KURZFASSUNG

der mit dem

Förderpreis 2023

„Statistical Science for the Society“

des Statistischen Bundesamtes aus-

gezeichneten Dissertation

zum Thema

„Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications“

von

Dr. Nora Würz

Otto-Friedrich-Universität Bamberg

Kurzfassung der Dissertation

Dr. Nora Würz

„Small-Area-Schätzung unter limitierten Zugang zu Hilfsinformationen aus der Population bei Modellverletzungen und ihre wirtschaftlichen Anwendungen“

Originaltitel: „Small Area Estimation under Limited Auxiliary Population Data Dealing with Model Violations and their Economic Applications”
Verteidigung: 19.12.2022
Bewertung: summa cum laude

Für evidenzbasierte Entscheidungsfindung sind zuverlässige Informationen über sozioökonomische Indikatoren unerlässlich. Stichprobenerhebungen ermöglichen eine kosteneffiziente Erhebung von Indikatoren und haben eine lange Tradition. Dabei besteht meist nicht nur ein Interesse an der quantitativen Erfassung dieser Indikatoren für die Gesamtpopulation, sondern insbesondere für Teilpopulationen (geografische Gebiete oder soziodemografische Gruppen). Um Einblicke in diese Teilpopulationen zu gewinnen, können disaggregierte direkte Schätzer verwendet werden, die ausschließlich auf Umfragedaten des jeweiligen Gebiets berechnet werden. In Small-Area-Forschung gilt ein Gebiet als "large", wenn die Stichprobengröße groß genug ist, um zuverlässige direkte Schätzungen für dieses Gebiet zu ermöglichen. Wenn die Genauigkeit der direkten Schätzungen nicht ausreichend ist 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 (SAE) beschäftigt sich mit der Überwindung dieses Problems ohne dass größere und damit teurere Umfragen erforderlich sind (Pfeffermann, 2013; Rao und Molina, 2015; Tzavidis et al., 2018). SAE-Techniken nutzen die Informationen von allen Gebieten gleichzeitig über ein statistisches Modell, um dadurch die Schätzungen für wiederum alle Gebiete zu verbessern. Dabei werden die Umfragedaten mit weiteren Hilfsdaten über ein Modell verknüpft und gebietsspezifische Strukturen ausgenutzt. Geeignete Hilfsdaten sind Verwaltungs- und Registerdaten und 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 SAE-Schätzern, die keine Hilfsdaten auf Mikrodaten-Ebene benötigen, sondern mit deutlich einfacher zugänglichen Aggregaten aus diesen Mikrodaten auskommen. In dieser Arbeit werden neue Methoden in Abwesenheit von Populations-Mikrodaten vorgestellt und mittels Anwendungen auf sozioökonomisch hoch relevante Indikatoren demonstriert.

Da verschiedene SAE-Modelle unterschiedliche Datenvoraussetzungen haben, wird diese kumulative Dissertation entsprechend dieser Voraussetzungen in zwei Bereiche geteilt. Der erste Teil betrachtet die Ausgangssituation von Umfragedaten auf Individualebene und gleichzeitigen limitierten Zugang zu Hilfsdaten, z.B. aggregierte Daten wie Mittelwerte. Diese Datensituation liegt besonders häufig vor. In dieser Arbeit wird das Nested-Error-Regression (NER) Modell von Battese et al. (1988) verwendet, da so die Individualdaten aus der Umfrage ohne Aggregation einbezogen werden können. Dieses Modell ist ein Spezialfall eines linearen gemischten Modells auf der Grundlage mehrerer Annahmen. Aber wie können Benutzer vorgehen, wenn die Modellannahmen nicht erfüllt sind? Im ersten Teil der Arbeit, welcher drei Publikationen umfasst, werden zwei neue Ansätze vorgestellt, die zur Lösung dieses Problems beitragen.
Ein vielversprechender Ansatz ist die Transformation der abhängigen Variable des Modells. Da mehrere sozioökonomisch relevante Variablen wie Einkommen eine schiefe Verteilung haben, ist die Log-Transformation eine bewährte Methode, um die Modell-Annahmen zu erfüllen (Berg und Chandra, 2014; Molina und Martín, 2018). Die datengetriebene Log-Shift-Transformation passt sich zusätzlich an die Daten an, indem die Logarithmus-Funktion um einen zusätzlichen Parameter erweitert und dadurch

