Dipl.-Mathematiker Martin Vogt

Small Area Estimation: Die Schätzer von Fay-Herriot und Battese-Harter-Fuller

Das Statistische Bundesamt hat im November 2008 zum zehnten Mal den Gerhard-Fürst-Preis für herausragende wissenschaftliche Arbeiten mit einem engen Bezug zur amtlichen Statistik verliehen.

Die von Herrn Professor Dr. Hans Wolfgang Brachinger (Université de Fribourg Suisse/Universität Freiburg Schweiz), dem Vorsitzenden des unabhängigen Gutachtergremiums, vorgetragenen Laudationes wurden in Ausgabe 12/2008 dieser Zeitschrift veröffentlicht.

Daran anknüpfend stellen nun die beiden Preisträger ihre Arbeiten in eigenen Beiträgen näher vor. Den Anfang macht Diplom-Mathematiker Martin Vogt, dessen bei Professor Dr. Ralf Münnich an der Universität Trier entstandene Diplomarbeit zum Thema „Small Area Estimation: Die Schätzer von Fay-Herriot und Battese-Fuller-Harter" von der Jury als herausragende Leistung bewertet und mit dem Gerhard-Fürst-Preis 2008 in der Kategorie „Diplom-/Magisterarbeiten" ausgezeichnet wurde.

1 Einleitung

Angenommen, es sei die Durchschnittsgröße der Einwohner der Stadt Berlin zu ermitteln. Dazu werden gemäß eines geeigneten Stichprobenplans einige Bewohner nach ihrer Größe befragt. Anschließend wird ein Schätzwert, wie etwa das arithmetische Mittel der erfassten Größen ermittelt. Dieser Schätzwert ist bei einer ausreichend großen Stichprobe hinreichend „gut". Soll nun zusätzlich die Durchschnittsgröße der Einwohner in den einzelnen Stadtteilen Berlins bestimmt werden, entsteht ein Problem. Die Stichprobe ist zwar ausreichend groß, um die Durchschnittsgröße der Bewohner der kompletten Stadt zu ermitteln, aber eventuell in einigen Bezirken sehr klein, im Grenzfall sogar null. Damit ist es schwer, verlässliche Schätzwerte für diese Bezirke zu bestimmen. Dies ist eine typische Problemstellung im Bereich der sogenannten Small-Area-Statistik. Eine Möglichkeit, an das obige Problem heranzugehen besteht darin, Verfahren zu entwickeln, die Hilfsinformationen benutzen, zum Beispiel Informationen aus benachbarten Bezirken oder aus Registern. Es könnte beispielsweise das Gewicht als Hilfsmerkmal zur Schätzung hinzugezogen werden, falls dieses – etwa aus einer vorherigen Schätzung – bekannt ist.

In der Diplomarbeit "Small Area Estimation: Die Schätzer von Fay-Herriot und Battese-Fuller-Harter"[1]) werden zwei Modelle behandelt. Im ersten Modell, dem Modell von Fay-Herriot[2]), werden Hilfsinformationen auf Bezirksebene hinzugezogen, zum Beispiel das Durchschnittsgewicht in jedem Bezirk. Im zweiten Modell, dem Modell von Battese-Harter-Fuller[3]), werden hingegen Hilfsinformationen auf Individualniveau benutzt. In diesem Modell ist also das Gewicht jedes Einwohners in der Stichprobe bekannt. Da hier Informationen auf Individualniveau vorausgesetzt sind, wird dieses Modell auch als Unit-Level-Modell bezeichnet. Das Modell

1) Vogt, M., 2007 (unveröffentlicht).
2) Fay, R. E./Herriot, R. A.: "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data" in Journal of the American Statistical Association, Vol. 74 (1979), No. 366, S. 269 ff.
3) Battese, G. E./Harter, R. M./Fuller, W. A.: "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data" in Journal of the American Statistical Association, Vol. 83 (1988), No. 401, S. 28 ff.

