Würz, Nora

Article

Schätzung regionaler Einkommensindikatoren unter Transformationen in Abwesenheit von Populations-Mikrodaten

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Würz, Nora (2024) : Schätzung regionaler Einkommensindikatoren unter Transformationen in Abwesenheit von Populations-Mikrodaten, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 2, pp. 107-116

This Version is available at:
https://hdl.handle.net/10419/294180

SCHÄTZUNG REGIONALER EINKOMMENSINDIKATOREN UNTER TRANSFORMATIONEN IN ABWESENHEIT VON POPULATIONS-MIKRODATEN

Nora Würz

Schlüsselwörter: Zensus – Kerndichteschätzung – amtliche Statistik – Unit-Level-Modelle – Small-Area-Schätzung

ZUSAMMENFASSUNG

Für Deutschland und andere entwickelte Länder werden Methoden zur Schätzung von sozioökonomischen Indikatoren auf räumlich disaggregierter Ebene benötigt, ohne dabei Populations-Mikrodaten zu verwenden, die meist nicht öffentlich verfügbar sind. Viele sozioökonomische Indikatoren, zum Beispiel Einkommen, sind schief verteilt, weswegen zur Erfüllung der Annahmen der Modelle (datengetriebene) Transformationen der abhängigen Variablen verwendet werden. Hierfür werden Verzerrungs-Korrekturen für die Small-Area-Vorhersagen benötigt. Die vorgestellte Methodik zur Verzerrungs-Korrektur basiert auf einer Kerndichte-Schätzung. Sie wird auf Daten des Sozio-oekonomischen Panels 2011 angewendet, um das durchschnittliche Bruttoeinkommen für 96 deutsche Raumordnungsregionen zu schätzen.

Keywords: *census – kernel density estimation – official statistics – unit-level models – small area estimation*

ABSTRACT

For Germany and other developed countries, methods are needed to estimate socio-economic indicators at a spatially disaggregated level without using population micro-data, which are usually not publicly available. Many socio-economic indicators, such as income, are skewed, and therefore (data-driven) transformations of the dependent variable are used to satisfy the model assumptions. This requires bias corrections for the small area predictions. The bias correction methodology presented in this article is based on kernel density estimation. It is applied to data from the Socio-Economic Panel 2011 to estimate the average gross income for 96 German spatial planning regions.

Dr. Nora Würz

ist akademische Rätin an der Otto-Friedrich-Universität Bamberg. Sie forscht zu Small-Area-Verfahren mit folgenden Schwerpunkten: Transformationen der abhängigen Variablen, Verwendung im Kontext von georäumlichen Daten, Poverty Mapping und Machine-Learning-Methoden. Für ihre Dissertation „Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications“ wurde sie mit dem wissenschaftlichen Nachwuchspreis „Statistical Science for the Society“ 2023 des Statistischen Bundesamtes ausgezeichnet.

1

Einleitung

Für eine evidenzbasierte Entscheidungsfindung sind zuverlässige Informationen über sozioökonomische Indikatoren unerlässlich. Stichprobenerhebungen ermöglichen eine kosteneffiziente Erhebung von Indikatoren und haben eine lange Tradition. Dabei sind neben der quantitativen Erfassung dieser Indikatoren für die Gesamtpopulation insbesondere auch die für Teilpopulationen (geografische Gebiete oder soziodemografische Gruppen) bedeutsam. Um Einblicke in diese Teilpopulationen zu gewinnen, können disaggregierte direkte Schätzer verwendet werden, die ausschließlich auf Umfragedaten des jeweiligen Gebiets berechnet werden. In der Small-Area-Forschung gilt ein Gebiet als „large", wenn die Stichprobe groß genug ist, um zuverlässige direkte Schätzungen für dieses Gebiet zu ermöglichen. Wenn die direkten Schätzungen nicht ausreichend genau sind oder in diesem Gebiet keine Einheit erhoben wurde, wird das Gebiet als „small" bezeichnet. Dies tritt besonders häufig bei hoher räumlicher oder soziodemografischer Auflösung auf. Small-Area-Schätzung (small area estimation – SAE) soll dieses Problem überwinden, ohne dass größere und damit teurere Umfragen erforderlich sind (Pfeffermann, 2013; Rao/Molina, 2015; Tzavidis und andere, 2018). SAE-Techniken nutzen die Informationen von allen Gebieten gleichzeitig mithilfe eines statistischen Modells, um dadurch die Schätzungen für wiederum alle Gebiete zu verbessern. Dabei werden die Umfragedaten mit weiteren Hilfsdaten über ein Modell verknüpft und regionenspezifische Strukturen ausgenutzt. Geeignete Hilfsdaten sind Verwaltungs- und Registerdaten sowie der Zensus. In vielen Ländern sind solche Daten durch Vertraulichkeitsvereinbarungen streng geschützt und der Zugang zu Individualdaten (Mikrodaten) ist selbst innerhalb der statistischen Ämter eine Herausforderung. Daher haben Anwendende ein großes Interesse an Small-Area-Schätzern, die keine Hilfsdaten auf Mikrodaten-Ebene benötigen, sondern mit deutlich einfacher zugänglichen Aggregaten aus diesen Mikrodaten auskommen. Würz und andere (2022) stellen eine neue Methode in Abwesenheit von Populations-Mikrodaten vor. Diese Methodik wird mittels einer Anwendung auf das Sozio-oekonomische Panel (SOEP) 2011 (Sozio-oekonomisches Panel, 2019) und Hilfsinformationen aus dem Zensus 2011 zur Schätzung des mittleren Bruttoeinkommens für die 96 regionalen Raumordnungsregionen in Deutschland demonstriert.

