Kommentare und Erwiderung zu: Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021

Sara Bleninger · Michael Fürnrohr · Hans Kiesel · Walter Krämer ·
Helmut Küchenhoff · Jan Pablo Burgard · Ralf Münnich ·
Martin Rupp

Eingegangen: 5. November 2019 / Angenommen: 11. November 2019 / Online publiziert: 2. Dezember 2019
© Der/die Autor(en) 2019

Zusammenfassung Burgard et al. (2020) stellen in ihrem Artikel zu *Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021* Erweiterungen der Stichproben- und Schätzmethoden des Zensus 2011 vor, die kleine Gemeinden unter 10.000 Einwohnern in den Entscheidungsprozess integrieren. Die Dringlichkeit

S. Bleninger · M. Fürnrohr
Bayerisches Landesamt für Statistik und Datenverarbeitung, München, Deutschland

S. Bleninger
E-Mail: sara.bleninger@statistik.bayern.de

M. Fürnrohr
E-Mail: michael.fuernrohr@statistik.bayern.de

H. Kiesel
Fakultät Informatik und Mathematik, Ostbayerische Technische Hochschule Regensburg,
Regensburg, Deutschland
E-Mail: hans.kiesel@oth-regensburg.de

W. Krämer
Fakultät Statistik, SFB 823, Technische Universität Dortmund, Dortmund, Deutschland
E-Mail: walterk@statistik.tu-dortmund.de

H. Küchenhoff
Statistisches Beratungslabor StaBLab, Institut für Statistik, Ludwig-Maximilians-Universität
München, München, Deutschland
E-Mail: kuechenhoff@stat.uni-muenchen.de

J. P. Burgard · R. Münnich ( ) · M. Rupp
FB IV, VWL, Wirtschafts- und Sozialstatistik, Universität Trier, Universitätsring 15, 54296 Trier,
Deutschland
E-Mail: muennich@uni-trier.de

J. P. Burgard
E-Mail: burgardj@uni-trier.de

M. Rupp
E-Mail: ruppm@uni-trier.de

zur Lösung dieses Problems wurde ebenso im Urteil des Bundesverfassungsgerichts zur Volkszählung 2011 festgestellt. Ziel dieser Erwiderung ist eine eingehende Diskussion der Ergebnisse des vorangegangenen Beitrags mit namhaften Experten auf diesem Gebiet. Insbesondere geht es um eine Einordnung des Artikels in den Wissenschaftskontext (Krämer), die Bedeutung von Nichtstichprobenfehlern für den Zensus (Küchenhoff), den Zensus aus Sicht der Amtsstatistik (Bleninger und Fürnrohr) sowie aus statistisch-methodischer Sicht (Kiesl). Darüber hinaus werden aktuelle Entwicklungen vorgestellt.

Schlüsselwörter Zensus · Ermittlung der Einwohnerzahl · Qualitätsmessung · Optimale Allokation · Total Survey Error

JEL-Klassifikation C83 · C15

Comments and rejoinder: quality measures respecting highly varying community sizes within the 2021 German Census

Abstract Burgard et al. (2020) present in their article on *Quality measures respecting highly varying community sizes within the 2021 German Census* extensions to the sampling and estimation framework of the census 2011 that integrate small communities under 10.000 inhabitants into the decision process. The urgency to solve this problem was also pointed out in the verdict of the Federal Constitutional Court on the census 2011. The aim of this rejoinder is a thorough discussion on the results of the preceding paper with well-known experts in this field. In particular, it deals with the classification of the article in the scientific context (Krämer), the relevance of non-sampling errors for the census (Küchenhoff), the census from the point of view of official statistics (Bleninger and Fürnrohr) and from a statistical-methodological point of view (Kiesl). Further, recent developments are presented.

Keywords Census · Determination of population size · Quality measurement · Optimal allocation · Total survey error

1 Kommentare

1.1 Kommentar 1: Walter Krämer

Das ist schon seltsam: Da bewegt ein Problem aus der Statistik sogar das Bundesverfassungsgericht, die Medien sind in Aufruhr, ganze Städte fühlen sich benachteiligt, und die ehrbare Wissenschaft von der Statistik gefällt sich wie üblich in oft selbstreferentiellen kleinlichen Methodendiskussionen, fügt da ein Epsilon hinzu, meldet dort an der Effizienz eines Testverfahrens Zweifel an und diskutiert wie seit Jahrzehnten an den wahren statistischen Problemen unserer Gesellschaft selbstverliebt vorbei. Kein Wunder, dass an einer deutschen Universität nach der anderen Statistiklehrstühle wegfallen und produktiveren Wissenschaften zugewendet werden.

Zum Glück gibt es aber Ausnahmen, und eine davon haben wir hier. Äußerst gründlich und von großer Detail- und Sachkenntnis geleitet setzen sich die Autoren mit der Problematik einer registergestützten Volkszählung auseinander, insbesondere damit, wie Informationen aus Registern und Stichproben optimal zu kombinieren sind. Der Fokus liegt dabei auf der Zielvariablen *Einwohnerzahl*, wie diese im Zensus 2011 ermittelt worden ist und wie dieses Verfahren in Zukunft verbessert werden kann. Bekanntlich gab es damals unterschiedliche Methoden für Gemeinden mit mehr und weniger als 10.000 Einwohnern. Hier existiert das Problem, dass die Grundgesamtheit in beiden Fällen nur per Umweg über Melderegister adressierbar ist und es einen hohen Anteil an Nullanschriften gibt, hinter denen sich keine lebenden Personen verbergen. Unabhängig davon wird sehr gründlich herausgearbeitet, welche Stichprobenumfänge zur Einhaltung vorgegebener Konfidenzintervallgrenzen nötig sind. Das ist zwar im Prinzip ein Lehrbuchproblem aus dem Grundstudium, aber im Kontext der hier vorliegenden institutionellen Rahmenbedingungen und Stichprobenpläne alles andere als trivial; das aus Simulationsstudien gewonnene Stichprobendesign ist als wichtiger Baustein künftiger registergestützter Zensen anzusehen. Hier wäre es vielleicht nützlich gewesen, die in Formel 4 aufgeführten Effizienzgewinne für das letztendlich vorgeschlagene Verfahren für ausgewählte Parameterwerte einmal konkret anzugeben. Immerhin ist es ja für Praktiker nicht irrelevant, ob man sich große Mühen für eine 1 % oder für eine 30 % Verbesserung der Schätzqualität auferlegen soll. Da sind die in Anschluss an Formel 6 getroffenen Aussagen doch etwas zu allgemein.

Der Schwerpunkt des Aufsatzes liegt natürlich bei der berüchtigten 10.000-Einwohner-Grenze und der Frage, wie die Präzisionsanforderungen für größere Gemeinden auf kleinere Gemeinden übertragbar sind. Zugleich soll das auch eine Empfehlung für den Zensus 2021 werden. Hier weisen die Autoren zunächst völlig richtig darauf hin, dass auch kleine absolute Standardfehler bei kleinen Gemeinden zu grotesken relativen Fehlern führen können, siehe die berüchtigten Jungen-Anteile in Flensburg im Zensus 2011. Die hierzu in Abschn. 3 unterbreiteten Anpassungsvorschläge sind innovativ und praktisch durchführbar; die Vorschläge für einen *glatten* Übergang von absoluten zu relativen Standardfehlern inklusive der Anforderungen an die Präzisionszielfunktion erscheinen mir plausibel und sachgerecht, die vorgeschlagenen Methoden zur Bestimmung der Parameter dieser Präzisionszielfunktion sind der Problemlage angepasst. Alles in allem liegt hier ein wertvoller Beitrag zur modernen Zensusproblematik vor.

1.2 Kommentar 2: Helmut Küchenhoff

Der Beitrag setzt sich mit Problemen auseinander, die über die Zielsetzung der Arbeit von Burgard et al. (2020) hinausgehen. Das Auftreten von Nicht-Stichprobenfehlern ist ein möglicher Grund für Unstimmigkeiten und unplausible Ergebnisse im Zensus 2011. Diese Probleme werden in dem Beitrag diskutiert. Es wird dargelegt, dass zum Thema Messfehler bei der Erhebung der Einwohnerzahl weitere Analysen nötig sind. Diese sind für die Sicherstellung der Qualität des Zensus 2021 und für eine korrekte Abschätzung der Genauigkeit von zentraler Bedeutung.

1.2.1 Einleitung

In dem Artikel von Burgard et al. (2020) wird ausführlich diskutiert, wie die Genauigkeitsanforderungen an den Zensus 2021 sinnvoll für kleinere Gemeinden angepasst werden können und damit ermöglicht wird, dass insgesamt eine einheitliche Methodik zur Anwendung kommen soll. Die von den Autoren gewählte Strategie ist im Rahmen der gemachten Annahmen sinnvoll, mit Simulationen ausführlich überprüft und stellt verschiedene Stichprobendesigns zur Auswahl. Dabei wird berechtigterweise davon ausgegangen, dass die mit einem Methodenwechsel bei der Grenze der Einwohnerzahl von 10.000 assoziierten unerwünschten Effekte (siehe dazu Christensen et al., 2015, und besonders Rendtel, 2015) vermieden werden können. Dabei soll das im Zensus 2011 verwendete Stichprobenverfahren in modifizierter Form auch bei kleineren Gemeinden zur Anwendung kommen. Dieses Vorgehen setzt implizit voraus, dass die Probleme und unerwünschten Effekte hauptsächlich in den im Zensus verwendeten Strategien bei den kleineren Gemeinden ihren Grund haben. Dafür spricht die tiefer gehende Analyse von Rendtel (2015), in der gezeigt wird, dass unplausible Werte bzgl. der Grenze bei 10.000 in Bundesländern mit zentral geführten Melderegistern nicht vorhanden sind. Allerdings ist bekannt, dass bei Stichproben neben dem in dem Artikel von Burgard et al. (2020) ausschließlich betrachteten Stichprobenfehler weitere Typen von Fehlern auftreten. In einem Gutachten für das Verwaltungsgericht Bremen (Küchenhoff 2014) wurde dargelegt, dass Messfehler bei der Erhebung der Einwohnerzahl eine Verzerrung zur Folge haben könnten. Da ich die dort eingebrachten Probleme für zentral halte, werden sie hier nochmal im Detail diskutiert. Zunächst wird die Problematik der Messfehler bei dem in dem Artikel skizzierten Design aufgezeigt. Anschließend werden mögliche Konsequenzen auf das Stichprobendesign für den Zensus 2021 diskutiert.

1.2.2 Messfehler

In dem Stichprobendesign nach Burgard et al. (2020) sind die Untersuchungseinheiten y_i die Einwohnerzahlen der i -ten Anschrift. Diese wird in der Erhebung vor Ort durch die Erhebungsbeauftragten festgestellt. Bei allen Berechnungen in Burgard et al. (2020) wird nun davon ausgegangen, dass diese Zahl korrekt erhoben wird. Allerdings können bei der Erhebung Fehler auftreten. Wir bezeichnen den in der Stichprobe erhobenen Wert mit y_i^* und die Differenz $y_i^* - y_i$ ist dann der Messfehler in der Erhebung. In der Survey-Literatur werden diese sehr breit diskutiert, siehe dazu z.B. Kauermann und Küchenhoff (2010). Zu beachten ist hier, dass die Untersuchungseinheiten die Anschriften und nicht die Personen sind. Daher führen Probleme, die sonst in der Survey-Literatur unter den Labels *Antwortverweigerung* und *Nicht-Erreichbarkeit* genannt werden, hier zu Messfehlern.

Die Ursachen für solche Messfehler sind im Allgemeinen sehr vielfältig. Im vorliegenden Design könnten einerseits Personen, die an dieser Adresse wohnen, nicht in der Zählung berücksichtigt werden. Andererseits könnten Personen, die nicht an der entsprechenden Adresse wohnen, fälschlicherweise in die Zählung kommen. Solche Fehler können durch die Erhebungsbeauftragten entstehen, wenn diese Personen nicht auffinden oder Personen angeben, die nicht an der Adresse wohnen. Letzteres

kann dadurch zu Stande kommen, dass die Erhebungsbeauftragten Listen von potentiellen Bewohnern zur Verfügung haben, die nicht genau geprüft werden (siehe Eckman und Kreuter 2017).

Im Rahmen des Zensus 2011 wurden vielfältige Maßnahmen getroffen, um solche Fehler zu vermeiden. Allerdings können mit diesen Maßnahmen solche Fehler nie ganz vermieden werden, siehe dazu auch Küchenhoff (2014). Eine systematische Untersuchung der Erhebungsfehler wurde nach meiner Kenntnis zum Zensus 2011 nicht durchgeführt. Geeignete Methoden sind dabei Kontrollzählungen (Post Enumeration Surveys) oder auch gezielte Prüfungen oder Heranziehen von Zusatzinformationen (Metadaten).

1.2.3 Auswirkungen von Messfehlern

Hierbei ist zwischen systematischen Messfehlern, d. h. dem Fall $E(y_i^* - y_i) \neq 0$, und zufälligen Messfehlern mit $E(y_i^* - y_i) = 0$ zu unterscheiden. Im letzteren Fall ist der Regressionsschätzer unverzerrt, wenn der Messfehler mit den Hilfsmerkmalen x unkorreliert ist. Allerdings ist die Varianzformel (2) aus Burgard et al. (2020) nicht mehr korrekt, wenn der Faktor (Endlichkeitskorrektur) $(1 - f_h)/n_h$ eine relevante Größenordnung hat. Ein einfacher Verzicht auf die Endlichkeitskorrektur führt zu konservativen Schätzungen der Varianz (siehe dazu auch Abschn. 7.2 in Kauermann und Küchenhoff 2010).

