

NEUES IMPUTATIONSVERFAHREN BEI ANWORTAUSFÄLLEN ZU GEBORENEN KINDERN IM MIKROZENSUS

Kerstin Lange, Olga Pöttsch

↳ **Schlüsselwörter:** Mikrozensus – Kinderlosigkeit – Kinderzahl je Frau – Antwortausfall – Imputation – Predictive Mean Matching

ZUSAMMENFASSUNG

Der Beitrag stellt ein neues Verfahren für die Behandlung der Antwortausfälle im Mikrozensus vor. Predictive Mean Matching hat sich in Simulationsstudien als geeignet herausgestellt, um die fehlenden Angaben bei freiwilligen Fragen zur Geburt von Kindern zu imputieren. Am Beispiel des Mikrozensus 2016 werden die Ergebnisse auf Basis des zuvor und des neu eingesetzten Imputationsverfahrens verglichen. Beide Verfahren liefern valide Ergebnisse, daher ist es nicht notwendig, frühere Daten und Ergebnisse zu revidieren. Der neue Ansatz ist insofern überlegen, als er eine vollständige Imputation und somit erweiterte Analysen der Paritätsverteilungen ermöglicht. Des Weiteren enthält der Beitrag die ersten Ergebnisse des Mikrozensus 2018 zur Fertilität und zur Kinderlosigkeit der Frauen, die auf dem Datensatz mit vollständig imputierten Antwortausfällen beruhen.

↳ **Keywords:** *microcensus – childlessness – number of children per woman – item nonresponse – imputation – predictive mean matching*

ABSTRACT

This article describes a new method for handling item nonresponse in the German microcensus. In simulation studies, predictive mean matching has proved to be a suitable method of item imputation for voluntary questions about childbirth. The 2016 microcensus is used to compare the results based on the former method with those based on the new imputation method. As both methods yield valid results there is no need for a revision of previously published data and results. The new approach is superior in that it enables a full imputation and therefore permits extended parity distribution analyses. The article also provides first results of the 2018 microcensus on women's fertility and childlessness which are based on the data file where all missing responses have been imputed.



Kerstin Lange

hat Statistik studiert und ist seit 2017 im Statistischen Bundesamt im Bereich mathematisch-statistische Methoden tätig. Im Referat „Maschinelles Lernen und Imputationsverfahren“ ist sie schwerpunktmäßig für Imputationsverfahren zuständig.



Olga Pöttsch

ist Diplom-Ökonomin und Referentin im Referat „Natürliche Bevölkerungsbewegungen, demografische Analysen, Vorausberechnungen“ des Statistischen Bundesamtes. Schwerpunkte ihrer Arbeit sind Bevölkerungs- und Haushaltsvorausberechnungen sowie Analysen der Fertilität.

1

Einleitung

In der deutschen amtlichen Statistik gibt es zwei Datenquellen für Indikatoren der Geburtenentwicklung: die laufende Geburtenstatistik und – seit 2008 – den Mikrozensus. Die Geburtenstatistik liefert vielfältige Angaben zu den Geborenen selbst. Dazu gehören Unterscheidungen nach Einzel- und Mehrlingsgeburt, nach der Geburtenfolge sowie demografische Merkmale der Eltern (zum Beispiel Wohnort, Familienstand, Alter und Staatsangehörigkeit). Aus dem Mikrozensus stammen dagegen Informationen zur Fertilität der Frauen. So wird aus den im Mikrozensus gewonnenen Angaben der Frauen zur Zahl der geborenen Kinder beispielsweise abgeleitet, wie hoch die Anteile der Mütter und der kinderlosen Frauen je Jahrgang sind. Des Weiteren zeigen sie, wie sich die Mütter nach der Zahl der geborenen Kinder (Parität) verteilen und welche sozioökonomischen Unterschiede zwischen diesen Gruppen bestehen. Die Daten aus den beiden Erhebungen beschreiben folglich unterschiedliche Aspekte der Geburtenentwicklung in Deutschland und tragen so zu einem umfassenderen Bild bei.

Im Hinblick auf Genauigkeit und Vollständigkeit der Angaben unterscheiden sich die beiden Datenquellen jedoch erheblich. Die Geburtenstatistik basiert auf den Standesamtsmeldungen zur Geburt und erfasst somit laufend alle geborenen Kinder. Diese Ergebnisse sind zuverlässig und haben eine hohe Qualität. Der Mikrozensus ist eine Stichprobe mit einer Auswahl von rund 1 % der privaten Haushalte. Die Fragen zur Geburt von Kindern werden im Mikrozensus alle vier Jahre an Frauen im Alter zwischen 15 und 75 Jahren gestellt. Die beiden Fragen nach den geborenen Kindern sind freiwillig zu beantworten (§ 13 Absatz 7 Mikrozensusgesetz).

Die im Mikrozensus gewonnenen Daten enthalten unterschiedliche Arten von Fehlern. Die wichtigsten sind dabei stichprobenbedingte Zufallsfehler und mögliche systematische Verzerrungen, die aufgrund von Antwortausfällen (Item-Non-Response) entstehen können.¹ Ein zufallsbedingter Stichprobenfehler wird mithilfe einer

Fehlerrechnung geschätzt und ist durch die Angabe eines Konfidenzintervalls nachvollziehbar (Statistisches Bundesamt, 2017a, hier: Tabelle 2). Mögliche systematische Verzerrungen der Stichprobenergebnisse sind dagegen tückisch: Ihr Einfluss auf das Ergebnis kann nicht gemessen und bei der Hochrechnung nur eingeschränkt berücksichtigt werden.

Für Stichprobenbefragungen gilt deshalb die Prämisse, bekannte Fehlerquellen für systematische Verzerrungen weitmöglichst auszuschließen. Bei einer gesetzlich vorgeschriebenen Freiwilligkeit der Angaben bleiben jedoch Unsicherheiten bestehen, da die Antwortausfälle in der Regel nicht zufällig auftreten. Auch für die Antwortausfälle bei den Fragen zur Geburt von Kindern hat sich gezeigt, dass zwischen den Frauen mit und Frauen ohne Angaben deutliche Unterschiede bestehen (siehe Kapitel 2). Die Antwortausfälle treten folglich systematisch auf und können die Ergebnisse – beispielsweise die Kinderlosenquote – beeinflussen.

Eine effektive Methode, die Verzerrungen aufgrund von Item-Non-Response zu reduzieren, ist eine sogenannte Imputation der fehlenden Angaben. Dabei ersetzen geeignete generierte Werte die fehlenden Angaben. Dieser Beitrag beschreibt, warum eine Imputation für die Angaben zur Geburt im Mikrozensus erforderlich ist und welche Verfahren dafür angewendet wurden. Die Angaben zur Geburt von Kindern wurden im Mikrozensus bisher in vier Mikrozensusbefragungen erhoben: 2008, 2012, 2016 und 2018². In den Datensätzen für die Jahre 2008, 2012 und 2016 wurde eine Imputation nur bei den Fällen mit einer sehr hohen Wahrscheinlichkeit für einen zu imputierenden Wert vorgenommen. Im Mikrozensus 2018 wurde zum ersten Mal ein neues Verfahren angewendet, das mittels einer stochastischen Imputation alle fehlenden Angaben durch imputierte Werte ersetzt.

1 Darüber hinaus ist bei einer primären Befragung mit weiteren nicht stichprobenbedingten Fehlern (zum Beispiel Erfassungs-, Mess- und Aufbereitungsfehler sowie Ungenauigkeiten bei Proxyangaben) zu rechnen. Diese können auch im Mikrozensus nicht ganz ausgeschlossen werden.

2 Die Fragen zu geborenen Kindern werden im Mikrozensus alle vier Jahre an Frauen gestellt. Aufgrund einer weitreichenden methodischen Umstellung des Mikrozensus ab dem Jahr 2020 beträgt der Abstand zwischen den Befragungen 2016 und 2018 ausnahmsweise nur zwei Jahre. Die nächste Befragung zu geborenen Kindern ist turnusgemäß für 2022 vorgesehen.

2

Antwortausfälle im Mikrozensus

Zur Geburt von Kindern gibt es im Mikrozensus zwei Fragen. Die erste Frage – „Haben Sie Kinder geboren?“ – liefert ein dichotomes Merkmal: Die Befragte ist Mutter von mindestens einem leiblichen³ Kind oder sie hat (noch) kein Kind geboren (im Weiteren als „kinderlos“ bezeichnet). Die zweite Frage richtet sich nur an die Mütter und liefert einen Zahlenwert: „Wie viele Kinder haben Sie insgesamt geboren?“

Die Antwortbereitschaft ist bei diesen freiwilligen Fragen insgesamt recht hoch. Die meisten Antwortausfälle entstehen bereits bei der ersten Frage. Der Anteil der Frauen ohne Angabe zur Geburt an allen Befragten der Zielgruppe (das heißt Frauen im Alter zwischen 15 und 75 Jahren) war mit knapp 10% im Jahr 2008 am höchsten und mit 7% im Jahr 2018 am geringsten. Der Anteil der Mütter ohne Angabe zur Kinderzahl bewegte sich zwischen 0,7 und 0,8%.

Eine Untersuchung der Antwortausfälle hat allerdings gezeigt, dass diese nicht unabhängig von anderen Merkmalen sind und bei bestimmten Merkmalskonstellationen viel häufiger als im Durchschnitt auftreten. Die ersten Hinweise darauf geben deutliche Unterschiede in den Antwortausfallquoten nach der Art der Befragung. Beim Mikrozensus kommen verschiedene Befragungsinstrumente zum Einsatz (Statistisches Bundesamt, 2017b): Face-to-Face-Interview (Computer Assisted Personal Interview, CAPI), schriftliche Befragung anhand eines Selbstausfüller-Fragebogens sowie zwei Arten von telefonischer Befragung (zum einen durch Interviewerinnen oder Interviewer, zum anderen durch Mitarbeiterinnen oder Mitarbeiter des zuständigen Statistischen Landesamtes, jeweils als sogenanntes Computer Assisted Telephone Interview, CATI).

