 2, Bildung und Kultur, Berufliche Schulen, Schuljahr 2008/2009. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (Hrsg.) (2009a): Handbuch zum Mikrozensus-Panel 2001-2004. Wiesbaden/Bonn. http://www.gesis.org/missy/fileadmin/missy/erhebung/Panel/2001-2004/MZP0104_Handbuch.pdf [24.04.2013].
Statistisches Bundesamt (2009b): Fachserie 11, Reihe 1, Bildung und Kultur, Allgemeinbildende Schulen, Schuljahr 2008/2009. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (2009c): Fachserie 11, Reihe 4.1, Bildung und Kultur, Studierende an Hochschulen, Wintersemester 2008/2009. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (2008a): Fachserie 11, Reihe 2, Bildung und Kultur, Berufliche Schulen, Schuljahr 2007/2008. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (2008b): Fachserie 11, Reihe 4.1, Bildung und Kultur, Studierende an Hochschulen, Wintersemester 2007/2008. Wiesbaden: Statistisches Bundesamt.
Statistisches Bundesamt (2007a): Qualitätsbericht. Statistik der allgemeinbildenden und beruflichen Schulen. Wiesbaden: Statistisches Bundesamt. https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/BildungForschungKultur/Schulen/AllgembBeruflSchulen.pdf?__blob=publicationFile [25.04.2013]
Statistisches Bundesamt (2007b): Qualitätsbericht. Studentenstatistik. Wiesbaden. https://www.destatis.de/DE/Publikationen/Qualitaetsberichte/BildungForschungKultur/Hochschulen/Studenten.pdf?__blob=publicationFile [25.04.2013]
Strack, F. and Martin, L.L. (1987): Thinking, judging, and communicating: A process account of context effects in attitude surveys. In: Hippler, H.-J./Schwarz, N./Sudman, S. (eds): Social Information Processing and Survey Methology. New York: Springer, 123-148.
Thomson, D. and Knight, T. (with Buscha, F./Urwin, P. and Sturgis, P.) (2010): Research into Measuring Adult Attainment Using the Labour Force Survey. Final Report. London: RM Data Solutions. http://www.thedataservice.org.uk/NR/rdonlyres/F573D7EA-8BD3-49B0-9D40-54DD7AD5F2E2/0/FinalReport.pdf [22.04.2013].

Tourangeau, R./Rips, L.J. and Rasinski, K. (2000): The psychology of survey response. Cambridge: Cambridge Univ. Pr.
Verordnung (EU) Nr. 99/2013 des Europäischen Parlaments und des Rates vom 15. Januar 2013 über das Europäische Statistische Programm 2013-2017. Brüssel. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:039:0012:0029:DE:PDF [22.04.2013].
Zühlke, S. (2008): Auswirkungen von Proxy-Interviews auf die Datenqualität des Mikrozensus. Statistische Analysen und Studien NRW 53, 3-10.

Tabelle A1: Schüler in allgemeinbildenden Schulen (in 1.000) nach Bildungsbereichen in der Bildungsstatistik (BS) und im Mikrozensus (MZ, jeweils 4. Quartal), relative Differenzen (RD) und Variationskoeffizienten (CV) zu den Gesamtwerten des Mikrozensus in Prozent

Bildungsbereich / Klassenstufen	BS 2007	MZ 2007	RD	CV	sig.	BS 2008	MZ 2008	RD	CV	sig.	BS 2011	MZ 2011	RD
Primarbereich / Klassenstufe 1-4	*3.260*	*3.134*	*-3,9*	*1,2*	*	*3.033*	*2.991*	*-1,4*	*1,2*		*2.832*	*2.860*	*1,0*
Sekundarbereich I / Klassenstufe 5-10	*4.997*	*4.952*	*-0,9*	*1,1*		*4.589*	*4.805*	*4,7*	*1,1*	*	*4.392*	*4.743*	*8,0*
Sekundarbereich II / Klassenstufe 11-13	*898*	*1.434*	*59,6*	*2,2*	*	*979*	*1.082*	*10,4*	*2,5*	*	*1.061*	*1.052*	*-0,8*
Förderschulen (keinem Bereich zugeordnet)[a]						*393*	*221*	*-43,8*	*6,3*	*	*366*	*229*	*-37,4*
Insgesamt	9.156	9.520	4,0	0,7	*	8.995	9.099	1,1	0,7		8.650	8.884	2,7

a) Schüler der Förderschulen wurden in der Bildungsstatistik [N = 400.399] 2007 nach Schätzungen auf die Bereiche aufgeteilt.
RD Relative Differenz der Gesamtwerte (%)
CV Variationskoeffizient (Standardfehlerler / Gesamtwert; in %)
*sig. * Gesamtwert der Bildungsstatistik liegt nicht im 95 %-Konfidenzintervall des Gesamtwertes des Mikrozensus*

Quellen zur Bildungsstatistik (BS)
- Statistisches Bundesamt (versch. Jahre): Fachserie 11, Reihe 1, Bildung und Kultur, Allgemeinbildende Schulen, Schuljahr 2007/2008, Schuljahr 2008/2009, Schuljahr 2011/2012; eigene Berechnungen.

Quellen zum Mikrozensus (MZ)
- MZ 2007 und 2008: FDZ der statistischen Ämter des Bundes und der Länder; eigene Berechnungen.
- MZ 2011: Sonderauswertung durch Statistisches Bundesamt, Referat F207; eigene Berechnungen

Tabelle A2: Schüler in allgemeinbildenden Schulen (in 1.000) nach Schulart in der Bildungsstatistik (BS, Schuljahre 2008/2009 und 2011/2012) und im Mikrozensus (MZ, jeweils 4. Quartal), relative Differenzen (RD) und Variationskoeffizienten (CV) zu den Gesamtwerten des Mikrozensus in Prozent

Schulart [Nummer der Antwortkategorie im MZ 2008 | 2011]	BS 2008	MZ 2008	RD	CV	sig.	BS 2011	MZ 2011	RD
Grundschule [1 | 1]	*2.997*	*3.030*	*1,1*	*1,2*		*2.790*	*2.872*	*2,9*
Hauptschule, Abendhauptschule [2 | 5]	*827*	*935*	*13,1*	*2,9*	*	*658*	*585*	*-11,1*
Realschule, Abendrealschule [3 | 6]	*1.284*	*1.603*	*24,9*	*2,1*	*	*1.150*	*1.499*	*30,3*
Orientierungsstufe 5./6. Klasse (z. B. an Grund- oder weiterführenden Schulen, Förderstufe [4 | 2]	*105*	*28*	*-73,4*	*16,3*	*	*101*	*76*	*-24,8*
Schule mit mehreren Bildungsgängen (z. B. Mittel-, Ober-, Regel-, Sekundar-, Regionale Schule) [5 | 4]	*302*	*65*	*-78,6*	*11*	*	*400*	*240*	*-40,1*
Gesamtschule, Waldorfschule [6 | 7, 8]	*583*	*553*	*-5,2*	*4,2*		*714*	*720*	*0,8*
Gymnasium, Abendgymnasium, Kolleg[a] [7 | 9, 11]	*2.505*	*2.665*	*6,4*	*1,7*	*	*2.471*	*2.664*	*7,8*
Förderschule [8 | 3]	*393*	*221*	*-43,8*	*6,3*	*	*366*	*229*	*-37,4*
Insgesamt	8.995	9.099	1,1	0,7		8.650	8.884	2,7

a) 2011 ohne berufliche Gymnasien, die in der Bildungsstatistik und im Mikrozensus 2008 den beruflichen Schulen zugeordnet sind.

Quellen und Erläuterungen: Siehe Tab. A.1

Tabelle A3: Schüler beruflicher Schulen (in 1.000) in der Bildungsstatistik (BS, Schuljahre 2007/2008, 2008/2009 und 2011/2012) und im Mikrozensus (MZ, jeweils 4. Quartal), relative Differenzen (RD) und Variationskoeffizienten (CV) zu den Gesamtwerten des Mikrozensus in Prozent

Schulart	BS 2007	MZ 2007	RD	CV	sig.	BS 2008	MZ 2008	RD	CV	sig.	BS 2011	MZ 2011	RD
Berufsschule, Berufsgrundbildungsjahr oder Berufsfachschule, die einen Berufsabschluss vermittelt, 1-jährige Schule des Gesundheitswesens	*2.087*	*2.121*	*1,6*	*1,8*		*2.098*	*2.112*	*0,6*	*1,8*		*1.850*	*1.961*	*6,1*
Berufsvorbereitungsjahr	*62*	*128*	*106,9*	*7,8*	*	*55*	*116*	*110,1*	*7,9*	*	*78*	*92*	*18,6*
Berufliche Schule, die einen mittleren Abschluss vermittelt (z. B. Realschulabschluss)	*101*	*104*	*3,0*	*8,5*		*89*	*167*	*88,1*	*6,7*	*	*65*	*85*	*30,9*
Berufliche Schule, die die Hochschul-/Fachhochschulreife vermittelt[a]	*407*	*178*	*-56,4*	*6,4*	*	*411*	*289*	*-29,7*	*4,9*	*	*441*	*302*	*-31,4*
Fachschule, Fach-/Berufsakademie, 2- oder 3-jährige Schule des Gesundheitswesens[b]	*301*	*378*	*25,3*	*4,4*	*	*310*	*319*	*3,1*	*4,7*		*352*	*317*	*-10,0*
Insgesamt	2.959	2.908	-1,7	1,6		2.964	3.005	1,4	1,5		2.785	2.758	-1,0

a) 2011 inklusive berufliche Gymnasien.
b) 2008 und 2011 inklusive Duale Hochschule Baden-Württemberg [2008 N = 23.013; 2011 N = 27.758], die im Mikrozensus 2011 und in der Bildungsstatistik ab 2008/2009 zu den Fachhochschulen zählt.

Quellen und Erläuterungen (siehe auch Tab. A1):
- Statistisches Bundesamt (versch. Jahre): Fachserie 11, Reihe 2, Bildung und Kultur, Berufliche Schulen, Schuljahr 2007/2008, Schuljahr 2008/2009, Schuljahr 2011/2012; eigene Berechnungen.