Handelt es sich um systematischen Messfehler, d. h. $E(y_i^* - y_i) = c \neq 0$ und ist dieser mit den Hilfsmerkmalen x unkorreliert, so wird dieser beim Regressionsschätzer dem konstanten Term zugeschlagen und es ergibt sich ein Bias in der Schätzung der Einwohnerzahl von $N \cdot c$, wobei N die Anzahl der Adressen in der jeweiligen Grundgesamtheit ist. Der Bias von $N \cdot c$ ist für den Fall, dass der Messfehler mit den Variablen x korreliert ist, nur noch approximativ gültig. Er kann sich je nach der Struktur erhöhen oder erniedrigen.

Entscheidend ist, dass der Bias auch bei sehr guten und genauen Hilfsvariablen x nicht verschwindet. Dies liegt an der Strategie, die erhobenen Größen der Einwohnerzahl als präzise und relevante Messung zu betrachten. Da die Regressions-schätzungen sinnvollerweise nach Sampling Points getrennt durchgeführt werden, kann es zu spezifischen systematischen Fehlern (z. B. bedingt durch Unterschiede in der Interviewer-Schulung) in einzelnen Gemeinden kommen. Man beachte, dass die Variablen aus dem Melderegister als Hilfsvariablen verwendet werden. Diese werden nicht zur Korrektur von Erhebungsfehlern, sondern nur zum Effizienzgewinn (dieser ist erheblich) und damit nur zur Verringerung des Stichprobenfehlers genutzt. Messfehler (sogar systematische) in den Hilfsvariablen führen nicht notwendig zu einem Bias in der Gesamtschätzung. Die Messfehler (z. B. verursacht durch Karteteilen und nicht gemeldete Personen) bei den Registerdaten wurden im Rahmen des Zensus 2011 ausführlich diskutiert, siehe z. B. Rendtel (2015), Tab. 2.

1.2.4 Diskussion

Aus meiner Sicht ergeben sich folgende Konsequenzen für die Vorbereitung zum Zensus 2021. Bei der Analyse der Genauigkeit der geschätzten Einwohnerzahlen

sollten neben der Betrachtung des Stichprobenfehlers auch andere Fehler betrachtet werden. Es sollte der in der Survey-Forschung bekannte Ansatz des Total Survey Errors verwendet werden, siehe z. B. Groves und Lyberg (2010) und Fuchs (2009) oder auch ein Sonderheft von Journal of Official Statistics, Eckman und Leeuw (2017). In Küchenhoff (2014) wurde ausführlich dargelegt, dass hier der Messfehler das Hauptproblem darstellt. Daher sind folgende Schritte sinnvoll:

1. Messfehleranalyse mit Hilfe der Erfahrungen aus dem Zensus 2011
2. Einbeziehen der Messfehler in die Simulationen zur Schätzgenauigkeit
3. Neubewertung der zu erreichenden Genauigkeit
4. Entwicklung von Strategien zur Korrektur der Messfehler insbesondere unter Nutzung der Registerdaten
5. Überdenken der Strategie, welche die Melderegistervariablen nur als Hilfsvariablen verwendet.

Durch die Wahl der neuen Methodik, wie sie in Burgard et al. (2020) beschrieben ist, werden sicher die Unstimmigkeiten und Unterschiede bezüglich der Grenze von 10.000 Einwohnern behoben. Das bedeutet aber nicht, dass dadurch systematische Fehler ausgeschlossen werden können. Daher sind die oben skizzierten Analysen im Vorfeld des Zensus 2021 von großer Bedeutung. Andernfalls könnten Ansprüche an die Genauigkeit gestellt werden, die einer detaillierten Prüfung nicht standhalten.

1.3 Kommentar 3: Sara Bleninger und Michael Fürnrohr

1.3.1 Vorbemerkung

Es war ein außerordentlich komplexes Problem, vor dem Burgard, Münnich und Rupp im Zuge der methodischen Vorbereitungsarbeiten für den Zensus 2021 standen und dessen Lösung in der vorliegenden Publikation *Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021* beschrieben ist. Es galt ein Stichprobendesign zu entwickeln, das nicht nur in Hinblick auf statistische Methodik höchsten wissenschaftlichen Ansprüchen zu genügen, sondern auch rechtlichen, politischen und finanziellen Restriktionen zu entsprechen hatte.

In der vorliegenden Kommentierung wird zunächst ausführlich auf das schwierige Spannungsfeld der bestehenden Randbedingungen bei dem von Burgard, Münnich und Rupp publizierten Stichprobendesign eingegangen. Dies unterstreicht die besondere Qualität des entwickelten Ansatzes, dessen fachliche Würdigung im zweiten Teil vorgenommen wird.

1.3.2 Das Stichprobendesign des Zensus 2021 im Spannungsfeld aus Präzision / Rechtssicherheit, Grundrechtsschonung / Kosten und Methodik

Nach dem Willen der Bundesregierung soll der nächste Zensus 2021 in seiner Grundstruktur weitgehend analog dem Zensus 2011 erfolgen. Der Zensus 2021 *knüpft an die bewährten Elemente des letzten Zensus im Jahre 2011 an und sieht dort, wo notwendig, methodische und organisatorische Fortentwicklungen vor* (Entwurf eines Gesetzes zur Durchführung des Zensus im Jahr 2021). Ein zentrales Element

des Zensusmodells 2011 war die Haushaltsbefragung auf Stichprobenbasis nach § 7 ZensG2011 (2009). Diese verfolgte primär den Zweck, durch Ermittlung von Fehlern in den Melderegistern die Gewinnung der amtlichen Einwohnerzahlen für Bund, Länder und Gemeinden qualitätszusichern (sog. Ziel 1). Daneben diente die Stichprobe der Gewinnung von soziodemographischen und erwerbsstatistischen Daten, für die es in Deutschland keine Verwaltungsregister gibt (Ziel 2). Da nach Erkenntnissen aus dem Zensustest 2001 die Qualität der Melderegister in Gemeinden unter 10.000 Einwohnern deutlich besser als in den größeren Gemeinden erschien (vgl. Statistische Ämter des Bundes und der Länder 2011), beschränkte der Gesetzgeber aus Kosten- und Belastungsgründen der Bürgerinnen und Bürger die Stichprobe entsprechend auf Gemeinden mit 10.000 oder mehr Einwohnern. Bei Gemeinden unterhalb dieser Grenze wurden lediglich die Melderegister und Befragungen zur Klärung von Unstimmigkeiten im Einfamilienhausbereich (§ 16 ZensG2011 2009) herangezogen. Insgesamt waren im Zensus 2011 bei der Haushaltsstichprobe und der Erhebung nach § 16 ZensG2011 (2009) rund 9,1 Mio. Bürgerinnen und Bürger in die Befragungen einbezogen.

Vor allem dieser Methodenbruch einhergehend mit einer vermeintlichen Ungleichbehandlung der Kommunen und letztlich auch der Bundesländer wurde in einer Reihe von Klagebegründungen in Verwaltungsgerichtsverfahren sowie den Normenkontrollanträgen des Senats von Berlin und des Senats des Freien und Hansestadt Hamburg moniert (vgl. Gößl 2018). Im BVerfG-Urteil vom 19. September 2018 wird das Vorgehen zur Stichprobenziehung im Zensus 2011 zwar als verfassungsgemäß bestätigt, gleichwohl fordert das BVerfG, das Verfahren der Stichprobenziehung in der Personenerhebung an neue Gegebenheiten und methodische Entwicklungen anzupassen (BVerfG, Urteil vom 19.09.2018, vgl. Gößl 2018).

Eine konsequente Auslegung dieser Vorgabe des BVerfG bedeutet somit, dass in Zukunft Gemeinden unabhängig von ihrer Größe bei der Einwohnerzahlermittlung gleich zu behandeln sind (vgl. Gößl 2018). Die beim Zensus 2011 noch auf einen Teil der Gemeinden beschränkte Stichprobe ist also im Zensus 2021 auf alle Gemeinden unabhängig von ihrer Größe auszuweiten.

Die Entwicklung eines Stichprobendesigns für alle Gemeinden Deutschlands ist für Stichprobenexperten wie Burgard, Münnich und Rupp per se noch kein großes fachliches Problem. Es galt aber einige teils sich widersprechende Randbedingungen angemessen zu berücksichtigen, welche die Schwierigkeit dieser Aufgabe erheblich erhöhten.

Beim Zensus 2011 hatte der Gesetzgeber für alle in der Stichprobe einbezogenen Gemeinden (mit mindestens 10.000 Einwohnern) ein anzustrebendes Präzisionsziel der amtlichen Einwohnerzahl in Höhe von 0,5 % einfacher relativer Standardfehler postuliert (§7 Abs.1 ZensG2011 2009). Ein Zurückbleiben hinter dieser Qualitätsnorm ist angesichts der finanziellen Bedeutung der amtlichen Einwohnerzahlen im horizontalen und vertikalen Finanzausgleich weder begründbar noch vermittelbar. Aber was ist mit den kleinen Gemeinden mit weniger als 10.000 Einwohnern? Sind hier die gleichen Maßstäbe anzuwenden?

Der Deutsche Städtetag und der Deutsche Städte- und Gemeindebund vertreten die Meinung, dass Rechtssicherheit nur dann erreicht werden kann, wenn das Präzisionsziel von 0,5 % einfacher relativer Standardfehler für alle Städte und Gemeinden

erreicht würde (vgl. Deutscher Städtetag 2019). Folgt man dieser juristisch intendierten Forderung, wäre eine erhebliche Ausweitung des Stichprobenumfangs die Folge. Nach Berechnungen des Statistischen Bundesamts beliefe sich der Befragungsumfang auf rund 15,2 Mio. Bürgerinnen und Bürger. Dies ist eine Steigerung gegenüber 2011 von rund 67 % (vgl. StBA 2018). Dies würde in etwa 2000 Gemeinden eine Vollerhebung bedeuten. Vor allem kleinere Gemeinden würden die große Last an der Ausweitung der Stichprobe unter der Maßgabe des einheitlichen Präzisionsziels von 0,5 % relativer Standardfehler zu tragen haben.

Die Zielsetzung nach umfänglicher Rechtssicherheit durch gleiche Präzision der Einwohnerzahlen in allen Gemeinden und dem damit verbundenen hohen Stichprobenumfang ist sorgfältig abzuwägen gegen den durch die Befragungen induzierten Eingriff in das Recht auf informationelle Selbstbestimmung der Bürgerinnen und Bürger. So hat das BVerfG zur Grundrechtsschonung im Zensus 2011 ausgeführt: *Das Stichprobenverfahren beim Zensus 2011 garantierte auf der einen Seite den erforderlichen Grad an Genauigkeit und Realitätsgerechtigkeit, begrenzte aber auf der anderen Seite die Zahl der zu befragenden Personen und die Eingriffe in das Grundrecht auf informationelle Selbstbestimmung auf ein möglichst geringes Maß.* (Göbl 2018, S. 44)

Gegen ein einheitliches relatives Präzisionsziel für alle Gemeinden sprechen auch Kostengesichtspunkte. Eine Stichprobe des Umfangs von 15,2 Mio. Befragten hätte bei den Ländern, die gemäß Art. 83 GG (Bundesrepublik Deutschland 1949) für die Durchführung der Erhebungen verantwortlich sind, außerordentlich hohe Erhebungskosten zur Folge. Darüber hinaus würden sich die Kosten aufgrund der unterschiedlichen Gemeindestrukturen der Bundesländer nicht gleichmäßig auf die Bundesländer verteilen (vgl. Burgard et al. 2020). Auch diesen Sachverhalt galt es bei der Gestaltung der Stichprobe angemessen zu berücksichtigen.

Zudem bestehen auch methodische Probleme, wenn ein einheitliches Präzisionsziel angesetzt wird: Der relative Standardfehler kann bei kleinen Gemeinden schnell dazu führen, dass gefordert wird, die Einwohnerzahl auf Bruchteile einer Person genau zu schätzen. Ferner macht die Forderung eines einfachen relativen Standardfehlers für die kleinen Gemeinden die Einhaltung der Box-Constraints unmöglich. In der Folge würde es bei vielen kleinen Gemeinden zu einer Vollerhebung kommen. Vollerhebungen haben methodisch gesehen den Nachteil, dass man Korrelationen zwischen den Stichprobeneinheiten nicht mehr ausschließen kann. In einer Stichprobe ist davon auszugehen, dass etwa räumliche Korrelationen zwischen Anschriften durch die Zufälligkeit der Auswahl ausgeschlossen werden können. Modellschätzungen basieren zumeist auf der Annahme, dass es keine Korrelationen der Stichprobeneinheiten gibt, ansonsten sind sie verzerrt (vgl. Broscheid und Geschwend 2003). Auch im Zensus wird mit dem GREG-Schätzer ein Schätzer verwendet, der von diesem Problem betroffen sein kann (vgl. Burgard et al. 2020).