von Fay-Herriot hingegen benutzt nur Durchschnittswerte als Hilfsinformationen, also Informationen auf Bezirksebene. Deshalb wird dieses Modell als Area-Level-Modell bezeichnet. Gründe für das Fehlen von Informationen auf Individualniveau können zum Beispiel der Datenschutz oder einfach Informationsmangel sein. Die obige Situation ist auf zahlreiche Fragestellungen der amtlichen Statistik übertragbar. Ein aktuelles Beispiel ist der Zensus 2011[4]), der erstmals registergestützt durchgeführt wird. Bei diesem neuen Verfahren werden hauptsächlich die in den Registern der Verwaltung vorhandenen Daten genutzt; zusätzlich wird eine Stichprobe erhoben. Der Einsatz von Stichproben stößt allerdings an Grenzen, wenn die Stichprobe sehr klein ist. Dies ist zum Beispiel bei tiefgegliederten Subpopulationen, wie etwa Landkreisen, Gemeinden oder Bezirken der Fall. Solche Subpopulationen müssen nicht geografisch abgegrenzt sein, sondern können auch inhaltlich gegliedert sein, zum Beispiel nach Alter, Geschlecht oder Nationalität. Allgemein wird eine „kleine" Subpopulation, die sowohl geografisch als auch inhaltlich motiviert sein kann, als Small Area bezeichnet. Klassische Schätzmethoden, wie der Horvitz-Thompson- oder der Generalized Regression-Schätzer (GREG), die nur Stichprobeninformationen in den jeweiligen Small Areas ausnutzen, besitzen bei solch kleinen Stichproben einen zu großen Stichprobenfehler. Im Gegensatz dazu verwenden die beiden Small-Area-Schätzer zusätzlich Modellvorstellungen über die Unterschiedlichkeit einer Menge von Areas. Einen Überblick über Small-Area-Methoden geben Jiang und Lahiri[5]), sowie Rao[6]). Die Basisidee hat Vogt[7]) an einem einfachen amüsanten Beispiel dargestellt.

2 Die Small-Area-Modelle und -Schätzer

Im Folgenden wird anhand des obigen Beispiels – der Schätzung der Duchschnittsgröße der Einwohner Berlins – zunächst die Modellbildung des Fay-Herriot- und Battese-Harter-Fuller-Schätzers dargestellt. Anschließend wird aufgezeigt, wie diese Situation auf den registergestützten Zensus übertragen werden kann.

Es sei die Durchschnittsgröße der Einwohner der Stadt Berlin zu schätzen. Dazu werde eine Stichprobe der Größe n = 1 erhoben, also eine sehr kleine Stichprobe. Das Stichprobenelement sei zudem ein Basketballspieler mit einer Körpergröße von 210 cm. Ist der Wert 210 cm als Schätzwert zur Schätzung der Durchschnittsgröße der Einwohner einer Stadt zu groß? Die meisten Leser werden das Gefühl haben, dass dies der Fall ist. Woher kommt dieses Gefühl? Es scheinen schon Vorinformationen zu existieren, bevor Daten erhoben wurden, also unabhängig von den erhobenen Daten. Im Folgenden wird gezeigt, wie diese Vorinformationen in die Schätzung integriert werden können. Die Größe der Einwohner sei dazu als Zufallsvariable Y aufgefasst. Eine mögliche Verteilungsannahme für Y ist die Normalverteilung mit einer hier als bekannt vorausgesetzten Varianz und einem unbekannten Erwartungswert θ:

$$Y \sim N(\theta, 15^2).$$

Hierbei ist θ die unbekannte, gesuchte Durchschnittsgröße der Einwohner der Stadt Berlin. Vorinformationen über θ können in die Modellierung eingebracht werden, indem wiederum eine Verteilung – etwa eine Normalverteilung – für θ spezifiziert wird, die sogenannte A-priori-Verteilung. Wenn zum Beispiel als Vorinformation angenommen wird, dass die Durchschnittsgröße ungefähr 170 cm ist mit einer Varianz 10^2, dann besitzt die Verteilung folgende Form:

$$\theta \sim N(170, 10^2).$$

Es gibt nun also eine Modellannahme für die Größe der Einwohner Y und eine für die Vorinformationen θ. Diese Verteilungsannahmen können mithilfe des Satzes von Bayes zu der A-posteriori-Verteilung

$$\theta \mid Y \sim N\left(\underbrace{\frac{\frac{210}{15^2}+\frac{170}{10^2}}{\frac{1}{15^2}+\frac{1}{10^2}}}_{\approx 182}, \frac{1}{\frac{1}{15^2}+\frac{1}{10^2}} \right)$$

verbunden werden. Wenn die beiden Ausgangsverteilungen, wie in diesem Beispiel, Normalverteilungen sind, dann gehört auch die A-posteriori-Verteilung zur Familie der Normalverteilungen. Dabei ist der A-posteriori-Erwartungswert ein gewichtetes Mittel aus dem A-priori-Erwartungswert 170 cm und dem Stichprobenwert 210 cm. Gewichtungsfaktoren sind die Varianzen. Die Notation $\theta \mid Y$ drückt dabei aus, dass in der A-posteriori-Verteilung Vorinformationen mit Stichprobeninformationen verbunden sind bzw. die Vorinformationen mithilfe der Stichprobe aktualisiert werden. Als Schätzwert bietet sich der A-posteriori-Erwartungswert von ungefähr 182 cm an.

Eine Stichprobengröße von n = 1 ist in den meisten Fällen nicht realistisch. Dieses Beispiel kann jedoch leicht auf eine Situation übertragen werden, in der eine kleine Stichprobe vorliegt, zum Beispiel, indem die Fragestellung erweitert wird und nicht die Durchschnittsgröße der Einwohner der kompletten Stadt, sondern die der Einwohner in den einzelnen Bezirken der Stadt gesucht wird. Je tiefer regional oder inhaltlich gegliedert wird, desto geringer wird die Stichprobengröße. Als Modell zur Schätzung auf Bezirksebene ergibt sich:

$$Y_i \sim N(\theta_i, D_i) \quad i = 1, \cdots, k$$
$$\theta_i \sim N(\mu_i, A) \quad i = 1, \cdots, k\,.$$

4) Münnich, R./Gabler, S./Ganninger, M.: "Some remarks on the register-based Census 2010/2011 in Germany", Southhampton 2007 (*www.s3ri/soton.ac.uk/isi2007/slides/Slides03.pdf*; Stand: 5. Februar 2009). Umfassende Informationen zum Zensus 2011 stellen die Statistischen Ämter des Bundes und der Länder auf ihrer gemeinsamen Internetseite *www.zensus2011.de* bereit.

5) Jiang, J./Lahiri, P.: "Mixed model prediction and small area estimation" in Test, Vol. 15 (2006), No. 1, S. 1 ff.

6) Rao, J. N. K.: "Small Area Estimation", New York 2003.

7) Siehe Vogt, M.: „Schlauer Wetten" in Die Wurzel: Zeitschrift für Mathematik, Heft 6/2008, S. 116 ff.

Hierbei steht *i* für einen der *k* Bezirke Berlins. Das obige Modell kann auf die Situation eines registergestützten Zensus übertragen werden. Anstatt der Größe der Einwohner kann die Zufallsvariable *Y* eine beliebige Zensusvariable darstellen (mit Normalverteilungsannahme) und die Bezirke Berlins können durch Gemeinden oder Verbandsgemeinden ersetzt werden. Da in einem registergestützten Zensus eine Stichprobe erhoben wird und die Stichprobengröße bei einer tiefen räumlichen Gliederung sehr klein ist, passt diese Situation zu dem obigen Beispiel. Es bleibt zu klären, woher die Vorinformationen kommen. Bisher sind diese als bekannt vorausgesetzt worden. Im Fall eines registergestützten Zensus können diese etwa aus den Registern gewonnen werden. Zum Beispiel indem μ_i, der A-priori-Erwartungswert des *i*-ten Bezirkes oder der *i*-ten Gemeinde/Small-Area, durch den Term $X_i\beta$ ersetzt wird. Hierbei ist X_i eine Matrix mit Hilfsvariablen aus den Registern und β ein unbekannter Vektor. Es ergibt sich dann allgemein:

$$Y_i \sim_{u.a.} N(\theta_i, D_i)$$
$$\theta_i \sim_{u.a.} N(X_i\beta, A) \quad i = 1, \cdots, k\,.$$