Tabelle A4: Hochschüler (in 1.000) in der Bildungsstatistik (BS, Wintersemester 2007/2008, 2008/2009 und 2011/2012) und im Mikrozensus (MZ, jeweils 4. Quartal), relative Differenzen (RD) und Variationskoeffizienten (CV) zu den Gesamtwerten des Mikrozensus in Prozent

Schulart	BS 2007	MZ 2007	RD	CV	sig.	BS 2008	MZ 2008	RD	CV	sig.	BS 2011	MZ 2011	RD
Verwaltungsfachhoch-schule	*26*	*21*	*-21,6*	*17,2*		*27*	*17*	*-36,2*	*19,2*	*	*32*	*18*	*-43,4*
Fachhochschule[a]	*546*	*460*	*-15,8*	*4,2*	*	*578*	*534*	*-7,5*	*3,9*	*	*716*	*588*	*-17,9*
Universität, wissen-schaftliche Hochschule, Kunsthochschule; inkl. Promotionsstudium	*1.369*	*1.592*	*16,3*	*2,8*	*	*1.398*	*1.619*	*15,8*	*3,4*	*	*1.605*	*1.779*	*10,8*
Insgesamt	1.941	2.073	6,8	2,4	*	2.002	2.171	8,4	2,4	*	2.353	2.385	1,4

a) 2008 und 2011 ohne Duale Hochschule Baden-Württemberg [2008 N = 23.013; 2011 N = 27.758].

Quellen und Erläuterungen (siehe auch Tab. A1):
- Statistisches Bundesamt (versch. Jahre): Fachserie 11, Reihe 4.1, Bildung und Kultur, Studierende an Hochschulen, Wintersemester 2007/2008, Wintersemester 2008/2009, Wintersemester 2011/2012; eigene Berechnungen.

Datenfusion: Theoretische Implikationen und praktische Umsetzung

Dr. Florian Meinfelder

Zusammenfassung

Datenfusionen gewinnen in der Praxis aus zweierlei Gründen an Bedeutung: Zum einen hat die Erkenntnis verstärkt eingesetzt, dass die Antwortbelastbarkeit von Befragten endlich ist und „Single Source"-Erhebungen – das heißt alle Informationen stammen aus einer Quelle – zu Problemen wie Unit Nonresponse Bias auf Grund von Teilnahmeverweigerung führen. Zum anderen erhöht sich kontinuierlich die Anzahl an ursprünglich voneinander unabhängigen Datenquellen, für die gemeinsame Analysewünsche bestehen. Dieser Artikel setzt sich daher mit der praktischen Umsetzung einer Datenfusion auseinander, wobei gleichzeitig ein Bezug zu Notation und Verfahren der statistischen Missing Data-Literatur hergestellt wird. Im Abschnitt zu Evaluierungsmöglichkeiten wird nochmals auf den aktuellen Stand der Forschung eingegangen und ein Gütekriterium vorgestellt, das auf der Annahme der bedingten Unabhängigkeit aufbaut, aber Vorzüge gegenüber den in der Praxis gängigen Gütekriterien aufweist.

Schlagwörter: Datenfusion, Statistical Matching, Nonresponse, Missing-data, Multiple Imputation, Missing at Random, Nearest Neighbour Imputation

Übersicht und Terminologie

Was ist eine Datenfusion? Es existieren unterschiedliche Definitionen dieses Begriffs, der vor allem in der Markt- und Medienforschung häufig gleichgesetzt wird mit dem „Verschmelzen" zweier (oder mehrerer) Datenquellen anhand von „Statistischen Zwillingen" (siehe z. B. Koschnick 1988: 189 ff.). Unserer Meinung nach ist diese Definition einerseits zu eng gefasst, da sie sich auf eine bestimmte Klasse von Methoden bezieht (Nearest-Neighbor-Matchingverfahren) und gleichzeitig nichts über die Analyseziele besagt. Wir folgen daher der in der Statistischen Literatur üblichen Definition, gemäß der Datenfusion ein spezifisches Datenausfallmuster bezeichnet, das entsteht, wenn (mindestens) zwei unabhängig entstandene Datenquellen „übereinandergestapelt" werden, sodass es eine Gruppe an Variablen (X) gibt, die in beiden Datenquellen vorkommen, sowie eine Gruppe an Variablen (Y), die nur in der ersten Datenquelle, und eine Gruppe an Variablen (Z), die nur in der zweiten Datenquelle existieren (vgl. Rässler 2002). Zudem bezieht sich das generelle Analyseziel auf die gemeinsame Verteilung von Y und Z – den Variablen, die nicht gemeinsam beobachtet wurden. Abbildung 1 beschreibt diese Datensituation. Auf die Verfahren, die eine gemeinsame Analyse von Variablen ermöglichen, die nicht gemeinsam beobachtet sind, wird nicht spezifisch eingegangen.

Ein konkretes Beispiel aus der Medienforschung soll unseren Standpunkt verdeutlichen: Daten des AGF/GfK Fernsehforschungspanels werden seit vielen Jahren mit Daten aus dem GfK ConsumerScan fusioniert und unter dem Namen t.o.m. (target optimizer for markets) vermarktet. Hierbei werden aus gemessenem Kaufverhalten gebildete Zielgruppen aus dem Consumer Tracking Panel der GfK (ConsumerScan) in das AGF/GfK Fernsehforschungspanel, der Datengrundlage für die offiziellen Einschaltquoten in Deutschland, übertragen. Ziel dieser Übertragung anhand von Variablen, die in beiden Panels vorliegen (X), ist die Schaffung einer Datengrundlage, in der gemessenes Kaufverhalten (Z) gemeinsam mit gemessener TV-Nutzung (Y) ausgewertet werden kann. Ursprünglich wurde diese Fusion durch ein Mahalanobis-Distanzmatching, später mittels Predictive Mean Matching (Rubin 1986; Little 1988) durchgeführt. In beiden Fällen handelt es sich um Nearest-Neighbor-Verfahren, doch ebenso wäre ein parametrisches Fusionsverfahren, beispielsweise über Regressionsschätzer, möglich.

Während das Datenausfallmuster designbasiert und somit nicht zufällig ist, ist der zugrundeliegende Datenausfallmechanismus im Idealfall absolut zufällig – nämlich dann, wenn die beteiligten Studien A und B Zufallsstichproben aus derselben Population sind. Der Datenausfallmechanismus wird in einem solchen Fall als Missing Completely at Random (MCAR) bezeichnet. Die Kategorisierung von Datenausfallmechanismen wurde zuerst von Rubin (1976) vorgestellt und später von Little und Rubin (1987) erweitert.

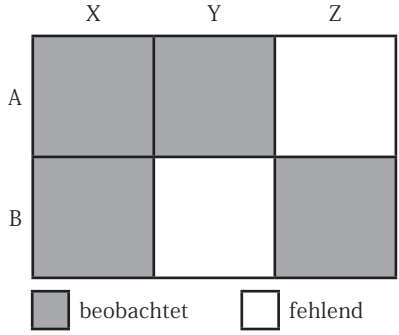


Abbildung 1: Schematische Darstellung einer Datenfusion als missing-by-design Datenausfallmuster

Zum Begriff der Datenfusion ist noch zu ergänzen, dass dieser oft fälschlicherweise gleichgesetzt wird mit Nearest-Neighbour-Algorithmen – der in der englischsprachigen Literatur häufig für „Datenfusion" verwendete Begriff Statistical Matching hat dazu beigetragen; allerdings handelt es sich hierbei nur um eine Gruppe zur Verfügung stehender Methoden, um die oben genannten Analyseziele zu erreichen.

Im Folgenden gehen wir zunächst auf zentrale Annahmen einer Datenfusion ein, die oft nicht explizit formuliert werden, aber dennoch notwendig sind. In den anschließenden Kapiteln wird ein (fiktives) Datenfusionsprojekt in der chronologischen Abfolge der Einzelschritte beschrieben, da viele Aspekte bei der Durchführung einer Fusion bisher kaum in der statistischen Literatur beleuchtet worden sind. Die Ausnahme hierzu bildet die Evaluierung der Güte beziehungsweise Aussagekraft einer Datenfusion, die den letzten Projektschritt darstellt und mit der dieser Artikel schließt.

Relevante Annahmen

Conditional Independence Assumption

Die zentrale Annahme einer Datenfusion ist die so genannte Conditional Independence Assumption, die besagt, dass, Y und Z gegeben, die gemeinsamen Merkmale X unabhängig sein müssen. Präziser ausgedrückt, besagt die Conditional Independence Assumption, dass die gemeinsame ‚künstliche' Verteilung für $(x, \tilde{y}, z)$ gegeben ist durch

$$\tilde{f}_{X,Y,Z}(x,\tilde{y},z) = f_{X,Z}(x,z) f_{Y|X}(y|x) = f_{X,Y,Z}(x,y,z) \frac{f_{Y|X}(y|x)}{f_{Y|X,Z}(y|x,z)}.$$

Die künstliche Fusions-Kovarianz zwischen Y und Z ist gegeben durch

$$\widehat{\mathrm{COV}}(Y,Z) = \mathrm{COV}(Y,Z) - \mathrm{E}((\mathrm{COV}(Y,Z|X)).$$

Die CIA besagt, dass dieser Erwartungswert gleich null ist (siehe Rässler 2002; d'Orazio/di Zio/Scanu 2006). Diese Annahme ist (so gut wie immer) nicht testbar, weswegen sich die Evaluierung einer Datenfusion in der Regel auf den Erhalt der auf die gemeinsamen Variablen bedingten Verteilung von Y vor und nach Fusion beschränkt. Es ist jedoch möglich, Grenzen für die Zusammenhänge zwischen Y und Z auf Grund deren jeweiliger Zusammenhänge zu den gemeinsamen Merkmalen X zu bestimmen.

Datenausfallmechanismus

Im einleitenden Kapitel wurde der Fall beschrieben, dass die beteiligten Stichproben Zufallsstichproben aus derselben Grundgesamtheit sind. In empirischen Datensituationen ist es jedoch häufig auf Grund von beispielsweise Teilnahmeverweigerung (allgemein *Unit Nonresponse*) so, dass die beteiligten Studien sich hinsichtlich ihrer Strukturen unterscheiden (siehe Abbildung 2). In diesem Falle wäre der Datenausfall nicht mehr MCAR. Auf Grund der Annahme der auf X bedingten Unabhängigkeit wird für den Ausfallmechanismus jedoch automatisch MAR angenommen, weshalb der Ausfall als ‚ignorierbar'[1] angesehen wird.

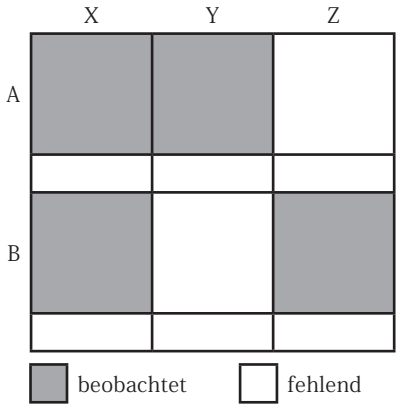


Abbildung 2: Schematische Darstellung einer Datenfusion als missing-by-design Datenausfallmuster unter Einbeziehung von Unit Nonresponse

1 Falls zusätzlich die schwächere Annahme der Distinctness erfüllt ist, die hier nicht thematisiert wird.

Konkret bedeutet dies, dass in Datenfusionssituationen die *Conditional Independence Assumption* die MAR-Annahme beinhaltet. An dieser Stelle sei noch darauf hingewiesen, dass die Annahme der bedingten Unabhängigkeit nicht bedeutet, dass Z (oder Y) perfekt durch X erklärt werden muss (obwohl in diesem Falle die CIA nachweislich erfüllt ist). Sollten beispielsweise Y und Z tatsächlich unabhängig voneinander sein, wäre die CIA ebenfalls erfüllt, ohne dass ein Bezug zur Erklärungskraft von X hergestellt werden muss.[2]

Umsetzung einer Datenfusion in der Praxis

Dieser Abschnitt behandelt Datenfusion als imaginäres Projekt, in dem in chronologischer Reihenfolge alle Teilschritte beschrieben werden. Hierbei sind zwei Situationen zu unterscheiden:

a) Die beteiligten Studien existieren bereits
b) Eine der beiden Studie ist noch in der Konzeptionsphase

In letzterem Falle kann noch Einfluss auf die Abfrage der gemeinsamen Variablen X genommen werden.