2

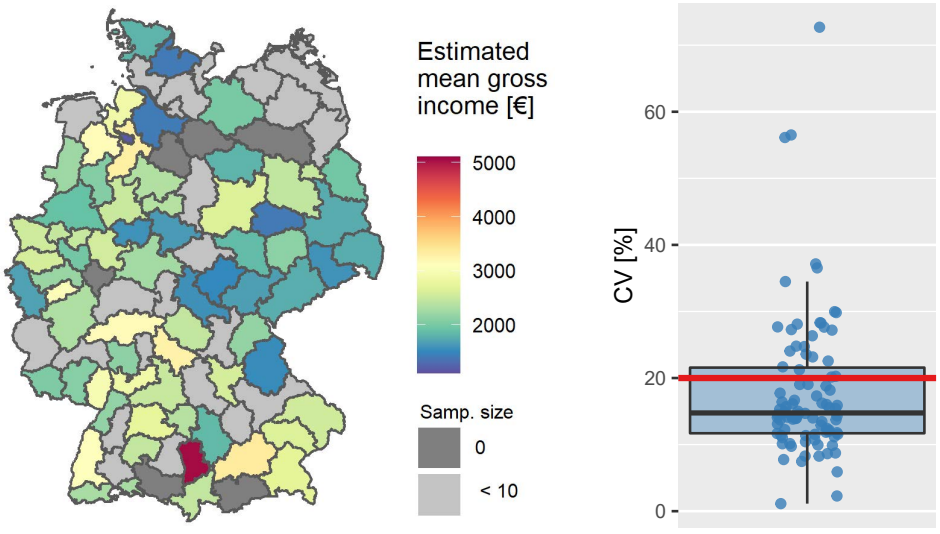
Räumliche Schätzung von Einkommen aus dem SOEP

Für die Einkommensschätzung in Deutschland wird hier das Sozio-oekonomische Panel 2011 verwendet. Diese repräsentative Langzeitstudie wird seit 1984 durchgeführt und ist am Deutschen Institut für Wirtschaftsforschung (DIW Berlin) angesiedelt. Das SOEP liefert Längsschnittdaten privater Haushalte in Deutschland für multidisziplinäre Themen und ist somit für Regierungsinstitutionen und Forschende aus verschiedenen Bereichen von großem Wert. Im SOEP wird im Gegensatz zu anderen wichtigen deutschen Umfragen (beispielsweise dem Mikrozensus) das individuelle Einkommen abgefragt.

In dieser Arbeit wird die Refreshment-Stichprobe aus dem Jahr 2011 für die Schätzungen verwendet, da in diesem Jahr auch der Zensus durchgeführt wurde. Die Zielvariable ist das individuelle Bruttoeinkommen in Euro im Monat vor dem Interview innerhalb des Jahres 2011. Die Zielbevölkerung ist die erwerbsfähige Bevölkerung im Alter von 15 bis 64 Jahren. Die Stichprobengrößen über den regionalen Raumordnungsregionen variieren von 0 bis 107 (1. Quartil: 8, Median: 16,5, Mittelwert: 20,83 und 3. Quartil: 26,5), wobei in 6 regionalen Raumordnungsregionen keine Daten erhoben worden sind. Zusätzlich dürfen für 23 regionale Raumordnungsregionen keine direkten Schätzergebnisse aufgrund von Vertraulichkeitsabkommen ausgegeben werden, da die Stichprobengrößen für diese Regionen geringer als 10 sind. Die direkten Schätzer werden mit dem Software-Paket emdi (Kreutzmann und andere, 2019) geschätzt und die Varianzen mittels einer kalibrierten Bootstrapmethode (Alfons/Templ, 2013) bestimmt. ↘ Grafik 1 stellt die direkten Schätzer in einer Karte dar. Das durchschnittliche Bruttoeinkommen[1] variiert von 1 173 Euro in Bremen bis 5 059 Euro in der Planungsregion Donau-Iller. Generelle Trends lassen sich auf der Karte erken-