Dies ist das sogenannte Modell von Fay-Herriot, ein Grundmodell der Small-Area-Schätzung. Um einen anderen Blickwinkel auf dieses Modell zu bekommen, werden im Folgenden zwei alternative Schreibweisen dargestellt. Aus der Verteilungsannahme für θ kann der Erwartungswert $X_i\beta$ herausgezogen werden. Dann ergibt sich:

$$Y_i \sim_{u.a.} N(\theta_i, D_i)$$
$$\theta_i = X_i\beta + u_i$$
$$u_i \sim_{iid} N(0, A) \quad i = 1, \cdots, k\,.$$

Derselbe Schritt kann mit der Verteilungsannahme von Y_i wiederholt werden:

$$Y_i = X_i\beta + u_i + e_i$$
$$u_i \sim_{iid} N(0, A)$$
$$e_i \sim_{u.a.} N(0, D_i) \quad i = 1, \cdots, k\,.$$

Somit ergibt sich ein Modell, welches stark an ein normales Regressionsmodell mit Regressionskomponente $X_i\beta$ und Fehlerterm *e* erinnert. Neu ist der Term *u*, für den auch eine Verteilungsannahme spezifiziert wird. Dieser Term wird random effect genannt und erfasst Schwankungen zwischen den Areas, die nicht von dem Regressionsterm aufgefangen werden.

In das Modell von Fay-Herriot wird für jede Area *ein* Wert Y_i gesteckt. Nun können in der Stichprobe für jede Area aber mehrere Elemente vorhanden sein. Dann müssten zwei Indizes verwendet werden: *i* für die Area und *j* für das Individuum. Das Modell von Fay-Herriot berücksichtigt dies nicht. Die Informationen kommen auf einer aggregierten Ebene in die Modellbildung hinein. Deshalb wird dieses Modell als Area-Level-Modell bezeichnet. Im Gegensatz dazu berücksichtigt das Modell von Battese-Harter-Fuller Informationen auf Individualniveau und wird somit als Unit-Level-Modell bezeichnet. Das Modell besitzt die folgende Form:

$$Y_{ij} = X_{ij}\beta + u_i + e_{ij}$$
$$u_i \sim_{iid} N(0, A)$$
$$e_{ij} \sim_{iid} N(0, D) \quad i = 1, \cdots, k; j = 1, \cdots, n_i\,.$$

Wie das Modell von Fay-Herriot besteht auch das Modell von Battese-Harter-Fuller aus drei Komponenten: einem Regressionsterm $X_{ij}\beta$, einem random effect u_i und einem Fehler e_{ij}. Letztendlich sind aber keine Modelle, sondern Schätzwerte gesucht. Diese werden im Folgenden nur angegeben. Für die mathematischen Hintergründe und Herleitungen sei auf die diesem Beitrag zugrunde liegende Diplomarbeit[8]) verwiesen, in der diese ausführlich dargestellt sind. Der Fay-Herriot-Schätzer ergibt sich als:

$$\hat{\theta}^{+}_{FH,i} = (1 - \hat{B}_i) \underbrace{Y_i}_{\text{Horvitz-Thompson}} + \hat{B}_i \underbrace{X_i\hat{\beta}}_{\text{Synthetischer Teil}}\,,$$

also als gewichtetes Mittel aus Horvitz-Thompson-Schätzer und einem synthetischen Teil, wobei $\hat{B}_i = \frac{D_i}{D_i + \hat{A}}$. Ferner sind $\hat{A}$ und $\hat{\beta}$ Schätzer für die Varianz *A* bzw. den Regressionskoeffizienten β.