Anpassung gemeinsamer Merkmale

Fall a) ist der Regelfall und zieht einen sehr arbeitsaufwendigen Arbeitsschritt nach sich, der in Publikationen – zumindest im Kontext „Datenfusion" – bislang weitgehend ignoriert worden ist: Je nach Umfang der Studien ist der Aufwand der Identifizierung gemeinsamer Merkmale nicht zu unterschätzen. Und selbst wenn Variablen identifiziert wurden, deren Beschreibung scheinbar dasselbe misst, besteht dennoch die Möglichkeit sehr unterschiedlicher Kategorisierungen.

Wenn man die Problematik unterschiedlicher Variablennamen außer Acht lässt, d. h. wenn beispielsweise die Variable „Geschlecht der befragten Person" in Studie A ‚sex' und in Studie B ‚geschl' heißt, ist der einfachste Fall einer Anpassung der, in der nur ein oder mehrere Werte umkodiert werden müssen. Für obiges fiktives Beispiel wäre somit folgende Situation gegeben:

- Studie A: Variablenname ‚Geschl' mit 0='männlich' und 1='weiblich'
- Studie B: Variablenname ‚Sex' mit 1='männlich' und 2='weiblich'

Durch Rekodierung einer der beiden Variablen entsteht so aus einem gemeinsamen Merkmal eine gemeinsame Variable.

2 Auch wenn diese hypothetische Situation die Sinnhaftigkeit einer Datenfusion von Y und Z in Frage stellt.

Ein etwas komplexerer Fall ist eine Situation, in der die Skalenniveaus der Merkmale unterschiedlich sind. Dies kann man sich am Beispiel „persönliches Netto-Einkommen“ vor Augen führen, dass z. B. in Studie A offen und in Studie B in Klassen abgefragt wird. In diesem Fall ist die Angleichung der Information sehr einfach[3], da aus dem (quasi-)stetigen Einkommen in Studie A das kategorisierte Einkommen in Studie B nachgebildet werden kann.

Generell lässt sich diese Form der Anpassung als genestete Situation bezeichnen, die auch bei gleichskalierten Merkmalen auftreten kann, wenn die Abfrage in einer Studie detaillierter als in der anderen erfolgte.
Wir betrachten im folgenden Beispiel eine Abfrage für „Erwerbstätigkeit“

- Studie A: 1=‘Vollzeit erwerbstätig‘, 2=‘Teilzeit erwerbstätig‘, 3=‘nicht erwerbstätig‘
- Studie B: 1=‘Vollzeit erwerbstätig‘, 2=‘Teilzeit erwerbstätig‘, 3=‘arbeitslos‘, 4=‘ausschließlich hauswirtschaftliche Tätigkeit‘, 5=‘berufsunfähig bzw. in Rente/Pension‘

Während im vorangegangenen Fall die Nestung durch Transformation auf ein untergeordnetes Skalenniveau die Bildung einer gemeinsamen Variable ermöglichte, liegt hier eine Nestung auf Grund feinerer Kategorisierung vor, da die Kategorien 3 bis 5 in Studie B der weniger spezifischen Kategorie ‚nicht erwerbstätig‘ in Studie A entsprechen.

In den nachfolgenden Fällen ist eine Anpassung der Information nur unter Hinzuziehung zusätzlicher Annahmen möglich, da keine Nestung vorliegt.

Wenn in beiden Studien gefragt wird, wie viel Sport man treibt und die Studien unterschiedlich viele Kategorien hierfür ausweisen (z. B. unetikettierte Skalen von 0-10 in Studie A und von 1-6 in Studie B), ist eine sinnvolle Zusammenfassung ebenso wenig möglich, wie bei kategorialen Abfragen, bei denen die Kategorien nur Schnittmengen haben oder nicht alle Antwortmöglichkeiten vorgegeben sind, wodurch Messfehler (bewusstes Wählen einer unpassenden Kategorie, die der eigentlichen Antwortabsicht nahekommt) oder Nonresponse (Verweigern der Antwort) entstehen können. Angenommen, Familienstand wird in Studie A und B auf folgende Arten abgefragt:

- Studie A: 1=‘Single‘, 2=‘verheiratet‘, 3=‘geschieden‘, 4=‘verwitwet‘
- Studie B: 1=‘Single‘, 2=‘verheiratet‘, 3=‘unverheiratet, aber mit Partner zusammenlebend‘, 4=‘verheiratet, aber getrenntlebend‘, 5=‘geschieden‘, 6=‘verwitwet‘

Studie B ist in den Ausprägungen präziser, aber wir wissen nicht, wie sich Befragte in Studie A entschieden haben, auf die die Merkmalswerte 3 oder 4 aus Studie B zugetroffen hätten.

3 Oder auch nicht: Eine ambitionierte Variante wäre tatsächlich, das klassierte Einkommen zu Einzelbeobachtungen aufzulösen (vgl. Heitjan und Rubin 1992).

Während im ersten Fall eine Lösung das Auflösen der diskreten Information durch eine latente stetige Variable sein kann, was technisch aufwendig ist und daher hier nicht weiter thematisiert werden soll, ist im letzteren Fall das Verwenden eindeutiger Kategorien als Dummyvariablen möglich (hier konkret für ‚Single' und ‚verwitwet').

Generell ist zu sagen, dass gemeinsame Merkmale häufig noch keine gemeinsamen Variablen darstellen und der Weg zu dem in Abbildung 1 dargestellten „gestapelten" Datensatz aus diesem Grund sich als schwierig und oft mit zusätzlich notwendigen Annahmen darstellt.

Wahl des Fusionsalgorithmus

Fusionsverfahren

Der nächste Projektschritt bezieht sich – wie alle nachfolgenden auch – bereits auf die späteren Analysen: Die Auswahl der Fusionsmethode, wobei wir davon ausgehen, dass nur Y-Variablen in Studie B ergänzt werden („eine Fusionsrichtung"). Neben den bereits erwähnten Nearest-Neighbour-Algorithmen können auch rein verteilungs- bzw. modellbasierte Algorithmen zum Einsatz kommen. Zudem sollte feststehen, ob auf den fusionierten Daten nur deskriptiv oder auch induktiv gearbeitet werden soll. Im letzteren Falle wird ein Verfahren benötigt, dass die durch die Ergänzung (die Fusion) induzierte Unsicherheit korrekt abbildet, wie etwa die ‚Multiple Imputation' (MI), bei der fehlende Information m > 1-mal ergänzt wird (Rubin 1987).

Ein weiteres Unterscheidungskriterium ist die Anzahl der gleichzeitig ergänzten Variablen: Viele moderne Imputationsalgorithmen basieren auf chained equations-Ansätzen und ergänzen jede Variable bedingt auf die restlichen Variablen. Mit diesen Verfahren wird tendenziell der Zusammenhangserhalt zwischen zu übertragenden Y-Variablen und den gemeinsamen Variablen X optimiert. „Zeilenweise" (Nearest-Neighbour-)Verfahren werden in Fusionsszenarien jedoch häufiger verwendet, da bei einer simultanen Übertragung mehrerer Variablen gewährleistet ist, dass der Vektor der Ausprägungen in sich konsistent bleibt. Dies ist vor allem dann sinnvoll, wenn nicht nur bivariate Assoziationen zwischen einer Y- und einer Z-Variable in die Analyse einbezogen werden sollen, sondern multivariate Zusammenhänge aus mehreren (simultan) fusionierten Y-Variablen eine Rolle spielen.

Neben ‚klassischen' Fusionsmethoden wie dem Mahalanobis-Distanzmatching ist ein Verfahren, das dies gewährleistet, Predictive Mean Matching (PMM) in der von R.J.A. Little (1988) vorgestellten Variante. Die Distanz ergibt sich hierbei aus $D_{i,j} = (\hat{y}_i - \hat{y}_j)^T S_{Y|X}^{-1} (\hat{y}_i - \hat{y}_j)$, mit $i = 1, \ldots, n$ und $j = n+1, \ldots, n+m$, wobei die ersten n

Beobachtungen der Donorenstichprobe und die verbleibenden m Beobachtungen der Rezipientenstichprobe entstammen. $\hat{y}_i$ und $\hat{y}_j$ sind Vektoren der Länge k (die Prädiktoren von k gemeinsam zu übertragenden Variablen aus der Donorenstichprobe) und $S_{Y|X}^{-1}$ ist die Inverse der Varianz-Kovarianz-Matrix der Residuen einer Regression von Y auf X. Diese Matrix fungiert als Gewichtungsmatrix, indem Distanzen bei Prädiktoren von Y-Variablen, die nicht gut durch die gemeinsamen Variablen X erklärt werden können, einen geringeren Einfluss auf die Gesamtdistanz haben. Obwohl dieses Verfahren eine Variante der Mahalanobis-Distanz verwendet, gibt es einen entscheidenden Unterschied zu einem klassischen Mahalanobis-Distanzmatching: Die Distanzen werden über die spezifischen Variablen und nicht über die gemeinsamen Variablen gebildet. Möglich ist dies, da PMM ein semi-parametrisches Verfahren ist, das auf ein Modell für Y zurückgreift.

Neben inhaltlichen und methodischen Überlegungen kann bei der Wahl der Fusionsmethode auch der Rechenaufwand eine Rolle spielen: Eine große Anzahl gemeinsamer Variablen bei geringer Anzahl spezifischer Variablen Y, würde PMM einen Vorteil gegenüber Mahalanobis-Distanzmatching verleihen und umgekehrt. Ein weiterer Vorteil dieser Nearest-Neighbour-Verfahren in der praktischen Umsetzung ist, dass die Fusion mit verdichteten (z. B. durch Faktorenanalyse gebildeten) Hilfsvariablen durchgeführt werden kann, wenn die eigentlichen spezifischen Variablen sehr zahlreich und/oder komplex sind, und man anschließend als Ergebnis eine Paarliste mit den Identifikatoren der Studien erhält, die es theoretisch erlaubt, weitere Informationen zu übertragen, die nur indirekt in das Modell eingeflossen sind.