Das am häufigsten genutzte Befragungsinstrument ist das persönliche Interview (CAPI). Im Jahr 2018 wurden auf diese Weise 63% der Frauen befragt. Zwar stieg der Anteil der telefonisch Befragten aufgrund von erweiterten Möglichkeiten beim Telefoninterview zwischen 2008

3 Das Interviewer-Handbuch und die Erläuterungen zum Selbstausfüllerfragebogen definieren, dass (1) Stief-, Adoptiv- oder Pflegekinder nicht berücksichtigt werden und (2) Kinder, die nicht lebend zur Welt gekommen sind, nicht mitgezählt werden.

und 2012. Mit einem Anteil von etwa 10% spielen die Telefonbefragungen im Mikrozensus jedoch immer noch eine untergeordnete Rolle. Der Anteil der Befragten, die den schriftlichen Fragebogen selbst ausfüllen, nahm dagegen seit 2008 deutlich zu, er belief sich 2018 auf 27%. In dieser Befragtengruppe ist zugleich die Antwortausfallquote besonders hoch – möglicherweise bedingt durch den Umfang und die Komplexität des Mikrozensusfragebogens mit seinen mehr als 200 Fragen. Deshalb spielen eine adäquate Platzierung und Gestaltung der freiwilligen Fragen eine besonders wichtige Rolle für die Antwortbereitschaft der Befragten. Im Mikrozensus 2008 wurde dies nicht ausreichend beachtet: Die Fragen zur Geburt wurden im letzten Teil des Fragebogens platziert, getrennt von den übrigen demografischen Informationen wie Alter und Familienstand. Dies hat offenbar zu einer besonders hohen Antwortausfallquote von 43% bei den Frauen beigetragen, die den Fragebogen selbst ausfüllten. Ab dem Mikrozensus 2012 wurde dieser Platzierungsfehler korrigiert. Infolgedessen ging die Antwortausfallquote der selbstausfüllenden Frauen bis auf 18% im Mikrozensus 2018 zurück. In der Struktur der Antwortausfälle hat jedoch diese Gruppe mit 67% nach wie vor den höchsten Anteil. [↘ Tabelle 1](#)

Tabelle 1
Struktur der Befragten und Antwortausfälle nach Art der Befragung bei der Frage nach der Geburt von Kindern im Mikrozensus

	2008	2012	2016	2018
%				
Verteilung der Befragten nach Art der Befragung				
CAPI	78	71	62	63
CATI	3	9	9	10
Selbstausfüller	19	20	29	27
Zusammen	100	100	100	100
Anteil der Antwortausfälle je Befragungsart				
CAPI	1	3	3	3
CATI	16	11	12	8
Selbstausfüller	43	28	19	18
Zusammen	10	8	9	7
Verteilung der Antwortausfälle nach Art der Befragung				
CAPI	11	23	23	22
CATI	5	12	13	11
Selbstausfüller	84	66	64	67
Zusammen	100	100	100	100

CAPI: Computer Assisted Personal Interview (persönliches Interview);
CATI: Computer Assisted Telephone Interview (telefonisches Interview).

Dabei unterscheiden sich die Selbstausfüllerinnen von den Teilnehmerinnen an einem Interview (CAPI/CATI) nach mehreren soziodemografischen Merkmalen: Sie sind durchschnittlich drei bis vier Jahre jünger als die CAPI/CATI-Teilnehmerinnen, höher gebildet (2018: mit akademischem Abschluss 19 gegenüber 15%), öfter erwerbstätig (2018: 67 gegenüber 60%) und seltener Mutter (2018: 61 gegenüber 67%).

Dies zeigt, dass die Antwortausfälle vermehrt bei einer Frauengruppe auftreten, die sich strukturell von den Frauen mit höherer Antwortbereitschaft unterscheidet. Ergebnisse, die ausschließlich auf den Angaben der antwortbereiten Frauen beruhen, sind deshalb mit hoher Wahrscheinlichkeit verzerrt. Um einen möglichen Bias (Verzerrung), zum Beispiel eine Unterschätzung der Kinderlosigkeit, zu reduzieren, wurden deshalb ab der ersten Befragungswelle 2008 fehlende Werte imputiert.

3

Erste Imputationserfahrungen

Bei den Datensätzen 2008, 2012 und 2016 wurde ein zweistufiges Imputationsverfahren angewendet (Bujard und andere, 2015).⁴ Im ersten Schritt wurden die Frauen ohne Angabe als Mutter oder als Kinderlose ein-

4 Die Autorinnen danken Sabine Köhne-Finster, die diese Imputationsmethode für den Mikrozensus 2008 entwickelt und dokumentiert hat.

gestuft. Im zweiten Schritt wurde die Zahl der geborenen Kinder bei Müttern geschätzt.

In der Stufe 1 erfolgte die Zuordnung einer Frau zu den Müttern oder zu den Kinderlosen nach zwei Ansätzen. Lebte eine Frau mit Kindern im Haushalt, die sie als ihre Mutter bezeichneten, wurde diese Frau als Mutter eingestuft. Bei dieser naheliegenden Zuordnung ist zu beachten, dass im Mikrozensus zwischen leiblicher Mutter einerseits und Pflege-, Adoptiv- oder Stiefmutter andererseits nicht unterschieden wird. Diese Unsicherheit lässt sich aber anhand der Angaben schätzen, die von den antwortbereiten Frauen erteilt wurden. Im Mikrozensus 2008 gaben 0,6 % der Frauen an, kein Kind geboren zu haben, obwohl sie von mindestens einem Kind im Haushalt als Mutter bezeichnet wurden. Diese geringe Unsicherheit wurde angesichts des Informationsgewinns durch die Imputation in Kauf genommen.

Für Frauen ohne Kind im Haushalt war dieses Vorgehen nicht möglich. Um von der „Kinderlosigkeit“ des Haushalts einer Frau auf die Kinderlosigkeit der Frau schließen zu können, bedarf es deshalb Annahmen auf Basis bekannter Merkmale der Frau. So zeigte sich für Frauen mit Angabe zur Geburt von Kindern, dass die Wahrscheinlichkeit, tatsächlich kinderlos zu sein, vor allem mit dem Bildungsniveau, dem Familienstand und dem Alter der Frau zusammenhängt. Mithilfe der logistischen Regression auf Basis der bekannten Angaben der Frauen ohne Kind im Haushalt wurden Wahrscheinlichkeiten, tatsächlich kinderlos zu sein, abgeleitet. In die Berechnung

Übersicht 1

Imputationsvorschrift für die Angabe „kein Kind geboren“ nach Familienstand, Bildungsstand und Alter

	Wahrscheinlichkeit, kein Kind geboren zu haben	
	≥ 90 % (Imputation)	< 90 % (keine Imputation)
Niedriger Bildungsstand ¹		
Unverheiratet und ohne Kind im Haushalt	unter 50 Jahre	50 Jahre oder älter
Verheiratet, geschieden, verwitwet und ohne Kind im Haushalt	unter 22 Jahre	22 Jahre oder älter
Mittlerer Bildungsstand ¹		
Unverheiratet und ohne Kind im Haushalt	unter 52 Jahre	52 Jahre oder älter
Verheiratet, geschieden, verwitwet und ohne Kind im Haushalt	unter 22 Jahre	22 Jahre oder älter
Hoher Bildungsstand ¹		
Unverheiratet und ohne Kind im Haushalt	unter 56 Jahre	56 Jahre oder älter
Verheiratet, geschieden, verwitwet und ohne Kind im Haushalt	unter 26 Jahre	26 Jahre oder älter

1 Nach der Internationalen Standardklassifikation im Bildungswesen – International Standard Classification of Education (ISCED-97/ISCED-2011).

sind die abhängige dichotome Variable „kinderlos“ (ja/nein) sowie die kategorisierten unabhängigen Variablen Bildungsstand, Familienstand und Alter eingegangen. Die Angabe „kein Kind geboren“ wurde nur dann imputiert, wenn eine Frau im Haushalt ohne Kind lebte und aufgrund ihres Bildungsstandes, ihres Familienstandes und ihres Alters eine über 90-prozentige Wahrscheinlichkeit hatte, kein Kind geboren zu haben. Der Schwellenwert von 90% sollte eine hohe Verlässlichkeit bei gleichzeitig befriedigender Ausschöpfung gewährleisten. [↘ Übersicht 1](#)

In der Stufe 2 wurde die Zahl der geborenen Kinder bei Müttern geschätzt. Auf Basis der Angaben der befragungsbereiten Mütter wurde untersucht, in welchen Fällen die Zahl der geborenen Kinder der Frau mit der Zahl der Kinder in der Familie übereinstimmt. Eine entscheidende Rolle spielte dabei das Alter der Mutter. Je älter sie ist, desto höher ist die Wahrscheinlichkeit, dass mindestens ein Kind bereits aus dem elterlichen Haushalt ausgezogen ist. Bei Frauen ab 40 Jahren ist deshalb davon auszugehen, dass die Zahl der geborenen Kinder größer ist als die Zahl der Kinder in der Familie. Neben dem Alter wurden Wohnort (früheres Bundesgebiet/neue Länder⁵) und Bildungsstand der Frau als Einflussfaktoren definiert. So verlassen die Kinder im früheren Bundesgebiet den elterlichen Haushalt später als in den ostdeutschen Bundesländern. Die höher gebildeten Frauen werden im Durchschnitt später Mutter. Auf dieser Basis wurden verschiedene Konstellationen abgeleitet, um die Auszugswahrscheinlichkeit für mindestens ein Kind zu schätzen und die Imputationsregeln zu formulieren. Ist die Auszugswahrscheinlichkeit aufgrund der Merkmale der Mutter (Alter, Bildungsstand und Wohnort) kleiner als 90%, wird die Zahl der geborenen Kinder mit der Zahl der Kinder in der Familie gleichgesetzt.

5 Neue Länder einschließlich Berlins.

Übersicht 2

Imputationsvorschrift für die Zahl der geborenen Kinder¹ nach Wohnort, Bildungsstand und Alter

	Auszugswahrscheinlichkeit	
	≥ 90% (keine Imputation)	< 90% (Imputation)
	Niedriger Bildungsstand ²	
Früheres Bundesgebiet	36 Jahre oder älter	unter 36 Jahre
Neue Länder	36 Jahre oder älter	unter 36 Jahre
	Mittlerer Bildungsstand ²	
Früheres Bundesgebiet	40 Jahre oder älter	unter 40 Jahre
Neue Länder	38 Jahre oder älter	unter 38 Jahre
	Hoher Bildungsstand ²	
Früheres Bundesgebiet	42 Jahre oder älter	unter 42 Jahre
Neue Länder	40 Jahre oder älter	unter 40 Jahre

1 Die Angabe zur Zahl der geborenen Kinder wird nur dann aus der Zahl der Kinder in der Familie übernommen, wenn die Wahrscheinlichkeit für den Auszug mindestens eines Kindes unter 90% beträgt.

2 Nach der Internationalen Standardklassifikation im Bildungswesen – International Standard Classification of Education (ISCED-97/ISCED-2011).