Auch der Battese-Harter-Fuller-Schätzer ergibt sich als ein gewichtetes Mittel, aber nicht zwischen dem Horvitz-Thompson-Schätzer und einem synthetischen Teil, sondern zwischen dem GREG-Schätzer und einem synthetischen Teil:

$$\hat{\theta}_{BHF,i} = (1 - \hat{B}_i)\underbrace{(\bar{y}_i + (\bar{x}_{N,i} - \bar{x}_i)\hat{\beta})}_{\text{GREG}} + \hat{B}_i \underbrace{\bar{x}_{N,i}\hat{\beta}}_{\text{Synthetischer Teil}}\,,$$

wobei $\hat{B}_i = \frac{\hat{D}/n_i}{\hat{D}/n_i + \hat{A}}$, n_i die Stichprobengröße der *i*-ten Area, $\bar{x}_i$ den Stichprobendurchschnitt der Hilfsvariablen sowie $\bar{x}_{N,i}$ den als bekannt vorausgesetzten Populationsdurchschnitt der Hilfsvariablen darstellen.

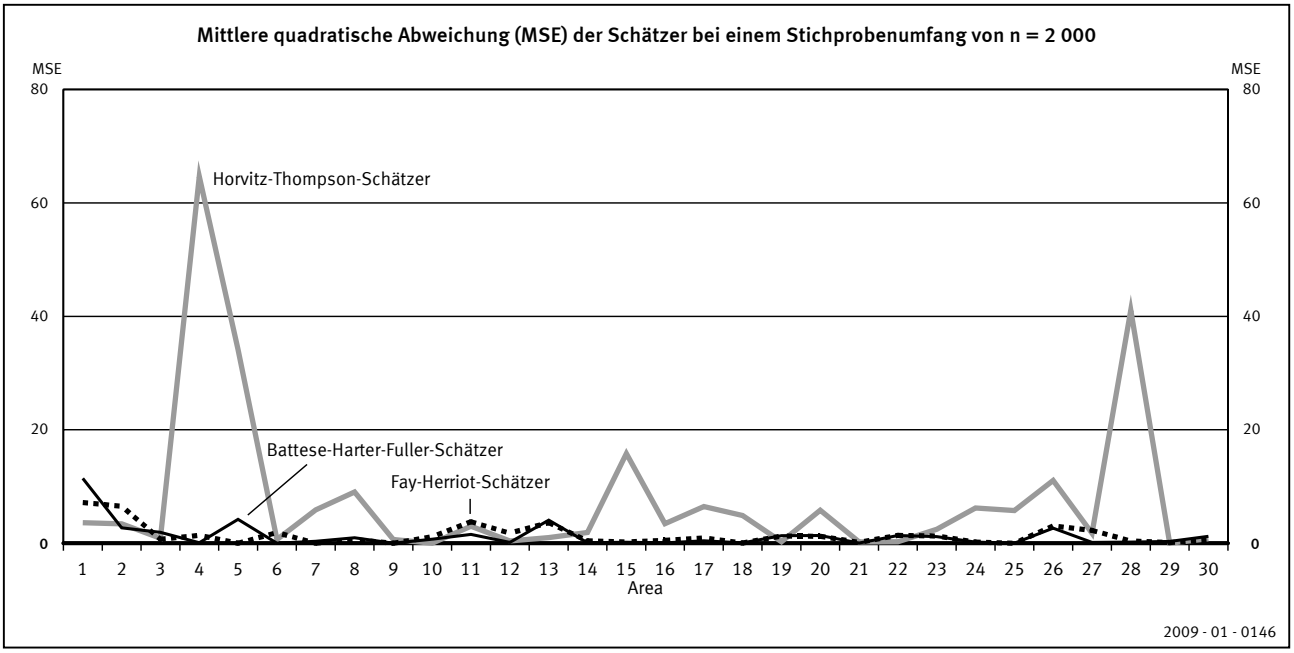
3 Ergebnisse einer Simulationsstudie

Um die Schätzer miteinander zu vergleichen, wurde in der diesem Beitrag zugrunde liegenden Arbeit eine Simulationsstudie durchgeführt. Dazu wurde eine künstliche Grundgesamtheit der Größe N = 30 000 konstruiert. Diese besteht aus 30 Areas, wobei jeweils 6 Areas eine Größe von 500, 750, 1 000, 1 250 beziehungsweise 1 500 besitzen. Dies könnte zum Beispiel eine Kleinstadt mit 30 000 Einwohnern darstellen, die in 30 verschiedene Bezirke untergliedert ist.