Das Stichprobendesign für den Zensus 2021 steht folglich in einem Spannungsfeld aus rechtlichen, kostenreduzierenden und methodischen Forderungen: Gleichbehandlung der Gemeinden im Sinne eines einheitlichen Präzisionsziels bedingt eine sehr umfangreiche Stichprobe. Grundrechtsschonung und Kostenreduktion eine Beschränkung des Erhebungsumfangs. Damit steht die Methodenwahl des Stichprobendesigns vor sich teilweise widersprechenden Vorstellungen:

- Die Stichprobe ist im Gegensatz zum Zensus 2011 auf alle Gemeinden auszuweiten.
- Für Gemeinden mit mindestens 10.000 Einwohnern ist wie beim Zensus 2011 ein relatives Präzisionsziel von 0,5 % relativer Standardfehler anzustreben.
- In Gemeinden mit weniger als 10.000 Einwohnern steht ein einheitliches relatives Präzisionsziel von 0,5 % relativer Standardfehler durch den außerordentlich hohen Stichprobenumfang im Widerspruch zum Postulat der Grundrechtsschonung, der Forderung nach Kostenbegrenzung und wirft zudem methodische Probleme auf.
- Ein Methodenbruch an der 10.000er Grenze ist weitestgehend zu vermeiden. Unter dem Gleichbehandlungsgrundsatz wäre es nur schwer vermittelbar, wenn die Präzision der Einwohnerzahl einer Gemeinde mit 9.999 Einwohnern sich merklich von der einer Gemeinde mit 10.000 Einwohnern unterscheiden würde.

1.3.3 Kein Kompromiss, sondern eine Lösung

Mit welchen Maßnahmen haben nun Burgard, Münnich und Rupp das Spannungsfeld aufgelöst? Zunächst galt es, den Gedanken aufzugeben, für alle Gemeinden den gleichen relativen Standardfehler zu fordern, ohne dabei eine Ungleichbehandlung nach Gemeindegröße herbeizuführen. Für kleine Gemeinden ist es sinnvoller, anstelle des einfachen relativen Standardfehlers einen absoluten Standardfehler anzusetzen. Für große Gemeinden macht wiederum der absolute Standardfehler keinen Sinn, weil die so formulierte Genauigkeit in keiner Relation zur Gemeindegröße stehen würde. Eine Lösung mit einer Sprungstelle ab einer gewissen Gemeindegröße würde aber dem rechtlich geforderten Gleichbehandlungsgebot widersprechen. Durch die harte Grenze an einer Gemeindegröße wird die Ungleichbehandlung zwischen Gemeinden mit mindestens 10.000 Einwohnern und Gemeinden unterhalb dieser Grenze zu eklatant.

Burgard, Münnich und Rupp lösen das Spannungsfeld zwischen der rechtlichen Voraussetzung der Gleichbehandlung und dem Kostendruck sowie methodischen Grenzen auf, indem sie vorschlagen, *unterschiedliche, teilweise flexible Formulierungen für relative Standardfehler* (Burgard et al. 2020, Kapitel 3) zu verwenden. Sie setzen damit auf einen weichen, fließenden Übergang vom einfachen relativen Standardfehler an der 10.000er Grenze zum absoluten Standardfehler für kleinere Gemeinden, bei denen der einfache relative Standardfehler nicht mehr genutzt werden kann. Der weiche Übergang wird durch eine Präzisionszielfunktion definiert, die zwischen der Gemeindegröße, bis zu welcher der absolute Standardfehler gelten soll, und der Gemeindegröße, ab welcher der relative Standardfehler gelten soll, den Übergang vom absoluten Standardfehler zum relativen Standardfehler beschreibt. Die konfligierenden Ziele aus Recht, Kosten und Methoden werden in methodische Anforderungen an eine Präzisionszielfunktion in Abhängigkeit von der Einwohnerzahl der Gemeinden übersetzt: Nachdem es einen glatten Übergang zwischen den Präzisionszielen geben soll, muss es sich um eine stetige Funktion handeln. Ferner ist festgelegt, dass ab der 10.000er Grenze der einfache relative Standardfehler von 0,5 % gelten soll. Offen sind zunächst noch: Die Untergrenze der Einwohnerzahl, bis zu der ein absoluter Standardfehler gelten soll, die Höhe des anzusetzenden ab-

soluten Standardfehlers und der genaue Verlauf der Präzisionszielfunktion zwischen absolutem und relativem Standardfehler.

Die Festlegung der Grenzen und der Höhe der dort anzusetzenden absoluten, bzw. relativen Standardfehler ist eine normative Entscheidung und hat einen wesentlichen Einfluss vor allem auf die Größe der Stichprobe im Zensus 2021. Die Untersuchungen und die methodische Diskussion mit verschiedenen Szenarien für die Höhe des Standardfehlers und der unteren Grenze, die Burgard, Münnich und Rupp unternehmen, liefern dabei das für die politischen Entscheidungen notwendige methodische Fundament.

Offen blieb die Festlegung des genauen Verlaufs der Funktion in ihrem Definitionsbereich. Die Entscheidung für eine Bézier-Kurve, die lediglich an den festen durch den absoluten und relativen Standardfehler definierten Grenzen stetig ist, löst das Problem, den Verlauf der Präzisionszielfunktion genau festlegen zu müssen. Die Geraden, die definiert werden durch den relativen Standardfehler bei 10.000 Einwohnern und durch den absoluten Standardfehler an der Untergrenze, werden als Tangenten für die Bézier-Kurve genutzt. Durch die Entscheidung, keine weiteren Tangenten zur Bestimmung des Verlaufs der Bézier-Kurve festzulegen, konnte man sich auf die noch ausstehenden kritischeren offenen Punkte konzentrieren: Eine Entscheidung über die Höhe des anzusetzenden absoluten Standardfehlers und die Gemeindegröße, bis zu der er gelten soll. Diese Festlegungen konnten vor dem Hintergrund des Spannungsfelds aus rechtlichen, methodischen und kostentechnischen Anforderungen politisch gefällt werden. Burgard, Münnich und Rupp schlussfolgern trocken: *Dies ermöglicht eine einfache Umsetzung als Präzisionszielfunktion unabhängig von a priori zu wählenden Parametern.* (Burgard et al. 2020, Abschn. 3.1).

Aus der Sicht der Verfasser dieser Kommentierung, beide Mitarbeiter eines Statistischen Landesamts, ist das von Burgard, Münnich und Rupp entwickelte und in der vorliegenden Publikation veröffentlichte Stichprobendesign für den Zensus 2021 zusammenfassend wie folgt zu würdigen:

- Mit dem Verfahren ist eine in allen Gemeinden weitestgehend einheitliche Ermittlung der amtlichen Einwohnerzahlen sichergestellt.
- Die formulierten Präzisionsziele tragen der hohen finanziellen Bedeutung der amtlichen Einwohnerzahlen hinreichend Rechnung.
- Das Verfahren begrenzt unter den gegebenen Randbedingungen die Belastung der Bürgerinnen und Bürger mit den Befragungen und erfüllt somit die Forderung des BVerfG nach Grundrechtsschonung.
- Die Kostenbelastung der Länder mit der Durchführung der Befragungen wird in vertretbaren Grenzen gehalten.

1.4 Kommentar 4: Hans Kiesel

1.4.1 Einleitung

Das Bundesverfassungsgericht hat in seinem Urteil zum Zensus 2011 (BVerfG 2018) die Methodik grundsätzlich für verfassungsgemäß erklärt, gleichzeitig das interkommunale Gleichbehandlungsgebot unterstrichen und schließlich eine kontinuierliche

Suche nach einer „grundrechtsschonenderen Datenerhebung“ eingefordert. Aus diesem Urteil folgt insbesondere der Auftrag an Wissenschaft und Amtsstatistik, durch Weiterentwicklung der 2011 angewandten Methodik im Jahr 2021 auf eine harte „10.000er-Grenze“ verzichten zu können.

Burgard et al. (2020) zeigen in ihrem ausführlichen und sehr überzeugenden Beitrag, in welche Richtung der Zensus 2021 gehen muss, um insbesondere dem Gleichbehandlungsgebot noch besser Rechnung tragen zu können. Ich halte die gemachten Vorschläge aus methodischer Sicht für exzellent.

Im Folgenden werde ich auf einzelne Aspekte der vorgeschlagenen Methodik spezieller eingehen und sie durch Kommentare ergänzen. Insbesondere möchte ich hervorheben, dass die hier zum ersten Mal in die Stichprobenmethodik eingeführten quadratischen Bézierkurven in der Tradition von Ideen aus der nationalen und europäischen amtlichen Statistik stehen, und ihren Zusammenhang mit dem bekannten Prinzip der vergleichbaren Präzision (Krug et al. 2001) diskutieren.

1.4.2 Anmerkungen zum Gleichbehandlungsgebot

Selbstverständlich darf im Zensus die interföderale und interkommunale Gleichbehandlung nicht in Frage stehen. Eine völlig identische Behandlung aller Kommunen im Rahmen der statistischen Methodik ist dennoch eine Chimäre. Falls etwa die Einwohnerzahl E einer Gemeinde stichprobentheoretisch geschätzt werden soll, besteht zwischen dem relativen und dem absoluten Standardfehler der Schätzfunktion (relSE bzw. absSE) die Beziehung $\text{absSE} = \text{relSE} \cdot E$. Würde man für alle Gemeinden denselben relativen Standardfehler vorschreiben, müssten die absoluten Standardfehler mit der Größe der Gemeinden wachsen; gleichzeitig müsste der Auswahlatz in kleinen Gemeinden sehr groß sein und vielerorts eine Vollerhebung durchgeführt werden. Von absoluter Gleichbehandlung aller Gemeinden könnte man also selbst bei konstantem relativen Standardfehler nicht sprechen.

Der Vorschlag von Burgard et al. (2020), für Gemeinden ab 10.000 Einwohnern (laut Register) einen konstanten *relativen* Standardfehler und für Gemeinden unter einer noch genau zu bestimmenden Schwelle E_0 (z. B. 1000 Einwohner) einen konstanten *absoluten* Standardfehler zu fordern, ist ein hervorragender Kompromiss zwischen dem Gleichbehandlungsgebot der Kommunen und einer gleichmäßigen Belastung der Bürgerinnen und Bürger. Zwischen den Werten E_0 und 10.000 muss ein gleitender Übergang mittels einer Präzisionszielfunktion gefunden werden. Burgard et al. (2020) schlagen hierfür eine Bézierkurve vor; ich werde in den nächsten Abschnitten noch ausführlich begründen, warum dieser Vorschlag aus methodischer Sicht sehr zu begrüßen ist.

Zuvor soll noch einmal betont werden, dass relative und absolute Standardfehler bei jeder Erhebung erst *nach* Vorliegen der Daten geschätzt werden können, weshalb eine strikte Vorgabe für einen maximalen Standardfehler *vor* der Erhebung nicht sinnvoll ist. Im Zensusgesetz 2011 war entsprechend auch von „angestrebten“ Standardfehlern die Rede; es zeigte sich aber, dass Nicht-Statistiker (und manche kommunalen Statistiker) mit dieser Formulierung überfordert waren. Um eine optimale Allokation des Gesamtstichprobenumfangs auf alle Schichten, insbesondere auf die einzelnen Gemeinden durchzuführen, die „angestrebte“ Fehler sicherzustellen

len versucht, müssen möglichst realistische Annahmen über Varianzen und Kovarianzen in den Schichten (die u.a. von der Qualität der Register abhängen) getroffen werden. Diese Annahmen können sich für den Zensus 2021 nur auf die Zensus-Stichprobe von 2011 stützen. Alle Modellrechnungen für 2021 enthalten also einerseits den Stichprobenfehler von 2011 und einen zeitlichen Versatz von 10 Jahren. Auch zu diesem Punkt machen Burgard et al. (2020) sehr gute Vorschläge (u.a. die Verwendung von Bootstrap-Verteilungen der geschätzten Varianzreduktionsfaktoren). Es liegt zwar in der Natur von Stichprobenerhebungen, dass eine Erreichung der angestrebten maximalen Fehler a priori nicht für alle Gemeinden garantiert werden kann, durch eine Umsetzung der vorliegenden Vorschläge wird aber alles getan, um die Zielerreichung weitestgehend sicherzustellen.

1.4.3 Präzisionszielfunktionen

Das Problem, die Stichprobenfehler in Schichten unterschiedlicher Größen aufeinander abzustimmen, ist in der amtlichen Statistik nicht neu. Im Statistischen Bundesamt wird seit Jahrzehnten das „Prinzip der vergleichbaren Präzision“ angewendet (Krug et al. 2001, S. 124). Es beruht auf einer Abstufung des relativen Standardfehlers in Abhängigkeit von den Totalwerten eines Merkmals in den Schichten; auf den Zensus bezogen lautet das Prinzip

$$\text{relSE}(E) = \frac{C}{E^a} \quad \text{mit } 0 \leq a \leq 1. \quad (1)$$

Der Exponent a muss dabei normativ gesetzt werden, C ist eine Proportionalitätskonstante, die vom Gesamtstichprobenumfang abhängt. Sollen alle Gemeinden denselben relativen Standardfehler erhalten, muss $a = 0$ sein. Sollen alle Gemeinden denselben absoluten Standardfehler erhalten, muss $a = 1$ gelten. In der Praxis der amtlichen Statistik werden für a üblicherweise Werte zwischen 0,1 und 0,3 gewählt (Krug et al. 2001, S. 125).

Um einen Eindruck zu bekommen, welchen Einfluss a auf den nötigen Stichprobenumfang hat, gehe ich vereinfacht von einer einfachen Stichprobe vom Umfang n aus N Anschriften in einer Gemeinde und ungebundener Hochrechnung aus; der relative Standardfehler für die Bevölkerungszahl ist dann

$$\text{relSE}(E) = \frac{N \sqrt{1-f} S_y}{E \sqrt{n}}, \quad (2)$$

wobei f der Auswahlsatz und S_y die Standardabweichung der Einwohnerzahl (in allen Anschriften) ist. Unter der vereinfachten Annahme, dass der Faktor $(1-f)$ vernachlässigbar und $N \cdot S_y / E$ ungefähr konstant ist, ergäbe sich durch Gleichsetzen von (1) und (2):

$$\text{relSE}(E) = \frac{C'}{\sqrt{n}} = \frac{C}{E^a},$$

also

$$n \propto E^{2a},$$

d.h. n wäre näherungsweise proportional zu E^{2a} .