Beträgt dagegen diese Wahrscheinlichkeit 90% oder mehr, wird kein Wert imputiert. [↘ Übersicht 2](#)

Um die Güte des Verfahrens zu bewerten, erfolgten Testrechnungen. Dabei wurden die Angaben der Frauen, die Fragen zur Geburt beantwortet haben, mit dem gleichen Imputationsverfahren geschätzt und die tatsächlichen Angaben anschließend mit den generierten Ergebnissen verglichen. Die dabei festgestellte geringe Fehlerquote attestiert dem Verfahren eine hohe Vorhersagegüte.⁶

Das beschriebene Verfahren hat allerdings einen wesentlichen Nachteil. Nach der Imputation der Angabe „Mutter versus Kinderlose“ in der ersten Stufe hat sich die Zahl der Antwortausfälle in allen drei Befragungswellen mehr als halbiert und ihr Anteil ist auf 4% gesunken. In der zweiten Stufe wurde dagegen nur ein geringer Teil der ursprünglichen Antwortausfälle durch eine Angabe zur Kinderzahl ersetzt. Zusätzlich zu den offen gebliebenen Fällen kamen die Frauen, die in der ersten Imputationsstufe den Müttern zugeordnet wurden, für die aber keine Angabe zur Kinderzahl geschätzt werden konnte. Dadurch hat sich die Anzahl der Mütter ohne Angabe zur Kinderzahl nach der Imputation im Vergleich zum Originaldatensatz sogar vergrößert. [↘ Tabelle 2](#)

Besonders für die Frage nach der Struktur der Frauenjahrgänge nach der Parität hat sich diese Diskrepanz zwi-

6 Auf eine ausführliche Darstellung der Ergebnisse der Testrechnungen wird verzichtet.

Tabelle 2

Antwortausfälle vor und nach der Imputation im Mikrozensus

	2008	2012	2016
Mutter versus Kinderlose			
Antwortausfälle vor Imputation	25 268	21 820	23 842
Imputierte Angaben	14 228	12 154	12 125
Antwortausfälle nach Imputation	11 040	9 666	11 717
Non-Response-Rate nach der Imputation in %	4	4	4
Zahl der geborenen Kinder			
Antwortausfälle vor Imputation	1 140	1 113	1 154
Imputierte Angaben	141	112	38
Antwortausfälle nach Imputation (einschließlich Frauen ohne Angabe zur Geburt, die den Müttern zugeordnet wurden)	5 685	5 385	6 494
Non-Response-Rate nach der Imputation in %	3	3	4

schen den Müttern insgesamt und Müttern mit Angabe zur Kinderzahl als problematisch erwiesen. Die Antwortausfälle verteilen sich nicht gleichmäßig nach dem Alter der Frau, sondern schwanken je nach Jahrgang zwischen 1 und 8 % (Statistisches Bundesamt, 2017a, hier: Seite 7 ff.). Es ist zudem anzunehmen, dass sie sich nicht proportional nach der Zahl der geborenen Kinder verteilen. Bei Auswertungen, die diese Unsicherheiten nicht berücksichtigen, kann dies zu Verzerrungen führen.

Für den Mikrozensus 2018 wurde deshalb ein alternatives Verfahren entwickelt, welches eine bessere Abstimmung zwischen den beiden Imputationsschritten gewährleistet und durch eine effektivere Imputation die Datengrundlage für die Fertilitätsanalysen verbessert.

4

Alternative Methoden zur Imputation der Antwortausfälle

Es gibt verschiedene Imputationsverfahren, die genutzt werden können, um fehlende Werte in den Daten komplett zu ersetzen. Die Wahl eines geeigneten Verfahrens hängt von der konkreten Datensituation ab. Bei den Daten im Mikrozensus wurde, wie in Kapitel 2 beschrieben, bei den Antwortausfällen eine Systematik erkannt. Dass die fehlenden Werte bei der Kinderanzahl von der Art der Befragung abhängen spricht dafür, dass der Antwortausfall Missing At Random (MAR) ist. Ein Antwortausfall wird als Missing At Random bezeichnet, wenn

der Ausfall innerhalb eines Merkmals zwar von den beobachteten Ausprägungen der anderen im Datensatz vorhandenen Variablen abhängig ist, nicht jedoch von der fehlenden Ausprägung des Merkmals selbst (Rubin, 1976). Einerseits ist davon aus fachlicher Sicht bei den Angaben zur Geburt im Mikrozensus nicht auszugehen. Andererseits können Verfahren, die für einen MAR-Ausfall konzipiert sind, den Ausfall über die Korrelation mit anderen Merkmalen zum größten Teil korrigieren. Es wird angenommen, dass die fehlenden Werte mithilfe zusätzlicher Variablen erklärt werden können.

Eine adäquate Methode bei einem MAR-Ausfall ist Predictive Mean Matching. Diese Methode gehört zu den sogenannten Hot-Deck-Verfahren. Die fehlenden Angaben werden dabei durch beobachtete Werte aus derselben Erhebung ersetzt.

Beim Predictive Mean Matching wird zunächst anhand der Datensätze mit vollständigen Angaben ein Regressionsmodell spezifiziert. Im vereinfachten Modell mit nur einer Prädiktorvariablen besitzt das Modell die Form

$$(1) y_{obs} = \beta_0 + x_{obs}\beta_1 + \epsilon$$

Dabei enthält $y_{obs} = \begin{pmatrix} y_{1obs} \\ \vdots \\ y_{Jobs} \end{pmatrix}$ die beobachteten Werte

der zu imputierenden Variable. y_{obs} enthält die Merkmalswerte der Prädiktorvariable für die Fälle mit beobachteten Werten in y . Mit β_0 und β_1 werden die Regressionskoeffizienten und mit ϵ der Störterm bezeichnet. Anschließend werden die durch dieses Modell geschätzten Modellparameter $\hat{\beta}_0$ und $\hat{\beta}_1$ berechnet. Im Fall einer

multiplen Imputation (siehe Ende des Kapitels) werden statt der Punktschätzer Ziehungen aus der A-Posteriori-Verteilung der Parameter verwendet (van Buuren, 2018). Mithilfe dieser geschätzten Modellparameter $\hat{\beta}_0$ und $\hat{\beta}_1$ und der definierten Einflussvariable werden zum einen

die prognostizierten Schätzwerte $\hat{y}_{mis} = \begin{pmatrix} \hat{y}_{1mis} \\ \vdots \\ \hat{y}_{Imis} \end{pmatrix}$ für

die Beobachtungen mit fehlenden Werten in y berechnet. Zum anderen werden auch die prognostizierten Schätzwerte \hat{y}_{obs} für die Beobachtungen mit bekannten y -Werten ermittelt. Dann werden für eine Beobachtung $i = 1, \dots, I$ mit einem fehlenden Wert \hat{y}_{imis} zum Beispiel die fünf potenziellen Spender ausgewählt, für die der Abstand $|\hat{y}_{imis} - \hat{y}_{obs}| (j=1, \dots, 5)$ minimal ist. Das sind also die Spender, deren Schätzwerte am nächsten an dem Schätzwert für den Datensatz mit der fehlenden Angabe bei der Zielvariablen liegen. Aus diesen Kandidaten wird dann ein Spender zufällig ausgewählt. Der bekannte Wert dieses ausgewählten Spenders wird dann als imputierter Wert y_{iimp} anstelle der fehlenden Angabe y_{imis} eingesetzt (Koller-Meinfelder, 2009). In [Grafik 1](#) ist Predictive Mean Matching für den einfachen Fall mit einer Prädiktorvariablen x und für einen fehlenden Wert y_{mis} dargestellt. Das Vorgehen beim Predictive Mean

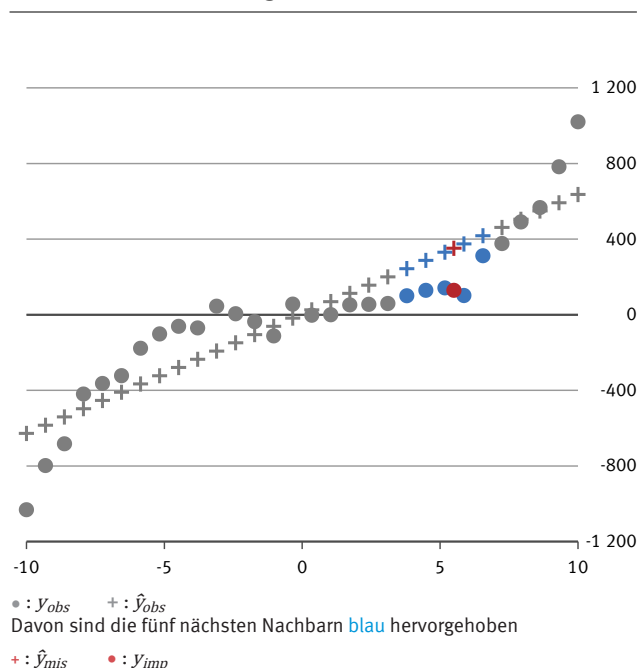
Matching setzt voraus, dass die Verteilung des fehlenden Werts identisch ist zur Verteilung der bekannten Angaben der Kandidaten (van Buuren, 2018).

Ein Vorteil dieses Imputationsverfahrens ist, dass die Verteilung der fehlenden Werte nicht parametrisiert werden muss. Predictive Mean Matching eignet sich gut für diskrete Merkmale wie die Anzahl geborener Kinder. Dabei werden nur plausible Werte imputiert, die auch tatsächlich beobachtet wurden.

Bei dem beschriebenen Vorgehen zur Methode Predictive Mean Matching wird allerdings anstelle jeder fehlenden Angabe bislang nur genau ein Wert imputiert. Dieser Ansatz wird als Single Imputation bezeichnet. Für anschließende Analysen werden die imputierten Werte dann wie beobachtete Werte behandelt. Die Unsicherheit, die mit diesen nicht beobachteten Werten einhergeht, wird nicht berücksichtigt. Das führt dazu, dass die berechnete Stichprobenvarianz kein unverzerrter Schätzer mehr für die Varianz des interessierenden Parameters (zum Beispiel Kinderlosenquote) ist. Der relative Standardfehler wird dadurch unterschätzt und die Präzision der Ergebnisse überschätzt.

Um dieses Problem zu beheben, kann eine Multiple Imputation angewendet werden. Dabei wird jeder fehlende Wert m -mal ersetzt, wobei $m > 1$ ist. Dieses Verfahren liefert m vollständige Datensätze. Diese werden separat analysiert und anschließend kombiniert: Der multiple Imputationsschätzer ist das arithmetische Mittel der Schätzer aus den m imputierten Datensätzen. Der Varianzschätzer der multiplen Imputation bezieht die Unsicherheit aufgrund des Fehlens der Werte mit ein. Er setzt sich zusammen aus der sogenannten Within-Varianz, welche als arithmetisches Mittel der Varianzen aus den m imputierten Datensätzen berechnet wird, und der sogenannten Between-Varianz, welche die Varianz zwischen den m Parameterschätzern aus den imputierten Datensätzen abbildet (Rubin, 1987).

Grafik 1
Predictive Mean Matching



2019 - 01 - 0604

5

Simulationsstudien

Um ein für den Mikrozensus passendes Imputationsverfahren auszuwählen, wurde eine Simulationsstudie durchgeführt. Anhand der Datenkonstellation und der Anforderungen an die Imputation erschienen zwei Ansätze aussichtsreich. Bei Methode A wird zuerst der Mutterschaftsstatus durch die Schätzwerte einer logistischen Regression ersetzt. Frauen, die dadurch den Status „kinderlos“ erhielten, wurden dementsprechend null Kinder zugeordnet. Für die anderen Frauen, also alle Mütter, wird danach das Verfahren Predictive Mean Matching angewandt, um eine Kinderzahl größer null zu imputieren. Bei Methode B werden die beiden Variablen in umgekehrter Reihenfolge imputiert. Dort wird als erstes das Verfahren Predictive Mean Matching angewendet, um die Anzahl geborener Kinder einschließlich null zu imputieren. Anschließend wird eine deterministische Imputation des Mutterschaftsstatus durchgeführt, wobei Frauen mit einer imputierten Anzahl von null Kindern zum Status „kinderlos“ zugeordnet werden und die anderen Frauen den Status „Mutter“ zugewiesen bekommen.