1 Es handelt sich hier um das arithmetische Mittel.

Grafik 1
Karte mit direktem geschätzten mittleren Bruttoeinkommen je Monat (in Euro) für die regionalen Planungsregionen in Deutschland (Stichprobengrößen unter 10 sind ausgegraut) und ihre dazugehörigen Variationskoeffizienten (CV)



nen: Der Osten weist ein niedrigeres durchschnittliches Bruttoeinkommen auf, während die Regionen um München, Stuttgart oder Frankfurt höhere Durchschnittseinkommen haben. Kleine Stichprobengrößen führen jedoch zu geschätzten durchschnittlichen Bruttoeinkommen mit hohen Varianzen. Diese hohen Varianzen spiegeln sich auch in hohen Variationskoeffizienten (CVs) für die Regionen wider, wobei 26 von 90 Regionen die Grenze von 20 % für verlässliche Variationskoeffizienten (Eurostat, 2013) überschreiten. Die kleinen Stichprobengrößen (zum Teil ohne erhobene Einheit) und die hohe Variabilität der gemeldeten individuellen Einkommen machen die Verwendung von Small-Area-Methoden erforderlich. Da einige der Hilfsvariablen des Sozio-oekonomischen Panels mit Variablen im deutschen Zensus übereinstimmen, können diese Zensus-Kovariaten-Daten als Hilfsinformationen in die Small-Area-Modelle einfließen. Informationen aus dem deutschen Zensus sind jedoch nur als Aggregate auf der Ebene der regionalen Raumordnungsregionen für Forschende verfügbar.

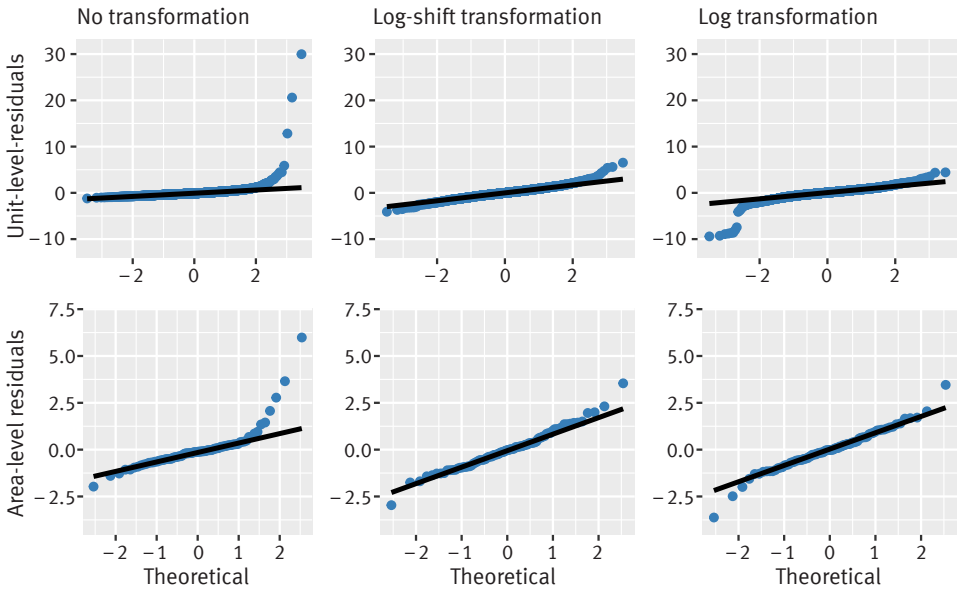
3

Hilfsinformationen aus dem Zensus und vorläufige Modellauswahl

SAE-Methoden verwenden Umfragedaten und populationsbezogene Hilfsinformationen, um die verfügbaren direkten Schätzungen zu verbessern. Insbesondere für kleine Stichprobengrößen, wie in der SOEP-Refreshment-Stichprobe von 2011, sind diese Methoden sehr hilfreich, um zuverlässiger zu schätzen. Wie in vielen Ländern ist der deutsche Zensus nicht auf Mikro-Ebene verfügbar, sodass nur aggregierte Hilfsinformationen (Mittelwerte und Kovarianzen) aus dem Zensus 2011 verwendet werden können, um das durchschnittliche Bruttoeinkommen zu schätzen.