Wählt man wieder $a = 0$, d.h. konstante relative Standardfehler, ergäbe sich ein konstanter Stichprobenumfang für alle Gemeinden. Wählt man $a = 1$, also konstante absolute Standardfehler, ergäbe sich ein Stichprobenumfang, der proportional zum Quadrat der Einwohnerzahl wäre. Für $a = \frac{1}{2}$ wäre n proportional zur Einwohnerzahl (was einer proportionalen Aufteilung des Gesamtstichprobenumfangs entspräche), für $a = \frac{1}{4}$ ergäbe sich $n \propto \sqrt{E}$, d.h. der Stichprobenumfang würde proportional zur Wurzel der Einwohnerzahl steigen.

Der von Burgard et al. (2020) zitierte Vorschlag von Verma (1991) lautet, wenn wir ihn auf den Zensus beziehen, wie folgt:

$$n \approx C \sqrt{E} + B.$$

Man kann dies als Kompromiss ansehen zwischen $a = 0$ (konstanter Summand B) und $a = \frac{1}{4}$ (erster Summand, der proportional zu $\sqrt{E}$ ist). Man sieht also, dass es sich bei Vermas Vorschlag im Wesentlichen „nur“ um eine Konkretisierung des Prinzips der vergleichbaren Präzision handelt.

1.4.4 Bézierkurven

Bézierkurven zur Interpolation Weil der Zensus 2021 nicht hinter die Genauigkeit des letzten Zensus zurückfallen soll, ist die Vorgabe eines anzustrebenden relativen Standardfehlers von 0,5 % für Gemeinden mit (laut Register) mindestens 10.000 Einwohnern sinnvoll. Ebenso sinnvoll erscheint es, für sehr kleine Gemeinden einen konstanten absoluten Standardfehler vorzugeben. Ab welcher Einwohnerzahl E_0 eine Gemeinde als „sehr klein“ gelten soll, ist eine politische und keine statistische Frage; statistische Gründe führen aber dazu, dass auch ein vorgegebener absoluter Standardfehler für manche Gemeinden zu einer Vollerhebung führen muss. Offen ist noch die Frage, welche Standardfehler in Gemeinden zwischen E_0 und 10.000 Einwohnern anzustreben sind. Burgard et al. (2020) haben den eleganten Vorschlag gemacht, eine Interpolation mit einer quadratischen Bézierkurve durchzuführen, weil sie unter den hier geltenden Rahmenbedingungen schöne mathematische Eigenschaften hat (differenzierbar, streng monoton, konvex). Die Autoren haben nicht explizit ausgeführt, ob die Interpolation durch eine Bézierkurve für die absoluten oder für die relativen Standardfehler geschehen sollte; tatsächlich sind die entstehenden Kurven nicht äquivalent. Ich werde im Folgenden argumentieren, dass die Verwendung der Bézierkurve im Kontext des Zensus äußerst sinnvoll erscheint, aber aus methodischer Sicht eine Interpolation der *absoluten* Standardfehlerfunktion mittels einer Bézierkurve vorzuziehen ist.

Bézierkurven sind in der Computergraphik und im CAD („computer aided design“) weit verbreitet, weil sie mit Hilfe weniger sogenannter Kontrollpunkte glatte Kurven in der Ebene erzeugen lassen. Die parametrisierte Darstellung einer qua-

dratischen Bézierkurve mit drei verschiedenen gegebenen Punkten $\underline{p}_0, \underline{p}_1, \underline{p}_2 \in \mathbb{R}^2$ ist

$$B : [0,1] \longrightarrow \mathbb{R}^2 \text{ mit } B(t) = (1-t)^2 \underline{p}_0 + 2t(1-t) \underline{p}_1 + t^2 \underline{p}_2. \quad (3)$$

Um den Graphen einer Funktion zwischen $\underline{p}_0 = (x_0, y_0)$ und $\underline{p}_2 = (x_2, y_2)$ zu interpolieren, kann man $\underline{p}_1 = (x_1, y_1)$ als Schnittpunkt der beiden Tangenten an den Graphen in x_0 und x_2 definieren. Wenn x_1 zwischen x_0 und x_2 liegt, liefert die Bézierkurve (3) eine auch an den Endpunkten differenzierbare Interpolation des Funktionsgraphen. Die explizite Funktionsdarstellung des interpolierten Graphen wird nachstehend hergeleitet.

Gegeben seien zwei x -Werte mit Funktionswerten und Ableitungen, $x_0, y_0 = f(x_0), y'_0 = f'(x_0)$ und $x_2, y_2 = f(x_2), y'_2 = f'(x_2)$. Der Schnittpunkt der beiden Tangenten an den Funktionsgraphen in x_0 und x_2 hat dann die Koordinaten

$$x_1 = \frac{y_0 - y_2 - x_0 y'_0 + x_2 y'_2}{y'_2 - y'_0},$$

$$y_1 = y_0 + y'_0(x_1 - x_0).$$

Ich nehme im Folgenden an, dass $x_0 < x_1 < x_2$ gilt; nur dieser Fall (der bei der Anwendung auf den Zensus gegeben ist) ist für die Funktionsinterpolation relevant. Aus (3) ergibt sich für die x -Koordinate der Bézierkurve die quadratische Gleichung

$$x = (1-t)^2 x_0 + 2t(1-t)x_1 + t^2 x_2,$$

die nun nach t aufgelöst werden soll. Hierfür müssen die beiden Fälle $x_1 \neq \frac{x_0+x_2}{2}$ und $x_1 = \frac{x_0+x_2}{2}$ unterschieden werden.

Für den Fall $x_1 \neq \frac{x_0+x_2}{2}$ ergibt sich

$$t = \frac{x_0 - x_1 + \sqrt{(x_0 - 2x_1 + x_2) \cdot x + x_1^2 - x_0 x_2}}{x_0 - 2x_1 + x_2}.$$

Eingesetzt in die y -Koordinate $y = (1-t)^2 y_0 + 2t(1-t)y_1 + t^2 y_2$ erhält man

$$y = ax + b + c\sqrt{dx + e} \quad (4)$$

mit

$$a = \frac{y_0 - 2y_1 + y_2}{x_0 - 2x_1 + x_2}, \quad d = x_0 - 2x_1 + x_2, \quad e = x_1^2 - x_0 x_2,$$

$$b = y_0 - \frac{2(y_1 - y_0)(x_1 - x_0)}{x_0 - 2x_1 + x_2} + \frac{((x_1 - x_0)^2 + x_1^2 - x_0 x_2)(y_0 - 2y_1 + y_2)}{(x_0 - 2x_1 + x_2)^2},$$

$$c = \frac{2(y_1 - y_0)}{x_0 - 2x_1 + x_2} - \frac{2(x_1 - x_0) \cdot (y_0 - 2y_1 + y_2)}{(x_0 - 2x_1 + x_2)^2}.$$

Der expliziten Darstellung (4) lässt sich entnehmen, dass es sich bei der quadratischen Bézierkurve um eine Linearkombination aus einer linearen Funktion und einem Wurzelterm handelt.

Im Fall $x_1 = \frac{x_0+x_2}{2}$ ergibt sich $t = \frac{x-x_0}{2(x_1-x_0)}$. Wiederum eingesetzt in die y -Koordinate $y = (1-t)^2y_0 + 2t(1-t)y_1 + t^2y_2$ ergibt sich

$$y = ax^2 + bx + c \quad (5)$$

mit

$$a = \frac{y_0 - 2y_1 + y_2}{4(x_1 - x_0)^2}, \quad b = \frac{2x_1(y_1 - y_0) + x_0(y_0 - y_2)}{2(x_1 - x_0)^2},$$

$$c = y_0 + \frac{x_0^2(y_2 + 2y_1 - 3y_0) + 4x_0x_1(y_0 - y_1)}{4(x_1 - x_0)^2}.$$

In diesem Spezialfall stellt die quadratische Bézierkurve also eine quadratische Funktion dar.

Bézierkurven als Präzisionszielfunktionen Interpoliert man den zulässigen absoluten Standardfehler (in Abhängigkeit der Einwohnerzahl E) zwischen E_0 und 10.000 mit einer Bézierkurve, ergeben sich (falls $x_1 \neq \frac{x_0+x_2}{2}$) aus (4) für den absoluten und den relativen Standardfehler Funktionen der Form

$$\text{absSE}(E) = aE + b + c\sqrt{dE + e},$$

$$\text{relSE}(E) = a + \frac{b}{E} + c\sqrt{d + \frac{e}{E}}.$$

Der relative Standardfehler ist also im Wesentlichen eine Linearkombination aus einem konstanten Term, einem Term der Ordnung $\frac{1}{\sqrt{E}}$ und einem Term der Ordnung $\frac{1}{E}$. Man könnte von einer Verallgemeinerung des Prinzips der vergleichbaren Präzision sprechen; es handelt sich um eine Linearkombination der Fälle $a = 0$, $a = \frac{1}{2}$ und $a = 1$. Wird E größer, wird der Term zu $a = 0$ bedeutsamer (d.h. wir nähern uns einem konstanten relativen Standardfehler); wird E kleiner, wird der Term zu $a = 1$ immer bedeutsamer (d.h. wir nähern uns einem konstanten absoluten Standardfehler).

Interpoliert man dagegen den relativen Standardfehler (in Abhängigkeit der Einwohnerzahl E) mit einer Bézierkurve, ergeben sich für den absoluten und den relativen Standardfehler Funktionen der Form

$$\text{relSE}(E) = aE + b + c\sqrt{dE + e}$$

$$\text{absSE}(E) = aE^2 + bE + c\sqrt{dE^2 + eE}.$$

Obwohl auch diese Interpolationen differenzierbare Funktionen liefern, lässt sich die funktionale Darstellung des relativen Standardfehlers nicht in den Kontext des Prinzips der vergleichbaren Präzision stellen (keine Linearkombination von Termen der Form $1/E^a$ mit $a \geq 0$). Die Definition der Bézierkurve für die absoluten

Standardfehler ist daher eindeutig vorzuziehen, weil sie sehr verwandt mit schon lange angewandten Prinzipien der Präzisionsabstufung in der amtlichen Statistik ist.

Falls $x_1 = \frac{x_0 + x_2}{2}$, ergeben sich aus (5) für den absoluten und den relativen Standardfehler bei Bézierkurven-Interpolation des absoluten Fehlers Funktionen der Form

$$\text{absSE}(E) = aE^2 + bE + c \quad \text{und} \quad \text{relSE}(E) = aE + b + \frac{c}{E},$$

während sich bei Interpolation des relativen Fehlers Funktionen der Form

$$\text{relSE}(E) = aE^2 + bE + c \quad \text{und} \quad \text{absSE}(E) = aE^3 + bE^2 + cE$$

ergäben. Auch hier wäre die Interpolation des absoluten Fehlers vorzuziehen, weil dann der relative Standardfehler eine größere Ähnlichkeit zu Funktionstermen nach dem Prinzip der vergleichbaren Präzision aufweist.

Ein Rechenbeispiel Werden der (maximale) absolute Standardfehler für Gemeinden bis 1000 Einwohnern auf 15 und der relative Standardfehler für Gemeinden ab 10.000 Einwohnern auf 0,5 % festgelegt, sind für die Bézierkurven-Interpolation des absoluten Standardfehlers $x_0 = 1000$, $y_0 = 15$, $y'_0 = 0$ und $x_2 = 10.000$, $y_2 = 50$, $y'_2 = 0,5\%$. Es ergibt sich die Funktion

$$\text{absSE}(E) = 0,007 \cdot E + 19,2 - 0,0056 \sqrt{5000 \cdot E - 1.000.000}$$

und daraus

$$\text{relSE}(E) = 0,007 + \frac{19,2}{E} - 0,0056 \sqrt{\frac{5000}{E} - \frac{1.000.000}{E^2}}.$$

Die Funktionsgraphen sind in Abb. 1 gestrichelt dargestellt.

Für die Bézierkurven-Interpolation des relativen Standardfehlers wären $x_0 = 1000$, $y_0 = 0,015$, $y'_0 = -1,5 \cdot 10^{-5}$ und $x_2 = 10.000$, $y_2 = 0,5\%$, $y'_2 = 0$. Es ergäbe sich die Funktion

$$\text{relSE}(E) = 1,304 \cdot 10^{-6} \cdot E + 0,0156 - 2,836 \cdot 10^{-6} \sqrt{7667 \cdot E - 7.222.222}$$

und daraus

$$\text{absSE}(E) = 1,304 \cdot 10^{-6} \cdot E^2 + 0,0156 \cdot E - 2,836 \cdot 10^{-6} \sqrt{7667 \cdot E^3 - 7.222.222 \cdot E}.$$

Die Funktionsgraphen sind in Abb. 1 gepunktet dargestellt. Deutlich ist zu sehen, dass die Interpolationsfunktion für den absoluten Standardfehler auf Basis der Bézierkurve für den relativen Standardfehler nicht konvex ist; ein weiterer Grund also, der für eine Interpolation auf Basis der Bézierkurve für den *absoluten* Standardfehler spricht.