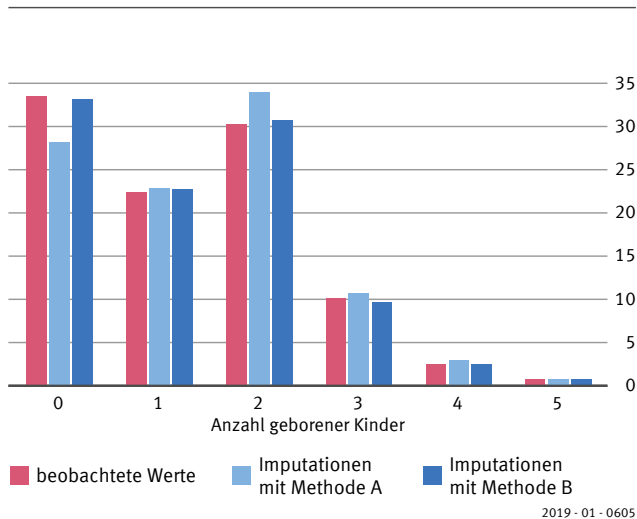
Für die Simulationsstudie wurden die Daten der letzten Mikrozensusbefragung aus dem Jahr 2016 als Testdaten genutzt. Bei rund 10% der vollständigen Datensätze ohne Item-Non-Response wurden die Angaben zu geborenen Kindern zunächst zufällig gelöscht. Diese wurden gemäß den Methoden A und B imputiert. Auf diese Weise wurden die Verteilungen anhand der imputierten und der beobachteten Werte verglichen. Als Prädiktorvariablen wurden Merkmale aus dem Mikrozensus für die Modellbildung verwendet, die vollständig verfügbar sind. Dabei handelte es sich um demografische und sozioökonomische Variablen mit Angaben zu Staatsangehörigkeit, Bildung, Erwerbstätigkeit und so weiter sowie die Art der Befragung, die aufgrund der Abhängigkeit zum Ausfallmuster in die Modellbildung eingeht.

Um die Wahrscheinlichkeit zufälliger extremer Ergebnisse zu reduzieren, wurde das Löschen und Ersetzen von Werten mehrfach durchgeführt. Schon die geringe Anzahl von zehn Simulationsdurchläufen hat gezeigt, dass eine der beiden Methoden deutlich überlegen ist, und erlaubte ein Urteil zur Qualität der Imputation. Die

Streuung zwischen den einzelnen Simulationsdurchläufen war gering, daher war durch eine größere Wiederholungszahl kein zusätzlicher Informationsgewinn zu erwarten.

Grafik 2

Frauen nach Zahl der geborenen Kinder 2016:
Vergleich der imputierten und der beobachteten Werte in %

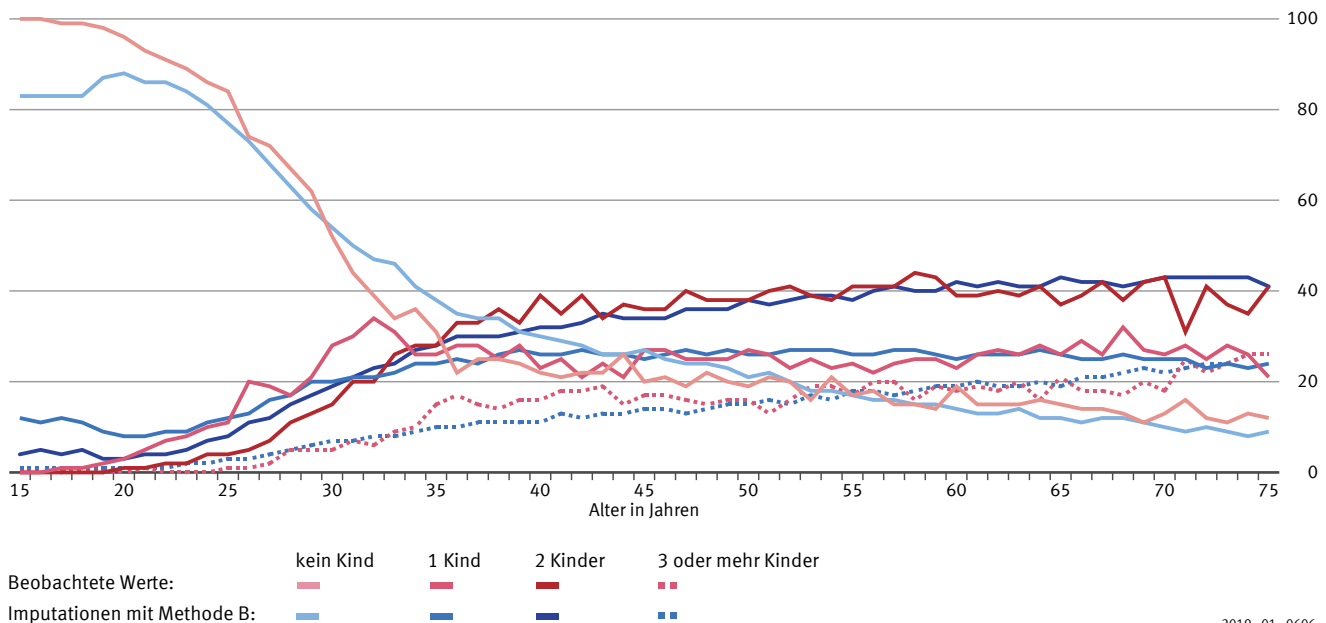


➤ Grafik 2 zeigt die Verteilung der Frauen nach Parität (kein Kind, ein, zwei, drei, vier, fünf Kinder)⁷. Dargestellt sind die gemittelten Ergebnisse über die zehn Imputationsdurchläufe. Die mit den beiden Methoden A und B imputierten Werte werden mit den entsprechenden beobachteten Anteilen verglichen. Hierbei wird deutlich, dass Methode B annähernd dieselben Anteile für die Anzahl geborener Kinder liefert, wie sie tatsächlich in den Ausgangsdaten beobachtet werden konnten. Methode A hingegen weicht insbesondere bei den Angaben zu null und zwei Kindern deutlich davon ab. Mit dem Vorgehen von Methode A wurden 5% weniger Frauen als kinderlos eingeordnet. Methode A, bei der zunächst der Mutterschaftsstatus mittels logistischer Regression und danach die Kinderzahl mit Predictive Mean Matching imputiert werden, erweist sich somit als nicht geeignet, um unverzerrte Schätzer zu liefern. Deshalb wird im Weiteren näher auf Methode B eingegangen, bei der als erstes die Kinderzahl mit Predictive Mean Matching imputiert wird und anschließend eine deterministische Imputation des Mutterschaftsstatus erfolgt.

7 Die Anteile der Frauen mit mehr Kindern sind gering und unterscheiden sich zwischen den beiden Methoden auch nur geringfügig.

Grafik 3

Frauen nach Alter und Zahl der geborenen Kinder: Vergleich der imputierten und der beobachteten Werte in %



2019 - 01 - 0606

Betrachtet man die Verteilung nach Kinderzahl insgesamt, müsste Methode B favorisiert werden. Eine tiefergehende Prüfung zeigt allerdings auch bei dieser Methode Probleme. Werden die Anteile der Frauen nach Alter und Kinderanzahl für die nach Methode B imputierten und für die beobachteten Daten miteinander verglichen, sind zwischen den Kurven mit der gleichen Kinderzahl deutliche Abweichungen zu erkennen. Diese Unterschiede können insbesondere für die jungen Frauen (unter 25 Jahre) festgestellt werden. Eine genauere Analyse der Imputationen zeigt, dass das Alter der Frau in einem verwendeten Spenderdatensatz teilweise um 40 Jahre höher oder niedriger liegen kann als im Datensatz mit einer fehlenden Angabe zur Kinderzahl. Es kann also vorkommen, dass mit dieser Methode in Einzelfällen für eine 17-Jährige die Kinderzahl einer 57-Jährigen imputiert wurde. [↘ Grafik 3](#)

Um diese Fälle auszuschließen, wurden sogenannte Imputationsklassen eingeführt. Diese Imputationsklassen unterteilen die Daten in homogene Untergruppen. Die Zugehörigkeit zu einer Imputationsklasse definiert sich über ein oder mehrere Merkmale oder Merkmalsausprägungen, die bei allen Datensätzen dieser

Imputationsklasse übereinstimmen müssen. In diesem Fall ist es sinnvoll, das Merkmal Alter für die Bildung der Imputationsklassen zu nutzen, um die Probleme von Methode B zu beheben. Mit der Ersetzung innerhalb einer Altersklasse werden dann nur Werte von Datensätzen eingesetzt, die derselben Altersklasse angehören. Die Grenzen der Altersklassen werden in Altersabständen von fünf Jahren gezogen, zum Beispiel 15 bis 19 Jahre, 20 bis 24 Jahre, 25 bis 29 Jahre.

Auch dieses Vorgehen wurde in einer zweiten Simulationsstudie anhand des Datensatzes aus dem Mikrozensus 2016 getestet. Diesmal wurden die fehlenden Werte jedoch nicht nur zufällig erzeugt. Um auch die Auswirkungen bei einem systematischen Ausfall zu analysieren – wie in den Voruntersuchungen bei den Daten von 2016 festgestellt –, wurden drei verschiedene Ausfallszenarien konstruiert:

- 1) zufälliges Ausfallmuster;
- 2) von der Art der Befragung abhängiges Ausfallmuster, das heißt eine erhöhte Ausfallwahrscheinlichkeit bei den Selbstausfüllenden;

3) fehlende Werte werden mit den vorhergesagten Wahrscheinlichkeiten für den Ausfall gezogen. Dafür wurde ein logistisches Modell für den Ausfall anhand des Mikrozensus 2016 erstellt.

Die so erzeugten fehlenden Werte wurden mit Methode B zum einen für alle relevanten Daten insgesamt und zum anderen innerhalb der Altersklassen imputiert. Die Imputation mit Methode B ohne Imputationsklassen liefert das gleiche Resultat wie in der ersten Simulationsstudie. Unabhängig davon, welches der drei Ausfallszenarien betrachtet wird, werden bei den unter 25-Jährigen sowie bei den ab 55-Jährigen mehr Mütter imputiert als bei den erhobenen Daten existieren. Im Gegensatz dazu werden bei den 25- bis 55-Jährigen zu viele Frauen mit dem Status kinderlos imputiert. Dieser Ansatz scheint also verzerrte Ergebnisse hervorzubringen, da der Anteil der Mütter je nach Altersklasse systematisch unter- beziehungsweise überschätzt wird.

