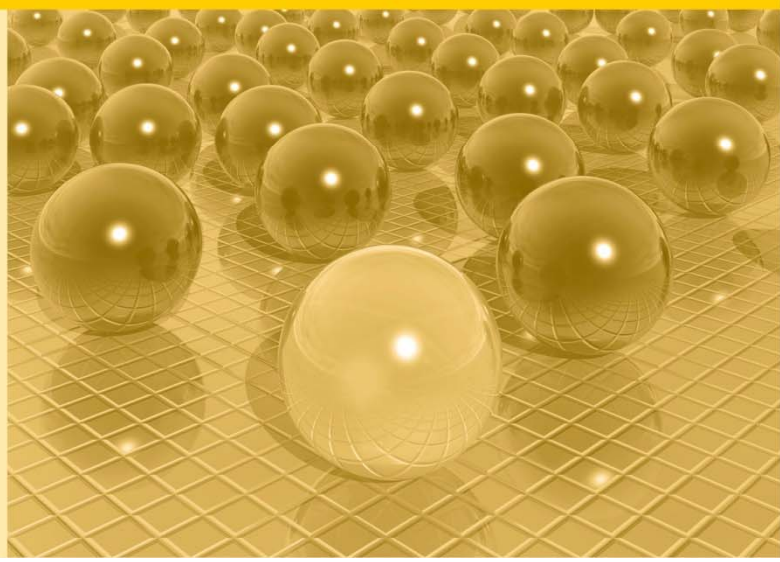
FDZ-Arbeitspapier
Nr. 52



Möglichkeiten zur Verknüpfung mehrerer Mikrozensus-Jahre zu einem Mikrozensus-Panel (2012–2015)

Sven Brocker, Hans-Ullrich Mühlenfeld

2020

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Landesamt Nordrhein-Westfalen
Mauerstraße 51, 40476 Düsseldorf • Postfach 10 11 05, 40002 Düsseldorf
Telefon 0211 9449-01 • Telefax 0211 9449-8000
Internet: http://www.it.nrw.de
E-Mail: poststelle@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2883
Fax.: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Informationen zum Datenangebot:

Forschungsdatenzentrum des
Statistisches Bundesamtes

Tel.: 0611 75-3277
Fax: 0611 75-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2883
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im November, 2020

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse der genannten Autoren im Zusammenhang mit der Nutzung der bereitgestellten Daten der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.

FDZ-Arbeitspapier
Nr. 52

Möglichkeiten zur Verknüpfung mehrerer Mikrozensus-Jahre zu einem Mikrozensus-Panel (2012–2015)

Sven Brocker, Hans-Ullrich Mühlenfeld

2020

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Einleitung

Bereits die beiden Mikrozensus-Panels 1996-1999 und 2001-2004 kamen den Wünschen der Wissenschaft und Politik nach der Verknüpfung von Mikrozensus-Jahrgängen nach. Beiden Projekten gemein war, dass sie sehr aufwändig waren, sodass sie u. a. nur durch „ein von amtlicher Statistik und Wissenschaft gemeinsam durchgeführtes Projekt […]“ (Handbuch Mikrozensus-Panel 1996-1999, S. 3) bewältigt werden konnten.

Erst mit der Einführung zeitstabiler Identifikatoren seit dem Jahrgang 2012 besteht faktisch für jede Nutzerin und jeden Nutzer die Möglichkeit, die Querschnittsdaten mehrerer Jahrgänge zu verknüpfen[1]. Grundpfeiler dieser Möglichkeit ist ein zeitstabiler Identifikator, der es ermöglicht, eine Person über die maximale Anzahl von vier Befragungen in den verschiedenen Datensätzen zu identifizieren.

Das vorliegende Arbeitspapier schildert die Möglichkeiten sowie die Vorgehensweise bei der Verknüpfung des formal anonymisierten On-Site-Files, so wie es am Gastwissenschaftsarbeitsplatz (GWAP) der Forschungsdatenzentren (FDZ) des Bundes und der Länder bzw. via kontrollierter Datenfernverarbeitung (KDFV) zur Verfügung gestellt wird. Aufgrund einer komplett neuen Stichprobe für den Mikrozensus-Jahrgang 2016 können die Daten nur von 2012 bis maximal 2015 verbunden werden. Dies allerdings – im Gegensatz zu den bereits oben genannten Panels – flexibel, sodass z. B. auch nur zwei oder drei Jahre verbunden werden können.

Das aus den vier Jahrgängen bestehende Produkt bietet 109 222 Personendatensätze, welche für Längsschnittanalysen verwendet werden können. Die Lektüre soll dem Datennutzerinnen und -nutzern einen verständlichen Einstieg in die Aufbereitung eines Mikrozensus-Panels ermöglichen.[2]

Neben allgemeinen Erläuterungen zum Mikrozensus wird gezeigt, welche Merkmale für die Erstellung eines Mikrozensus-Panels notwendig sind und welche Rolle hierbei der zeitstabile Personenidentifikator spielt. Ferner wird erläutert, welche methodischen Probleme sich aus der Verknüpfung der im Grunde als Querschnittserhebung angelegten Haushaltebefragung ergeben und wie mit den sogenannten Jahresüberhängen (Erläuterung hierzu im Kapitel „Umgang mit Jahresüberhängen“) umgegangen wird. Des Weiteren werden einige Nutzungshinweise für Panelanalysen in Stata angeführt. Es wird erläutert, welche Zusammenführungsquoten sich für die verschiedenen Möglichkeiten der Verknüpfung ergeben, welche Ausfallmuster für methodische Fragestellungen betrachtet werden können. Zudem

[1] Bei den beiden vorhergehenden Panels konnten die Daten nicht von Personen außerhalb der Projekte verknüpft werden, weil dazu Daten benötigt wurden, die Wissenschaftlerinnen und Wissenschaftlern bei einer normalen Nutzung nicht zur Verfügung gestellt wurden.