Da mehrere sozioökonomisch relevante Variablen wie Einkommen eine schiefe Verteilung haben, ist die Log-Transformation eine bewährte Methode, um die Modell-Annahmen von Small-Area-Modellen zu erfüllen (Berg/Chandra, 2014; Molina/Martín, 2018). Mit den Stichprobendaten wurden drei verschiedene linear gemischte Modelle geschätzt, die sich in der angewandten Transformation auf die abhängige Variable unterscheiden:

Grafik 2
QQ-Plots für die Residuen auf Individualebene (1. Reihe) und die regionenspezifischen Residuen (2. Reihe) unter den Modellen mit verschiedenen Transformationen



› ohne Transformation,

› mit Log-Shift-Transformation,

› mit Log-Transformation.

Die datengetriebene Log-Shift-Transformation passt sich an die Daten an, indem die Logarithmus-Funktion um einen zusätzlichen Parameter (λ) erweitert und dadurch flexibler wird: log $(y + \lambda)$. Die Validität der Normalitätsannahmen für die Fehlerterme der zugrunde liegenden Modelle wird mit QQ-Plots überprüft. Basierend auf dieser Untersuchung wird eine Log-Shift-Transformation verwendet, um das durchschnittliche Bruttoeinkommen für deutsche regionale Raumordnungsregionen zu schätzen, da die QQ-Plots hier am wenigsten von den Normalverteilungsannahmen abweichen. ↘ Grafik 2

4

Kurzüberblick der entwickelten Methodik

Die entwickelte Methodik baut auf dem viel genutzten Nested-Error-Regressionsmodell (NER-Modell) von Battese und anderen (1988) auf. Mittels der Stichprobendaten (Zielvariable und Kovariat-Daten) wird das Modell gefittet und die Populationsdaten (nur Kovariat-Daten) werden für die Prädiktion verwendet.

Die Population U von Länge N ist unterteilt in D Regionen $U_1, U_2, ..., U_D$ bestehend aus $N_1, N_2, ..., N_D$ Individuen. Der Index $i = 1, ..., D$ wird verwendet, um auf die jeweilige Region und der Index $j = 1, ..., N_i$, um auf die jeweiligen Individuen zu indizieren. Die Stichprobe s besteht aus n Individuen mit regionenspezifischen Stichprobengrößen $n_1, n_2, ..., n_D$. Mit s_i werden die Individuen bezeichnet, welche in Region i innerhalb der Stichprobe sind. $\bar{s}_i$ beschreibt für Region i die Individuen außerhalb der Stichprobe. Für alle Individuen in der Stichprobe wird die stetige Zielvariable y_{ij} beobachtet. Die Hilfsinformationen aus der Stichprobe für Individuum j in Region i sind als Vektor $x_{ij} = (1, x_{1ij}, x_{2ij}, ..., x_{pij})^T$ gegeben, der

neben dem Intercept p erklärende Variablen enthält. Das Modell von Battese und anderen (1988) modelliert den linearen Zusammenhang zwischen den Kovariaten x_{ij} und der Zielvariable y_{ij} folgendermaßen:

$$y_{ij} = x_{ij}^T\beta + u_{ij} + e_{ij},$$

$$u_i \sim N(0, \sigma_u^2) \text{ und } e_{ij} \sim N(0, \sigma_e^2),$$

wobei $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ der Vektor mit den Regressionskoeffizienten, e_{ij} die individuellen Fehler und u_i die regionenspezifischen Fehler sind. Dabei wird angenommen, dass die Fehlerterme unabhängig verteilt sind. Der beste linear verzerrte Schätzer für jedes Individuum außerhalb der Stichprobe $j \in \bar{s}_i$ ist gegeben durch:

$$\mu_{ij} = x_{ij}^T\beta + u_i = x_{ij}^T + \gamma_i\left(\sum_{j\epsilon s_i} y_{ij} - x_{ij}^T\beta\right),$$

wobei $\gamma_i = \dfrac{\sigma_u^2}{\sigma_u^2 + {\sigma_e^2}/{n_i}}$ der Gewichtungsfaktor ist. Daraus lässt sich dann die empirisch beste lineare Schätzung für den Populations-Mittelwert $\bar{y}_i$ jeder Region i ermitteln:

$$\widehat{\bar{Y}}_i^{BHF} = \frac{1}{N_i}\left(\sum_{j\,\epsilon s_i} y_{ij} + \sum_{j\,\epsilon \bar{s}_i} \hat{\mu}_{ij}\right)$$

$$= \hat{\gamma}_i\left(\frac{1}{n_i}\sum_{j\,\epsilon s_i} y_{ij} + \left(\bar{x}_i - \frac{1}{n_i}\sum_{j\,\epsilon s_i} x_{ij}\right)^T\hat{\beta}\right) + (1-\hat{\gamma}_i)\,\bar{x}_i^T\hat{\beta},$$