Ziel der Untersuchung ist es, den Area-Mittelwert jeder Area (Stadtteil) zu schätzen. Denkbar ist zum Beispiel, dass der Area-Mittelwert wie in dem obigen Beispiel der Durch-

8) Siehe Fußnote 1.

Schaubild 1



schnittsgröße entspricht. Zunächst wird eine uneingeschränkte Zufallsstichprobe ohne Zurücklegen vom Umfang n = 2 000 gezogen, anschließend eine von n = 60. Die Beobachtungen bestehen jeweils aus dem Paar x und y. Dabei repräsentiert x die Hilfsinformation (z. B. das Gewicht) und y das Merkmal, dessen Area-Mittelwert das Untersuchungsziel darstellt (z. B. die Größe). Genauer werden die Variablen wie folgt konstruiert: Es wird von normalverteilten Hilfsinformationen der Form

$$x_{ij} \sim N(m_i, m_i/6), i = 1,2,\cdots,30, j = 1,2,\cdots,n_i$$

ausgegangen, wobei

$$m = (72,154,31,139,106,22,146,117,49,56,73,145,66,164,189,195,96,120,24,110,37,194,58,185,101,122,84,186,182,29)$$

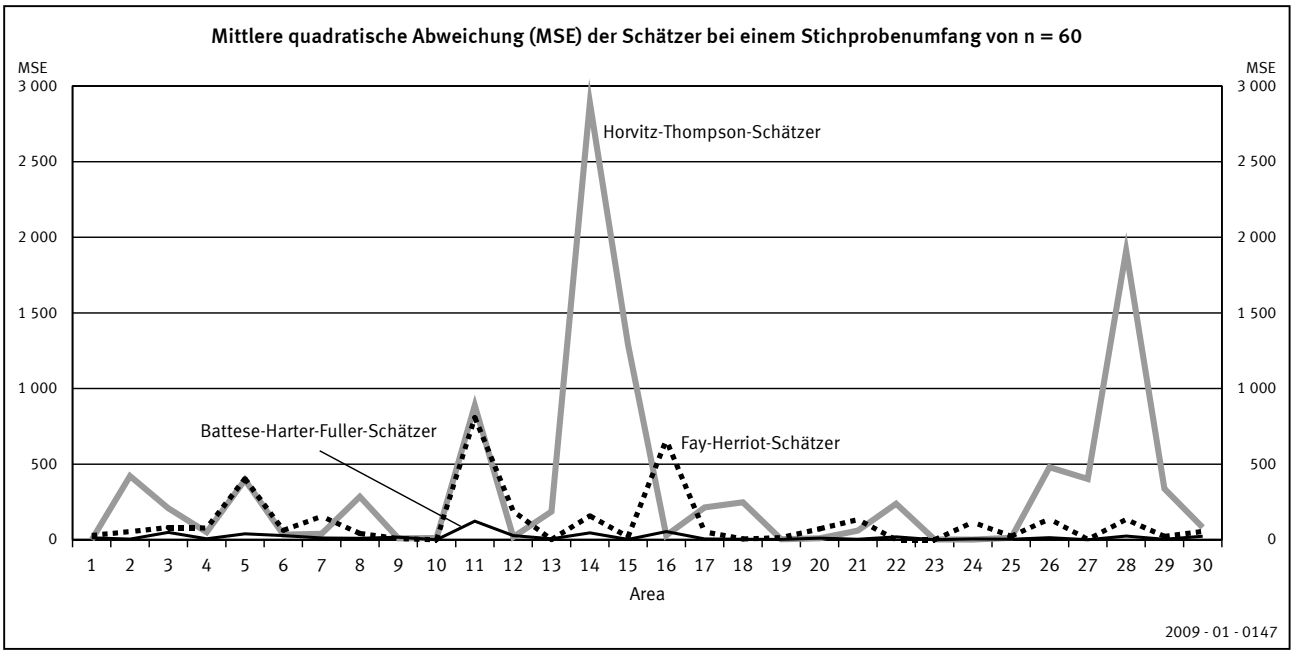
und n_i der Größe der i-ten Area entspricht. Außerdem sind die Individualfehlerkomponenten normalverteilt gemäß