[2] Im Folgenden werden – soweit nicht anders angegeben – Variablennamen kursiv dargestellt; Befehle in Stata stehen in eckigen Klammern [ ].

werden die Ergebnisse einer Güteprüfung anhand theoretisch zeitkonsistenter Angaben wie Geschlecht und Geburtsjahr dargestellt.

Stichprobendesign

Der Mikrozensus ist eine einstufige Klumpenstichprobe. Es werden nicht primär Personen, sondern Anschriften bzw. Teile von Anschriften ausgewählt. Dadurch ist die Anzahl an Personen, welche in die Stichprobe gelangen, im Vorhinein nicht festgelegt, sondern aufgrund der an einer Anschrift gemeldeten Personen abhängig vom Zufall. Der Erwartungswert liegt hier bei ca. 1% der Bevölkerung. Die angestrebte Grundgesamtheit stellt die Wohnbevölkerung im Gebiet der Bundesrepublik Deutschland dar und umfasst alle Personen, die in Privathaushalten und Gemeinschaftsunterkünften leben.

1962 wurde die Rotation der Auswahlbezirke eingeführt; dies bedeutet, dass ein Auswahlbezirk bis zu vier Jahre lang hintereinander im Mikrozensus verbleibt und alle in einem Auswahlbezirk wohnberechtigten Personen befragt werden (Herter-Eschweiler 2019: 210). Rechtlich ermöglicht das Mikrozensusgesetz 1996 die Zusammenführung der Querschnittsdaten zu einem Mikrozensuspanel. Durch die jährliche Rotation von einem Viertel der Auswahlbezirke (rotierende Panelstichprobe) verkleinert sich folglich mit jedem Jahr die Größe der Personen oder Haushalte, welche verknüpft werden kann. Hinzu kommt, dass Befragte, die aus dem Auswahlbezirk fortziehen, nicht weiter befragt werden können, sondern stattdessen durch nachziehende Personen ersetzt werden. Somit ist es praktisch nicht möglich, die jährlich um 25% sinkende Startmenge an Befragten zu erreichen.

Ein Nachteil des Stichprobendesigns besteht darin, dass für Personen, welche den Auswahlbezirk verlassen, keine Informationen nach dem Wegzug vorliegen. Analog besteht für die neu hinzukommenden Haushalte und Personen das Problem, dass in der Regel ebenfalls keine Informationen vor dem Zuzug vorliegen. Herter-Eschweiler und Schimpl-Neimanns (2018: 2f.) weisen zusätzlich darauf hin, dass es zu Merkmalen, welche im Abstand von vier Jahren erhoben werden (Zusatzerhebungen), aufgrund des Rotationsschemas keine Panelangaben gibt. Trotz dieser angeführten Aspekte ist die Erfassung zeitlicher Veränderungen von Personen im Längsschnitt interessant, da somit die Probleme von Querschnittsstudien bei Kausalitätsaussagen abgeschwächt werden (Schnell, Hill und Esser 2018: 212f.). Durch die großen Fallzahlen (aufgrund der Auskunftspflicht) können vor allem auch „seltene Ereignisse“ mit ausreichender Anzahl an Merkmalsträgern abgedeckt werden.[3]

[3] Nach Porter (2008: 689) sind jene Merkmale „selten“, die bei weniger als 10% der Bevölkerung vorliegen.

Der Aspekt der Rotation[4] und die sich daraus ergebenden Überschneidungen werden in der folgenden Abbildung hypothetischer Szenarien deutlich. Jede Rotationsgruppe verbleibt für ein anderes Intervall in der Stichprobe; die vier Jahre von 2012–2015 überlappen nur partiell mit den Gruppen, die in den folgenden Jahren ausgewählt wurden. Dasselbe gilt auch für Personen, welche vor 2012 erstmalig in der Erhebung waren und entsprechend früher auch wieder aus der Stichprobe ausgeschieden sind. Nur für die Personen aus Stichprobennummer 8/Rotationsgruppe 2 ergibt sich eine Verknüpfung über die gesamten vier Jahre. Die nicht angezeigten Jahre liegen entsprechend vor 2012 bzw. nach 2015.

Stichproben-nummer	Rotations-gruppe	2012	2013	2014	2015
7	3				
7	4				
8	1				
8	2				
8	3				
8	4				
9	1				

Abbildung 1: nach Herter-Eschweiler und Schimpl-Neimanns (2018: 2)
Rotation im Mikrozensus nach Stichprobennummer und Rotationsgruppe für die Jahre 2012–2015

Es sei darauf hingewiesen, dass durch die komplett neue Ziehung der Stichprobe des Mikrozensus keine Verknüpfung von Daten vor 2016 mit den Daten ab 2016 möglich sind. Die methodischen Details für die neue Stichprobe finden sich bei Bihler und Zimmermann (2016). Des Weiteren sind Verknüpfungen ab 2016 nur bis einschließlich 2019 möglich, da der Mikrozensus ab 2020 in ein neues System von Haushaltserhebungen überführt wird.