mit $\hat{\gamma}_i = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + {\widehat{\sigma_e^2}}/{n_i}}$. Der Vektor $\bar{x}_i^T = \dfrac{1}{N_i}\sum_{j\in U_i} x_{ij}^T$ enthält die Mittelwerte in der Population zu den p erklärenden Variablen für die jeweilige Region i. Um die beiden Varianzkomponenten (σ_u^2, σ_e^2) zu schätzen, können unter anderem Restricted-Maximum-Likelihood-Methoden verwendet werden (Rao/Molina, 2015). Aus obiger Formel ist direkt ersichtlich, dass es für die Populationsdaten ausreichend ist, wenn diese nur als regionenspezifische Mittelwerte zu allen Kovariaten vorliegen.

Das NER-Modell kann in der besonders häufig vorliegenden Datensituation von Umfragedaten auf Individuen-Ebene bei gleichzeitig limitiertem Zugang zu Hilfsdaten (zum Beispiel aggregierte Daten wie Mittelwerte) angewandt werden. Es kann jedoch keine Transformationen der abhängigen Variablen berücksichtigen. Hierzu sind Erweiterungen notwendig (Berg/Chandra, 2014; Molina/Martín, 2018; Rojas-Perilla und andere, 2020). Die zitierten Publikationen haben gemeinsam, dass zur Small-Area-Schätzung Populations-Mikrodaten benötigt werden. Dies hängt mit der Rücktransformation des synthetischen Parts zusammen, welche nur unter der Verwendung von Populations-Mikrodaten nicht verzerrt ist. Wird eine konvexe Rücktransformation, wie die Exponential-Funktion (verwendet für die Rücktransformation von zuvor Log-transformierten Einkommensverteilungen) angewandt, so führt eine naive Rücktransformation ($\mu_{ij}^{\text{trans,naive}}$) zu einer Unterschätzung von μ_{ij} auf rücktransformierter Ebene. Grund hierfür ist die Jensensche Ungleichung:

$$\mu_{ij}^{\text{trans,naive}} = \exp\left(x_{ij}^T\beta + u_i\right) < E\left[\exp\left(y_{ij}^*\right)|y_s, X_s\right],$$

wobei $(y_{ij}^*$ die log-transformierte Zielvariable, y_s die untransformierte Zielvariable und X_s die Kovariat-Daten-Matrix aus der Stichprobe ist. Aufgrund der Eigenschaften der Log-Transformation lässt sich für die Log- und Log-Shift-Transformation diese Verzerrung $\left(\alpha_i = \dfrac{\sigma_u^2(1-\gamma_i) + \sigma_e^2}{2}\right)$ analytisch bestimmen und ausgleichen, wenn Populations-Mikrodaten vorliegen:

$$\mu_{ij}^{\text{trans}} = \exp\left(x_{ij}^T\beta + u_i + \alpha_i\right)$$

In entwickelten Ländern wie Deutschland sind diese Mikrodaten nicht leicht zugänglich. Daher wird dringend Methodik benötigt, die ohne Populations-Mikrodaten auskommt und gleichzeitig Transformationen einbeziehen kann. Für die Log- und die Log-Shift-Transformation kann nun das Problem darauf reduziert werden, den mittleren rücktransformierten synthetischen Part $\left(\sum_{j\,\epsilon\bar{s}_i} exp\left(x_{ij}^T\hat{\beta}\right)\right)$ nur mit aggregierten Kovariat-Daten zur Population zu schätzen $\left(\exp\left(\bar{x}_i^T\hat{\beta}\right)\right)$. Auch hier führt die Jensensche Ungleichung zu einer Unterschätzung für die naive Rücktransformation des geschätzten synthetischen Parts:

$$\exp\left(\bar{x}_i^T\hat{\beta}\right) < \sum_{j\,\epsilon\bar{s}_i} exp\left(x_{ij}^T\hat{\beta}\right)$$

Die in Würz und andere (2022) vorgeschlagene Methodik korrigiert diese Verzerrung unter ausschließlicher Verwendung von Populations-Aggregaten (Mittelwerte und Kovarianzen). Um die Verteilung des synthetischen Teils zu schätzen, wird die Kerndichteschätzung vorge-

schlagen. Dieses Vorgehen bietet zwei wesentliche Vorteile:

- Der synthetische Part ist eine univariate Größe, sodass eine Kerndichteschätzung auch unter der Berücksichtigung von sehr vielen Kovariaten (metrisch und kategorial) nicht an die Grenzen der technischen Umsetzbarkeit stößt.
- Es sind keine parametrischen Annahmen an die Kovariaten erforderlich und es werden ausschließlich aggregierte Hilfsinformationen zur Population benötigt.