Ein weiteres Unterscheidungskriterium betrifft auf Matchingverfahren basierende Fusionsmethoden generell: Die Optimierung der Einzeldistanzen oder der Distanzsumme (bei „gleichmäßiger" Donorenverwendung). Ersteres wird in der Literatur als greedy matching bezeichnet und erlaubt entweder unrestringierte Verwendung der Donoren oder führt ansonsten zu tendenziell größeren Distanzen mit zunehmender Abnahme verbleibender Donoren; von dieser Variante ist jedoch generell abzuraten – insbesondere falls m > n, d. h. es liegen mehr Rezipienten als potentielle Donoren vor. Bei einer Optimierung der Distanzsumme wird Mehrfachverwendung unterbunden (zumindest, falls m < n) und die Matches so gebildet, dass die Summe der Einzeldistanzen möglichst gering ist. Diese Klasse an Verfahren hat gegenüber einem unrestringierten Matchingverfahren den Vorteil, dass „Falldoppelungen" nach Möglichkeit nicht auftauchen und somit die effektive Fallzahl der Stichprobe größer ist. Der Nachteil dieser Verfahren ist jedoch, dass sie die Strukturen aus der Donorenstichprobe auf die Rezipientenstichprobe übertragen, selbst wenn diese sich unterscheiden (siehe Abbildung 2 und die Diskussion zu MAR in Kapitel 2). Zudem führt – trotz der Bezeichnung „Distanzsummenoptimierung" – eine Optimierung der Einzeldis-

tanzen bei zugelassener Mehrfachverwendung zur minimalen Distanzsumme, wodurch auch Zusammenhänge zwischen den gemeinsamen und den fusionierten Variablen besser erhalten bleiben. Verfahren mit Distanzsummenoptimierung wie Optimal Matching (Hansen 2004) sind daher vor allem zu empfehlen, wenn Donoren- und Rezipientenstichprobe Zufallsstichproben aus der gleichen Grundgesamtheit sind und mehr Donoren als Rezipienten zur Verfügung stehen.

Propensity Score-Matching als ungeeignete Fusionsmethode

Neben den oben genannten Kriterien sollte bei der Wahl des Algorithmus vor allem beachtet werden, dass „Matchingalgorithmus" nicht mit „Imputationsverfahren" gleichzusetzen ist. Insbesondere die zunehmende Popularität des Rubin Causal Model (Rubin 1974) hat Propensity Score-Matching-Verfahren in den Vordergrund gerückt (Rosenbaum und Rubin 1983), die für Datenfusionsproblemstellungen genutzt werden, aber gänzlich ungeeignet sind, da sie für den Zweck entwickelt wurden, strukturgleiche Stichproben zu erzeugen. Grob erklärt wird bei diesen Verfahren zunächst eine Zugehörigkeitswahrscheinlichkeit („Propensity") zu einer Gruppe versus einer zweiten modelliert. Anschließend werden Beobachtungen aus dieser zweiten Gruppe anhand des Propensity Scores (typischerweise durch ein binomiales verallgemeinertes lineares Modell geschätzt) einem nächsten Nachbarn aus der ersten Gruppe zugewiesen. In Problemstellungen zu Kausalinferenz handelt es sich bei den beiden Gruppen um eine Testgruppe, die ein bestimmtes „Treatment" erhalten hat und eine Kontrollgruppe. Da bei Beobachtungsstudien die Zuweisung des „Treatments" nicht zufällig erfolgt, versucht man über das Propensity Score Matching den Einfluss von anderen Variablen (so genannten „Confounders") zu eliminieren. Würde man das Verfahren auf Daten aus einem Zufallsexperiment anwenden, hätte das Modell keinerlei Erklärungskraft und das Propensity Score Matching entspräche einer zufälligen Zuweisung der „Matched Controls" zu den Rezipienten (mit „Treatment"). Dies entspricht exakt der Situation, wie man sie bei einer Datenfusion vorfindet, in der beide Studien Zufallsstichproben der gleichen Grundgesamtheit sind und der Datenausfallmechanismus somit MCAR ist. Die Fälle der Donorenstichprobe nehmen die Rolle der Beobachtungen der Kontrollgruppe ein, die Fälle der Rezipientenstichprobe die der Beobachtungen der Testgruppe, und die gemeinsamen Variablen entsprechen den „Confounders" im Propensity Score-Modell (das auf Grund der identischen Stichprobenstrukturen auch hier keinerlei Erklärungskraft hat). Die zufällige Zuweisung spielt bei einer Kausalanalyse keine Rolle, aber bei einer Fusion ist sie fatal, da es ja genau die Zusammenhänge zwischen den spezifischen Merkmalen sind, die von Auswertungsinteresse sind; bei Verwendung eines Propensity Score Matchings sind Y und Z in einer solchen Situation jedoch unabhängig voneinander.

Segmentierung

Um zudem sicherzustellen, dass die „gespendete" Information des „Donors" zu der des „empfangenden" „Rezipienten" konsistent ist, kann die Fusion in Segmenten (Unterstichproben) vollzogen werden. Aus der Perspektive eines Matchingverfahrens, bei dem über die gemeinsamen Merkmale gematcht wird, ist dies gleichbedeutend mit einem unendlich hohen Gewicht für die in die Segmentierung einbezogenen X-Variablen, da bei diesen eine perfekte Zuordnung erzwungen wird. Auch wenn eine Segmentierung aus Konsistenzgründen sinnvoll sein kann und den Rechenaufwand bei matchingbasierten Fusionsmethoden verringert, sollte auf ein übermäßiges Segmentieren verzichtet werden, da es einen Eingriff in die vom Algorithmus auf Gesamtebene optimierte Lösung bedeutet. Im Beispiel des Mahalanobisdistanz-Matchings ist die Segmentierung gleichbedeutend mit der Vergabe eines unendlich hohen Gewichts für die in die Segmentierung einbezogenen X-Variablen.

Fusionsdurchführung

Bei der Fusionsdurchführung selbst ist im Grunde nur zu unterscheiden, ob Y (bzw. Z) oder beide Gruppen Y und Z der spezifischen Merkmale ergänzt werden, ob die Fusion mit einem MI-Verfahren durchgeführt wurde und ob ein (zeilenweises) Nearest-Neighbour-Verfahren zum Einsatz kam, da diese Faktoren auf die Datenhaltung Einfluss haben. Beispielsweise wäre das konkrete Ergebnis einer Fusion mit Nearest-Neighbour-Verfahren, bei der multipel in beide Richtungen ergänzt wird, $m \times 2$ Paarlisten. Über diese Paarlisten können – wie bereits im vorangegangenen Abschnitt erwähnt – weitere Variablen übertragen werden, die im Fusionsmodell selbst keine Rolle gespielt haben. Diese mit Nearest-Neighbour-Verfahren mögliche „passive" Datenfusion ist insbesondere bei der Fusion von komplexen Datenstrukturen sinnvoll. Beispielsweise lassen sich die sekundengenau gemessenen Umschaltvorgänge im AGF/GfK Fernsehpanel zu Aggregaten verdichten (z. B. durchschnittliche Gesamt-TV-Nutzung in Sekunden pro Tag über einen bestimmten Zeitraum errechnet), die einfacher in die Modellierung einbezogen werden können als die Rohdaten. Mit einer rein modellbasierten Fusionsmethode (z. B. Regression) wäre dies nicht möglich. Des Weiteren lassen sich über diese Paarlisten im Falle einer Fusion von Paneldaten in eine andere Stichprobe auch aktualisierte Daten ohne Durchführung einer erneuten Fusion überführen – sofern sich alle Donoren der Vorperiode noch im Panel befinden. Ansonsten besteht die Möglichkeit einer kompletten Neufusion oder einer Fusion für die Rezipienten, deren Donor sich nicht mehr in der Stichprobe befindet. Ist die Rezipientenstudie ebenfalls ein Panel, so muss dort umgekehrt der Fall berücksichtigt werden, dass Neurekrutrierungen noch keinen Donor haben.

Evaluierung

Die abschließende Evaluierung einer Datenfusion ist der projektchronologisch gesehen letzte, aber wichtigste Schritt, zu welchem auch die meisten Publikationen vorliegen. Die wichtigsten wissenschaftlichen Beiträge beschäftigen sich mit einer möglichen Verletzung der Conditional Independence Assumption (siehe beispielsweise Moriarty und Scheuren 2001) und deren Auswirkungen auf die Verteilung von $(\tilde{y}, z)$. Kiesl und Rässler (2009) haben eine Lösung für die Berechnung der Korrelationsgrenzen im multivariaten Fall entwickelt (wobei entweder Y oder Z ein Vektor sein muss), für die die Varianz-Kovarianz zwischen X, Y und Z positiv definit ist. In einem einfachen Beispiel ausgedrückt bedeutet dies, wenn X und Y beziehungsweise X und Z jeweils stark positiv miteinander korrelieren, dann kann die echte (unbekannte) Korrelation zwischen Y und Z nicht negativ sein – und ebenso wenig perfekt positiv. In der Praxis werden diese Korrelationsbandbreiten aus unterschiedlichen Gründen kaum verwendet: Neben der derzeitig nur für metrische Variablen verfügbaren Lösung ist unseres Erachtens das größte Problem, dass die Bandbreiten bei in empirischen Studien üblichen Korrelationen sehr groß ausfallen, so dass die Null fast immer enthalten ist und es schwerfällt einzugestehen, dass man nicht einmal feststellen kann, ob der (lineare) Zusammenhang zwischen Y und Z in Wahrheit positiv oder negativ ist.