Der Scientific-Use-File enthält eine faktisch anonymisierte 70% Substichprobe der Auswahlbezirke des jeweiligen Mikrozensus. Hierbei sind Variablen kategorisiert bzw. die vorhandenen Kategorien vergröbert, um die faktische Anonymität der Daten zu gewährleisten. Dies impliziert den Nachteil, dass kleinräumige Analysen mithilfe dieser Daten nicht mehr durchführbar sind (Schimpl-Neimanns und Herwig 2014: 8). Bei der On-Site-Nutzung ist u. a. die Arbeit an Gastwissenschaftsarbeitsplätzen bei den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder möglich. Zusätzlich besteht

[4] Rotiert werden de facto die Auswahlbezirke. Das bedeutet, dass davon auch natürlich die Haushalte in den Auswahlbezirken und die Personen in den Haushalten betroffen sind.

die Möglichkeit der kontrollierten Datenfernverarbeitung, bei der ein Analyseskript eingereicht wird. In beiden Fällen werden die Ergebnisse überprüft, um die Geheimhaltung zu gewährleisten. Bei beiden Zugangswegen werden den Nutzenden formal anonymisierte Daten zur Verfügung gestellt (Voshage, Janke und Malchin 2018: 20f.).

Erstellung eines Panels mittels der Mikrozensus-Jahrgänge 2012–2015

In diesem Abschnitt werden die einzelnen Schritte zur Erstellung eines Panels mit Hilfe mehrerer Mikrozensus-Jahrgänge (2012–2015) beschrieben. Zuerst wird die allgemeine Idee erläutert und damit verbundene Fragestellungen diskutiert.

Ordnungsnummern zur Erzeugung eines Identifikators

Um die zusammengehörenden Beobachtungen aus den einzelnen Querschnittsdatensätzen zusammenzuspielen und anschließend den entsprechenden Personen zuordnen zu können, ist es notwendig, diese zu identifizieren. Durch die hierarchische Datenstruktur (Auswahlbezirk → Haushalte → Personen) sind jene Analyseeinheiten, in denen eine Person enthalten ist, durch identifizierende Variablen klar zuordenbar. Diese Merkmale bzw. Ordnungsnummern sind aus Gründen des Datenschutzes systemfrei.

Als Variablen im Mikrozensus, welche längsschnittkonsistent sind und in Kombinationen zu einem Identifikator verknüpft werden können, eignen sich:

EF1: Land der Bundesrepublik

EF3: Systemfreie Auswahlbezirksnummer

EF4: Systemfreie laufende Nummer des Haushalts im Auswahlbezirk

EF63: Feste Personennummer über alle vier Berichtsjahre

Die Variable *EF63* wird seit 2011 von den Statistischen Landesämtern zugewiesen.

Jeder Person kann anhand dieser Variablen idealerweise eine eineindeutige Identifikationsnummer zugewiesen werden, so dass die Kombination der vier Variablen für zwei verschiedene Personen nie dieselbe, sondern immer eine andere Abfolge an Zahlen ergibt. Ausnahmen ergeben sich durch sogenannte Jahresüberhänge. Auf die dadurch entstehenden Herausforderungen wird später noch eingegangen.

Die Variable *EF1* weist Werte zwischen 01-16 auf, *EF3* rangiert von 00001-99999, *EF4* von 01-98. Die feste Personennummer ergibt sich aus der Reihenfolge in den Haushaltsmerkmalen und muss nicht durchgängig sein. Laut Schlüsselverzeichnis liegen die Werte zwischen 01-99.

Indem man eine Verkettungsoperation der vier Variablen durchführt [concatenate] wird eine ID-Variable *idpers* erstellt, die man auf Duplikate[5] überprüft, um fehlerhafte Identifikatoren aufzuspüren.

Eine Überprüfung aller vier Querschnittsdatensätze zeigte, dass nur bei weniger als 10 Fällen ein solches Duplikat und Klärungsbedarf vorlag. Da keine Systematik in den Variablen der verdächtigen Fälle gefunden wurde, ist davon auszugehen, dass es sich um fehlerhafte Einträge handelt; bei einer Gesamtfallzahl von 2,76 Millionen Personen über vier Jahre ist dies jedoch marginal und weist auf eine hohe Datenqualität hin. Die doppelt vorhandenen Einträge wurden entfernt.

Umgang mit Jahresüberhängen

Bedingt durch das Erhebungsdesign des Mikrozensus kommt es bei den Erhebungen jährlich zu sogenannten Jahresüberhängen. Gemeint sind Haushalte, die nicht im laut Stichprobenplan vorgesehenen Jahr, sondern erst im Folgejahr befragt werden konnten. Beispielsweise liegen also im Querschnittsdatensatz 2013 noch Angaben vor, welche Personen betreffen, die laut Stichprobenplan Angaben zum Jahr 2012 hätten machen sollen. Aufgrund der jährlichen Erhebung des Mikrozensus ist dieser Aspekt stets zu beachten.