Anders sieht es bei der Imputation mit Methode B innerhalb der Altersklassen aus. Bei dem zufälligen Ausfallmuster und dem von der Art der Befragung abhängigen Ausfallmuster (Ausfallszenarien 1 und 2) liegen die Abweichungen zwischen den beobachteten und den imputierten Anteilen für die Frauen nach Zahl der geborenen Kinder und Altersklassen bei maximal einem Prozentpunkt. Außerdem befindet sich hier der imputierte Anteil nicht systematisch über oder unter dem Anteil der beobachteten Werte für die einzelnen Altersklassen. Beim Ausfallmuster mit dem logistischen Modell (Ausfallszenario 3) werden ebenfalls Ergebnisse erzeugt, die deutlich besser sind als bei der Imputation ohne Imputationsklassen. Hier kommen Abweichungen um einen Prozentpunkt für eine bestimmte Kinderzahl in einzelnen Altersklassen häufiger vor. Dennoch werden auch unter diesem Ausfallszenario zufriedenstellende Ergebnisse erzielt.

Die Ergebnisse dieser Simulationsstudie zeigen, dass die Imputation der Kinderzahl mit Predictive Mean Matching innerhalb von Altersklassen und die anschließende Übertragung auf den Mutterschaftsstatus ein geeignetes Imputationsverfahren für die fehlenden Angaben zu geborenen Kindern im Mikrozensus ist. Gegenüber dem bisherigen, in den Erhebungen 2008, 2012 und 2016 verwendeten Verfahren bietet das Vorgehen den Vorteil, dass alle fehlenden Werte ersetzt werden.

6

Anwendung des Imputationsverfahrens im Mikrozensus 2016 und 2018

6.1 Vergleich mit den veröffentlichten Ergebnissen 2016

Beim Vergleich der Ergebnisse auf Basis des alten und des neuen Imputationsverfahrens werden die Kinderlosenquote und die Verteilung der Mütter nach Zahl der geborenen Kinder separat betrachtet. Der Grund dafür liegt in der Besonderheit des früheren zweistufigen Imputationsverfahrens. Wie in Kapitel 3 erläutert, wurden dabei nicht alle fehlenden Werte imputiert. Während die Eigenschaft „Mutter“ oder „kinderlos“ in den meisten Fällen zugeordnet werden konnte, war eine Schätzung für die fehlende Kinderzahl bei Müttern in vielen Fällen nicht möglich. Eine Berechnung der Anteile der Mütter nach Zahl der geborenen Kinder (Parität) muss sich in diesem Fall nur auf die Mütter mit Angabe zur Zahl der geborenen Kinder beziehen, da die Verteilung der verbleibenden Antwortausfälle unbekannt ist.

Die Kinderlosenquoten 2016 auf Basis des alten und des neuen Imputationsverfahrens weisen keine gravierenden Unterschiede auf. Die jahrgangswise Abweichungen betragen weniger als einen Prozentpunkt und sind somit vernachlässigbar gering. Auch die regionalisierten Ergebnisse (für westliche und östliche Flächenländer sowie für die Stadtstaaten) nach Altersgruppen weisen nur geringe Abweichungen auf. Der Anteil der kinderlosen Frauen nach der neuen Imputationsmethode liegt dabei tendenziell geringfügig – bis zu einem Prozentpunkt – höher. Diese Tendenz ist – ausgehend von der in Kapitel 2 beschriebenen Zusammensetzung der Antwortausfälle – nicht überraschend. Da die meisten Antwortausfälle bei Selbstausfüllerinnen entstehen, die zugleich zu einem höheren Anteil kinderlos sind, führt eine vollständige Imputation zu geringfügig mehr kinderlosen Frauen als das bisherige Verfahren.

↪ **Tabelle 3, Grafik 4**

Eine ähnlich gute Übereinstimmung zeigt der Vergleich der Mütteranteile nach der Parität. Bei den Geburtsjahrgängen bis 1996 betragen die Abweichungen

Neues Imputationsverfahren bei Antwortausfällen zu geborenen Kindern im Mikrozensus

Tabelle 3

Kinderlosenquote, Mütter nach Zahl der geborenen Kinder und durchschnittliche Kinderzahl je Mutter 2016
Ergebnisvergleich auf Basis des alten und des neuen Imputationsverfahrens¹

	Kinderlosenquote		Verteilung der Mütter nach Zahl der geborenen Kinder						Durchschnittliche Kinderzahl je Mutter	
	alt	neu	mit 1 Kind		mit 2 Kindern		mit 3 oder mehr Kindern		alt	neu
			alt	neu	alt	neu	alt	neu		
Deutschland										
15 bis 19 Jahre	99	99	/	/	/	/	/	/	/	/
20 bis 24 Jahre	91	91	74	73	22	22	5	5	1,3	1,4
25 bis 29 Jahre	70	71	62	61	29	30	9	9	1,5	1,5
30 bis 34 Jahre	42	43	47	48	39	39	13	13	1,7	1,7
35 bis 39 Jahre	25	26	35	35	46	46	19	19	1,9	1,9
40 bis 44 Jahre	21	22	31	31	48	48	21	21	2,0	2,0
45 bis 49 Jahre	20	21	32	32	48	48	20	20	2,0	2,0
50 bis 54 Jahre	19	20	31	31	48	48	21	21	2,0	2,0
55 bis 59 Jahre	17	17	28	28	49	49	23	23	2,0	2,0
60 bis 64 Jahre	16	16	30	30	48	48	22	22	2,0	2,0
65 bis 69 Jahre	14	14	31	31	48	48	22	22	2,0	2,0
70 bis 75 Jahre	12	12	29	28	46	46	26	26	2,1	2,1
Westdeutsche Flächenländer										
15 bis 19 Jahre	99	99	/	/	/	/	/	/	/	/
20 bis 24 Jahre	92	91	73	72	23	23	4	5	1,3	1,4
25 bis 29 Jahre	71	72	60	59	31	31	9	9	1,5	1,6
30 bis 34 Jahre	42	43	46	46	40	40	14	14	1,7	1,7
35 bis 39 Jahre	25	26	33	33	47	47	20	20	1,9	2,0
40 bis 44 Jahre	22	22	29	29	48	48	22	22	2,0	2,0
45 bis 49 Jahre	21	22	30	30	49	49	22	21	2,0	2,0
50 bis 54 Jahre	20	21	29	29	49	48	22	22	2,0	2,0
55 bis 59 Jahre	18	19	27	28	48	48	24	24	2,1	2,1
60 bis 64 Jahre	17	17	30	30	47	47	23	23	2,1	2,1
65 bis 69 Jahre	14	15	30	30	47	47	23	23	2,0	2,1
70 bis 75 Jahre	13	13	27	26	46	46	27	28	2,2	2,2
Ostdeutsche Flächenländer										
15 bis 19 Jahre	98	98	/	/	/	/	-	/	/	/
20 bis 24 Jahre	86	85	80	79	/	16	/	5	1,3	1,3
25 bis 29 Jahre	60	60	68	68	23	24	9	9	1,4	1,5
30 bis 34 Jahre	32	32	52	52	38	38	10	10	1,6	1,6
35 bis 39 Jahre	18	19	39	38	45	45	16	16	1,8	1,9
40 bis 44 Jahre	15	16	39	38	46	46	15	15	1,8	1,8
45 bis 49 Jahre	12	12	43	43	43	43	13	14	1,8	1,8
50 bis 54 Jahre	10	11	37	36	49	49	14	15	1,8	1,8
55 bis 59 Jahre	8	8	29	29	53	53	17	18	1,9	2,0
60 bis 64 Jahre	7	7	31	31	53	52	17	17	1,9	1,9
65 bis 69 Jahre	7	7	31	31	52	52	17	17	1,9	1,9
70 bis 75 Jahre	8	8	32	32	45	45	23	24	2,0	2,0

noch Tabelle 3

Kinderlosenquote, Mütter nach Zahl der geborenen Kinder und durchschnittliche Kinderzahl je Mutter 2016
Ergebnisvergleich auf Basis des alten und des neuen Imputationsverfahrens¹

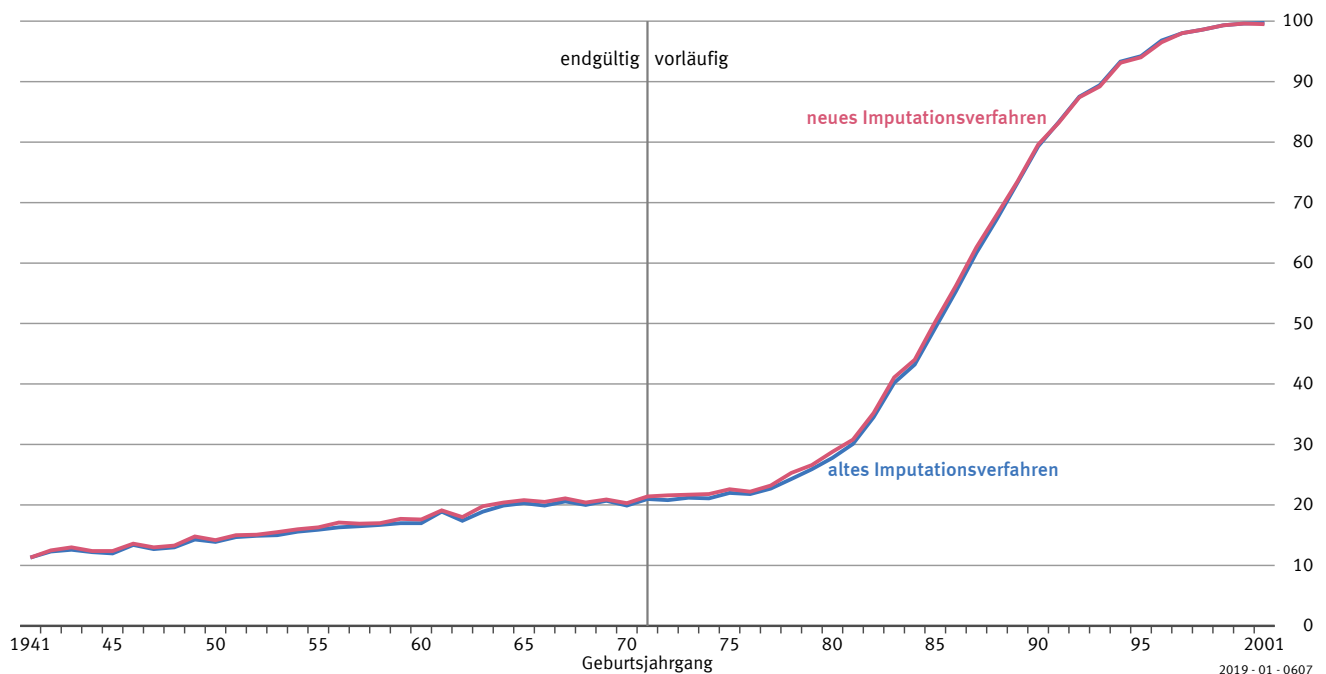
	Kinderlosenquote		Verteilung der Mütter nach Zahl der geborenen Kinder						Durchschnittliche Kinderzahl je Mutter	
			mit 1 Kind		mit 2 Kindern		mit 3 oder mehr Kindern			
	alt	neu	alt	neu	alt	neu	alt	neu	alt	neu
Stadtstaaten										
15 bis 19 Jahre	99	99	/	/	/	/	-	/	/	/
20 bis 24 Jahre	89	89	67	66	/	24	/	10	1,4	1,5
25 bis 29 Jahre	76	76	62	60	30	32	/	8	1,5	1,5
30 bis 34 Jahre	54	55	49	50	37	35	14	14	1,7	1,7
35 bis 39 Jahre	37	38	41	40	42	42	18	19	1,9	1,9
40 bis 44 Jahre	29	29	34	35	43	43	22	22	2,0	2,0
45 bis 49 Jahre	28	29	39	39	44	44	17	17	1,9	1,9
50 bis 54 Jahre	25	26	37	37	44	44	19	19	1,9	1,9
55 bis 59 Jahre	26	26	35	35	46	46	20	20	1,9	1,9
60 bis 64 Jahre	25	26	34	35	47	47	18	18	1,9	1,9
65 bis 69 Jahre	21	22	35	36	47	46	18	18	1,9	1,9
70 bis 75 Jahre	17	18	38	37	44	44	19	19	1,9	1,9

1 Ergebnisse des Mikrozensus 2016 nach dem alten Imputationsverfahren: Statistisches Bundesamt 2017a, Tabellen 1.2 und 1.4.