In der Praxis beschränkt man sich daher vorwiegend auf Evaluierungen, die von einem Zutreffen der Annahme der bedingten Unabhängigkeit ausgehen. Die meisten dieser Evaluierungsmethoden betrachten den Verteilungserhalt von $\tilde{y}|x$ im Vergleich zu $y|x$, obwohl dies nicht dem Analyseziel entspricht und im Falle unterschiedlicher Stichprobenstrukturen (MAR-Mechanismus) sogar grundsätzlich falsch ist. Die häufig verwendeten t- und χ^2-Tests, die zur Evaluierung von empirischen Datenfusionen durchgeführt werden, haben zudem einen noch viel schwerwiegenderen Makel. Bei den t-Tests werden auf Ausprägungen gemeinsamer Variablen bedingte Mittelwerte von spezifischen Variablen der Donorenstichprobe mit den Pendants der fusionierten Stichprobe verglichen, bei den χ^2-Tests gemeinsame relative Häufigkeiten von gemeinsamen und (kategorisierten) spezifischen Variablen der Donorenstichprobe mit denen der fusionierten Rezipientenstichprobe. In beiden Fällen ist die den Tests zu Grunde liegende Nullhypothese, dass es keinen Unterschied zwischen den Mittelwerten (Zellhäufigkeiten) der Donoren- und der fusionierten Stichprobe gibt. Die Fusionsdurchführung wird als erfolgreich angesehen, wenn der Anteil abgelehnter (signifikant unterschiedlicher) Tests bei unter $\alpha \times 100\%$ liegt. Die Tests sind natürlich nicht voneinander unabhängig, aber das weitaus größere Problem ist die falsch spezifizierte Nullhypothese, die das „Wunschergebnis" (kein Unterschied vor und nach Fusion) enthält. Die Essenz der von Jerzy Neyman (1937) entwickelten Hypothesentests ist, dass gegen ein Ereignis (formuliert durch die Nullhypo-

these) getestet wird mit dem Ziel, dieses Ereignis mit einer vorab definierten Irrtumswahrscheinlichkeit ablehnen zu können. Ziel dieser Tests ist es jedoch, die Nullhypothese nicht abzulehnen (und wird leider in der Praxis fälschlicherweise mit einem Annehmen der Nullhypothese gleichgesetzt); diese Umkehrung des Testkonzepts führt dazu, dass Unsicherheit auf Grund kleiner Fallzahlen oder großer Varianzen „belohnt" wird, da in diesen Fällen die Nullhypothese mit geringerer Wahrscheinlichkeit abgelehnt wird. Dies ist nicht nur wissenschaftstheoretisch nicht erwünscht, sondern bietet auch Möglichkeiten zur Manipulation, etwa wenn bei einem zu hohen Anteil an Ablehnungen ein erneutes Testen mit zusätzlichem Bedingen auf die Segmente erfolgt.[4]

Wir schlagen als sinnvollere Alternative zu einem Gütemaß, das aus einer unzulässigen Aggregation von Einzeltests und einer „verkehrtherum" spezifizierten Nullhypothese besteht, grafische Diagnostiken auf Basis von bivariaten Korrelationen vor. Diese sind bei unterschiedlichen Stichprobenstrukturen robuster als bedingte Mittelwertsvergleiche und geben gleichzeitig auch etwas Auskunft bezüglich der Erklärungskraft der gemeinsamen Merkmale.

Die Abbildungen 3a und 3b sind Beispiele für eine derartige Diagnostik. Die erste Abbildung enthält einfach ein Histogramm der Differenzen von bivariaten absoluten Korrelationen zwischen gemeinsamen und spezifischen Merkmalen in der Donorenstichprobe zu den entsprechenden Korrelationen in der fusionierten Stichprobe. Beträgt z. B. die Korrelation zwischen der gemeinsamen Variable X_k und der spezifischen Variable Y_l in der Donorenstudie 0,3 und die gleiche Korrelation in der fusionierten Stichprobe 0,1, so wird die Differenz von 0,2 im Histogramm festgehalten. Ein positiver Mittelwert würde somit bedeuten, dass die Originalkorrelationen durchschnittlich größer ausfallen. In Abbildung 3b sind die Einzelkorrelationen abgetragen – sortiert nach absoluter Abweichung, die durch die gestrichelte Diagonallinie dargestellt wird.

Wenn man die (absoluten) Korrelationen zwischen gemeinsamen und spezifischen Variablen vor und nach Fusion in einem Streudiagramm abträgt, erhält man einen zusätzlichen Eindruck vom Erhalt der Zusammenhänge zwischen gemeinsamen und spezifischen Merkmalen (siehe Abbildung 4).

Zudem kann man nun das Ganze zu einem Gütemaß zusammenfassen, wenn man eine entsprechende lineare Regression (durch den Ursprung) anpasst und das R^2 hierfür ausweist.

4 Durch die dadurch verringerten Fallzahlen fallen die Konfidenzintervalle breiter aus und die Wahrscheinlichkeit für signifikante Differenzen sinkt.

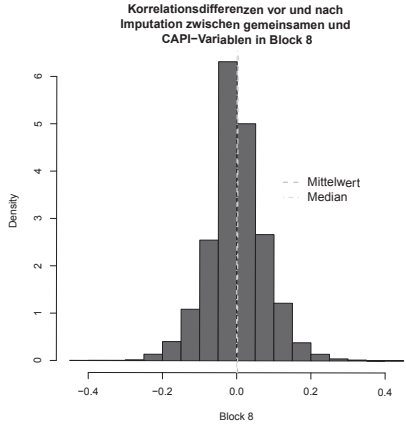


Abbildung 3a: Histogramm mit den Differenzen der absoluten Korrelationen

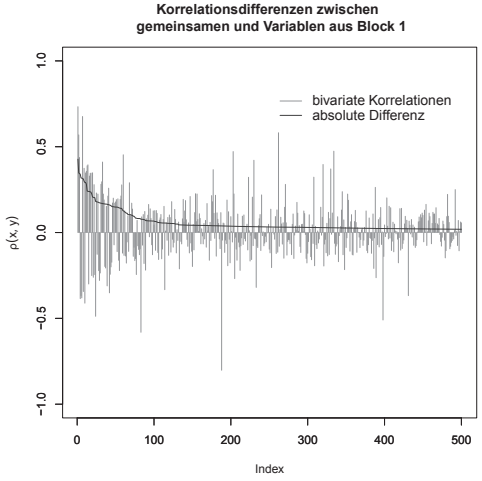


Abbildung 3b: Korrelationen sortiert nach absoluter Abweichung

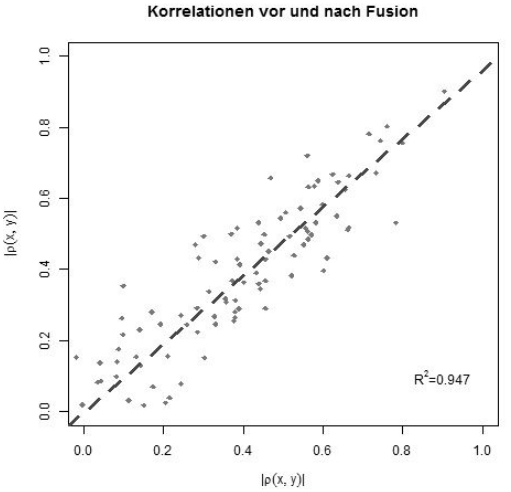


Abbildung 4: Streudiagramm und OLS-Regressionslinie der Korrelationen vor und nach Fusion

Wie bereits erwähnt, testet diese Art der Fusionsevaluierung nicht die Gültigkeit der Conditional Independence Assumption. Eine näherungsweise simulative Überprüfung der CIA ist aber zumindest teilweise möglich, wenn z. B. in der Rezipientenstichprobe eine oder mehrere Variablen existieren, die inhaltlich spezifischen Variablen der Donorenstichprobe ähneln – etwa eingefragte statt gemessener TV-Nutzung als „Proxyvariable".

Wenn man nun nur in dieser Rezipientenstichprobe durch Löschung einzelner Beobachtungen das Datenausfallmuster einer Datenfusion imitiert und anschließend ergänzt, lässt sich ein Vergleich für den Zusammenhang zwischen Z-Variablen und Proxyvariable (als Ersatz für Y) vor Löschung gegenüber nach Ergänzung in der künstlichen Rezipientenstichprobe ermitteln. Für diese Art Test hat Friedrich Wendt den Begriff „Split-half-Experiment" geprägt. Allerdings bleibt hierbei unberücksichtigt, dass die Richtgröße aus der Stichprobe vor Löschung ebenfalls nur ein Schätzer ist und die Interpretation von Unterschieden nicht in eine systematische Komponente und den Stichprobenfehler trennbar ist.

Zusammenfassung

Datenfusionen wurden in Deutschland zwar bereits in den 1970-er Jahren von Friedrich Wendt durchgeführt, allerdings hat erst die Verfügbarkeit von notwendiger Rechenleistung sowie die Verfügbarkeit von Imputations-, Matching- und Data Mining-Algorithmen zu einem großen Popularitätsschub geführt. Wie in der Einleitung bereits angedeutet, ist der Begriff „Datenfusion" diffus und bezieht sich bisweilen eher auf die Methode, als dass er der eingangs gegebenen Definition entspricht.

Die nach unserem Wissen ersten größeren, kontinuierlichen Datenfusionsprojekte in Deutschland stammen aus der kommerziellen Medienforschung, da dort der Wunsch nach einer Datenquelle, die alle relevanten Medien abbildet, auf die Eigenständigkeit der einzelnen „Währungsstudien"[5] sowie medienpolitische Widerstände trifft.

Auf Grund der Methodenvielfalt und unterschiedlichen Auswertungsziele ist es schwierig, ein Verfahren als universell am geeignetsten zu bezeichnen. In Kapitel drei wurde erläutert, nach welchen Kriterien eine Einteilung der Verfahren möglich ist. Die seit langem eingesetzten Matchingverfahren werden weiterhin eine wichtige Rolle spielen, allerdings ist davon auszugehen, dass diese zukünftig verstärkt aus der Perspektive der Statistik-Literatur zu fehlenden Daten untersucht und somit auch die Erkenntnisse im Bereich der Inferenz von unvollständigen Daten zunehmend eine Rolle spielen werden. Ein erster Schritt ist es zu erkennen, dass Datenfusion eine Datenergänzung und der durch die Ergänzung induzierten Unsicherheit durch größere Konfidenzintervalle Rechnung zu tragen ist. Wir gehen daher davon aus, dass varianzkorrigierende Verfahren wie Multiple Imputation zukünftig häufiger in empirischen Datenfusionsprojekten zu finden sein werden.

Der schwierigste Transfer aus der Wissenschaft in die Anwendung ist unserer Meinung nach der Umgang mit der Annahme der bedingten Unabhängigkeit, da insbesondere in Surveydaten die Zusammenhänge oft nicht ausreichend sind, um die Bandbreiten der möglichen Korrelationen zwischen Y und Z auf ein für die Praxis befriedigendes Niveau zu bringen. Hier könnten zukünftig Bayesianische Ansätze Abhilfe schaffen, über die Hilfsinformation bezüglich der gemeinsamen Verteilung von Y und Z in die Modellierung aufgenommen wird, die in den Datenmatrizen selbst nicht abbildbar ist.

5 Der Begriff wird typischerweise für die Studie gewählt, auf deren Basis Preise für Werbung im jeweiligen Medium ausgehandelt werden.

Literatur

D'Orazio, M./Di Zio, M. and Scanu, M. (2006): Statistical Matching – Theory and Practice. New York: John Wiley & Sons.

Hansen, B. B. (2004): Full Matching in an Observational Study of Coaching for the SAT. Journal of the American Statistical Association 99, 609–618.

Kiesl, H. and Rässler, S. (2009): How Valid Can Data Fusion Be? Journal of Official Statistics. In Erscheinung.

