Maria Dieterle, M. A.

Schätzung regionaler Daten mithilfe von Small Area-Schätzmethoden

1 Einleitung

Deutschland hat sich im Rahmen verschiedener internationaler Verträge wie dem Genfer Luftreinhalteabkommen und dessen acht Protokollen, der Klimarahmenkonvention der Vereinten Nationen und deren Nachfolgeprotokollen (zum Beispiel Kyoto-Protokoll) verpflichtet, regelmäßig über die Emissionen klimarelevanter Gase (Kohlendioxid, Lachgas, Methan) und anderer Luftschadstoffe (zum Beispiel Ammoniak) zu berichten. Die Emissionen aus der Landwirtschaft werden dabei vom Institut für Agrarrelevante Klimaforschung des Johann Heinrich von Thünen-Instituts (vTI-AK) in Zusammenarbeit mit dem Kuratorium für Technik und Bauwesen in der Landwirtschaft e. V. (KTBL) und dem Statistischen Bundesamt berechnet.

In das Emissionsmodell gehen unter anderem Tierbestände, insbesondere die Anzahl von Rindern und Schweinen, auf Kreisebene ein.[1] Die Verfügbarkeit dieser Daten auf Kreisebene nimmt aber tendenziell ab. Während die Viehbestände (Rinder, Schweine, Schafe) bis zum Jahr 1999 noch alle zwei Jahre total erhoben wurden, fanden danach nur noch in den Jahren 2001, 2003 und 2007 im Rahmen der Agrarstrukturerhebung sowie im Jahr 2010 im Rahmen der Landwirtschaftszählung totale Erhebungen der Viehbestände statt. Bei Rindern ist dies unproblematisch – die entsprechenden Merkmale werden seit dem Jahr 2008 der HIT-Datenbank (Herkunftssicherungs- und Informationssystem für Tiere) entnommen, in der alle Rinder erfasst sind. Merkmale zu Schweinebeständen hingegen werden nach der Landwirtschaftszählung 2010 erst wieder im Jahr 2016 total erhoben. In den Jahren zwischen den Totalerhebungen werden zweimal jährlich im Mai und November repräsentative Viehbestandserhebungen durchgeführt; diese sind für verlässliche freie Hochrechnungen auf Länderebene, nicht jedoch auf Kreisebene ausgelegt.

Da es für die Emissionsberechnungen wünschenswert wäre, auf Basis der repräsentativen Viehbestandserhebungen regelmäßig über kleinräumige Daten zu verfügen, untersucht dieser Beitrag am Beispiel der Schweinebestände die Möglichkeit, mithilfe von Small Area-Schätzverfahren verlässliche Kreisdaten zu berechnen.

2 Theoretische Grundlagen und Vorgehensweise

2.1 Übersicht über Small Area-Schätzverfahren

Den Begriff *Small Area* (auch *small domain*) definiert Rao[2] als inhaltlich oder geografisch abgegrenzte Subpopulation, für die eine direkte Schätzung[3] interessierender Merkmale aufgrund einer zu niedrigen Anzahl von Stichprobeneinheiten zu inakzeptabel hohen Standardfehlern führt. Der Grund

1 Die Rinder- und die Schweineproduktion sind für 86 % beziehungsweise 11 % der Treibhausgasemissionen und für etwa 45 % beziehungsweise 26 % der Ammoniak-Emissionen aus der landwirtschaftlichen Tierhaltung verantwortlich (berechnet mit Zahlen aus Rösemann, C., und andere: „Berechnung von gas- und partikelförmigen Emissionen aus der deutschen Landwirtschaft 1990 – 2009“, Landbauforschung, Sonderheft 342, hier: Seiten 366 und 369).