Grafik 4

Kinderlosenquote nach Jahrgängen 2016

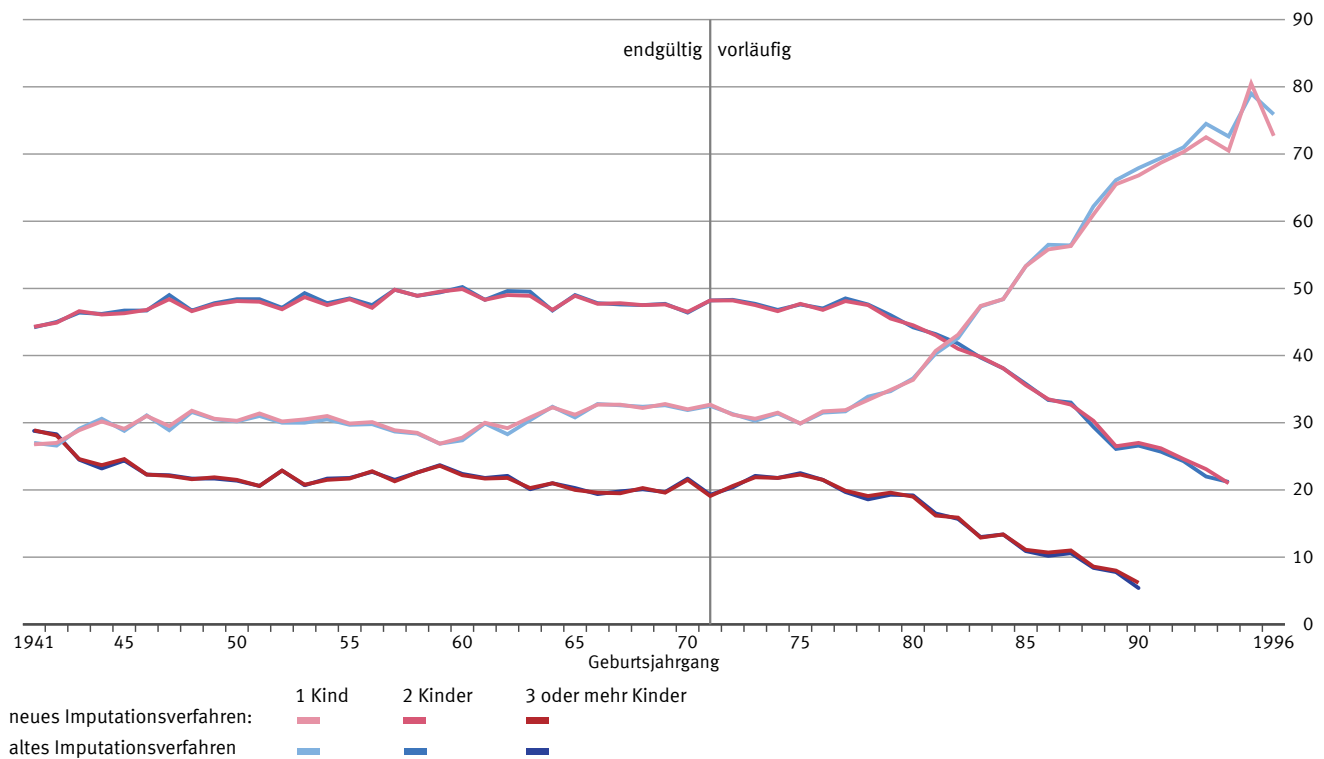
Ergebnisse auf Basis des alten und des neuen Imputationsverfahrens im Vergleich, in %



Grafik 5

Mütter nach Zahl der geborenen Kinder und Jahrgängen 2016

Ergebnisse auf Basis des alten und des neuen Imputationsverfahrens im Vergleich, in %



2019 - 01 - 0608

durchschnittlich weniger als einen Prozentpunkt. Lediglich bei sehr jungen Frauen, die im Jahr 2016 erst zwischen 15 und 19 Jahren alt und zu einem sehr geringen Anteil Mutter waren, weichen die Ergebnisse beider Imputationsverfahren stärker voneinander ab. Aufgrund der niedrigen Fallzahl sind aber die Mütteranteile bei den unter 20-Jährigen ohnehin unsicher und werden nicht weiter thematisiert. [↘ Grafik 5](#)

Dieser Vergleich zeigt, dass auch das frühere Imputationsverfahren valide Ergebnisse geliefert hat und dass kein Revisionsbedarf für die Geburtenindikatoren der Jahre 2008, 2012 und 2016 besteht (Statistisches Bundesamt, 2009, 2013 sowie 2017a). Das neue Imputationsverfahren liefert jedoch eine größere verwertbare Fallzahl und erweitert zudem die Analysemöglichkeiten. Da es eine vollständige Imputation fehlender Items ermöglicht, schließt sich die Lücke zwischen der Zahl der Mütter einerseits und der Summe der Mütter mit Angabe zur Zahl der geborenen Kinder andererseits. Damit können – unter Beachtung üblicher Stichprobeneinschränkun-

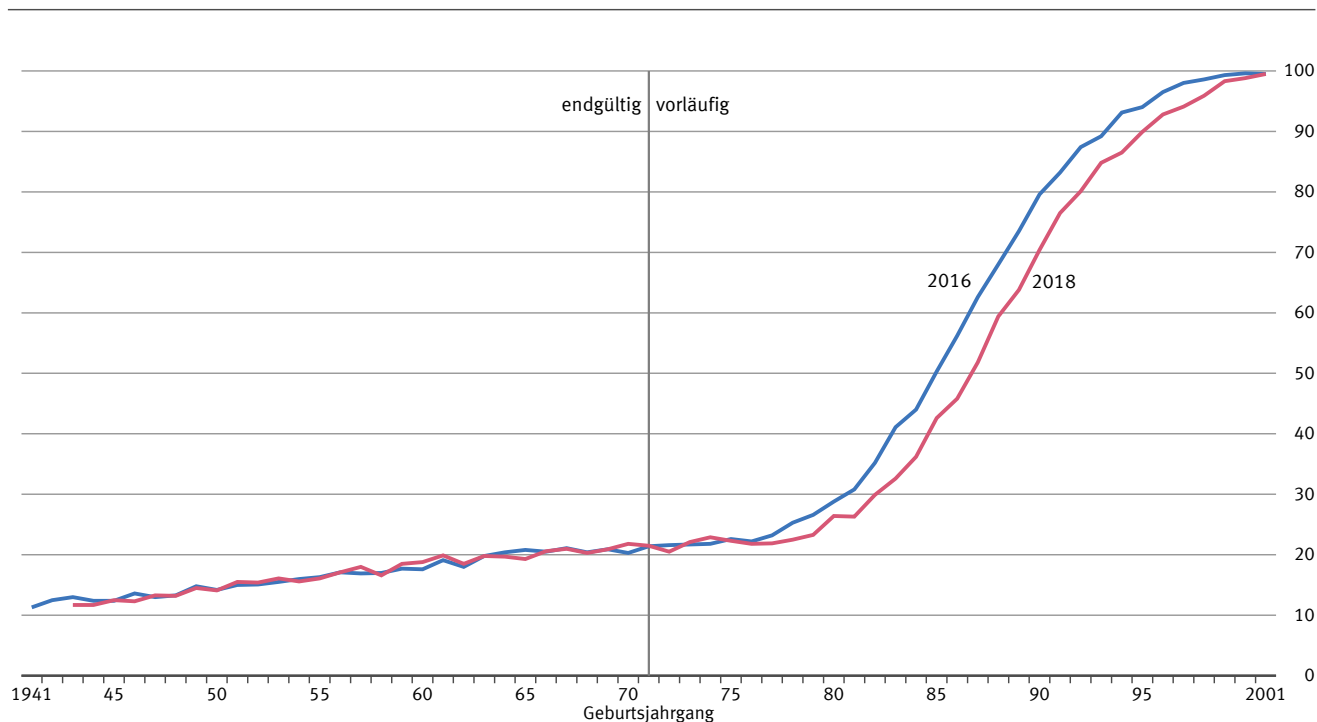
gen – die Paritätsverteilung einschließlich kinderloser Frauen vollständig nachgewiesen und die durchschnittliche Kinderzahl je Frau problemlos berechnet werden.

6.2 Ergebnisse 2018

Für den Datensatz des Mikrozensus 2018 wurde nur das neue Imputationsverfahren angewendet. Die auf Basis des Mikrozensus 2018 ermittelte sogenannte endgültige Kinderlosenquote für die Geburtsjahrgänge 1943 bis 1976, die im Jahr 2018 zwischen 42 und 75 Jahren alt waren, weicht nur geringfügig von den Ergebnissen des Mikrozensus 2016 ab (der Vergleich bezieht sich hier auf die neu imputierten Ergebnisse 2016; [↘ Grafik 6](#)). Kleinere Schwankungen erscheinen plausibel, da zwischen den Befragungen zwei Jahre liegen, in denen sich die Frauenjahrgänge durch starke Migrationsbewegungen und Sterblichkeit veränderten. Auch erhebungsbedingte Fehler können nicht ausgeschlossen werden (siehe Kapitel 1).

Grafik 6

Kinderlosenquote nach Jahrgängen
in %



Ergebnisse des Mikrozensus 2016 und 2018 auf Basis der neuen Imputationsmethode.

2019 - 01 - 0609

Die aktuelle Kinderlosenquote des Jahrgangs 1976, der im Jahr 2018 sein 43. Lebensjahr erreicht hat, beträgt 22%. Dieser Wert wird sich bis zum Ende der gebärfähigen Phase (statistisch gesehen im Alter ab 50 Jahre) voraussichtlich nicht mehr spürbar verändern, da nur wenige Frauen im Alter ab 42 Jahre ihr erstes Kind bekommen.