Ein weiterer Faktor ist, dass abhängig vom Befragungszeitpunkt die Haushalte aus den Jahresüberhängen das Frageprogramm des aktuellen bzw. des vergangenen Jahres erhalten. Dies betrifft nur Selbstausfüller, die den Fragebogen erst nach einem spezifischen Termin an ihr Statistisches Landesamt schicken. Lediglich in 2012 liegen einige Fälle vor, in denen ein Interviewer ein telefonisches Interview mit dem Fragebogen aus 2011 durchgeführt hat. Zu beachten ist demnach, dass bei Fragestellungen, die sich auf das Zusatzprogramm beziehen, welches alle vier Jahre rotiert, bei Haushalten aus dem Jahresüberhang keine Informationen vorliegen (Herter-Eschweiler und Schimpl-Neimanns 2017: 11f.).

In den Querschnittsdatensätzen des Mikrozensus liegen Informationen über den Erhebungszeitpunkt vor; das „Vorjahreskennzeichen“ für Überhänge ist in *EF5u2* (*EF5b* im SUF) gespeichert:

Es wurde mit 1 kodiert, wenn es sich um ein Überhanginterview handelt und die Erhebung mit dem Frageprogramm des Vorjahres stattfand, mit 2, wenn es ein Überhanginterview ist und die Erhebung

[5] Zumeist entstehen solche Duplikate durch Jahresüberhänge.

das Frageprogramm des aktuellen Jahres verwendet und mit Missing (.), wenn es sich nicht um einen Jahresüberhang handelt (0 im SUF).

Bei der Verknüpfung der Datensätze für das Mikrozensus-Panel wurden die Überhänge – sofern möglich – in die Datensätze inkludiert, um alle Beobachtungen zu nutzen und keine Ressourcen – wie bei einer Löschung – zu verschwenden. Die passenden Interviews können für die Erhebungszeitpunkte zusammengespielt werden, indem man – wie von Herter-Eschweiler und Schimpl-Neimanns (2018) vorgeschlagen – eine Jahres-ID dafür generiert, welche die Daten und Beobachtungen klar identifiziert.

Jahrgang	Anzahl Fälle	Jahrgang	Anzahl Überhang[6]
2012	671 667	2012 → 2013	20 359
2013	662 660	2013 → 2014	26 134
2014	685 232	2014 → 2015	27 817
2015	663 186	-	-

Tabelle 1: Verteilung von Jahresüberhängen über Erhebungsjahr und Datenquelle

Tabelle 1 zeigt die Verteilung der Beobachtungen über die vier Querschnittsdatensätze und wie sich diese auf die Jahrgänge und Jahrgangskombinationen verteilen. Jahresüberhänge sind mit Pfeilen markiert und es zeigt sich, dass diese etwa 20 000 bis 28 000 Beobachtungen umfassen. Wie zuvor erwähnt, sind Verknüpfungen mit diesen Daten nur bis 2015 möglich, da ab 2016 eine neue Stichprobe verwendet wurde. Dies betrifft auch die Verwendung von Jahresüberhängen aus 2015, die sich im Datensatz 2016 befinden. Es war nicht sinnvoll, diese per Identifikator an den Hauptdatensatz anzuspielen. Beispielsweise wurden im MZ 2012 671 667 Personen befragt; im Folgejahr 662 660 Personen, welche 2013 erhoben werden sollten, sowie 20 359 weitere Befragte, die als Überhänge noch zum Vorjahr 2012 gehören.

Die Frage der Überhänge sollte also beachtet werden, wenn Analysen geplant sind, welche speziell auf bestimmte Zeitpunkte gerichtet sind.

Verknüpfen [append] und Kennzeichnung der Jahrgänge

Die Querschnittsdatensätze werden untereinandergestellt bzw. aneinandergehängt [append], wobei für

[6] Ist eine Darunter-Position der Gesamtzahl.

jeden Schritt mittels der Variablen *file* ein Label vergeben wird, das identifiziert, zu welchem Jahrgang die Daten gehören und ob es sich um einen Jahresüberhang handelt.

Um klarer darzustellen, wie die Antwortpatterns der Personen ausfallen, wird in der Variablen *ptyp* gespeichert, zu welchen Zeitpunkten Antworten vorliegen. Ein (Trennstrich) symbolisiert dabei einen Ausfall, eine Zahl (12, 13, 14, 15) steht für vorhandene Informationen für die Person. *ptyp* nimmt hier für Personen, die über alle vier Jahre im Panel vorhanden sind, die Ausprägung 8 an. Für entsprechende Analysen anderer Subgruppen (z. B. 2- oder 3-Jahrespanel) können die Patterns innerhalb der Variablen konsultiert werden.

Abschließend wurden Variablen, die nicht zum Grundprogramm des Mikrozensus gehören, für die Jahre 2012–2015 entfernt und in Stata der Datensatz bezüglich der relevanten Variablen als Panel deklariert. Mittels der Variable t (Zeitpunkt) wird das jeweilige Erhebungsjahr markiert. Dies kann für eine Umstellung der Datenstruktur [long/wide; reshape] genutzt werden. Eine Variable verändert sich in der Benennung beispielsweise für das erste Erhebungsjahr 2012 zu *variable1_1*; für die Daten zum Zeitpunkt 2015 ist dies analog *variable1_4*.