$$e_{ij} \sim N(0,10), i = 1,2,\cdots,30, j = 1,2,\cdots,n_i$$

genau wie die Areafehlerkomponenten u_i

$$u_i \sim N(0,2), i = 1,2,\cdots,30.$$

Schaubild 2



Die abhängige Variable *y* wird konstruiert gemäß

$$y_{ij} = 5 + x_{ij} + u_i + e_{ij}\ , i = 1,2,\cdots,30,\ j = 1,2,\cdots,n_i\ .$$

Damit sind die Voraussetzungen des Battese-Harter-Fuller-Modells

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij}\ , i = 1,2,\cdots,30,\ j = 1,2,\cdots,n_i$$

erfüllt. Da der Fay-Herriot-Schätzer Informationen nur auf Area-Level benutzt, wird das arithmetische Mittel x_i bzw. y_i der Werte x_{ij} bzw. y_{ij} für jede Area berechnet und der Schätzer unter Verwendung nur dieser Daten konstruiert. Als Vergleichsmaßstab wird die mittlere quadratische Abweichung (Mean Squared Error – MSE) verwendet.

Die Schaubilder 1 und 2 zeigen die mittlere quadratische Abweichung des Horvitz-Thompson-, des Fay-Herriot- und des Battese-Harter-Fuller-Schätzers für die 30 Areas. Dabei ist zu erkennen, dass bei einer großen Stichprobengröße von n = 2 000 alle drei Schätzer gut abschneiden, während dies bei der kleinen Stichprobengröße (n = 60) nicht mehr der Fall ist. Auch bei weiteren Untersuchungen (in der diesem Beitrag zugrunde liegenden Diplomarbeit) erwiesen sich die Small-Area-Schätzer insgesamt als eine sehr gute, robuste Alternative.

4 Fazit

Zusammenfassend ist festzuhalten, dass die Methodik der Small-Area-Schätzungen bisher in Anwendungen – und hier insbesondere in der amtlichen Statistik – noch wenig verbreitet ist. Allerdings zeigen jüngere Entwicklungen, dass diese Methodik in naher Zukunft Einzug in einige statistische Ämter Europas halten wird. Auch in Deutschland wird diese Methodik, insbesondere im Rahmen des Zensus 2011, intensiv diskutiert. In der vorliegenden Arbeit wurden die beiden wesentlichen Schätzverfahren der Small-Area-Statistik eingehend von der statistisch-mathematischen Herkunft bis zur Anwendung untersucht. Zudem wurde eine vergleichende Überprüfung der Effizienz der Verfahren anhand einer Simulationsstudie durchgeführt. Hierbei zeigte sich, dass diese Schätzer eine sinnvolle Alternative zu klassischen Schätzverfahren darstellen und ihr Einsatz insbesondere bei erfüllten Modellannahmen und einer kleinen Stichprobe die Schätzergebnisse wesentlich verbessern kann. Bei nicht erfüllten Modellannahmen können die Small-Area-Schätzer jedoch verzerrte Schätzwerte liefern, sodass von einer naiven Anwendung der Small-Area-Modelle abzusehen ist.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Roderich Egeler
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden
- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:
- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- www.destatis.de/kontakt