Koschnick, W. J. (1988): Standard-Lexikon für Mediaplanung und Mediaforschung. München: Saur K.G. Verlag GmbH.

Little, R. J. A. (1988): Missing-Data Adjustments in Large Surveys. Journal of Business and Economic Statistics 6, 287-310.

Little, R. J. A. and Rubin, D. B. (1987): Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Moriarity, C and Scheuren, F. (2001): Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. Journal of Official Statistics 17, 407-422.

Neyman, J. (1937): Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 236, 333–380.

Rässler, S. (2002): Statistical Matching – A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 70, 41-55.

Rubin, D. B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology 66, 688-701.

Rubin, D. B. (1976): Inference and Missing Data. Biometrika 63, 581-591.

Rubin, D. B. (1986): Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Econometric Statistics 4, 87-95.

Rubin, D. B. (1987): Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Wie genau sind die Kreisergebnisse des Mikrozensus – Einsatzmöglichkeiten von Small Area-Verfahren

Prof. Dr. Ralf Münnich
Jan Pablo Burgard
Thomas Zimmermann

Zusammenfassung

Der zunehmende Informationsbedarf an regionalen Statistiken und Indikatorwerten durch Politik, Verbände und Gesellschaft verursacht einen immensen Druck auf den statistischen Produktionsprozess der amtlichen Statistik. Bei immer geringeren Budgets sollen immer genauere Statistiken erzeugt werden, welche als Grundlage für politische und ökonomische Entscheidungsprozesse dienen. Gerade bei nichtlinearen Indikatoren, wie sie etwa in der Armutsmessung verwendet werden, wirkt sich die Qualität der Daten in besonderer Weise aus. In den letzten Jahren wurde daher immer mehr nach modernen statistischen Methoden verlangt, welche mit Unterstützung von geeigneten Modellierungen zu effizienteren Schätzungen führen können. In der vorliegenden Arbeit sollen die Probleme klassischer Methoden am Beispiel der Armutsmessung dargestellt werden. Ferner sollen Möglichkeiten zur Verwendung von sogenannten Small Area-Methoden dargelegt werden, mit deren Hilfe genauere Schätzungen für Indikatoren ermöglicht werden sollen.

Einleitung

Politik, Verbände und Wirtschaft verlangen immer mehr nach regional und inhaltlich tief untergliederten Indikatorwerten. Geradezu als selbstverständlich wird eine hohe Qualität der Statistiken vorausgesetzt, mit deren Hilfe unter anderem regionale Fördermaßnahmen begründet werden. Vielfach werden diese Daten oder Statistiken jedoch mittels Stichprobenerhebungen gewonnen, so dass die Genauigkeit der Ergebnisse durch den Datenerhebungsprozess stark beeinflusst wird. Zusätzlich kommt erschwerend dazu, dass das Stichprobendesign, zumindest in der amtlichen Statistik, auf den eigentlichen Verwendungszweck zugeschnitten ist, der in der Regel tief untergliederte Auswertungen nicht umfasst. Stichprobendesigns, die nicht für tief untergliederte Auswertungen gedacht sind, berücksichtigen im Allgemeinen nicht die Aufteilung des Stichprobenumfangs auf die untergliederten Subgruppen. Dies führt dazu, dass in diesen Untergliederungen mit designbasierten Schätzern oft keine hinreichend genauen statistischen Aussagen getroffen werden können. Zusätzlich erschwerend ist, dass die Subgruppen nicht immer a priori festgelegt sind, so dass sich sogar zufällige Teilstichprobenumfänge ergeben können, welche teilweise ex-trem gering sind. Auch hier können dann wiederum mit den designbasierten Schätzern keine präzise Aussagen getroffen werden.

Einen möglichen Ausweg bieten hierbei modellbasierte Methoden. Diese ermitteln die Schätzwerte über eine Prädiktion aus einem geschätzten Modell heraus. Allerdings werden dabei für die Modellierung zum Teil sehr detaillierte Registerinformationen gebraucht. Für die Prädiktion von nicht linearen Statistiken, wie zum Beispiel der meisten Armutsmaße, sind sogar personen- oder haushaltsspezifische Registerinformationen notwendig. Weiterhin lässt sich die Genauigkeit der Schätzung durch mit der abhängigen Variable möglichst hoch korrelierte Register-Kovariablen erhöhen, sodass für ein breites Spektrum an interessierenden Variablen ein ebenso breites Spektrum an Register-Kovariablen hilfreich ist.

Bei einer Neukonstruktion der Haushaltsstichproben sollten diese Schwierigkeiten der Stichprobenaufteilung und der Verfügbarkeit von Register-Kovariablen unbedingt berücksichtigt werden. Neben diesen Aspekten muss bei einer Stichprobendesign-Optimierung auch noch die klassische Längsschnittauswertung im Mikrozensus in Form eines Rotationspanels berücksichtigt werden. Ebenso sind möglicherweise auch noch die Koordination verschiedener Teilstichproben in Quer- und Längsschnitt zu beachten. Aktuelle Forschungsarbeiten zeigen aber, dass Stichproben-Optimierungen bei designbasierten und modellbasierten Verfahren unterschiedliche Ergebnisse liefern.

In dieser Arbeit sollen kurz die Probleme und Chancen bei der Implementierung modellbasierter Verfahren, zu denen auch Small Area-Verfahren gehören,

dargelegt werden. Besonders aktuell in der Diskussion ist hierbei eine regionale Darstellung der Armut, welche auch hier als Beispiel dient.

Qualität von Schätzergebnissen auf Kreisniveau

Seit langem und doch aktuell wird diskutiert, ob im deutschen Mikrozensus verlässliche Aussagen über Teilpopulationen getroffen werden können, die auf oder gar unterhalb der Kreisebene liegen. In diesem Abschnitt soll dargestellt werden, inwieweit aktuelle Haushaltsstichproben in der Lage sind, qualitativ hochwertige Ergebnisse auf Kreisebene zu liefern. Dies wird beispielhaft anhand der Schätzung der Armutsgefährdungsquote eruiert. Dieser Indikator spielt im Rahmen der EU-Armutsberichterstattung eine zentrale Rolle (siehe bspw. Destatis 2012, sowie die Berichte zum Forschungsprojekt AMELI auf http://ameli.surveystatistics.net), und ist definiert als der Anteil der Personen deren Haushaltsnettoäquivalenzeinkommen unterhalb der Armutsschranke liegt. Die Armutschranke wiederum wird auf 60% des Medianeinkommens gesetzt.

In Deutschland gibt es prinzipiell verschiedene Datenquellen, welche zur Armutsmessung herangezogen werden können, wie etwa der Mikrozensus, SILC oder SOEP. Da die wahren Werte Deutschlands nicht vorliegen, wird die nachfolgende exemplarische Analyse auf Basis eines teilsynthetischen Datensatzes durchgeführt. Dieser basiert auf den Daten des Zensus-Stichprobenforschungsprojekts (siehe Münnich et al. 2012) und wurde zu diesem Zweck um Einkommensvariablen mit Hilfe statistischer Modelle erweitert. Dieser Datensatz ist personenscharf, allerdings werden hier, gemäß einer Haushaltsstichprobe, ganze Haushalte gezogen, von denen es in Deutschland etwa 40 Mio. gibt. Die größte Haushaltsstichprobe in Deutschland ist der Mikrozensus, in welchem 1% der Haushalte für die Stichprobe ausgewählt werden. Der Auswahlsatz, der für die EU-Armutsmessung herangezogen wird, D-SILC (Statistics on Income and Living Conditions), liegt demgegenüber nur bei etwa 0,036% der Haushalte, also etwa 1/30 der Mikrozensus-Stichprobe.

Im Rahmen einer kleinen Simulationsrechnung wurde in Anlehnung an das Mikrozensus-Stichprobenschema eine Stichprobe vom Umfang 0,1% in Rheinland-Pfalz gezogen. Anschließend wurde aus den Daten die Armutsgefährdungsquote geschätzt. Auf eine eingehende Darstellung der Schätzmethodik wird an dieser Stelle verzichtet und auf die eingehenden Ausführungen im AMELI-Projekt, insbesondere Workpackage 3, verwiesen (http://ameli.surveystatistics.net).

In der nachstehenden Abbildung sind die Schätzwerte der Armutsgefährdungsquote für die Kreise in Rheinland-Pfalz für zwei zufällig ausgewählte Stichproben ausgewiesen.

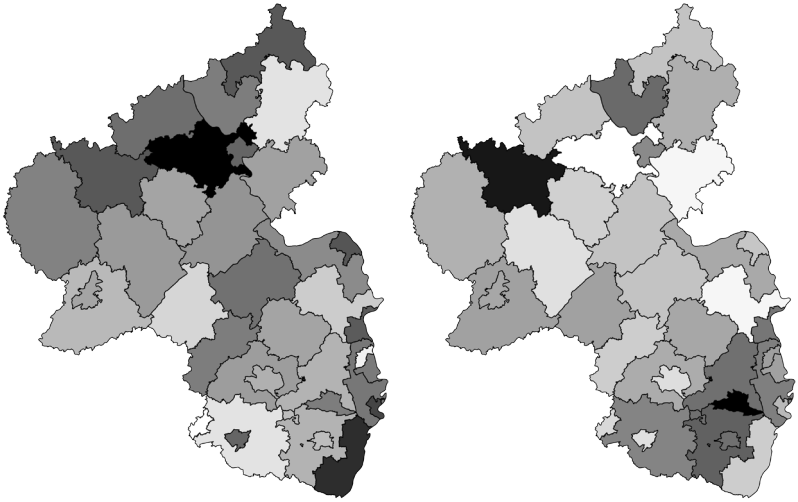


Abbildung 1: Armutsgefährdungsquoten in Kreisen in Rheinland-Pfalz geschätzt aus zwei zufällig gewählten Stichproben

Da es sich um eine teilsynthetische, aber trotzdem realitätsnahe Simulation handelt, wurde auf eine genaue Skalierung verzichtet. Die unterschiedliche Farbintensität kennzeichnen unterschiedliche Armutsgefährdungsquoten. Man erkennt sehr deutlich, dass sich bei unterschiedlichen Stichproben auf Basis von Kreisauswertungen erhebliche Unterschiede in den Armutsquoten ergeben. Gerade in der Pfalz resultieren zum Teil gravierend unterschiedliche Rangordnungen von arm und nicht arm. Dabei dominiert der reine Stichprobeneffekt die Rangfolge der Kreise in Bezug auf die Armustgefährdungsquote. Mit anderen Worten: Es ist hauptsächlich von der realisierten Stichprobe abhängig, ob ein Kreis einen höheren oder niedrigeren Schätzwert für die Armustgefährdungsquote aufweißt und nicht von den wahren Begebenheiten. Bei anderen Paaren von Stichproben hätten ebenso andere Regionen als die Pfalz in den Vordergrund rücken können. Derart dominante Stichprobeneffekte resultieren sehr schnell, wenn die für die Regionen zugeordneten Teilstichprobenumfänge zu gering werden. Ergebnisse mit derartig hohen Stichprobenfehlern qualifizieren sich sicher für keinerlei Veröffentlichung oder gar als Grundlage für politische Entscheidungsprozesse. Im Gegensatz zum Mikrozensus wurde in diesem Beispiel von einer präzisen Angabe des Einkommens ausgegangen und nicht von klassierten Daten. Die Klassierung der Einkommen würde die Schätzung der Armutsgefährdungsquoten erheblich verkomplizieren und in der Regel zu noch stärker variierenden Schätzwerten führen.