Abb. 1 Absolute (*links*) und relative Standardfehler (*rechts*) aus Interpolation der absoluten Standardfehler (*gestrichelt*) und aus Interpolation der relativen Standardfehler (*gepunktet*) mit Bézierkurven

1.4.5 Berücksichtigung kleinster Gemeinden

Burgard et al. (2020) machen in Abschn. 3.6 das Problem deutlich, das in Bundesländern mit vielen Kleinstgemeinden besteht. Selbst ein vorgegebener konstanter absoluter Standardfehler für Gemeinden unter 1000 Einwohnern führt zur Notwendigkeit von Vollerhebungen in Kleinstgemeinden. Aus statistischer Sicht spräche nichts gegen ein solches Vorgehen zur Einwohnerzahlbestimmung, aber es besteht ein Zielkonflikt mit der vom Bundesverfassungsgericht geforderten *schonenden* Befragung. Daher ist es hinnehmbar, Kleinstgemeinden zu gemeinsamen Schichten zusammenzufassen und die Präzisionsziele nur auf der zusammengefassten Ebene zu formulieren. Trotzdem müssen auf Gemeindeebene Schätzungen der Einwohnerzahl produziert werden. In der Small-Area-Literatur sind zusammengesetzte Schätzer bekannt, die eine Konvexkombination aus einem direkten und einem synthetischen Schätzer bilden:

$$\hat{E}_{\text{zus}} = \lambda \hat{E}_d + (1 - \lambda) \hat{E}_{\text{syn}} \quad \text{mit } \lambda \in [0, 1].$$

Im Rahmen des Arbeitskreises für mathematisch-statistische Methodik beim Statistischen Bundesamt (im Juni 2019) haben die Autoren Burgard, Münnich und Rupp die Anwendung eines solchen synthetischen Schätzers bereits diskutiert; als direkter Schätzer wäre die Schätzung auf Gemeindeebene mit den GREG-Gewichten zu verwenden, als synthetischen Schätzer könnte man die Registerzahl sehen. λ würde man umso größer wählen, je größer der Stichprobenumfang in der betrachteten Gemeinde ist. Allerdings muss zum einen noch die Kohärenz mit den Schätzungen auf Verbandsgemeindeebene sichergestellt sein, und zum anderen müsste die Schätzung des Standardfehlers von $\hat{E}_{\text{zus}}$ design- und modellbasierte Komponenten verbinden. Hierzu sind in jedem Fall noch weitere methodische Überlegungen notwendig.

1.4.6 Ausblick

Mit den Überlegungen aus Burgard et al. (2020) ist der Zensus 2021 im Zusammenspiel aus Stichprobenziehung und Hochrechnung aus methodischer Sicht optimal aufgestellt. In Verwaltung und Politik scheint es bereits ausgemacht, dass der Zensus 2031 vollständig registerbasiert durchgeführt werden wird. Zweifellos ist es aus Kostengründen und auch zur Entlastung der Bürgerinnen und Bürger geboten, alle Informationen zu nutzen, die in administrativen Daten und/oder Registern vorliegen, bevor eine zusätzliche Primärerhebung durchgeführt wird. Ob wenigstens die Einwohnerregister schon die nötige Qualität aufweisen, um in Zukunft ganz auf ergänzende Stichproben verzichten zu können, werden die Ergebnisse des Zensus 2021 zumindest teilweise zeigen. Mir scheint jedenfalls auch in Zukunft eine (möglicherweise kontinuierliche und im Stichprobenumfang bedeutend reduzierte) stichprobengestützte Qualitätsüberprüfung der Registerinformationen geboten, um insbesondere den Umfang von Karteileichenbeständen regelmäßig zu kontrollieren. Ob dies schon im Vorfeld oder erst während des Zensus 2031 stattfindet, kann diskutiert werden. Dass eine ergänzende Stichprobenerhebung auch für den Zensus 2031 noch nötig und sinnvoll sein wird, halte ich aber für sehr realistisch.

2 Beantwortung der Kommentare

2.1 Einleitung

Wir danken den Autoren der vier Kommentare zu unserem Artikel sehr für die konstruktive Auseinandersetzung mit der Thematik. Diese öffentliche Auseinandersetzung mit dem Zensus 2021, und hier insbesondere der Qualitätsmessung, soll zur Transparenz der Verfahren beitragen. Gerade angesichts der Diskussion um die 10.000er Grenze beim Bundesverfassungsgericht erachten wir es für unabdingbar, eine plausible und insbesondere in der Fachwelt diskutierte und geeignete Lösung vorzustellen, wie man eine Qualitätsmessung unter den Rahmenbedingungen für den Zensus 2021 durchführen kann.

2.2 Kommentar 1: Walter Krämer

Die Einleitung von Walter Krämer erinnert ein wenig an ein kleines gallisches Dorf. In der Tat ist eine Spezialisierung auf die Survey-Statistik mit Zensen, Stichprobenerhebungen und Registern im Zeitalter der Effizienzmessung durch *Impactfaktoren* nicht ganz einfach. Gerade aber der aktuelle Hype um Data Science macht deutlich, dass das Verständnis um Qualität und Qualitätsmessung sowie die Herkunft der Daten in Form des datengenerierenden Prozesses, auch statistischer Produktionsprozess in der amtlichen Statistik genannt, von besonderer Bedeutung ist und bei Anwendungen von statistischen Methoden unbedingt Berücksichtigung finden muss. Wie Walter Krämer anmerkt, sind diese Themen so bedeutsam für Staat und Gesellschaft, dass sie vor dem Bundesverfassungsgericht diskutiert werden. Hierzu wurden mit der Thematik einschlägig vertraute Statistiker als Gutachter bestellt. Die

Abb. 2 Effizienzgewinne in Abhängigkeit von Schätzer und Design

Wichtigkeit dieses Wissenschaftszweiges wurde auch vom *European Statistical System* erkannt, unter deren Dach der *European Master in Official Statistics* (EMOS; <https://ec.europa.eu/eurostat/web/european-statistical-system/emos>) entstanden ist.

Der Hinweis auf die theoretische Aufarbeitung des Effizienzgewinnes zwischen dem Horvitz-Thompson Schätzer und einem GREG-Schätzer ist durchaus berechtigt. Er lässt sich im einfachen Fall einer einfachen Zufallsstichprobe durch Formel (4) darstellen. Wie im Anhang A dargestellt, ist dieser Vergleich unter einem stratifizierten Design nicht mehr so trivial, jedoch trotzdem analytisch beweisbar. Um allerdings die tatsächlichen Effizienzgewinne zu messen, die durch die Optimierung des Designs resultieren, müsste der Zusammenhang zwischen GREG-Schätzer unter einer einfachen Zufallsstichprobe und dem GREG-Schätzer unter einem stratifizierten Design dargestellt werden (siehe Abb. 2).

Dies ist jedoch nicht ohne weiteres explizit durch einen analytischen Ausdruck darzustellen. Betrachtet man die Varianz des HT von SRS auf StrRS, dann redet man vom reinen Schichtungseffekt. Dieser lässt sich gut im Falle einer proportionalen Aufteilung des Stichprobenumfangs auf die Schichten darstellen, da sich eine Varianzreduktion proportional zur Zwischenschichtenvarianz ergibt. In Analogie hierzu kann man vergleichbare Formeln für den GREG ableiten, welche jedoch im Modell mit Zurücklegen beide Varianzreduktionsfaktoren (vom HT zum GREG) enthalten:

$$\text{var}(\hat{\tau}_{\text{GREG}}) = \text{var}(\hat{\tau}_{\text{HT}}) \cdot \frac{1 - \varrho^2}{1 - \vartheta^2} + N^2 \cdot \frac{\sigma_b^2}{n} \cdot (1 - \varrho^2) \quad .$$

Für andere Allokationen, und dies gilt auch für die verwendete optimale Aufteilung mit Unter- und Obergrenze, kann keine analoge Darstellung in einfacher Weise hergeleitet werden.

Somit besteht die einzige Möglichkeit einer Analyse unter Verwendung eines geeigneten synthetischen Datensatzes. Allerdings basieren die Berechnungen dann auf einem synthetischen Datensatz, in welchem die Karteileichen und Fehlbestände vordefinierten Modellen unterliegen. Diese Modelle wurden vor der Durchführung des Zensus 2011 entwickelt. Leider war es uns aufgrund fehlender Informationen nicht möglich, diese Modelle geeignet zu aktualisieren. Daher entsprechen die Strukturen in dem synthetischen Datensatz unter Umständen nicht vollständig der Realität, wodurch eine Angabe von Effizienzgewinnen auf Basis einer Simulationsstudie nur unzureichend auf reale Verhältnisse übertragbar wäre. In Abb. 3 wurden für Rheinland-Pfalz unter Verwendung der Zensus-Simulationsgesamtheit von 2011 bei aktueller Schichtung in Bezug auf Gemeindeverbände die Varianzreduktionsfaktoren in Bezug auf acht Größenklassen der Gemeindeverbände dargestellt. Die sieben Trenn-

Abb. 3 Varianzreduktion im Vergleich vom HT bei einfacher Zufallsstichprobe im Vergleich zu Schichtung oder Verwendung des GREG

punkte der Größenklassen sind bei 10.000, 12.500, 15.000, 17.500, 20.000, 30.000 und 50.000 Einwohnern. Der Einfachheit halber wurde jeweils die Verbandsgemeinde (bzw. Stadt) in der Klasse genommen, welche den Median der Populationsgröße aufweist. Zur besseren Übersicht wurde die Kombination HT / SRS jeweils auf 100% gesetzt.

Man erkennt eine klare Struktur der Varianzreduktion sowie die Tatsache, dass in allen Fällen der GREG bei geschichteter Zufallsstichprobe das beste Verfahren in diesem Vergleich ist. Ebenso ersichtlich sind unterschiedliche, von der Größenklasse abhängige Varianzreduktionen, die zwischen den Verfahren unterschiedlich ausfallen.

Wie im Kommentar ausführlich dargestellt, sind die meisten methodischen Überlegungen eine Folge der Frage, ob die Präzisionsanforderungen für größere Gemeinden auch auf kleinere übertragbar sind. Als Statistiker stand für uns die allgemeine Problematik von relativen Maßen in kleinen Gemeinden im Mittelpunkt. Dies führte letztendlich zur Anpassung der gesamten Methodik durch die Einführung einer nicht-linearen, aber glatten Präzisionszielfunktion. Die im Kommentar ausgeführten *berüchtigten Jungen-Anteile in Flensburg im Zensus 2011* sind nach unseren Analysen jedoch nicht ausschließlich auf die Größe der Gemeinde zurückzuführen. Eine entscheidende Rolle dabei nehmen die verwendeten Hilfsmerkmale für die GREG-Schätzung ein. Einerseits kann eine zu geringe Anzahl von verwendeten Regressoren zu einer Unterkorrektur der Designgewichte führen, andererseits führen eine zu hohe Zahl von Regressoren – insbesondere in kleinen Regionen – zu einer Überkorrektur der Gewichte. Falls eines der beiden Phänomene auftritt, kann es sehr leicht zu verzerrten Ergebnissen insbesondere für Untergruppen in kleinen Regionen kommen. Wie bereits in Burgard et al. (2020) in der Einleitung erläutert, kann dieses Problem insbesondere durch die Verwendung einer verallgemeinerten Kalibrierung (siehe Burgard et al. 2019) erheblich abgeschwächt werden. Diese erlaubt die Einschränkung der Kalibrierungsgewichte unter Verwendung von nicht exakt zu erfüllenden Nebenbedingungen. Die genaue Vorgehensweise im Rahmen der ge-

meindespezifischen Schätzungen des Zensus 2021 werden bei der Beantwortung des Kommentars 4 genauer erläutert.

Über die Bewertung des Artikels als *wertvoller Beitrag zur modernen Zensus-problematik* freuen wir uns sehr, möchten allerdings gleichzeitig erwähnen, dass der Artikel nicht die vollständige stichprobentheoretische Methodik für den Zensus abdeckt. Beispielsweise steht die Analyse der Qualität der Register, die als Auswahlgrundlage dienen, sowie die Schätzmethodik auf Gemeindeebene nicht auf der Agenda dieses Artikels.

2.3 Kommentar 2: Helmut Küchenhoff

Helmut Küchenhoff weist zurecht darauf hin, dass neben Stichprobenfehlern auch Nichtstichprobenfehler von besonderem Interesse bei einem Zensus sind. Als wohl größte Erhebung der amtlichen Statistik unterliegt sie selbstverständlich auch dem *European Statistics Code of Practice* (siehe <https://ec.europa.eu/eurostat/documents/4031688/9394019/KS-02-18-142-DE-N.pdf/27ca19ca-e349-45f8-bbd4-4d78a33601ae>), und insbesondere auch dem Grundsatz 12.2, welcher sich genau mit diesen beiden Arten von Fehlern befasst.

Eine intensive und ausführliche Auseinandersetzung mit Nichtstichprobenfehlern übersteigt die Möglichkeiten in dieser Antwort, da konkrete Empfehlungen tief gehende Untersuchungen erfordern würden, welche mit geeigneten Daten untermauert werden müssten. Für den Zensus 2011 wurde aber bedauerlicherweise gesetzlich festgelegt, dass jedwede Hilfsmerkmale in angemessener Zeit gelöscht werden müssen. Paradata zum Zensus 2011, aus welchen man möglicherweise hätte lernen können, standen den Autoren des Artikels auch nicht zur Verfügung. Damit besteht keine geeignete Basis für Nachuntersuchungen zu Nichtstichprobenfehlern im Zensus 2011. Neben den von Küchenhoff näher dargelegten Messfehlern spielt insbesondere der inkorrekte Rahmen eine herausragende Rolle. Nonresponse und Interviewereffekte können zwar auf Adressebene als Messfehler modelliert werden. Allerdings gibt es zahlreiche Methoden, um mit diesen Fehlern auf Individualebene umzugehen.