Für die nach 1976 geborenen Frauen liefert der Mikrozensus 2018 erwartungsgemäß niedrigere vorläufige Kinderlosenquoten als der Mikrozensus 2016. Diese Frauen waren zwischen beiden Befragungen im gebärfähigen Alter und viele von ihnen haben in der Zwischenzeit eine Familie gegründet. Am stärksten, um bis zu 10 Prozentpunkte, sank der Kinderlosenanteil bei den Frauen der Jahrgänge 1986 bis 1989, die im Jahr 2016 zwischen 27 und 30 Jahre alt waren. Auch bei Frauen im Alter zwischen 31 und 35 Jahren nahm der Kinderlosenanteil erheblich um 9 bis 5 Prozentpunkte ab. Ab dem Alter von 36 Jahren (Jahrgänge 1976 bis 1980) veränderten sich dagegen die Kinderlosenquoten deutlich weniger. Dies entspricht den Befunden in anderen Studien

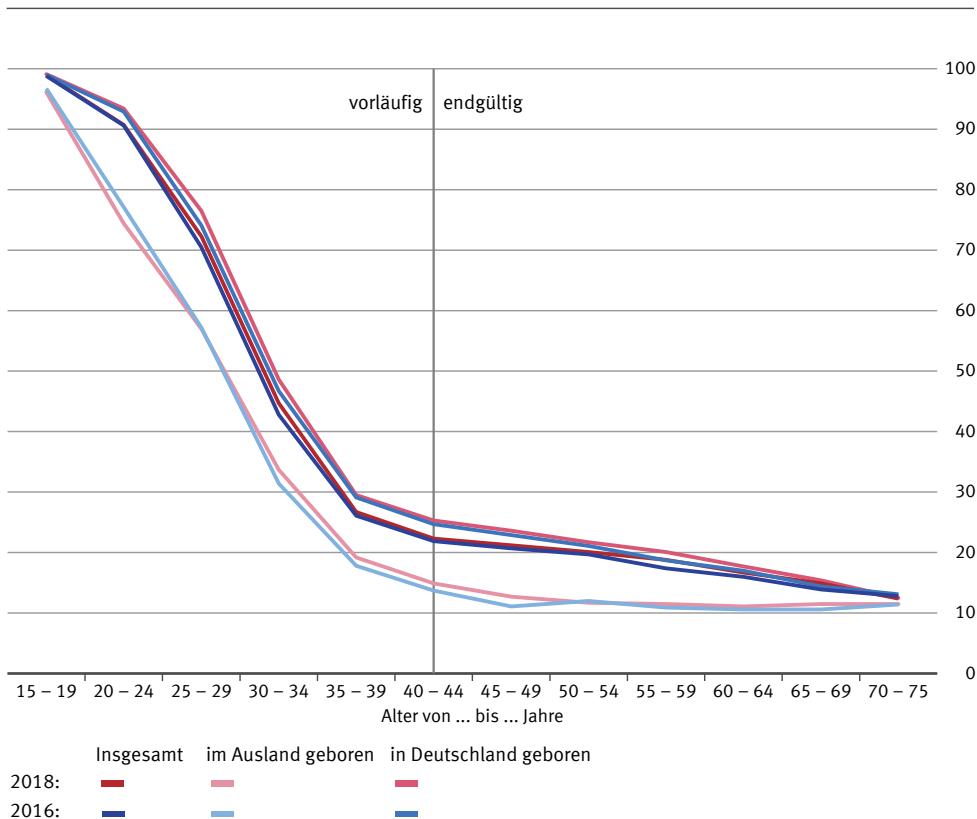
(te Velde und andere, 2012; Taupitz und andere, 2012), die auf eine zunehmende Verlangsamung des Übergangs zum ersten Kind im Verlauf des vierten Lebensjahrzehnts der Frau hinweisen.

Ein Vergleich der Kinderlosenquoten nach Altersgruppen zeigt für 2018 geringfügig höhere Werte als im Jahr 2016. Bei den in Deutschland geborenen Frauen liegt die Kinderlosenquote 2018 in allen Altersgruppen entweder gleich hoch oder leicht über den Werten von 2016. Bei den zugewanderten Frauen trifft das auf die Altersgruppen ab 30 bis 34 Jahre ebenfalls zu. Bei den 20- bis 29-jährigen im Ausland geborenen Frauen war dagegen der Anteil der Kinderlosen 2018 bis zu 3 Prozentpunkte niedriger als im Jahr 2016. [↪ Grafik 7](#)

Die etwas höheren altersspezifischen Kinderlosenquoten 2018 gegenüber dem Stand 2016 können auf die Fortsetzung des langjährigen Trends zu einer höheren Kinderlosigkeit hindeuten. Eine niedrigere Kinderlosigkeit der jüngeren im Ausland geborenen Frauen kann Folge der jüngsten Zuwanderung sein. Dieser Befund

Grafik 7

Kinderlosenquote nach Altersgruppen und Geburtsort der Frauen
in %



Ergebnisse des Mikrozensus 2016 und 2018 auf Basis der neuen Imputationsmethode.

2019 - 01 - 0610

steht im Einklang mit den Ergebnissen der laufenden Geburtenstatistik. Diese zeigte in den Jahren 2016 und 2017 eine gestiegene Geburtenhäufigkeit bei Frauen mit ausländischer Staatsangehörigkeit, insbesondere im Alter von Anfang 20 Jahren.

Die durchschnittliche Kinderzahl je Frau war 2018 nach Ergebnissen des Mikrozensus in fast allen Altersgruppen etwas niedriger als 2016. Nur bei den 40- bis 44-jährigen Frauen war die durchschnittliche Kinderzahl höher als bei den Gleichaltrigen im Jahr 2016 und als bei den fünf Jahre älteren Frauen im Jahr 2018. Darin spiegelt sich eine bereits aus der Geburtenstatistik bekannte Änderung des Trends zu immer niedrigerer endgültiger Kinderzahl je Frau wider (Pöttsch, 2018). Frauen, die 2018 zwischen 40 und 44 Jahre alt waren, sind in den 1970er-Jahren geboren. Sie haben eine etwas höhere Fertilität als die Ende der 1960er- bis Anfang der 1970er-Jahre geborenen Frauen. [↗ Tabelle 4](#)

Tabelle 4

Durchschnittliche Kinderzahl je Frau¹ nach Altersgruppen und Geburtsort der Frauen

	Alle Frauen		In Deutschland geboren		Im Ausland geboren	
	2016	2018	2016	2018	2016	2018
20 bis 24 Jahre	0,13	0,13	0,09	0,08	0,35	0,39
25 bis 29 Jahre	0,45	0,42	0,38	0,34	0,72	0,74
30 bis 34 Jahre	0,98	0,97	0,87	0,86	1,31	1,31
35 bis 39 Jahre	1,43	1,42	1,31	1,30	1,74	1,73
40 bis 44 Jahre	1,56	1,57	1,44	1,43	1,90	1,90
45 bis 49 Jahre	1,55	1,55	1,46	1,45	1,96	1,92
50 bis 54 Jahre	1,59	1,57	1,52	1,49	2,02	2,01
55 bis 59 Jahre	1,69	1,64	1,62	1,58	2,02	2,01
60 bis 64 Jahre	1,69	1,68	1,61	1,61	2,13	2,07
65 bis 69 Jahre	1,74	1,71	1,67	1,64	2,17	2,11
70 bis 75 Jahre	1,86	1,79	1,82	1,74	2,07	2,06

Ergebnisse des Mikrozensus 2016 und 2018 auf Basis des neuen Imputationsverfahrens.

¹ Die zweite Nachkommastelle im Schätzer für die durchschnittliche Kinderzahl je Frau ist aufgrund des stichprobenbedingten Fehlers mit Unsicherheiten behaftet. Sie dient hier ausschließlich dem Vergleich zwischen zwei Jahren und verdeutlicht die Änderungsrichtung.

Bei den in Deutschland geborenen Frauen war 2018 die durchschnittliche Kinderzahl in allen Altersgruppen niedriger als 2016. Bei den 20- bis 29-jährigen zugewanderten Frauen nahm sie dagegen zwischen 2016 und 2018 zu. In den älteren Altersgruppen war 2018 auch bei diesen die durchschnittliche Kinderzahl niedriger als 2016.

6.3 Fehlerrechnung für Ergebnisse 2018

Die Auswertungen der Mikrozensusergebnisse 2008 bis 2016 basieren auf einer sogenannten Single Imputation, wobei jeder fehlende Wert genau einmal ersetzt wird. Dieses Vorgehen hat den Vorteil, dass Auswertungen nach Standardroutinen durchgeführt werden können und keine zusätzlichen methodischen Kompetenzen im Hinblick auf Multiple Imputation erforderlich sind. Nachteil bei der Single Imputation ist, dass dabei die zusätzliche Unsicherheit, die aufgrund der ursprünglich fehlenden Werte besteht, nicht berücksichtigt wird. Dies resultiert in einer Unterschätzung der Varianz der Ergebnisse, was wiederum zu schmale Konfidenzintervalle zur Folge hat, deren tatsächliche Überdeckungsrate dann unter der durch das Signifikanzniveau festgelegten Überdeckungsrate liegt. Eine Folge davon kann sein, dass Ergebnisse fälschlicherweise als signifikant eingestuft werden.

Für die Kennzahlen Kinderlosenquote und durchschnittliche Kinderzahl je Mutter, die auf Basis der Mikrozensusangaben einschließlich Imputationen berechnet werden, wurde bisher eine Fehlerrechnung durchgeführt, deren Ergebnisse auch veröffentlicht wurden (Statistisches Bundesamt, 2017a, hier: Tabelle 2). Beide Kennzahlen und deren absolute Standardfehler wurden jeweils für die Gesamtpopulation sowie für die einzelnen Altersklassen und Regionen berechnet.

Um eine adäquate Varianzschätzung für die Ergebnisse 2018 auf Basis der neuen Imputation anhand von Predictive Mean Matching zu erhalten, erfolgt die Fehlerrechnung für diese beiden Kennzahlen aufbauend auf der Multiplen Imputation. Wie in Kapitel 4 beschrieben, wird jeder fehlende Wert mehrfach imputiert. Die hier gewählte Anzahl an Imputationen beträgt $m=20$. Anschließend erfolgt die Berechnung der Standardfehler unter Berücksichtigung der Between-Varianz. Diese zusätzliche Varianz, welche die Streuung zwischen den einzelnen Anteilswertschätzern misst, bewegt sich für

die Kinderlosenquote im Bereich 10^{-6} und darunter und für die durchschnittliche Kinderzahl je Mutter im Bereich bis zu 10^{-4} . Damit ist sie sehr gering und verursacht nur einen minimalen Anstieg der Totalvarianz, der nach der Rundung der Werte nicht einmal sichtbar wird. Die geringe Between-Varianz zeigt, dass die Imputationen der einzelnen Durchläufe ähnliche Parameterschätzer liefern.

➤ Tabelle 5 zeigt eine Übersicht der Totalvarianzen auf Bundesebene. Auch im Vergleich zu den veröffentlichten Varianzen des Mikrozensus 2016 hat sich die Totalvarianz, wenn überhaupt, nur im hinteren Nachkommastellenbereich verändert.