Die Kerndichteschätzung wird angewandt auf die Kovariat-Daten aus der Stichprobe, welche zuvor mittels der Populations-Aggregate adjustiert wurden. In einem ersten Schritt werden dazu die Kovariat-Daten aus der Stichprobe für jede Region standardisiert. Im Anschluss werden diese an die regionenspezifischen Populations-Aggregate angepasst. Ist die Stichprobengröße für Region i sehr klein, so werden als Input für die Adjustierung die standardisierten Kovariat-Daten von allen Regionen verwendet und an die für Region i spezifischen Populations-Aggregate angepasst. Ist die Stichprobengröße von i hingegen groß, so werden nur die standardisierten Stichprobendaten von Region i mittels ihrer Populations-Aggregate adjustiert. Durch dieses Vorgehen werden für alle Regionen Verteilungen erzeugt, die mit den Eigenschaften der Population (in Bezug auf den bekannten Wert des Populations-Aggregats) übereinstimmen. Aus den Verteilungen wird in einem letzten Schritt der Erwartungswert für den rücktransformierten synthetischen Part mittels numerischer Integration bestimmt. Die Schätzung kann dann verwendet werden, um die regionenspezifischen SAE-Mittelwerte zu erhalten. Diese sind nun so korrigiert, dass ausschließlich Populations-Aggregate verwendet wurden.

Die Schätzung der Unsicherheit für die zugehörigen Punktschätzer ist von großer Bedeutung. In der Publikation wird ein parametrisches Bootstrapverfahren vorgeschlagen, welches den Ideen von González-Manteiga und anderen (2008) folgt.

5

Validierung der Methodik

Durch modell- und designbasierte Simulationen ist es möglich, eine neue Methodik zu testen und Vorteile sowie Schwächen zu ermitteln. In solchen Settings sind die wahren Werte bekannt, sodass Qualitätskriterien für Punkt- und Unsicherheitsschätzung ermittelt werden können.

Der vorgeschlagene Schätzer sowie seine Unsicherheitsschätzung wurden in verschiedenen modellbasierten Settings getestet: in einem idealen Normalitäts-Setting, in zwei Log-Settings und in einem für Einkommen mit Bezug auf die Verteilung realistischen GB2[2]-Setting. Die vorgeschlagene Methodik erzielt in diesen Settings vergleichbar gute Ergebnisse wie die EBP[3]-Methode. Diese Methode kann als Gold-Standard angesehen werden, obwohl sie Populations-Mikrodaten verwendet. Andere Methoden, welche nur aggregierte Daten verwenden, konnten in den Settings, in denen Transformationen nötig sind, vom vorgeschlagenen Schätzer bezüglich der Gütemaße übertroffen werden.

Zusätzlich zu den modellbasierten Simulationen lief eine designbasierte Simulation mit realen Daten aus Mexiko, um den Nutzen der vorgeschlagenen Methodik auf Basis von realen Daten zu bestätigen. Wiederholt zeigte sich, dass die Qualität der vorgeschlagenen Methodik ähnlich hoch ist wie die Qualität der EBP-Schätzungen unter Verwendung von Populations-Mikrodaten.

6

Ergebnisse und Simulationen mit der vorgeschlagenen Methodik

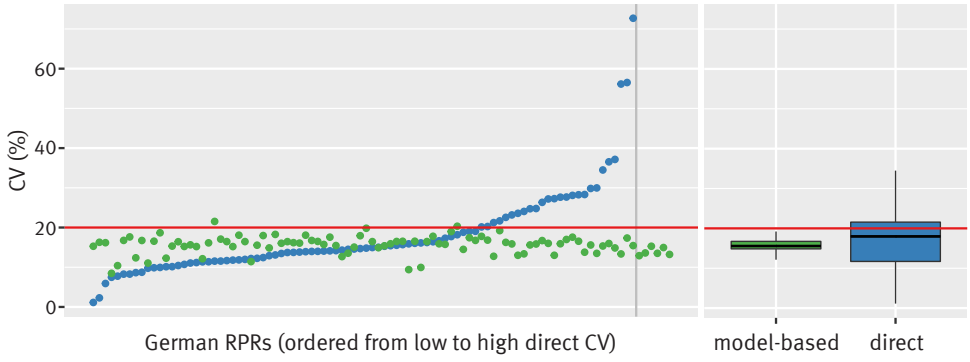
Die vorliegenden SOEP-Refreshment-Daten benötigen die vorgestellte Methodik, da einerseits nur Zensus-Aggregate als Kovariate zur Verfügung stehen und andererseits gleichzeitig die Anwendung einer Transformation notwendig ist (siehe Grafik 2). Im Vergleich zur direkten Schätzung (Design wurde berücksichtigt; Varianz-

2 Verallgemeinerte Beta-Verteilung zweiter Art.

3 „Empirical best prediction“ nach Molina/Rao (2010).