Im Vorfeld des Zensus 2011 wurden zahlreiche Untersuchungen zu möglichen Auswirkungen der Karteileichen- und Fehlbestandsmodelle auf die Methodik des Zensus 2011 und insbesondere auf das optimale Stichprobendesign durchgeführt (siehe Münnich et al. 2012). Dabei zeigt sich, dass selbst eine Kenntnis der Verteilung von Karteileichen und Fehlbeständen kaum zu einer Verbesserung des Designs beitragen kann, aber inkorrekte Annahmen durchaus zu Verschlechterungen. Im Rahmen zweier Nachuntersuchungen zum Zensus 2011 wurden weitere Nichtstichprobenfehler untersucht. Zum einen wurden eine Reihe von inkorrekten Nullanschriften gefunden, welche nicht zum Ziehungsrahmen des Zensus gehören sollten, wie etwa Garagen und Elektrohäuschen (siehe Münnich et al. 2013). Es ergab sich bei diesen Untersuchungen, dass die empfohlene Methodik sehr robust ist und trotz dieser Fehler keine anderen Empfehlungen hätten getroffen werden müssen. Allerdings zeigt sich im Einzelfall auch, dass bei Häufungen dieser eigentlich nicht zu berücksichtigenden Nullanschriften mit einem Qualitätsverlust zu rechnen ist. Ein ähnlicher Effekt wird von Küchenhoff in Bezug auf Messfehler erwähnt. Des

Weiteren wurde untersucht, wie fehlende Werte beim Zensus statistisch korrigiert werden können. Diese waren aber in Bezug auf Ziel 2, also weitere interessierende Variablen neben der im Vordergrund stehenden Bevölkerungszahl (siehe Münnich et al. 2015). Die Ergebnisse zeigen, dass eine Korrektur für Nonresponse durchaus erfolgreich ist, allerdings zu einer Verwendung sehr computerintensiver Methoden führt.

Ebenso korrekter Weise wird von Helmut Küchenhoff auch auf den *Post Enumeration Survey* hingewiesen. Die Literatur dazu ist in Bezug auf Zensen umfangreich. Mit Hilfe dieser Art von Wiederholungsbefragungen sollen insbesondere mögliche Rahmen- und Messfehler aufgedeckt werden. Für den Zensus 2011 wurde eine solche Erhebung tatsächlich durchgeführt (siehe Klink und Bihler 2015). Idee dieser Wiederholungsbefragung ist es, mit besonders geschultem Personal in kleinerem Umfang präzisere Ergebnisse zu erhalten und mögliche Abweichungen zwischen beiden Befragungen aufzudecken. Diese Art der Untersuchungen sind außerordentlich wichtig. Gleichwohl bleibt die Frage, wie mit derartigen Ergebnissen umzugehen ist. Eine Korrektur der eigentlichen Zensusergebnisse auf Basis einer kleineren Stichprobe ist nicht unumstritten.

Abschließend erlauben wir uns nochmals auf einige Prämissen hinzuweisen. Unsere Untersuchungen waren insbesondere auf das Stichprobendesign inklusive geeigneter Schätzmethodik ausgerichtet. Geeignete Informationen, um die von Helmut Küchenhoff angesprochenen Nichtstichprobenfehler mit zu untersuchen, konnten uns nicht zur Verfügung gestellt werden. Somit hatten wir keine geeignete Basis, diese mit in die Simulationsuntersuchungen einzubeziehen. Aus unserer Erfahrung ist der GREG, so wie er im Zensus verwendet werden soll, sehr robust. Kleinere Qualitätsverluste können jedoch bei spürbaren systematischen Fehlern nicht ausgeschlossen werden. Wie bereits von Küchenhoff angemerkt, wäre es für zukünftige Zensen sinnvoll, die Qualitätsziele des Zensus aus Sicht des *Total Survey Errors* (siehe Groves und Lyberg 2010) zu betrachten.

2.4 Kommentar 3: Sara Bleninger und Michael Fürnrohr

Der Kommentar erörtert sehr ausführlich die Hauptproblematik, mit der sich unser Artikel auseinandersetzt. Durch die Betrachtung wechselseitiger Anforderungen, vielseitiger Restriktionen und allgemeingültiger Zielvorgaben entsteht bei der Entwicklung der Methodik für den Zensus 2021 ein Spannungsfeld aus Schätzpräzision, Rechtssicherheit, Grundrechtsschonung, Kostenkontrolle und der allgemeinen Stichprobenmethodik. Diesbezüglich galt es in erster Instanz die rechtlich vorgegebenen Rahmenbedingungen einzuhalten, die sich aus dem Zensusvorbereitungsgesetz (siehe ZensusvorbereitungsgG2021 2017) und dem Urteil des Bundesverfassungsgerichtes (siehe BVerfG 2018) ergaben. Auf dieser Grundlage wurden innovative Präzisionsanforderungen in Abhängigkeit von der Gemeindegröße definiert, um dem Spannungsfeld aus Schätzpräzision, Grundrechtsschonung und Kostenkontrolle gerecht zu werden. Dabei teilen wir vollends die Bemerkung der Kommentatoren über die Wichtigkeit der Gewährleistung der *informationellen Selbstbestimmung der Bürgerinnen und Bürger*. Wie auch im Kommentar bereits erläutert, lässt sich dies nur mit einer nicht-linearen Präzisionszielfunktion sowie einer teilweisen Betrachtung von

Gemeindeverbänden (bzw. Gemeindeverbandsresten) als Aggregat mehrerer kleiner Gemeinden vereinbaren. Gleichzeitig ist uns die Problematik einer über alle Gemeinden uneinheitlichen Definition des maximal zulässigen relativen Standardfehlers bewusst. Diese Skepsis, vorgetragen durch den Deutschen Städtetag und den Deutschen Städte- und Gemeindebund (siehe Deutscher Städtetag 2019), ist durchaus legitim, allerdings lässt der Grundsatz der *schonenden Befragung* keine Alternative zu.

Der Kommentar würdigt das entwickelte Verfahren als: *Kein Kompromiss, sondern eine Lösung*. Über diese Einordnung der präsentierten Verfahren freuen wir uns sehr. Dies kann auch als Leitsatz für die Forschungsarbeiten im Rahmen des Zensus 2021 angesehen werden. Das Ziel der Forschungsarbeiten war es jederzeit, die Fundiertheit der Methodik des Zensus 2011 nicht durch das Urteil des Bundesverfassungsgerichtes einzuschränken, sondern auf Basis des Urteils und der Methodik des Zensus 2011 ein möglichst optimales neues Verfahren zu entwickeln. Dabei stand die Erstellung eines *für die politischen Entscheidungen notwendige[n] methodische[n] Fundament[s]* immer im Vordergrund, diese Aussage teilen wir vollständig mit den beiden Autoren des Kommentars. Zudem freut es uns, dass die Autoren des Kommentars die Komplexität der Forschungsarbeiten in dem zu Beginn erwähnten Spannungsfeld zu würdigen wissen.

2.5 Kommentar 4: Hans Kiesel

In vielerlei Hinsicht stimmen wir mit den Kommentaren und Gedanken von Hans Kiesel überein, in denen er ausführlich die Hauptaussagen des Artikels erörtert. Im Folgenden werden wir insbesondere auf die Präzisionszielfunktion, die Methodik der gemeindespezifischen Schätzungen und den Ausblick auf den Zensus 2031 eingehen. Darüber hinaus möchten wir uns für die detaillierte Ausarbeitung der Definition der Bézierkurve sowie dessen Einsatz in der von uns gewählten Präzisionszielfunktion für den Zensus 2021 bedanken. Dies war uns aufgrund der Komplexität und Länge des Artikels nicht möglich.

2.5.1 Präzisionszielfunktion unter stark variierenden Gemeindegrößen

Hans Kiesel geht zu Beginn ausführlich auf die Definition und Rolle der Präzisionszielfunktion ein. Er nennt den Vorschlag der *glatten* Verbindung von absolutem und relativem Standardfehler einen *hervorragenden Kompromiss zwischen dem Gleichbehandlungsgebot der Kommunen und einer gleichmäßigen Belastung der Bürgerinnen und Bürger*. Die Lösung dieses Spannungsfeldes, wie es im Kommentar 3 genannt wird, war eine der Hauptzielsetzungen bei der Methodenentwicklung. Wir freuen uns, dieser Aufgabe gerecht geworden zu sein. Der in Kommentar 3 genannten Notwendigkeit einer fixen a priori-Definition eines Präzisionsziels sowie der Skepsis des Deutschen Städtetages und des Deutschen Städte- und Gemeindebundes gegenüber einer über alle Gemeinden uneinheitlichen Definition des maximal zulässigen relativen Standardfehlers (vgl. Deutscher Städtetag 2019) begegnet der Autor des Kommentars mit der Aussage, dass auch beim Zensus die Standardfehler erst *nach Vorliegen der Daten geschätzt werden können, weshalb eine strikte Vor-*

gabe für einen maximalen Standardfehler vor der Erhebung nicht sinnvoll ist. Wir stimmen als Stichprobenmethodiker dieser Meinung vollends zu – der zweite Autor des diskutierten Artikels hat hierauf auch bereits bei der Anhörung zum Zensus vor dem Bundesverfassungsgericht ausdrücklich darauf hingewiesen. Allerdings verstehen wir die Sorge der amtlichen Statistik über die Rechtssicherheit der Methodik und haben deshalb versucht, die Methodik unter der Verwendung der oben genannten Präzisionszielfunktion möglichst effizient, aber gleichzeitig robust zu gestalten.

Wie in Burgard et al. (2020, Kapitel 3.1) bereits erläutert, entsprechen die *Draft Regulations* des European Statistical System Committee (siehe European Statistical System Committee 2015, S. 28 f.) bezüglich der Präzisionsanforderungen für Haushaltserhebungen in der Europäischen Union grundsätzlich den Formulierungen von Verma (1991). Im Artikel wird argumentativ gezeigt, dass die für den Zensus 2021 entwickelte Präzisionszielfunktion auf Basis einer Bézierkurve grundsätzlich auch mit diesem Modell vereinbar ist. Dies wird durch den Autor des Kommentars nochmals aufgegriffen und mit dem in der deutschen amtlichen Statistik seit Jahrzehnten angewendeten *Prinzip der vergleichbaren Präzision* (Krug et al. 2001, S. 124) verglichen, welches auf einer Abstufung des relativen Standardfehlers in Abhängigkeit von den Totalwerten eines Merkmals in den Schichten beruht. Es wird gezeigt, dass es sich beim Vorschlag von Verma (1991) im wesentlichen „nur“ um eine Konkretisierung des Prinzips der vergleichbaren Präzision handelt. Somit ist die für den Zensus 2021 entwickelte Präzisionszielfunktion nicht nur durch die *Draft Regulations* des European Statistical System Committee (siehe European Statistical System Committee 2015, S. 28 f.), sondern auch durch die deutsche amtliche Statistik legitimiert. Wir hatten über diese Verknüpfung während der Forschung bereits diskutiert, allerdings fehlt deren Erklärung und deren Beweis innerhalb des Artikels.

Im Zusammenhang mit der in der Präzisionszielfunktion verwendeten Bézierkurve wird die theoretische Herleitung der Bézierkurve im Kommentar detailliert erläutert. Dabei wird insbesondere auf die Grundlage eingegangen, auf der die Bézierkurve definiert wird. Für diese mathematischen Erläuterungen bedanken wir uns sehr, auch dies ist innerhalb des Artikels nur sehr kurz abgehandelt worden. Es wird gezeigt, dass es einen Unterschied macht, ob die Interpolation durch die Bézierkurve auf Basis des absoluten oder des relativen Standardfehler definiert wird. Im Fall der Interpolation auf Basis des relativen Standardfehlers führt dies wie in Abb. 1 zu sehen (gepunktete Kurven) zu einer nicht konvexen Funktion für den absoluten Standardfehler. Unter anderem dieses mathematische Argument führte zu einer Verwendung der Interpolation auf Basis des absoluten Standardfehlers (gestrichelte Kurven). Dieser Aspekt wird im Kommentar insbesondere nochmals durch ein analytisches Beispiel untermalt. Zudem wird gezeigt, dass sich nur die Interpolation auf Basis des absoluten Standardfehlers mit dem Prinzip der vergleichbaren Präzision und dem Prinzip nach Verma (1991) vereinbaren lässt, was ein weiteres Argument für die verwendete Vorgehensweise ist.