Im folgenden Abschnitt werden weitere Hinweise gegeben, wie die so vorbereiteten Datensätze zur Analyse genutzt werden können.

Analysen und vorbereitende Maßnahmen

Hinweise für Panelanalysen in Stata

Der vom Forschungsdatenzentrum der Statistischen Ämter des Bundes und der Länder zur Verfügung gestellte Datensatz für das Mikrozensus-Panel der Jahre 2012–2015 wird standardmäßig im Long-Format zur Verfügung gestellt. Dies ist zweckmäßig, da das in Stata implementierte Programmpaket [xt-suite], welches Werkzeuge zur Verfügung stellt, mit denen Paneldaten analysiert werden können, das Vorliegen der Daten im Long-Format erfordert.

Wide- und Long-Format sind zwei verschiedene Darstellungsformen für tabellarische Informationen. Eine übliche Darstellungsweise ist das Wide-Format, in welchem für jede Variable zu einem Zeitpunkt eine Spalte verwendet wird. Würde beispielsweise das Haushaltsnettoeinkommen zu zwei Zeitpunkten erhoben werden, wären diese Informationen in den Variablen *income_2012* und *income_2013* gespeichert. Im Long-Format hingegen würde die Variable lediglich *income* heißen, so dass die wiederholten Messungen in einer Spalte gespeichert würden. Paneldaten an verschiedenen Personen zu unterschiedlichen Zeitpunkten lassen sich in diesem Format also so darstellen, dass eine Zeile eine Person

[i, *group identifier*] zu einem Messzeitpunkt [j *within-group identifier*] beschreibt. Die Darstellungsart ist insofern relevant, da nicht alle statistischen Verfahren für Panelanalysen im intuitiven Wide-Format unterstützt werden.

Mittels [xtset]-Befehl wird Stata mitgeteilt, welche Variable im Datensatz zur Identifizierung der Person *personid*[7] und welche zur Identifizierung des Zeitpunkts [t, year] verwendet wird. Diese Information wird im Datensatz gespeichert und erleichtert die folgenden Aufrufe für Tabellen oder beispielsweise Regressionen (Fixed-, Between- bzw. Random-Effects). Für spezifische Panelanalysen (wie den zuvor genannten Regressionen) benötigt Stata zwingend das Long-Format. Möchte man keine Längsschnittanalysen mit dem Datensatz durchführen, sondern lediglich Werte verschiedener Jahrgänge miteinander vergleichen, kann dies auch im Wide-Format erfolgen.[8]
Der Datensatz ist bereits durch [xtset] so vorbereitet, dass [xt]-Befehle sofort genutzt werden können. Es empfiehlt sich das Lesen der entsprechenden Dokumentation der zur Analyse geplanten Befehle im xt-Manual von Stata.[9]

Sofern man die Darstellungsformen für die Daten wechseln möchte, empfiehlt es sich, bereits im Vorfeld eine Auswahl an Variablen zu treffen, damit nicht unnötigerweise der gesamte Datensatz transformiert werden muss und Rechenzeit gespart wird. Um zudem zu verhindern, dass durch [reshape] ein Verlust von Variablenlabeln entsteht, kann anstelle des [reshape]-Befehls der Befehl [*reshape8]*[10] verwendet werden. Dieser eignet sich für eine zuverlässige Transformation von ca. 100 Stammvariablen gleichzeitig (s. [help reshape8] bzw. [help reshape] in Stata).

Zusammenführungsquote

Um den Erfolg der Verknüpfung der Jahrgänge 2012 bis 2015 zu überprüfen, wurde eine Konsistenzprüfung anhand konstanter Merkmale (s. folgender Abschnitt) durchgeführt. Ein Gradmesser ist die Zusammenführungsquote. In Abhängigkeit davon, wie erfolgreich eine Zuordnung der Jahresüberhänge ist, können hier kleine Variationen zwischen den Jahrgängen aufkommen; im Schnitt sind etwa 20 000 Personen in einem Überhang enthalten und können noch an den eigentlichen Jahrgangsdatensatz angespielt werden.

[7] Diese ID muss eine numerische Variable sein und kein String.
[8] Dies wird im Abschnitt zur Konsistenz erneut aufgegriffen.
[9] Eine schnelle Einführung in den jeweiligen Befehl lässt sich auch mittels [help xt], [help xtregt] o. Ä. aufrufen. Alternativ kann das Manual auch online aufgerufen werden: https://www.stata.com/manuals/xt.pdf.
[10] [reshape8] wurde 2003 von Bill Rising zur Verfügung gestellt: https://ideas.repec.org/c/boc/bocode/s436202.html.

Die Art, wie die Zusammenführungsquote berechnet wird, soll an dieser Stelle expliziert werden. Es ist noch einmal darauf hinzuweisen, dass durch das rotationsbedingte Design in jedem nachfolgenden Jahr 25% der Stichprobe wegfallen und ersetzt werden. Das bedeutet, die theoretisch maximal zu erreichende Gesamtzahl, die zu Beginn in einem Datensatz enthalten ist, wird in jedem folgenden Jahr, das zusätzlich betrachtet wird, um 25% kleiner.