Grafik 3

Regionenspezifische Variationskoeffizienten für die direkten (blauen) und die vorgeschlagenen modellbasierten (grünen) Schätzungen, geordnet von niedrigen zu hohen Variationskoeffizienten der direkten Schätzungen, sowie die zugehörigen Boxplots



Die graue Linie im linken Diagramm trennt die Regionen, die innerhalb der Stichprobe sind, von den nicht eingeschlossenen Regionen.
Die rote Linie markiert den 20-%-Schwellenwert zur Definition verlässlicher Schätzungen.

schätzung mittels kalibrierten Bootstrap-Verfahrens) wurde gezeigt, dass die Unsicherheit der Punktschätzer reduziert werden konnte. Dazu werden in ↘ Grafik 3 die Varianzkoeffizienten der direkten Schätzung mit denen der vorgeschlagenen modellbasierten Methodik verglichen: 26 der direkten Schätzungen überschreiten den 20-%-Schwellenwert und für 6 Regionen sind keine Daten vorhanden. Im Gegensatz dazu überschreiten nur 2 der auf dem vorgeschlagenen modellbasierten Schätzer basierenden Varianzkoeffizienten den 20-%-Schwellenwert. Die Varianzkoeffizienten der modellbasierten Schätzungen sind im Durchschnitt kleiner als die der direkten Schätzungen und weisen eine geringere Streuung auf. Besonders in Gebieten mit unzuverlässigen direkten Schätzungen aufgrund kleiner Stichproben sind die modellbasierten Schätzungen genauer.

7

Bereitstellung der Methodik als R-Paket

Um weitere Anwendungen zu ermöglichen, wird diese neue Methodik im R-Paket saeTrafo (R Core Team, 2022; Würz, 2022) zur Verfügung gestellt. Die Funktionen des Pakets werden anhand öffentlich verfügbarer Einkommensdaten illustriert. Um die Benutzungsfreundlichkeit des Pakets zu erhöhen, werden weitere etablierte SAE-Modelle für Stichprobendaten auf Individualebene mit Transformation angeboten. Auch Unsicherheitsschätzer sind direkt verfügbar. Auf Basis der eingegebenen Datenstruktur wird zudem die geeignetste Methode automatisiert ausgewählt.

8

Fazit

Die vorgestellte Arbeit untersucht die Schätzung von SAE-Mittelwerten unter Transformationen in Abwesenheit von Populations-Mikrodaten. Viele relevante sozioökonomische Indikatoren sind schief verteilt, sodass Transformationen es ermöglichen, nötige Modellannahmen zu erfüllen. Die verwendete Literatur zu Transformationen für das Modell von Battese und anderen (1988) geht davon aus, dass Populations-Mikrodaten als Hilfsinformationen zur Verfügung stehen (Karlberg, 2000; Chandra/Chambers, 2011; Molina/Martín, 2018), was eine starke Einschränkung in der Anwendung darstellt. Meist ist es nicht möglich, solche Mikrodaten zu erhalten. Aus diesem Grund ist die Kombination aus Transformation der abhängigen Variablen, wenn nur Populations-Aggregate vorliegen, für Anwendende besonders hilfreich. Aus methodischer Sicht wird eine Verzerrungskorrektur im Fall der Log- sowie Log-Shift-Transformation untersucht. Für diesen Fall wird eine Methodik basierend auf der Kerndichteschätzung vorgeschlagen, um die

Verzerrung aufgrund der Rücktransformation der aggregierten Hilfsinformationen zu korrigieren. Dabei werden keine parametrischen Annahmen getroffen. Außerdem wird eine dazugehörige Unsicherheitsschätzung bereitgestellt. Modell- und designbasierte Simulationen zeigen, dass die vorgeschlagene Methodik vergleichbare Ergebnisse liefert wie bekannte Methoden, die auf Populations-Mikrodaten angewiesen sind. Hervorzuheben ist, dass die entwickelte Methodik im R-Paket saeTrafo open-source zur Verfügung steht, um ihre Anwendung zu erleichtern. Dieses Paket ist benutzerfreundlich und automatisiert die Auswahl geeigneter SAE-Modelle für unterschiedliche Datensätze. Die vorgestellte Methodik wird angewandt, um die Zuverlässigkeit der Schätzung von Einkommen in Deutschland auf Ebene der regionalen Raumordnungsregionen zu verbessern.