2.5.2 Der gemeindespezifische Schätzer

Zum Zeitpunkt der Einreichung des Artikels war die Forschung bezüglich der gemeindespezifischen Schätzungen noch nicht weit fortgeschritten, deshalb wurde die-

ser Aspekt nicht konkretisiert. Kiesel verweist allerdings korrekterweise auf die Wichtigkeit dieses Teilforschungsbereichs für die Gesamtmethodik des Zensus 2021 hin. Daher wird dessen methodische Betrachtung im Folgenden im Rahmen dieser Antwort auf den Kommentar nachgeholt:

Wie bereits erwähnt, ist der vom Bundesverfassungsgericht angemahnte Grundsatz einer *schonenden Befragung* nur mit einer Zusammenfassung von Kleinstgemeinden innerhalb der Schichtung vereinbar. Bei den in Burgard et al. (2020) vorgeschlagenen Schichtungen und Allokationen nach Gemeindeverbandsresten (Alternative B) bzw. Gemeindeverbänden (Alternative C) gelten daher die aus der Präzisionszielfunktion vorgegebenen Qualitätsvorgaben nur für die im Design explizit berücksichtigten Schichten. Diese Schichten werden im Folgenden als *Sampling Point (SMP)* bezeichnet. Diese können einer einzelnen Gemeinde, mehrerer Gemeinden oder einem Gemeindeverband entsprechen. Für die SMPs wird die klassische Schätzmethodik via GREG-Schätzer (bzw. auch verallgemeinerter Kalibrierungsschätzer mit GREG-Zielfunktion) zur Punktschätzung angewendet. Für die Gemeinden, die explizit einem SMP entsprechen, ist somit die Punktschätzung $\hat{\tau}_{\text{SMP}}$ unproblematisch. Für die übrigen, nicht explizit als SMP berücksichtigten Gemeinden wird eine alternative Vorgehensweise vorgeschlagen, die sich einerseits (in Abhängigkeit der Stichprobengröße der Gemeinde) an den Registerwerten orientiert und andererseits zur Gemeindeverbands-(Rest-)Ebene kohärente Schätzungen liefert. Die genaue Methodik wird im folgenden Kapitel detailliert dargestellt.

Zuerst einmal sind die Herausforderungen aufzuzeigen, die bei der Erzeugung gemeindespezifischer Schätzungen zu berücksichtigen sind. Bei der Stichprobenziehung entstehen für die nicht als SMP berücksichtigten Gemeinden teils sehr kleine und insbesondere zufällige gemeindespezifische Stichprobenumfänge. Dies kann bei der Punktschätzung zu einer Instabilität der klassischen GREG-Schätzung führen, was im Allgemeinen für gemeindespezifische Stichprobenumfänge von kleiner als 100 Anschriften erwartbar ist. Des Weiteren muss die Kohärenz der gemeindespezifischen Schätzungen zu den übergeordneten Aggregationsebenen gewährleistet sein. Um diesen Herausforderungen gerecht zu werden, schlagen wir eine zusammengesetzte Schätzung aus einer direkten und indirekten Komponente vor, wobei für die indirekte Schätzung das Konzept der verallgemeinerten Kalibrierung (siehe Burgard et al. 2019) unter Verwendung von relaxierten Nebenbedingungen zum Einsatz kommt.

Wir betrachten nun im Folgenden einen SMP, der aus mehreren Gemeinden besteht. Der SMP ist also ein Gemeindeverband (Alternative C) oder ein Gemeindeverbandsrest (Alternative B). Zur Notation bezeichnen wir mit $g = 1, \dots, G$ alle Gemeinden innerhalb dieses SMP. Wir erinnern nochmals daran, dass dieses SMP einer Schicht im Design entspricht. Des Weiteren bezeichnen wir die Stichprobe des SMP mit S_{SMP} und der Gemeinden mit S_g . Damit gilt $\bigcup_{g=1}^G S_g = S_{\text{SMP}}$. Dann setzt sich die kombinierte Schätzung aus direktem und indirektem Anteil für eine Gemeinde g wie folgt zusammen:

1. **Direkter Schätzer:** $\hat{\tau}_{\text{dir},g}$ Der direkte Schätzer wird berechnet als gemeindespezifischer Punktschätzer mit der Methodik der Schätzung auf SMP-Ebene mit dem Regressionskoeffizienten β_{SMP} aus dem zugehörigen SMP, also dem Gemeinde-

verband bzw. Gemeindeverbandsrest. Für den in Burgard et al. (2020) beschriebenen Fall der Verwendung des SMP-separaten GREG-Schätzers gilt dann:

$$\hat{\tau}_{\text{dir},g} = N_g \cdot \sum_{h=1}^H \gamma_h \cdot (\bar{y}_h + (\bar{X}_h - \bar{x}_h) \cdot \beta_{\text{SMP}})$$

für alle Gemeinden $g = 1, \dots, G$.

2. **Indirekter Schätzer:** $\hat{\tau}_{\text{ind},g}$ Der indirekte Schätzer wird berechnet unter der Verwendung der verallgemeinerten Kalibrierung nach Burgard et al. (2019):

$$\hat{\tau}_{\text{ind},g} = \sum_{k \in S_g} d_k \cdot g_k \cdot y_k,$$

wobei d_k das Designgewicht und y_k die Anzahl der Einwohner der Anschrift k aus der Stichprobe S_g ist. Des Weiteren ist g_k ($k \in S_g$) Teil der Lösung des folgenden Kalibrierungsproblems, welches *einmalig für jeden SMP* gebildet und gelöst wird:

$$\min_{(g_k \ (k \in S_{\text{SMP}}), \ \epsilon \in \mathbb{R}^G)} \sum_{k \in S_{\text{VBG}}} d_k D(g_k) + \sum_{g=1}^G \delta_g D(\epsilon_g) \quad (6)$$

$$\text{s.t.} \quad \sum_{k \in S_{\text{SMP}}} d_k g_k y_k = \hat{\tau}_{\text{SMP}} \quad (\text{exakte Kalibrierung gegen GREG-Schätzer des SMP})$$

$$\sum_{k \in S_g} d_k g_k y_k = \epsilon_g \tau_g \quad \forall g = 1, \dots, G \quad (\text{rel. Kalib. gegen Registerwert aller Gemeinden})$$

$$L_{g_k} \leq g_k \leq U_{g_k} \quad \forall k \in S_{\text{SMP}} \quad (\text{Box-Constraints der Kalibrierungsgewichte})$$

$$L_{\epsilon_g} \leq \epsilon_g \leq U_{\epsilon_g} \quad \forall g = 1, \dots, G \quad (\text{Box-Constraints der Abweichungen, abh. von } \lambda)$$

In dem Optimierungsproblem (6) sind folgende Funktionen und Parameter zu setzen:

- Exakte Kalibrierung gegen den schon berechneten Punktschätzer des SMPs. Dadurch wird die Kohärenz gesichert.
- Relaxierte Kalibrierung gegen die Registerwerte $\tau_1, \dots, \tau_G$ der Gemeinden. Dabei wird die erlaubte Toleranz (Abweichung von $\hat{\tau}_{\text{ind},g}$ vom Registerwert τ_g) für eine Gemeinde g geringer, je weniger Stichprobenelemente auf die Gemeinde entfallen.

Damit gilt für alle Gemeinden

$$g=1, \dots, G \quad g = 1, \dots, G$$

- $D(\cdot) : \mathbb{R} \rightarrow \mathbb{R}_+$ ist die entsprechende Penalty-Funktion (z. B. quadratisch beim GREG)
- $\delta_g \in \mathbb{R}_+$ ist ein Systemparameter, der durch die Optimierungsroutine der generalisierten Kalibrierung automatisiert festgelegt wird (inhaltlich unbedeutend)
- $L_{\epsilon_g}, U_{\epsilon_g} \in \mathbb{R}$ für alle $g = 1, \dots, G$ sind die Box-Constraints der erlaubten Abweichungen zwischen indirekter Schätzung der Einwohnerzahl und dem Registerwert, definiert als Quotient $\epsilon_g = \hat{\tau}_{\text{ind},g} / \tau_g$. Sie ist abhängig von der

Stichprobengröße n_g . Für die genaue Definition der Werte wird auf das Ende des Kapitels verwiesen.

Die kombinierte Schätzung (VBG:CAL) ergibt sich dann als Konvexkombination von direktem und indirektem Schätzer:

$$\hat{\tau}_g = \lambda_g \hat{\tau}_{\text{dir},g} + (1 - \lambda_g) \hat{\tau}_{\text{ind},g}$$

für alle $g = 1, \dots, G$, wobei die Werte $\lambda_1, \dots, \lambda_G$ in Abhängigkeit der Anzahl der Stichprobeneinheiten pro Gemeinde ($n_1, \dots, n_G$) eine streng monoton steigende Funktion bilden. Dabei gelten folgende Festlegungen:

- $\lambda_g = 0$ für Gemeinden mit $n_g < A$ (Ausweis des indirekten Schätzers $\hat{\tau}_{\text{ind},g}$)
- $\lambda_g = 1$ für Gemeinden mit $n_g > B$ (Ausweis des direkten Schätzers $\hat{\tau}_{\text{dir},g}$)
- $\lambda_g \in (0,1)$ für Gemeinden mit $n_g \in [A, B]$ (Ausweis einer Konvexkombination aus direktem und indirektem Schätzer) mit

$$\lambda_g = \frac{1}{B - A} \cdot n_g - \frac{A}{B - A}.$$

Somit ist λ_g eine Funktion in Abhängigkeit des Stichprobenumfangs n_g in Gemeinde g mit Funktionsvorschrift

$$\lambda_g(n_g) = \text{Proj}_{[0,1]} \left(\frac{1}{B - A} \cdot n_g - \frac{A}{B - A} \right).$$

Im Folgenden werden wir nochmals explizit auf die Wahl der erlaubten Abweichungen zwischen indirekter Schätzung der Einwohnerzahl und dem Registerwert, L_{ϵ_g} und U_{ϵ_g} , eingehen. Da die Festlegung dieser einen starken Einfluss auf die Größe der zulässigen Menge des Optimierungsproblems (6) hat und daher maßgeblich die Lösbarkeit beeinflusst, müssen diese Parameter gemeinde-spezifisch gesetzt werden. Dabei werden diese zuerst sehr strikt gewählt. Falls die Lösungsmenge von (6) dann leer sein sollte, werden sie iterativ gelockert, bis das Problem lösbar ist. Es wird folgendes iteratives Verfahren angewandt:

1. Setze $c = 10$
2. Berechne $\gamma_g = \frac{\lambda_g}{c}$; setze $L_{\epsilon_g} = 1 - \gamma_g$ und $U_{\epsilon_g} = 1 + \gamma_g$.
3. Falls Optimierungsproblem lösbar, STOP; sonst setze $c = c - 1$ und gehe zu 2.

Die Iteration kann so lange durchgeführt werden, bis $c = 1$. Sollte dann immer noch keine Lösbarkeit von (6) gewährleistet sein, kann auf $c = 1/2$, $c = 1/3$ usw. gesetzt werden. An der Definition der Box-Constraints L_{ϵ_g} und U_{ϵ_g} sowie der Parameter A und B ist zu erkennen, dass sie immer so gewählt sind, dass für Gemeinden mit weniger als A Stichprobeneinheiten der indirekte Schätzer $\hat{\tau}_{\text{ind},g}$ dem Registerwert τ_g entspricht. In diesem Fall ist $\lambda_g = 0$, also auch $\gamma_g = 0$ und damit $L_{\epsilon_g} = U_{\epsilon_g} = 1$, was eine exakte Kalibrierung gegen den Registerwert τ_g impliziert.

Abb. 4 Relative Verzerrung der Schätzung von Gemeinden in Rheinland-Pfalz

Am Beispiel einer Simulation zum Bundesland Rheinland-Pfalz, in welchem viele kleine Gemeinden in Verbandsgemeinden zu berücksichtigen sind, werden die Schätzergebnisse verdeutlicht. In den Abb. 4 und 5 sind relative Verzerrung und relativer Standardfehler der Gemeindeschätzungen insgesamt sowie nach Gemeindegroßen dargestellt. GEM:GREG bezeichnet dabei den Goldstandard, wenn man die Gemeinden selber in der Optimierung berücksichtigt hätte, was allerdings zu erheblich höheren Stichprobenumfängen geführt hätte. VBG:GREG führt zu einer Schätzung auf Gemeindeverbänden, welche anschließend ohne jede weitere Korrektur auf die Gemeinden runtergebrochen wird. Wie zu erwarten war, ist die Schätzqualität für Gemeinden unterhalb von 2000 Einwohnern erheblich schlechter als in den größeren Gemeinden. Im Gegensatz dazu führen die vier verschiedenen Versionen zu VBG:CAL mit unterschiedlichen Parametersätzen A/B zu einer Stabilisierung der Schätzqualität gerade bei kleinen Gemeinden.

Auffällig ist, dass für Gemeinden unterhalb 2000 Einwohnern bei den kalibrierten Verfahren (CAL) die Verzerrung das Gesamtbild dominiert. Dies liegt an einer resultierenden Gewichtung, die sich an den Registerwerten orientiert. Dies ist prinzipiell sinnvoll, da in sehr kleinen Gemeinden kaum Stichprobeninformationen zur Verfügung stehen. Ein Kompromiss zwischen Verzerrung und Varianz bevorzugt unter den vier Möglichkeiten das Modell 10/100. Bei den dargestellten Simulationen muss aber nochmals darauf hingewiesen sein, dass die verwendeten Karteileichen- und Fehlbestandsmodelle aus dem Forschungsprojekt zum Zensus 2011 stammen. Damit lassen sich zwar die Verfahren sehr gut veranschaulichen und auch Vergleiche zwischen den Verfahren machen, allerdings ist davon auszugehen, dass die Qualität kleiner Gemeinden zu schlecht eingeschätzt wird, während sehr große Gemeinden möglicherweise zu gut dargestellt werden. Dies resultiert daraus, dass die Modelle für Karteileichen und Fehlbestände ohne weitere Informationen auf sehr kleine Ge-

Abb. 5 Relativer Standardfehler der Schätzung von Gemeinden in Rheinland-Pfalz

meinden übertragen werden mussten, und in großen Gemeinden die dort im Zensus beobachteten Ausreißer kaum vorhanden sind.