Jahrgänge	Fallzahl[11]	Theoretisch erreichbar	Quote
2012	692 026	692 026 (1,0)	-
2012/13	428 688	519 019,5 (0,75)	82,60%
2012/13/14	250 867	346 013 (0,5)	72,50%
2012/13/14/15	109 222	173 006,5 (0,25)	63,13%

Tabelle 2: Tabelle mit de facto Fallzahl, theoretisch möglicher Maximalzahl und daraus resultierender Zusammenführungsquote.

Wie in Tabelle 2 zu erkennen ist, verbleiben nach vier Jahren noch 109 222 Personen, für welche vier vollständige Personensätze der einzelnen Jahre (2012–2015) vorliegen. Relativiert man diese an der Gesamtzahl, welche 2012 im Panel vorlag (671 667+20 359 = 692 026), so ergibt sich ein recht geringer Anteil von 15,78%. Dies berücksichtigt jedoch nicht die erhebungsbedingte Rotation; mit jedem Erhebungsjahr können demnach weniger Personen überhaupt zugeordnet werden. Für eine Verknüpfung von 2012–2015 bedeutet dies, dass von 692 026 lediglich 25% maximal vorliegen; damit würde der Anteil 109222/173006 = 63,13% betragen. Dass der theoretisch maximale Wert der Verknüpfung in keiner Jahreskombinationen (2-Jahre, 3-Jahre, 4-Jahre) erreicht werden kann, begründet sich durch Ausfälle, die durch Fortzüge und Todesfälle zustande kommen.

Die von Herter-Eschweiler und Schimpl-Neimanns (2018) angeführte Zusammenführungsquote für die Verknüpfung von zwei aufeinander folgenden Jahren (Scientific-Use-File des MZ 2012 und 2013) liegt hingegen bei 67,7%; diese berechnet sich allerdings anders als im zuvor genannten Beispiel der 4-Jahres-Verknüpfungen: Von 482 127 Personen lassen sich 314 334 zuordnen; berücksichtigt man die Rotation (Faktor 0,75) dann liegt die eigentliche Zusammenführungsquote bei 86,9%. Würde man die

[11] Inklusive Jahresüberhänge.

Berechnungsweise der Autoren auf die genannte 4-Jahres-Verknüpfung anwenden, läge diese statt bei etwa 63% bei lediglich 15,8%.

Für die On-Site-Files des Mikrozensus liegen die Zusammenführungsquoten für die 2-Jahres-Verknüpfungen bei etwa 86%; die für die 3-Jahres-Verknüpfungen bewegen sich zwischen 68,7% und 70,2%. Diese umzugs- und sterbebedingten Ausfälle sind neben der Variation der Jahresüberhänge maßgeblich für unterschiedliche Werte verantwortlich.

Patterns der Ausfälle und Beantwortung

Je nach inhaltlicher Fragestellung kann für eine Analyse die Art des jeweiligen Ausfalls relevant sein. Um hier nähere Informationen darüber zu erhalten, ob Ausfälle durch Geburten und Zuzüge, Fortzüge und Todesfälle bzw. anderweitig bedingt sind, lassen sich die Patterns der Auswahlbezirke und Personensätze nebeneinander betrachten. Die durch das Mikrozensusgesetz geregelte Auskunftspflicht für den Mikrozensus führt vermutlich zu geringeren Befragungsausfälle (Nonresponse). Die zuvor erwähnten Interviews aus den Jahresüberhängen liegen im Datensatz des Folgejahrs vor.

Interessiert man sich beispielsweise für Fragestellungen, die eng mit Mobilität, Mortalität oder Fertilität verbunden sind, kann es sinnvoll sein, den Fokus auf bestimmte Subgruppen zu richten. Es ist daher zu unterscheiden, ob die Ausfälle zwischen den Jahren durch das stichprobenbedingte Rotationsdesign verursacht werden, oder ob zuvor genannte Gründe in Frage kommen. Die Idee besteht darin, zu vergleichen, ob bei einem Ausfall auf Personenebene ebenfalls der Auswahlbezirk ausfällt, was dann darauf hinweist, dass dieser nicht länger in der Stichprobe ist. Liegen Diskrepanzen vor, dass zum Beispiel der Auswahlbezirk konstant vorhanden war, jedoch erst zu einem späteren Zeitpunkt Personendaten, liegt die Annahme einer Geburt oder eines Zuzugs nahe. Herter-Eschweiler und Schimpl-Neimanns (2018: 7f.) weisen auf diese verschiedenen Muster in der 2-Jahres-Verknüpfung hin; die folgende Tabelle überträgt diese Idee auf exemplarische Szenarien der Verknüpfung der Jahre 2012–2015.

Personensätze	Auswahlbezirke	Interpretation
12 - - -	12 - - -	Ausfälle nach 2012; Auswahlbezirk war 2012 **letztmalig in der Stichprobe** vorgesehen
12 - - -	12 13 14 15	**Ausfälle bedingt durch Fortzüge oder Todesfälle**

- 13 14 15	- 13 14 15	**Ab 2013 erstmalig in der Stichprobe**; daher ein Ausfall für das erste Jahr
- 13 14 15	12 13 14 15	**Ausfall 2012 durch Geburt oder Zuzug**[12]; keine Angabe zum ersten Zeitpunkt, jedoch zu späteren Zeitpunkten
12 13 14 15	12 13 14 15	**Keine Panelmortalität**; sowohl für Personenangaben als auch für die Auswahlbezirke liegen zu allen Zeitpunkten Daten vor.