Tabelle 5
Standardfehler für ausgewählte Indikatoren

	Kinderlosenquote		Durchschnittliche Kinderzahl je Mutter	
	Ergebnis der Hochrechnung in %	absoluter Standardfehler in Prozentpunkten	Ergebnis der Hochrechnung	absoluter Standardfehler
15 bis 19 Jahre	99,0	0,1	/	/
20 bis 24 Jahre	92,8	0,2	1,4	0,02
25 bis 29 Jahre	74,7	0,3	1,5	0,01
30 bis 34 Jahre	47,3	0,3	1,7	0,01
35 bis 39 Jahre	27,7	0,3	1,9	0,01
40 bis 44 Jahre	22,3	0,3	2,0	0,01
45 bis 49 Jahre	21,3	0,3	2,0	0,01
50 bis 54 Jahre	20,1	0,2	2,0	0,01
55 bis 59 Jahre	18,6	0,2	2,0	0,01
60 bis 64 Jahre	16,6	0,2	2,0	0,01
65 bis 69 Jahre	15,1	0,2	2,0	0,01
70 bis 75 Jahre	12,5	0,2	2,0	0,01

Ergebnisse des Mikrozensus 2018.

7

Fazit und Ausblick

Der Beitrag stellt das neue Imputationsverfahren Predictive Mean Matching für die Behandlung der Antwortausfälle bei zwei freiwilligen Fragen zur Geburt von Kindern im Mikrozensus vor. Die hier erstmals veröffentlichten Ergebnisse des Mikrozensus 2018 zur Fertilität der Frauen beruhen auf einem neuen Imputationsverfahren, welches eine vollständige Ersetzung aller fehlenden Angaben ermöglicht.

Die Fragen zur Geburt von Kindern und zur Anzahl geborener Kinder im Mikrozensus sind freiwillig zu beantworten. Dadurch kommen in den erhobenen Daten Antwortausfälle vor. Antwortausfälle können die Ergebnisse verzerren, wenn sie nicht zufällig auftreten und es Unterschiede zwischen den Frauen mit und ohne Angaben gibt. Bei den fehlenden Angaben zur Geburt von Kindern lassen sich Abhängigkeiten zur Art der Befragung und zu anderen Merkmalen der Frauen erkennen. Das Nichtberücksichtigen solcher systematischer Ausfälle kann zu verzerrten Parameterschätzungen führen, zum Beispiel in Form von Anteilswerten. Deshalb müssen die fehlenden Angaben durch statistisch erzeugte Werte so ersetzt werden, dass aus dem vervollständigten Datensatz möglichst unverzerrte Ergebnisse resultieren.

Das in den ersten Befragungswellen 2008, 2012 und 2016 eingesetzte Imputationsverfahren hat zwar die Datengrundlage deutlich verbessert und die Ableitung von belastbaren Ergebnissen ermöglicht. Da aber dabei nicht alle fehlenden Werte ersetzt wurden, ergaben sich weiterführende Probleme für die Auswertung zum Beispiel der durchschnittlichen Kinderzahl je Frau. Auf der Suche nach einem effizienteren Imputationsverfahren wurden in Simulationsstudien verschiedene Vorgehen untersucht. Die besten Ergebnisse lieferte die Methode, bei der im ersten Schritt die Kinderzahl und im zweiten Schritt der Mutterschaftsstatus imputiert wird. Im ersten Schritt wird das Imputationsverfahren Predictive Mean Matching durchgeführt, um die fehlenden Angaben zur Kinderzahl (einschließlich null) vollständig zu befüllen. Dieser Imputationsschritt erfolgt getrennt nach Altersklassen, sodass nur die Kinderzahl einer Frau aus der gleichen Altersspanne eingesetzt werden kann. Im zweiten Schritt wird danach der Mutterschaftsstatus ergänzt, wobei Frauen mit einer imputierten Anzahl von null Kindern als „kinderlos“ und die anderen Frauen als „Mutter“ definiert werden.

Am Beispiel des Mikrozensus 2016 wurden die Ergebnisse auf Basis des alten und des neuen Imputationsverfahrens verglichen. Es zeigte sich, dass auch das frühere Imputationsverfahren valide Ergebnisse geliefert hat und damit kein Revisionsbedarf besteht.

Die Ergebnisse des Mikrozensus 2018 weisen im Vergleich zum Mikrozensus 2016 (auf Basis des neuen Imputationsverfahrens) einen geringen Anstieg der Kinderlosenquote und einen leichten Rückgang der

durchschnittlichen Kinderzahl je Frau in fast allen Altersgruppen nach. Lediglich im Alter zwischen 40 und 49 Jahren hat sich die durchschnittliche Kinderzahl je Frau stabilisiert beziehungsweise sogar zugenommen. Außerdem weisen die Mikrozensusergebnisse auf eine gestiegene Fertilität bei zugewanderten Frauen im Alter unter 30 Jahren hin, was den Befunden der Geburtenstatistik entspricht. Ausführliche Ergebnisse des Mikrozensus 2018 zu Geburten und Kinderlosigkeit werden im November 2019 veröffentlicht.

In Verbindung mit dem neuen Imputationsverfahren für die Mikrozensusdaten 2018 wurde auch die Fehlerrechnung für zentrale Schätzer modifiziert. Bei dem ausgewählten Imputationsverfahren mittels Predictive Mean Matching handelt es sich um ein stochastisches Verfahren. Um die Unsicherheit über die fehlenden Werte mit zu berücksichtigen, wird für die Fehlerrechnung eine Multiple Imputation durchgeführt. Hierbei wird bei der Fehlerrechnung die Between-Varianz mit einbezogen. Für 2018 lässt sich nur eine sehr geringe Between-Varianz feststellen, wodurch die Totalvarianz der Schätzer auf Grundlage der imputierten Datensätze nur minimal ansteigt.

Neben dem hier beschriebenen Vorgehen zur Imputation mittels Predictive Mean Matching für die Anzahl geborener Kinder und der ebenfalls in der Simulationsstudie getesteten logistischen Regression für die Imputation der Mutterschaftsangabe gibt es noch weitere Möglichkeiten zur Ersetzung der fehlenden Werte.

Ein weiteres Verfahren, das für die vorliegende Datensituation ebenfalls in Betracht kommen kann, ist die Imputation basierend auf CART-Modellen. Die Abkürzung CART steht für Classification and Regression Trees. Dabei handelt es sich um ein maschinelles Lernverfahren, bei dem basierend auf den vollständigen Fällen mit den Kovariablen ein Entscheidungsbaum erstellt wird, der die Datensätze in möglichst homogene Gruppen unterteilt. Die unvollständigen Datensätze werden daraufhin gemäß den Regeln des Entscheidungsbaums klassifiziert. Datensätze aus derselben Gruppe dienen dann als Spender. Interessant wären weitere Tests mit dem Ziel, mögliche Verbesserungspotenziale durch CART-Modelle im Vergleich zum Predictive Mean Matching zu erforschen. [📄](#)

LITERATURVERZEICHNIS

Bujard, Martin/Dorbritz, Jürgen/Herter-Eschweiler, Robert/Lux, Linda. *Das unterschätzte Potenzial hoher Fallzahlen – Stärken und Limitierungen des Mikrozensus am Beispiel von Fertilitätsanalysen*. In: Zeitschrift für Familienforschung. 27. Jahrgang. Heft 3/2015, Seite 343 ff.

Koller-Meinfelder, Florian. *Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. Dissertation. Bamberg 2009.

Pöttsch, Olga. [Aktueller Geburtenanstieg und seine Potenziale](#). In: WISTA Wirtschaft und Statistik. Ausgabe 3/2018, Seite 72 ff.

Rubin, Donald B. *Inference and missing data*. In: Biometrika. Jahrgang 63. Ausgabe 3/1976, Seite 581 ff.

Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. New York 1987.

Statistisches Bundesamt. *Kinderlosigkeit, Geburten und Familien – Ergebnisse des Mikrozensus 2016, Ausgabe 2017*. 2017a.

Statistisches Bundesamt. *Qualitätsbericht Mikrozensus 2016*. 2017b.

Statistisches Bundesamt. *Geburtentrends und Familiensituation in Deutschland, Ausgabe 2012*. 2013.

Statistisches Bundesamt. *Mikrozensus 2008: Neue Daten zu Kinderlosigkeit in Deutschland*. 2009.

Taupitz, Jochen/Beier, Henning M./Daele, Wolfgang van den/Diedrich, Klaus/Dudenhäusen, Joachim W./Felberbaum, Ricardo/Gigerenzer, Gerd/Gille, Gisela/Habenicht, Friederike-Ursula/Hinderberger, Philipp/Holzgreve, Wolfgang/Ledger, William/Nieschlag, Eberhard/Ritzinger, Petra/te Velde, Egbert R. *Medizinische und biologische Aspekte der Fertilität*. In: Stock, Günter/Bertram, Hans/Fürnkranz-Prskawetz, Alexia/Holzgreve, Wolfgang/Kohli, Martin/Staudinger, Ursula M. (Herausgeber). *Zukunft mit Kindern: Fertilität und gesellschaftliche Entwicklung in Deutschland, Österreich und der Schweiz*. Frankfurt am Main 2012, Seite 294 ff.

van Buuren, Stef. *Flexible Imputation of Missing Data. 2. Auflage*. Boca Raton 2018.

te Velde, Egbert/Habbema, Dik/Leridon, Henri, Eijkemans, Marinus. *The effect of postponement of first motherhood on permanent involuntary childlessness and total fertility rate in six European countries since the 1970s*. In: Human Reproduction. Jahrgang 27. Ausgabe 4/2012, Seite 1179 ff. [doi: 10.1093/humrep/der455].

RECHTSGRUNDLAGEN

Gesetz zur Durchführung einer Repräsentativstatistik über die Bevölkerung und die Arbeitsmarkt-beteiligung sowie die Wohnsituation der Haushalte (Mikrozensusgesetz - MZG) vom 7. Dezember 2016 (BGBl. I Seite 2826).

Herausgeber

Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung

Dr. Daniel Vorgrimler

Redaktionsleitung: Juliane Gude

Redaktion: Ellen Römer

Ihr Kontakt zu uns

www.destatis.de/kontakt

Erscheinungsfolge

zweimonatlich, erschienen im Oktober 2019

Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Print

Einzelpreis: EUR 19,- (zzgl. Versand)

Jahresbezugspreis: EUR 114,- (zzgl. Versand)

Bestellnummer: 1010200-19005-1

ISSN 0043-6143

ISBN 978-3-8246-1084-6

Download (PDF)

Artikelnummer: 1010200-19005-4, ISSN 1619-2907

Vertriebspartner

IBRo Versandservice GmbH

Bereich Statistisches Bundesamt

Kastanienweg 1

D-18184 Roggentin

Telefon: +49 (0) 382 04 / 6 65 43

Telefax: +49 (0) 382 04 / 6 69 19

destatis@ibro.de

Papier: Design Offset, FSC-zertifiziert

© Statistisches Bundesamt (Destatis), 2019

Vervielfältigung und Verbreitung, auch auszugsweise, mit Quellenangabe gestattet.