Insgesamt lassen sich also bei der Anwendung des oben beschriebenen Verfahrens gemeindespezifische Schätzungen auch für kleine Gemeinden unter Berücksichtigung eines kohärenten Gesamt-Schätzsystems erzielen.

2.5.3 Ausblick auf den registerbasierten Zensus 2031

2031 soll ein registerbasierter Zensus durchgeführt werden. Ziel ist es, die notwendigen Daten, die aktuell aus den Melderegistern und der zusätzlichen Stichprobe stammen, dann vollständig aus Registern zu gewinnen. Somit entfällt auch die aktuell auf Basis der Stichprobe durchgeführte Korrektur der Bevölkerungszahl in Folge von Karteileichen und Fehlbeständen. Derartige Verfahren werden schon seit längerem in Skandinavien und den Niederlanden erfolgreich durchgeführt.

Grundsätzlich ist ein derartiges Verfahren sehr zu begrüßen. Zunächst erlaubt es eine schnellere und bei gut geführten Registern recht präzise Ermittlung der Zensus-Kennwerte. Darüber hinaus liefert eine solche Datenbasis eine wesentliche geeignetere Grundlage für alle Haushaltsstichproben, wie z. B. den Mikrozensus bzw. das neue System der Haushaltsstichproben, wodurch sicherlich erhebliche bessere Schätzergebnisse, insbesondere auch auf regionaler Ebene, ermöglicht werden.

Es gibt aber eine Reihe von Sonderproblemen zu beachten – ganz im Sinne von *there is no free lunch*. Das Zusammenführen von Registern erfordert, so es präzise sein soll, eine eindeutige Identifikationsnummer. Ansonsten müssen aufwändig Verfahren des Record Linkage (siehe bspw. Schnell 2019, Kapitel 12.6.5) angewendet werden, bei welchen auch die entstehenden Fehler genau analysiert werden müssen. Eine derartige Identifikationsnummer steht für derartige statistische Zwecke jedoch

noch nicht zur Verfügung. Gleichwohl erlaubt das Bundesstatistikgesetz von 2016 die Verwendung von Registern.

Ganz im Sinne des Kommentars von Helmut Küchenhoff wird aber auch hier dann von korrekten Registern ausgegangen. Spätestens seit der Anhörung zum Zensus 2011 vor dem Bundesverfassungsgericht darf jedoch die Fehlerfreiheit der Melderegister bezweifelt werden. Solange jedoch die Behörden, welche die Aufsicht über Register haben, auch direkt finanziell von einer zu hohen Zahl von gemeldeten Personen profitieren können, besteht die Möglichkeit eines *Principal-Agent-Problems*. Überdies darf nicht davon ausgegangen werden, dass Menschen fehlerfrei arbeiten sowie Computerhardware und Software ohne Fehler sind. Insofern bedarf es in jedem Falle eines Auditierungssystems, welches mögliche Registerfehler aufdeckt und dann auch korrigieren lässt. Idealerweise auch unter Verwendung eines zentral geführten Melderegisters.

Für alle weiteren Variablen, die nicht in den Melderegistern enthalten sind, müssen aber weitere Register herangezogen werden. Der Aufbau eines Bildungsregisters befindet sich in der Vorbereitung. Somit lassen sich neue Bildungsverläufe gut abbilden. Ein vollständiger fehlerfreier Aufbau vor dem Zensus 2031 scheint aber sehr ambitioniert. Die Bundesagentur für Arbeit stand im Frühjahr 2019 in der Kritik, die Arbeitslosen möglicherweise nicht korrekt zu zählen, was durch eine Untersuchung des Bundesrechnungshofes aufgeworfen und von der Presse medial umgesetzt wurde. Auch hier wurde von einer zu jedem Zeitpunkt korrekten Durchführung der Registrierung ausgegangen. Verfahrenstechnische Abweichungen, etwa durch die Geschwindigkeit des Registrierungsprozesses bei unterschiedlichen Gruppen wurden (möglicherweise) nicht berücksichtigt.

Wie bei allen Statistiken spielt auch bei einem rein auf Registern basierten Zensus die Qualität die entscheidende Rolle. Zwar entfallen die aktuell so diskutierten Stichprobenfehler, alle andere Merkmale einer Qualitätsmessung, wie sie auch im European Statistics Code of Practice formuliert werden, bleiben jedoch erhalten. In Deutschland bisher noch kaum berücksichtigt, muss unbedingt eine Methodenforschung zu Registern und deren Qualität bzw. Qualitätsmessung etabliert werden. Und noch viel wichtiger, ein Qualitätsaudit muss bei allen Registern durchgeführt werden, insbesondere wenn aus ihren Zahlen Statistiken abgeleitet werden, die möglicherweise zu Transferzahlungen herangezogen werden. Die Bedeutung einer hohen Qualität von Statistiken ist aber über die Betrachtung von Registern hinweg ganz essentiell für eine evidenzbasierte Politik sowie deren Nachvollziehbarkeit und damit auch für eine Demokratie.

2.6 Zusammenfassung und Ausblick

Ursprung des Artikels Burgard et al. (2020) war es, eine Vorgehensweise der Schätzung und Qualitätsmessung zu entwickeln, welche die sehr heterogene Größe der Gemeinden in Deutschland geeignet berücksichtigt. Ziel dieser sehr ausführlichen Darstellung war es schließlich, eine eingehende wissenschaftliche Diskussion zu motivieren, um für den Zensus 2021 eine geeignete Grundlage zu schaffen. Neben Diskussionen auf Tagungen und Workshops, welche implizit Berücksichtigung fanden, trugen vier Kommentare hier zu einer vertieften Diskussion bei. Diese um-

fassten das gesamte Repertoire der Methodik inklusive Nichtstichprobenfehlern. Da der ursprüngliche Artikel früh zur Diskussion bereitgestellt werden musste und die Forschungsarbeiten noch andauerten, haben wir letzte Entwicklungen und Ergebnisse noch in unsere Antwort zu den Kommentaren einarbeiten können.

Der Zensus 2021 wird sicher letztmalig ein registergestützter Zensus mit Zusatzstichprobe sein. Die vom Bundesverfassungsgericht angemahnten Methodenentwicklungen auf Grund von neuen Erkenntnissen sind mit den vorgeschlagenen Methoden sicher umfassend umgesetzt. Detailfragen werden sich möglicherweise mit den Ergebnissen des nächsten Zensus ergeben. Es bleibt aber in jedem Fall mehr als wünschenswert, die Hilfsmerkmale und Paradata umfassend auswerten zu können. Sicher lassen sich daraus viele Kenntnisse, auch für den reinen Registerzensus sowie auch für das neue Haushaltsstichprobensystem gewinnen. Insofern erachten wir den Mehrwert für Forschung und amtliche Praxis für besonders hoch.

Grundsätzlich ist ein Registerzensus die Zukunft, sofern man Qualität und Qualitätsmessung geeignet in und für Register umsetzt. Hierzu ist noch einiges an Forschung nötig. Sobald die Methodik vollständig und mit hoher Qualität umgesetzt ist, können stets aktuelle Zahlen, etwa auch für Quartale und nicht nur jede 10 Jahre abgeleitet werden. Für eine wissensbasierte Gesellschaft ist das eine essentielle Grundlage.

Danksagung Wir danken explizit den Kolleginnen und Kollegen in zahlreichen Diskussionen zu unseren Vorträgen sowie insbesondere Sara Bleninger, Michael Fürnrohr, Hans Kiesl, Walter Krämer und Helmut Küchenhoff, die in vier umfangreichen Kommentaren maßgeblich zur Erstellung des diskutierten Artikels beigetragen haben. Die Forschungsarbeiten hierzu wurden im Rahmen des Forschungsprojektes *Research Innovation for Official and Survey Statistics (RIFOSS)* durchgeführt, das vom Statistischen Bundesamt gefördert wird. Des weiteren danken wir dem DFG Graduiertenkolleg *Algorithmische Optimierung (ALOP – RTG 2126)*.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

- Broscheid A, Geschwend T (2003) Augäpfel, Murmeltiere und Bayes. Zur Auswertung stochastischer Daten aus Vollerhebungen. MPiFG Working Paper 03/7
- Bundesrepublik Deutschland (1949) Grundgesetz für die Bundesrepublik Deutschland vom 23. Mai 1949. Zuletzt geändert durch Gesetz vom 13. Juli 2017
- Burgard JP, Münnich R, Rupp M (2019) A generalized calibration approach ensuring coherent estimates with small area constraints. Res Pap Econ 10/19: https://www.uni-trier.de/fileadmin/fb4/prof/VWL/EWF/Research_Papers/2019-10.pdf
- Burgard JP, Münnich R, Rupp M (2020) Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021. AStA Wirtsch Sozialstat Arch. <https://doi.org/10.1007/s11943-019-00256-6>
- BVerfG (2018). Urteil des Zweiten Senats vom 19. September 2018 – 2 BvF 1/15 – Rn. (1-357). ECLI:DE:BVerfG:2018:fs20180919.2bvf000115
- Christensen B, Christensen S, Hoppe T, Spandel M (2015) Everything counts! AStA Wirtsch Sozialstat Arch 9(3-4):215–232
- Deutscher Städtetag (2019) Anhörung zur Durchführung des Zensus im Jahr 2021. Stellungnahme des Deutschen Städtetages und Deutschen Städte- und Gemeindebundes zum Gesetzentwurf der Bundesregierung zur Durchführung des Zensus im Jahr 2021. Berlin 02.05.2019

- Eckman S, Kreuter F (2011) Confirmation bias in housing unit listing. *Public Opin Q* 75:1–12
- Eckman S, de Leeuw E (2017) Editorial – special issue on total survey error (TSE). *J Off Stat* 33(2):301. <https://doi.org/10.1515/JOS-2017-0015>
- European Statistical System Committee (2015) Draft regulation on statistics relating to persons and households. EEA ESSC 2015/27/7/EN
- Fuchs M (2009) (Optimal) governance of research support by survey methodology. German Council for Social and Economic Data (RatSWD). Working paper Nr. 98
- Göbl T (2018) Der Zensus vor dem Bundesverfassungsgericht. *Bayern Zahlen* 10:685–696
- Groves RM, Lyberg L (2010) Total survey error: past, present, and future. *Public Opin Q* 74(5, SI):849–879
- Kauermann G, Küchenhoff H (2010) Stichproben: Methoden und praktische Umsetzung mit R. Springer, Heidelberg, Berlin, New York
- Klink S, Bihler W (2015) Die Wiederholungsbefragung beim Zensus 2011. *ASTa Wirtsch Sozialstat Arch* 9:191–214
- Krug W, Nourney M, Schmidt J (2001) Wirtschafts- und Sozialstatistik. Gewinnung von Daten, 6. Aufl. Oldenbourg, München
- Küchenhoff H (2014) Gutachten zur Verwaltungsrechtssache Stadt Bremerhaven gegen die Freie Hansestadt Bremen. https://www.stablab.stat.uni-muenchen.de/projekte/stichprobenziehung/files/gutachten_zensus2011.pdf. Zugegriffen: 7. Mai 2019
- Münnich R, Gabler S, Bruch C, Burgard JP, Enderle T, Kolb J-P, Zimmermann T (2015) Tabellenauswertungen im Zensus unter Berücksichtigung fehlender Werte. *ASTa Wirtsch Sozialstat Arch* 3(4):269–304
- Münnich R, Gabler S, Ganninger M, Burgard JP, Kolb J-P (2012) Stichprobenoptimierung und Schätzung im Zensus 2011. *Statistik und Wissenschaft* Bd. 21
- Münnich R, Gabler S, Ganninger M, Burgard JP, Kolb J-P (2013) Validierungsprojekt zum deutschen Zensus 2011 – Abschlussbericht. https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsatz_Archiv/2013_03_Validierungsprojekt_zum_deutschen_Zensus_2011_Abschlussbericht.html?sessionid=0A66526E8368BE0244D0F440C24C05AA.2_cid389?nn=3065474#=#1. Zugegriffen: 19. Okt. 2019
- Rendtel U (2015) Warum im Zensus die Ergebnisse der Stichprobenmethode keine Benachteiligung der großen Gemeinden darstellen. Eine Detektivarbeit. *ASTa Wirtsch Sozialstat Arch* 9(3-4):233–249
- Schnell R (2019) Survey-Interviews. Methoden standardisierter Befragungen, 2. Aufl. VS, Wiesbaden
- Statistische Ämter des Bundes und der Länder (2011) Das registergestützte Verfahren beim Zensus 2011. Destatis, Wiesbaden
- StBA (2018) Simulationsrechnungen der Stichprobe der Personenerhebung. Daten zur internen Verwendung
- Verma V (1991) Sampling methods training handbook. Statistical Institute for Asia and the Pacific (SIAP), Tokyo (Handbook, Tokyo: Statistical Institute for Asia and the Pacific (SIAP))
- ZensusG2011 (2009) Gesetz über den registergestützten Zensus im Jahre 2011 vom 8. Juli 2009 (BGBl. I S. 1781). http://www.gesetze-im-internet.de/zensg_2011/. Zugegriffen: 19. Sept. 2019
- Zensusvorbereitungsgesetz 2021 (2017) Zensusvorbereitungsgesetz 2021 vom 3. März 2017 (BGBl. I S. 388), geändert durch Artikel 1 des Gesetzes vom 27. November 2018 (BGBl. I S. 2010). http://www.gesetze-im-internet.de/zensusvorb_2021/. Zugegriffen: 19. Sept. 2019

Hinweis des Verlags Der Verlag bleibt in Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutsadressen neutral.