Tabelle 3: Ausfallmuster und Interpretation

Es bietet sich selbstverständlich bei einem Interesse an ausschließlich durchgehend vorhandenen Personenauskünften (keine Panelmortalität) an, diese Subgruppen entsprechend zu filtern, so dass der Datensatz nur noch solche Fälle enthält. Zum derzeitigen Zeitpunkt sind alle Daten im zur Verfügung gestellten Datensatz enthalten, so dass solche Informationen genutzt werden können. Über die Variablen *ptyp* und *ptyp_psu* sind diese Patterns abruf- und vergleichbar.

Klärungsmöglichkeit besteht ebenfalls über die bereits von Herter-Eschweiler und Schimpl-Neimanns (2018: 8) vorgeschlagenen Variablen *EF45* (In den letzten 12 Monaten in den Haushalt eingezogen) und *EF451* (Wohnsitz vor 12 Monaten wie zur Zeit der Erhebung? *freiwillige Beantwortung*). Hier böte es sich an, die Daten des Folgejahres des interessierenden Zeitpunkts (t+1) zu kontrollieren und dies mit den betrachteten Patterns abzugleichen. Weiterführende Empfehlungen für den Umgang mit Ausfällen finden sich im Datenhandbuch des MZ-Panels (Statistisches Bundesamt 2006).

Konsistenzangaben für Geschlecht und Geburtsjahr

Wie die meisten Panels ist auch der verknüpfte Mikrozensus nicht frei von zeitlich inkonsistenten Angaben. Zur Überprüfung der Konsistenz über die 4 Jahre hinweg ist es angeraten, Merkmale auf Konsistenz zu überprüfen, die sich grundsätzlich nicht ändern sollten. Herter-Eschweiler und Schimpl-Nei-

[12] Ein Ausfall im Jahr 2012 kann auch durch eine leerstehende Wohnung oder dem Ausfall des Haushalts bzw. der Gemeinschaftsunterkunft verursacht sein.

manns (2018) schlagen hierzu in einer ersten einfachen Prüfung die beiden Merkmale Geschlecht sowie Geburtsjahr vor[13]; diese werden für das On-Site-File ebenfalls verwendet. Grundsätzlich sind auch andere Variablen denkbar, von denen man annehmen kann, in jedem Jahr zeitstabile Angaben zu erhalten.

MZ-Panel	Geschlecht	Geburtsjahr	Kombi
2012-2013	99,66	97,21	96,90
2013-2014	99,68	97,24	97,04
2014-2015	99,62	97,17	96,94
2012-2014	99,51	95,98	95,71
2013-2015	99,47	95,94	95,62
2012-2015	99,31	94,83	94,48

Tabelle 4: Konsistenzprüfung für die Merkmale Geschlecht und Geburtsjahr

Wie sich Tabelle 4 entnehmen lässt, weisen alle möglichen Verknüpfungen auf eine hohe Datenqualität bei der Verknüpfung hin. Betrachtet man die 2-Jahres-Verknüpfungen, so liegt die Inkonsistenz hinsichtlich des Geschlechts lediglich bei 0,34%; die des Geburtsjahrs bei etwa 2,8%. Kombiniert man beide Merkmale, steigt die Inkonsistenz noch einmal geringfügig, um etwa 0,3%-Punkte.

Wie zu erwarten, ist die Konsistenz am stärksten, je kürzer die betrachtete Zeitperiode ist. Da das Fehlerpotenzial mit jedem zusätzlich betrachteten Jahr steigt, ist es inhaltlich naheliegend, dass dieses linear geringfügig steigt (bzw. die Konsistenz entsprechend sinkt). Insgesamt scheinen diese beiden Parameter in verschiedenen Jahrgängen vergleichbar zu sein.

Man kann verschiedene Erklärungen für den leichten Unterschied der Konsistenz zwischen Geburtsjahr und Geschlecht heranziehen; einerseits besteht die Möglichkeit, dass bei Proxyangaben leicht Verwechslungen des Geburtsjahres zustande kommen – vor allem dann, wenn die Person im Haushalt, für die man eine Angabe macht, einem entweder nicht sehr nahesteht, bzw. bereits deutlich älter ist. Geschlechtsangaben sollten – selbst bei Proxyangaben – weniger fehlerbelastet sein. Herter-Eschweiler und Schimpl-Neimanns (2018) geben als zusätzliche Erklärung an, dass bei der Betrachtung der Fehler in den Scientific Use Files vor allem „Zahlendreher“ eine Rolle spielen.

[13] Die Variablen müssen nicht zwingend zeitstabil sein. Auch sinnvoll interpretierbare Veränderungen, wie z.B. die Veränderung des höchsten Bildungsabschlusses, können Hinweise auf die Konsistenz der Daten geben.