Aus designbasierter Perspektive ist der Stichprobenumfang die Hauptstellschraube zur Verbesserung der Qualität der Schätzung. Eine Erhöhung der (Teil-) Stichprobenumfänge wirkt sich jedoch unmittelbar auf die Kosten der Erhebung und die Belastung der Bürger aus. Es kann gezeigt werden, dass die Stichprobenvarianz umgekehrt proportional zum Stichprobenumfang ist, die relative Schwankungbreite sogar nur zur Wurzel des Stichprobenumfangs. Das bedeutet, dass eine Halbierung der Schwankungsbreite, welche üblicherweise mit Hilfe von Konfidenzintervallen dargestellt wird, eine Verfierfachung des Stichprobenumfangs erfordert.

Nachfolgend sollen Möglichkeiten dargestellt werden, wie mit modernen statistischen Methoden möglicherweise derartig hohe und damit teure Stichprobenumfänge vermieden werden können.

Einsatzmöglichkeiten von Small Area-Verfahren

In diesem Abschnitt sollen die grundlegenden Verfahren der Small Area-Statistik kritisch gewürdigt werden. Da der Fokus auf Einsatzmöglichkeiten dieser Verfahren liegt, wird auf eine formale Darstellung der Verfahren verzichtet. Eine deutschsprachige Übersicht über die Verfahren der Small Area-Statistik wurde in Münnich at al. (2013) veröffentlicht. Ausgehend von einem designbasierten, also Amtsstatistik-nahen Ansatz, der auch im letzten Abschnitt verwendet wurde, werden die verschiedenen Verfahren der Small Area-Statistik unter Verwendung der aktuellen Zensus-Diskussion vorgestellt. In diesem ausführlichen Artikel wird auch ein Überblick über die aktuelle Literatur gegeben. Das Standardwerk der Small Area-Statistik, Rao (2003), wird 2014 in der zweiten, stark überarbeiteten Auflage erscheinen. Dies zeigt die aktuelle Bedeutung der Small Area-Statistik und ihre rasche Entwicklung in den letzten Jahren.

Klassische Small Area-Schätzungen

Ein wesentlicher Grund Small Area-Methoden anzuwenden entsteht aus der zuvor bereits erörterten Problematik, dass bei der Aufteilung eines Gesamtstichprobenumfangs, etwa der deutschen SILC-Stichprobe, bei der 2011 13.512 Haushalte befragt wurden (siehe Destatis, 2012), auf über 400 Kreise bzw. kreisfreie Städte im Durchschnitt knapp 34 Haushalte gezogen werden würden. Ein Einsatz designbasierter Methoden ist bei derartig kleinen Stichprobenumfängen nicht mehr sinnvoll, da diese ausschließlich die in der jeweiligen Teilpopulation ermittelten Informationen der Stichprobe verwenden können. Auch die Verwendung von Randinformationen aus Registern, wie beim Regressionsschätzer, liefert selten

akzeptable Ergebnisse. Insbesondere bei der Schätzung von Armutsmaßen oder anderen (nichtlinearen) Indikatoren liefern die designbasierten Schätzverfahren kaum noch reliable Schätzwerte.

Im Allgemeinen kann davon ausgegangen werden, dass bestimmte Zusammenhänge zwischen Variablen nicht nur in einem Kreis Gültigkeit besitzen, sondern generell in der Population auf ähnliche Weise vorliegen. Es entsteht nun die Frage, inwieweit die Informationen aller oder zumindest vieler Regionen zu einer Stabilisierung der Schätzungen in den einzelnen Regionen beitragen können. Dieses Vorgehen wird borrowing strength genannt, also das Borgen von Informationen zur Verbesserung der eigentlichen Schätzungen.

In der Small Area-Statistik dominieren zwei Grundmodelle. Beide basieren auf der Schätzung eines Random Intercept-Modells. Dies ist im Prinzip ein klassisches Regressionsmodell, bei dem ein weiterer Zufallsterm die zufällige Abweichung der Area-Achsenabschnitte vom Gesamt-Achsenabschnitt angibt. Beide Modelle unterscheiden sich dabei in Form der verwendeten Information:

1. Das Modell von Fay und Herriot (1979) verwendet aggregierte Informationen zur Modellschätzung, also die Durchsschnittswerte der Areas. Somit werden bei der Schätzung keinerlei Individualinformation der Einheiten benötigt. Dies erweist sich gerade bei Geheimhaltungsproblemen als vorteilhaft. Allerdings muss die Unterschiedlichkeit der Areas über die Informationen der Area-spezifischen Mittelwerte ableitbar sein.
2. Im Gegensatz dazu verwendet das Modell von Battese et al. (1988) Individualinformationen.

Die Hochrechnung erfolgt dann in beiden Fällen über $\hat{\mu}_d = \overline{X}_d\hat{\beta} + \hat{u}_d$ (1).
Dabei ist $\hat{\mu}_d$ der geschätzte Mittelwert der *d*-ten Area, $\hat{\beta}$ der geschätzte Parametervektor des Modells und $\overline{X}_d$ der Vektor der Mittelwerte der Hilfsvariablen, welche als Aggregate in der Grundgesamtheit, meistens Registerwerte, als bekannt vorausgesetzt werden. $\hat{u}_d$ ist die Ausprägung der Area-spezifischen zufälligen Abweichung vom Gesamtachsenabschnitt, welcher sich aus dem Produkt eines Shrinkage-Faktors und der Differenz der beobachteten Mittelwerte zu den aus dem Modell vorhergesagten Mittelwerten ergeben. Details können in Münnich et al. (2013) und der dort zitierten Literatur nachgelesen werden.

Durch Umformung können beide Modelle als zusammengesetzte Schätzer dargestellt werden, also als Linearkombination aus zwei Komponenten, wobei der Gewichtungsfaktor dem zuvor erwähnten Shrinkage-Faktor entspricht. Das Modell von Battese et al. (1988) ergibt sich dann als Linearkombination eines Regressionsschätzers und dem regressionssynthetischen Schätzer, wobei eine Random-Intercept-Regression verwendet wurde. Beim Schätzer von Fay und Herriot (1979) resultiert eine ähnliche Linearkombination mit dem Horvitz-Thompson-Schätzer.

Bei kleinen Stichprobenumfängen werden durch den Shrinkage-Faktor die designbasierten Kompontenten geringer und die synthetische Komponente höher gewichtet. Damit werden die Vor- und Nachteile beider Komponenten gegeneinander abgewägt. Die designbasierten Komponenten sind design-erwartungstreu, weisen bei kleinen Stichprobenumfängen aber möglicherweise sehr hohe Varianzen der Schätzung auf. Die synthetischen Komponenten weisen oft sehr kleine Schätzvarianzen auf, sind aber nicht design-unverzerrt. Falls das Modell jedoch perfekt passen würde, wäre der Schätzer nahezu unverzerrt. In der Regel ist aber solch ein Modell nicht erstellbar, womit in der Praxis Verzerrungen auftreten. Dieser Zielkonflikt zwischen geringer Variabilität und geringer Verzerrung wird anhand der Abbildung 2 veranschaulicht.

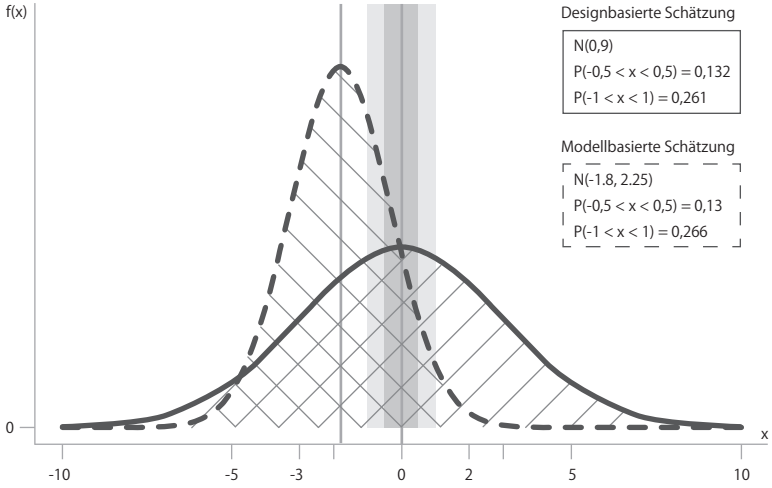


Abbildung 2: Vergleich von Schätzverteilung von Design- und Modell-basierten Verfahren in einer Area

In dieser Abbildung sind zwei Normalverteilungen abgetragen, welche die Punktschätzverteilung eines designbasierten Schätzers (durchgezogene Linie) und eines modellbasierten Schätzers (gestrichelte Linie) repräsentieren. Der designbasierte Schätzer hat eine hohe Variabilität und ist unverzerrt. Als wahrer Wert wird in diesem Beispiel die 0 gewählt. Der modellbasierte Schätzer hingegen ist verzerrt, weist dafür aber eine kleinere Variabilität auf. Das Ziel ist, dass der Schätzwert möglichst nahe am wahren Wert liegt. Für diese Betrachtung kann zum Beispiel ein Intervall um den wahren Wert definiert werden, in dem der Schätzwert liegen sollte. Hiermit kann berechnet werden, wie hoch die Wahrscheinlichkeit ist, dass der Schätzwert des designbasierten bzw. modellbasier-

ten Schätzers in dem akzeptablen Bereich liegt. Wird nun der Bereich von -0,5 bis 0,5 gewählt, so liegt hier die Wahrscheinlichkeit, den gewünschten Bereich zu treffen beim designbasierten Schätzer bei etwa 13,2%, beim modellbasierten Schätzer aber bei etwa 13%. Aufgrund der Verzerrung ist die Wahrscheinlichkeit, dass der modellbasierte Schätzer einen Wert im eng gewählten akzeptablen Bereich einnimmt geringer als für den designbasierten Schätzer. Wird der akzeptable Bereich zum Beispiel zwischen -1 und 1 gewählt, also verlängert, dreht sich das Ergebnis um. Nun ist die Wahrscheinlichkeit, dass der designbasierte Schätzer in diesem Bereich liegt, bei 26,1%, beim modellbasierten Schätzer dagegen bei 26.6%. Mit anderen Worten, trotz einer Verzerrung kann ein Schätzer mit höherer Wahrscheinlichkeit in einem akzeptablen Bereich um den wahren Wert liegen als ein unverzerrter. Gerade bei kleinen Stichprobenumfängen ist ein zu eng gewählter akzeptabler Bereich nicht realistisch, wodurch ein Schätzer mit geringerer Variabilität trotz einer möglichen Verzerrung vorzuziehen ist. Somit empfiehlt es sich, insgesamt eine kompensatorische Norm zwischen Verzerrung und Variabilität für die Beurteilung der Qualität der Schätzung zu verwenden. Üblicherweise wird hierbei der MSE genommen, der die Summe der Variabilität und der quadrierten Verzerrung ist. Die Small Area-Schätzer von Fay-Herriot und Battese-Harter-Fuller sind per Konstruktion MSE minimale Schätzer, falls das Modell zutreffend ist.