LITERATURVERZEICHNIS

Alfons, Andreas/Templ, Matthias. *Estimation of Social Exclusion Indicators from Complex Surveys: The R Package laeken*. In: Journal of Statistical Software. Band 54. Ausgabe 15/2013, Seite 1 ff. DOI: 10.18637/jss.v054.i15

Battese, George E./Harter, Rachel M./Fuller, Wayne A. *An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data*. In: Journal of the American Statistical Association. Jahrgang 83. 1988. Ausgabe 401, Seite 28 ff. DOI: 10.2307/2288915

Berg, Emily/Chandra, Hukum. *Small area prediction for a unit-level lognormal model*. In: Computational Statistics & Data Analysis. Band 78. Ausgabe Oktober 2014, Seite 159 ff. DOI: 10.1016/j.csda.2014.03.007

Chandra, Hukum/Chambers, Ray. *Small area estimation under transformation to linearity*. In: Survey Methodology. Jahrgang 37. Ausgabe 1/2011, Seite 39 ff. [Zugriff am 29. Februar 2024]. Verfügbar unter: www150.statcan.gc.ca

Eurostat (Statistisches Amt der Europäischen Union). *Handbook on precision requirements and variance estimation for ESS households surveys*. Luxemburg 2013. [Zugriff am 7. März 2024]. Verfügbar unter: ec.europa.eu

González-Manteiga, Wenceslao/Lombardia, María J./Molina, Isabel/Morales, Domingo/Santamaría, Laureano. *Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model*. In: Computational Statistics & Data Analysis. Band 52. Ausgabe 12/2008, Seite 5242 ff. DOI: 10.1016/j.csda.2008.04.031

Karlberg, Forough. *Population total prediction under a lognormal superpopulation model*. Metron – International Journal of Statistics. Band 58. Ausgabe 3-4/2000, Seite 53 ff. [Zugriff am 29. Februar 2024]. Verfügbar unter: www.researchgate.net

Kreutzmann, Ann-Kristin/Pannier, Sören/Rojas-Perilla, Natalia/Schmid, Timo/Templ, Matthias/Tzavidis, Nikos. *The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators*. In: Journal of Statistical Software. Band 91. Ausgabe 7/2019, Seite 1 ff. DOI: 10.18637/jss.v091.i07

Molina, Isabel/Martín, Nirian. *Empirical best prediction under a nested error model with log transformation*. In: The Annals of Statistics. Band 46. Ausgabe 5/2018, Seite 1961 ff. DOI: 10.1214/17-AOS1608

Molina, Isabel/Rao, J. N. K. *Small area estimation of poverty indicators*. In: The Canadian Journal of Statistics, Band 38. Ausgabe 3/2010, Seite 369 ff. DOI: 10.1002/cjs.10051

Pfeffermann, Danny. *New Important Developments in Small Area Estimation*. In: Statistical Science. Band 28. Ausgabe 1/2013, Seite 40 ff. DOI: 10.1214/12-STS395

Rao, J. N. K./Molina, Isabel. *Small Area Estimation*. Zweite Auflage. Hoboken 2015.

LITERATURVERZEICHNIS

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Wien 2022.

Rojas-Perilla, Natalia/Pannier, Sören/Schmid, Timo/Tzavidis, Nikos. *Data-Driven Transformations in Small Area Estimation*. In: Journal of the Royal Statistical Society Series A: Statistics in Society. Band 183. Ausgabe 1/2020, Seite 121 ff. DOI: 10.1111/rssa.12488

Socio-Economic Panel. *Data from 1984-2017. SOEP-Core v34 (data 1984-2017)*. Berlin 2019. DOI: 10.5684/soep.v34.

Tzavidis, Nikos/Zhang, Li-Chun/Luna, Angela/Schmid, Timo/Rojas-Perilla, Natalia. *From Start to Finish: A Framework for the Production of Small Area Official Statistics*. In: Journal of the Royal Statistical Society Series A: Statistics in Society. Band 181. Ausgabe 4/2018, Seite 927 ff. DOI: 10.1111/rssa.12364

Würz, Nora. *saeTrafo: Transformations for Unit-Level Small Area Models*. R package version 1.0.0. 2022.

Würz, Nora/Schmid, Timo/Tzavidis, Nikos. *Estimating Regional Income Indicators under Transformations and Access to Limited Population Auxiliary Information*. In: Journal of the Royal Statistical Society Series A: Statistics in Society. Band 185. Ausgabe 4/2022, Seite 1679 ff. DOI: 10.1111/rssa.12913

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im April 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24002-4, ISSN 1619-2907