flexibler wird (Berg und Chandra, 2014; Molina und Martín, 2018). Das erste Projekt (veröffentlich im Journal of the Royal Statistical Society: Series A) führt Methodik für die Log- und Log-Shift-Transformation ein bei ausschließlicher Verfügbarkeit von aggregierten Populations-Hilfsinformationen und gleichzeitigem Vorliegen der Stichprobe auf Individualebene. Eine besondere Herausforderung besteht darin, die auf der transformierten Ebene geschätzten Small-Area-Mittelwerte wieder auf die ursprüngliche Skala zurückzuführen. Hierbei werden geeignete Verzerrungs-Korrekturen für die Small-Area-Vorhersagen benötigt. Der vorgeschlagene Ansatz kombiniert aggregierte Statistiken (Mittelwerte und Kovarianzen) und Kerndichte-Schätzungen, um das Problem des fehlenden Zugangs zu Populations-Mikrodaten zu adressieren. Zudem wird die Schätzung des mittleren quadratischen Fehlers mit einem parametrischen Bootstrap-Verfahren vorgestellt. Umfangreiche modellbasierte und designbasierte Simulationen werden verwendet, um die vorgeschlagene Methode mit alternativen Methoden zu vergleichen. Die vorgeschlagene Methode wird angewendet, um das regionale Einkommen für die 96 Raumordnungsregionen in Deutschland unter Verwendung des Sozio-Ökonomischen-Panels (Socio-Economic Panel, 2019) und Zensusdaten (Statistisches Bundesamt, 2015) zu schätzen. Sie erzielt eine klare Verbesserung der Zuverlässigkeit und demonstriert damit die Vorteile dieser Methode.
Um weitere Anwendungen zu ermöglichen, wird diese neue Methodik im R-Paket *saeTrafo* (R Core Team, 2022; Würz, 2022) Anwendenden zur Verfügung gestellt. Das zweite Kapitel der Dissertation stellt die verschiedenen Funktionen des Pakets anhand öffentlich verfügbarer Einkommensdaten dar. Um Benutzungsfreundlichkeit des Pakets zu erhöhen, werden gleichzeitig weitere etablierte SAE-Modelle für Stichprobendaten auf Individualebene mit Transformation angeboten. Auch Unsicherheitsschätzer können direkt erhalten werden und auf Grundlage der eingegebenen Datenstruktur wird die geeignetste Methode automatisiert ausgewählt.
Für einige Anwendungen ist es jedoch herausfordernd, eine geeignete Transformation zu finden oder, allgemeiner ausgedrückt, ein Modell in Gegenwart komplexer Wechselwirkungen (Interaktionen) zu spezifizieren. In diesem Fall sind Machine-Learning-Methoden wertvoll, da eine Transformation nicht unbedingt erforderlich ist und ein Modell nicht explizit spezifiziert werden muss (Hastie et al., 2009; Varian, 2014). Der semi-parametrische Konzept von Mixed-Effects-Random-Forest (MERF) kombiniert die Vorteile von Random-Forests (Robustheit gegenüber Ausreißern und implizite Modellselektion) mit der Fähigkeit hierarchische Abhängigkeiten wie bei SAE-Ansätzen zu modellieren (Krennmair und Schmid, 2022). Das dritte Kapitel führt MERFs in Abwesenheit von Mikrodaten der Population ein. Da Random-Forest-Algorithmen Populations-Mikrodaten benötigen, wird eine alternative Strategie zur Schätzung eingeführt. Diese Strategie verwendet Kalibrierungsgewichte, welche mittels aggregierten Hilfsinformationen bestimmt werden, um den Zugriff auf Populations-Mikrodaten zu umgehen. Diese Methodik wird zur Schätzung von Opportunitätskosten von Pflegearbeit in Deutschland aus dem Sozio-Ökonomischen-Panels (Socio-Economic Panel, 2019) und Zensusdaten (Statistisches Bundesamt, 2015) angewandt. Dabei zeigt sich, dass eine höhere Genauigkeit im Vergleich zur direkten Schätzungen und dem klassischen NER-Modell erzielt wird.

Im Gegensatz zu Methoden, die Stichprobendaten auf Individualebene verwenden, konzentriert sich Teil II der Dissertation auf die bekannte Klasse der Area-Level-SAE-Modelle (Fay und Herriot, 1979), die eine direkte Schätzungen aus Umfragedaten mit (erneut) nur aggregierte Populations-Hilfsinformationen über Modelle verbinden. Diese Dissertation zeigt zwei besonders relevante Anwendungen dieser Modellklasse. Kapitel 4 untersucht regionale Verbraucherpreisindizes (VPIs) im Vereinigten Königreich (UK) und trägt somit zum Monitoring der Inflation auf räumlicher Ebene bei (Fenwick und O'Donoghue, 2003). Die SAE-Herausforderung besteht in der modellbasierten Schätzung von Ausgabenanteilen, um daraus die regionalen Waren- und Dienstleistungskörbe und ihrer Wägung für die zwölf Regionen in UK zusammenzustellen. Die dazu verwendete Umfrage ist die Einkommens- und Verbrauchsstichprobe (Defra und ONS, 2019). Darüber hinaus werden Preisdaten (ONS, 2020) mit den SAE-geschätzten Warenkörben auf regionaler Ebene verknüpft, um regionale VPIs zu erstellen. Die regionalen VPI-Zeitreihen konnten hierdurch deutlich verbessert werden, sie sind allerdings immer noch zu volatil für die politische Entscheidungsfindung. Diese Forschung dient als wertvoller Ausgangspunkt für die Entwicklung eines zukünftigen regionale VPIs für UK.
Die zweite Anwendung untersucht ebenfalls einen politisch und wirtschaftlich hochrelevanten Indikator, nämlich die Arbeitslosenquote. Das regionale Zielniveau sind die Functional-Urban-Areas im deutschen Bundesland Nordrhein-Westfalen. In Kapitel 5 (u.a. in Zusammenarbeit mit Frau Sandra Hadam, Statistisches Bundesamt) werden zwei Arten von Arbeitslosenquoten - die traditionelle und eine alternative Definition, die den Pendlerverkehr berücksichtigt (Grözinger, 2018) - geschätzt und