Die Ausführungen zeigen, dass die Anwendbarkeit von Small Area-Modellen sehr von der Modellierung abhängt. Da die Hilfsinformationen vollständig vorhanden sein müssen, spielen Registervariablen bei Haushaltsstichproben eine wesentliche Rolle. Je umfangreicher diese vorhanden sind, desto besser werden die Modellierungen, desto geringer werden die Verzerrungen und desto präziser werden die Schätzungen. Im Gegensatz zur klassischen designbasierten Welt, in der insbesondere die Optimierung im Sinne der Stichprobenziehung dominierte, spielt nun Modellauswahl und -validierung eine zentrale Rolle. Ebenso darf nicht mehr nur von einem zu optimierenden Ziel in einer Gesamtheit ausgegangen werden. Die Optimierung muss nun simultan in vielen Areas gleichzeitig erfolgen und möglicherweise sogar bezüglich verschiedener Variablen durchgeführt werden. Im Rahmen der Zensusforschung sind eine Reihe von Möglichkeiten zur Erreichung dieser Ziele in Münnich et al. (2012) aufgeführt.

Während bei Mittel- und Totalwerten für die Schätzungen in (1) im allgemeinen die aggregierten Registerinformationen ausreichen, reichen diese bei nichtlinearen Indikatoren nicht aus. Dies gilt zum Beispiel bei den zuvor erwähnten Armutsindikatoren der EU, insbesondere der Armutsgefährdungsquote. Ein viel versprechender Ansatz wurde von Molina und Rao (2010) entwickelt. Hierbei wird mit Hilfe eines Small Area-Modells die Verteilung der Einkommen modelliert. Über einen prädiktiven Monte-Carlo-Ansatz lassen sich mit Hilfe des

Modells sehr unterschiedliche nichtlineare Statistiken und vor allem Armuts- und Disparitätsmaße auf Regionalniveau schätzen. Für diesen Ansatz werden aber personen- oder haushaltsspezifische Registerinformationen benötigt.

Stichprobendesigns für Small Area-Verfahren

Bereits in der Kritik von Gelman (2007) kommt zum Ausdruck, dass die klassischen Design-Gewichte, wie sie in Stichprobenverfahren auch von der amtlichen Statistik verwendet werden, mit statistischen Modellierungen nur bedingt vereinbar sind. In der klassischen Vorgehensweise der Stichprobenoptimierung werden stets designbasierte Schätzverfahren zugrundegelegt. Ziel ist die Optimierung eines Varianzfunktionals einer Schätzfunktion, meist eines Mittel- oder Totalwertes, in Abhängigkeit vom Stichprobendesign. In der Modellwelt wird jedoch hauptsächlich von unabhängig und identisch verteilten Beobachtungen ausgegangen. In einem einfachen Design, dem Clustersampling, wird diese Annahme jedoch bereits offensichtlich verletzt, sofern nicht alle Cluster nahezu identisch sind. Es gibt einige Vorschläge, wie Stichprobendesigns für die Verwendung von Small Area-Methoden angepasst werden können. Costa et al. (2004) schlagen zum Beispiel vor eine Kombination aus proportionaler und Gleich-Aufteilung über die interessierenden Teilpopulationen hinweg durchzuführen. Longford (2006) hingegen schlägt ein Verfahren vor, anhand dessen optimale Stichprobenumfänge auch für modellbasierte Schätzungen unter bestimmten Voraussetzungen für Teilpopulationen berechnet werden können. Ebenso schlagen Gabler et al. (2012) sowie Münnich et al. (2012b) ein Verfahren zur Erlangung optimaler Stichprobenumfänge vor, unter Berücksichtigung von Box-Constraints auf die Spannweite der Designgewichte. In einer Monte-Carlo Simulationsstudie zeigen Münnich et al. (2012a), dass diese Designs funktionierende Alternativen zu den herkömmlichen Stichprobendesigns darstellen. Dies wird nicht zuletzt durch die erfolgreiche Verwendung des Designs von Gabler et al. (2012) im deutschen Zensus 2010 bewiesen.

Zusammenfassung und Ausblick

Ziel der Ausführungen war es, die Probleme einer Erfassung von Schätzwerten und insbesondere Indikatorwerten auf Kreisen darzulegen, welche für eine Regionalpolitik von erheblicher Bedeutung sind. Klassische Methoden, die sogenannten designbasierten Verfahren, können hier in vielen Fällen nur noch unzureichende Ergebnisse liefern. Dies führt zu einem Paradigmenwechsel in der amtlichen Statistik, in dem komplexe Modelle unter Verwendung von Registerdaten

oder ähnlichen administrativen Quellen in die Schätzungen integriert werden müssen, um eine hinreichende Genauigkeit der Schätzungen zu erhalten. Viele andere Länder, insbesondere in Skandinavien, verwenden derartige Systeme bereits oder diskutieren deren Verwendung intensiv. Nach Ansicht der Autoren wird in Zukunft auch kaum noch auf die Verwendung von Small Area-Methoden verzichten werden können, wenn valide regionale Auswertungen von Statistiken im Vordergrund stehen und die sonst dafür benötigten sehr hohen Stichprobenumfänge zu kostenintensiv sind.

Angesichts der Neustrukturierung der Haushaltsstichproben der deutschen amtlichen Statistik bietet sich die Chance, diese Argumente in die Neukonzeption einzubeziehen. Damit sollten in Zukunft auch neuere, moderne Methoden zur Ermittlung von Indikatoren und anderen Schätzwerten herangezogen werden. Eine rein designbasierte Vorgehensweise könnte hierbei unnötig einschränkend sein. Am Beispiel des deutschen Zensus 2011 ist erkenntlich, dass bei einer vorsichtigen Berücksichtigung sowohl der designbasierten wie auch der modellbasierten Welt mehr Chancen als Risiken entstehen. Allerdings kann auch nicht verschwiegen werden, dass eine Design-Optimierung von Quer- und Längsschnitt (Rotationspanel) und dessen Verwendbarkeit für modellbasierte Verfahren bisher noch wenig untersucht wurde. Bei allen Konzepten und Optimierungsvorschlägen sollte jedoch möglichst vorher festgelegt werden, welche Fragen mit einem neuen Survey gelöst werden sollen. Dies trifft umso mehr auf die Frage zu, wie tief disaggregtierte Schätzwerte produziert und benötigt werden.

Danksagung

Die Autoren danken insbesondere Herrn Riede, Statistisches Bundesamt, sowie dem Rat für Sozial- und Wirtschaftsdaten für die Möglichkeit, langjährige Forschungsergebnisse zu Design-Optimierung im Spannungsfeld zwischen designbasierten und Small Area-Methoden im Rahmen der Neukonzeption der deutschen Haushaltsstichproben vorstellen zu dürfen.

Literatur

Battese, G. E./Harter, R. M. and Fuller, W. A. (1988): An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association 83 (401), 28-36.

Costa, À./Satorra, A. and Ventura, E. (2004): Improving both domain and total area estimation by composition. SORT 28 (1), 69-86.

Destatis (2012): Wirtschaftsrechnungen. Leben in Europa (EU-SILC), Einkommens- und Lebensbedingungen in Deutschland und der Europäischen Union. Fachserie 15, Reihe 3.

Fay, R.E. und Herriot, R.A. (1979): Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 74 (366), 269-277.

Gabler, S./Ganninger, M. and Münnich, R. T. (2012): Optimal allocation of the sample size to strata under box constraints. Metrika 75 (2), 151-161.

Gelman, A. (2007): Struggles with Survey Weighting and Regression Modeling. Statistical Science 22 (2), 153-164.

Longford, N.T. (2006): Sample Size Calculation for Small Area Estimation. Survey Methodology 32 (1), 87-96.

Molina, I. and Rao, J. N. K. (2010): Small area estimation of poverty indicators. Canadian Journal of Statistics 38 (3), 369-385.

Münnich, R. T./Burgard, J. P .und Vogt, M. (2013): Small Area-Statistik: Methoden und Anwendungen. AStA Wirtschafts- und Sozialstatistisches Archiv 6 (3/4), 149-191.

Münnich, R.T./Gabler, S./Ganninger, M./Burgard, J.P. und Kolb, J.-P. (2012): Stichprobenoptimierung und Schätzung im Zensus 2011. Statistik und Wissenschaft 21. Statistisches Bundesamt, Wiesbaden.

Münnich, R. T./Burgard, J. P. and Zimmermann, T. (2012a): Small area modelling under complex survey designs for business data. Contributed paper at the Fourth International Conference on Establishment Surveys, Montreal.

Münnich, R. T./Sachs, E. W. and Wagner, M. (2012b): Numerical solution of optimal allocation problems in stratified sampling under box constraints. AStA Advances in Statistical Analysis 96 (3), 435-450.

Rao, J. N. K. (2003): Small Area Estimation of Wiley series in survey methodology. John Wiley and Sons, New York.

Amtliche Haushaltsstatistiken unterliegen aufgrund von europäischen wie nationalen Anforderungen einem kontinuierlichen Anpassungsprozess. Um auf zukünftige Entwicklungen flexibel und bedarfsgerecht reagieren zu können, hat das Statistische Bundesamt 2012 ein Projekt zur Weiterentwicklung des Systems der Haushaltsstatistiken, also den Erhebungen der Statistischen Ämter des Bundes und der Länder bei den privaten Haushalten, initiiert. In Kooperation mit den Statistischen Ämtern der Länder soll ein Gesamtsystem etabliert werden, in dem die bisher weitgehend unabhängig voneinander durchgeführten Haushaltsstatistiken als eine gemeinsame Erhebung umgesetzt wird.

In dem vorliegenden Band kommen unterschiedliche Experten zu Wort und geben einen Überblick über Stand und neue Ansätze zur Weiterentwicklung der Haushaltsstatistiken. Das Buch liefert damit für alle Interessierten einen spannenden Einblick in die Methodendiskussion zur künftigen Ausgestaltung von haushaltsstatistischen Erhebungen.