Schluss/Nutzungshinweise

Abschließend sei noch auf weiterführende Literatur verwiesen. Rendtel (2005) bietet einen Überblick über die Möglichkeit der Auswertung der Daten des Mikrozensus im Längsschnitt. Für weitere methodische Details kann das Handbuch zum Mikrozensus-Panel herangezogen werden (Methodenverbund „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“, 2006). Einen Einstieg in die Analyse mit Paneldaten liefern Giesselmann und Windzio (2013). Für ausführlichere Analysen im Längsschnitt mit Stata siehe Rabe-Hesketh und Skondral (2011).

Addendum: Verwaltungshinweise / Kauf bei On-Site-Nutzung

Im Gegensatz zu den fertig erstellten Produkten der Mikrozensus-Panels 1996–1999 und 2001–2004, besteht bzgl. der vorgenannten Möglichkeit zur Verknüpfung der Mikrozensus-Jahrgänge die Flexibilität der Nutzung der Daten auch im Querschnitt. Dies bedeutet, dass nicht ein fertiges Produkt erworben wird, sondern dass einzelne Zeitscheiben erworben werden, von denen maximal vier Produkte verbunden werden können (aber nicht verbunden werden müssen). Es müssen auch nicht zwingend vier Jahrgänge verbunden werden, sondern es können auch nur zwei oder drei Jahrgänge verbunden werden. Antragstellenden, die entsprechende Jahrgänge verknüpfen wollen, können dies anhand zur Verfügung gestellter Syntax entgeltfrei selbst durchführen. Zu beachten ist, dass das Vorhaben, die Daten verknüpfen zu wollen, im Antrag beschrieben werden muss.

Literatur

Bihler, W. & Zimmermann, D. (2016). Die neue Mikrozensus-Stichprobe ab 2016. *Wirtschaft und Statistik*, 6/2016, 20-29.

Emmerling, D. & Riede, T. (1997). 40 Jahre Mikrozensus. Wirtschaft und Statistik, 3.

Giesselmann, M. & Windzio, M. (2019): *Regressionsmodelle zur Analyse von Paneldaten*. 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Herter-Eschweiler, R. & Schimpl-Neimanns, B. (2017). Wichtige Informationen zur Nutzung des Mikrozensus Scientific Use Files 2013. Dokumentation und Datenaufbereitung. Zugriff 20.05.2019 unter https://www.gesis.org/missy/files/documents/MZ/readme/readme_suf2013.pdf

Herter-Eschweiler, R. & Schimpl-Neimanns, B. (2018). Möglichkeiten der Verknüpfung von Mikrozensus-Querschnittserhebungen ab 2012 zu Panels. GESIS: Leibniz-Institut für Sozialwissenschaften. Zugriff 20.05.2019 unter https://www.gesis.org/missy/files/documents/MZ/panelbildung_suf2012.pdf

Herter-Eschweiler, R. (2019). Der Mikrozensus und die Möglichkeiten seiner Regionalisierung. *GESIS Schriftenreihe Band 23. Regionale Standards, Ausgabe 2019, 3. überarbeitete und erweiterte Auflage.*

Methodenverbund „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“. (2006). Handbuch Mikrozensus-Panel 1996-1999.

Porter, S. R. (2008). Rare Populations. In P. J. Lavrakas (Hrsg.), *Encyclopedia of Survey Research Methods Volume 1 & 2* (S. 689-691). Thousand Oaks: Sage.

Rabe-Hesketh, S. & Skrondal, A. (2011). *Multilevel and longitudinal modeling using Stata. Volume I: Continuous Responses*. 3. Auflage. College Station, TX: Stata Press.

Rendtel, U. (2005). *Wie geeignet ist der Mikrozensus für Längsschnittanalysen?* (Arbeitspapier Nr. 7). Methodenverbund „Aufbereitung und Bereitstellung des Mikrozensus als Panelstichprobe“.

Rising, B. (2003). RESHAPE8: Stata module to reshape while preserving variable labels, *Statistical Software Components* S436202, Boston College Department of Economics. Zugriff 20.05.2019 unter https://ideas.repec.org/c/boc/bocode/s436202.html

Schimpl-Neimanns, B. & Herwig, A. (2014). *Mikrozensus Scientific Use File 2011: Dokumentation und Datenaufbereitung* (Technical Report Nr. 2014|08). GESIS: Leibniz-Institut für Sozialwissenschaften.

Schnell, R., Hill, P. B. & Esser, E. (2018): *Methoden der empirischen Sozialforschung* (11. Auflage). München: Oldenbourg.

StataCorp, (2016): Stata Longitudinal-Data/Panel-Data Reference Manual, Release 15. Zugriff 11.06.2019 unter https://www.stata.com/manuals/xt.pdf

Voshage, R., Janke, D. & Malchin, A. (2018): Entwicklungen in der amtlichen Statistik. Der FDZ Standort im AfS – Angebot und Nachfrage. Zeitschrift für amtliche Statistik Berlin Brandenburg 3+4. Zugriff 14.06.2019 unter https://www.statistik-berlin-brandenburg.de/publikationen/aufsaetze/2018/HZ_201803-12.pdf

Statistische Ämter des Bundes und der Länder,
Arbeitspapier Nr. 52 – Möglichkeiten zur Verknüpfung mehrerer Mikrozensus-Jahre zu einem Mikrozensus-Panel (2012–2015)