verglichen. Direkte Schätzungen aus der Arbeitskräfteerhebung (Eurostat, 2019) werden mit SAE-Methoden verknüpft, um passiv erhobene Mobilfunkdaten zu verwenden. Diese alternative Datenquelle ist in Echtzeit verfügbar, bietet räumlich flexible Auflösungen und ist dynamisch (Toole et al., 2015; Marchetti et al., 2015; Steele et al., 2017; Schmid et al., 2017) und dient als Hilfsinformation im SAE-Modell. So wird die Zuverlässigkeit der Arbeitslosenquoten-Schätzern verbessert, und die resultierenden Vorhersagen zeigen, dass alternative Arbeitslosenquoten in den deutschen Stadtzentren niedriger sind als die von den offiziellen Arbeitslosenquoten angegebenen traditionellen Schätzungen.

Übersicht der einzelnen Publikationen dieser kumulativen Dissertation

- Würz, N., Schmid, T., and Tzavidis, N. (2022) Estimating regional income indicators under transformations and access to limited population auxiliary information, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4), pp. 1679-1706, doi: https://doi.org/10.1111/rssa.12913.
- Würz, N. (2022) The R package saeTrafo for estimating unit-level small area models under transformations. *R-Paket-Vignette*, https://CRAN.R-project.org/package=saeTrafo.
- Krennmair, P., Würz, N., and Schmid, T. (2022) Analysing opportunity cost of care work using mixed effects random forests under aggregated census data. *Arbeitspapier*, https://arxiv.org/abs/2204.10736.
- Dawber, J., Würz, N., Smith, P., Flower, T., Thomas, H., Schmid, T., and Tzavidis, N. (2022) Experimental UK regional consumer price inflation with modelbased expenditure weights. *Journal of Official Statistics*, 38(1), pp. 213-237, doi: https://doi.org/10.2478/jos-2022-0010.
- Hadam, S., Würz, N., Kreutzmann, A.-K., and Schmid, T. (2022) Estimating regional unemployment with mobile network data for functional urban areas in Germany. Im Begutachtungsprozess (major revision) bei *Statistical Methods and Applications*.

Literaturverzeichniss

- Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*(401), 28–36.
- Berg, E. and H. Chandra (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis 78*, 159–175.
- Defra and ONS (2019). Living costs and food survey, 2008-2014. https://doi.org/10.5255/UKDA-SN-7992-4 (DOI for 2014 only). 3rd Edition. UK Data Service. Data collection. SN: 7992 and also SN: 6385, 6655, 6945, 7272, 7472, 7702.
- Eurostat (2019b). EU labour force survey database user guide. https://ec.europa.eu /eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf. [abgerufen: 08.2019].
- Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*(366), 269–277.
- Fenwick, D. and J. O'Donoghue (2003). Developing estimates of relative regional consumer price levels. *Economic Trends 599*, 72–83.
- Grözinger, G. (2018). Regionale Arbeitslosigkeit: Falsche Eindrucke von Stadt-Land-Differenzen. *Wirtschaftsdienst 98*(1), 68–70.
- Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics 31*(2), 263 – 281.
- Molina, I. and N. Martin (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics 46*(5), 1961–1993.
- ONS (2020). Consumer price inflation item indices and price quotes. https://www.ons. gov.uk/economy/inflationandpriceindices/datasets/consumerpriceindicescpiandretailprices indexrpiitemindicesandpricequotes. [abgerufen: 04.2021].
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28*(1), 40–68.
- Rao, J. N. K. and I. Molina (2015). *Small Area Estimation (Second Edition)*. Hoboken: John Wiley & Sons.
- Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(4), 1163–1190.
- Socio-Economic Panel (2019). Data for years 1984-2017, version 34i, SOEP. Socio-Economic Panel, Berlin. doi: https://doi.org/10.5684/soep.v34.
- Statistisches Bundesamt (2015). Zensus 2011 Methoden und Verfahren. Statistisches Bundesamt, Wiesbaden. https://www.zensus2011.de/SharedDocs/Downloads/DE/Publikationen/Aufsaetze_Archiv/2015_06_MethodenUndVerfahren.pdf?__blob=publicationFile&v=6. [abgerufen: 12.2020].
- Steele, J. E., P. R. Sundsoy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, K. Engo-Monsen, Y.-A. de Montjoye, A. M. Iqbal, K. N. Hadiuzzaman, X. Lu, E. Wetter, A. J. Tatem, and L. Bengtsson (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society Interface 14*(127), 20160690.
- Toole, J. L., Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. Gonzalez, and D. Lazer (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface 12*(107), 20150185.
- Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(4), 927–979.