2 Rao, J. N. K.: “Small Area Estimation”, New Jersey 2003.

3 Eine direkte Schätzung nutzt nur Merkmalsausprägungen derjenigen Stichprobeneinheiten, die sich in der betreffenden Subpopulation befinden (siehe Fußnote 2, hier: Seite 9).

für die hohen Standardfehler ist meist, dass die Stichprobenerhebung auf gute Ergebnisse für eine übergeordnete Population abzielt und der Stichprobenplan somit nur eine ausreichend große Stichprobe auf der übergeordneten Ebene garantiert, während die Stichprobenumfänge in den Subpopulationen zufällig sind.[4] Um trotzdem Merkmale auf einer tiefer gegliederten Ebene schätzen zu können, wurden verschiedene Methoden entwickelt, die unter dem Begriff *Small Area Estimation* zusammengefasst werden. Diese Methoden bedienen sich meist sogenannter Hilfsinformationen (*auxiliary information*), um die effektive Stichprobe zu vergrößern und somit die Genauigkeit der Schätzung zu erhöhen. Mögliche Hilfsinformationen sind zum Beispiel Merkmalsausprägungen aus der Vergangenheit, Werte aus benachbarten oder übergeordneten Subpopulationen und/oder Werte von Hilfsvariablen, die stark mit dem interessierenden Merkmal korrelieren (siehe Fußnote 4). Methoden, die Hilfsinformationen verwenden, werden auch indirekte Schätzverfahren genannt.

Indirekte Schätzverfahren können in zwei Gruppen aufgeteilt werden, je nachdem, ob sie auf impliziten oder expliziten Modellen aufbauen. Zur ersten Gruppe gehören unter anderem synthetische und zusammengesetzte Schätzer. Synthetische Schätzer nutzen meist Informationen aus einer übergeordneten Einheit, zusammengesetzte Schätzer kombinieren die Informationen eines direkten und eines synthetischen Schätzers. Die expliziten Schätzverfahren basieren auf statistischen Regressionsmodellen, die je nach Verfügbarkeit Hilfsinformationen auf der Ebene der Erhebungseinheiten (*unit level model*) oder auf der Ebene der Subpopulation (*area level model*) verwenden und zusätzlich zu den erklärenden Variablen einen Fehlerterm für zufällige Unterschiede (*random effects*) zwischen den Subpopulationen beinhalten können.[5]

2.2 Methodik und Daten

Die Entscheidung für ein bestimmtes Small Area-Schätzverfahren hängt in erster Linie von den verfügbaren Hilfsinformationen und deren Qualität ab. Zudem spielen auch die Zielgruppe der Daten und deren Verwendung eine wichtige Rolle, da sie den Anspruch an die Genauigkeit der Schätzung bestimmen.[6] Die verfügbaren Ressourcen an Zeit und Know-how sind weitere Faktoren, die die Auswahl einer Methode beeinflussen können.

In diesem Beitrag wird ein zusammengesetzter Schätzer beschrieben und angewendet, mit dem die Schweinebestände auf Kreisebene im Jahr 2007 mithilfe von Ergebnissen aus der totalen Agrarstrukturerhebung aus dem Jahr 2003 geschätzt werden. Die Vorgehensweise ist wie folgt: Zunächst werden die Schweinebestände auf Kreisebene im Jahr 2007 anhand des Stichprobenmaterials der Agrarstrukturerhebung 2007 direkt geschätzt (freie Hochrechnung), um die Notwendigkeit einer Small Area-Schätzmethode zu bestätigen. Anschließend wird ein synthetischer Schätzer angewendet, der als Hilfsinformationen die Anteile der Schweinebestände je Kreis aus dem Totalmaterial der Agrarstrukturerhebung 2003 verwendet. Da die „wahren" Anteile im Jahr 2007 bekannt sind[7], kann die Verzerrung des synthetischen Schätzers und somit der mittlere quadratische Fehler (*mean squared error*, MSE) genau quantifiziert werden. Auf der Basis der mittleren quadratischen Fehler der direkten und der synthetischen Schätzer für die Schweinebestände auf Kreisebene werden anschließend die Gewichte für die zusammengesetzten Schätzer ermittelt, die eine Kombination der direkten und synthetischen Ergebnisse darstellen. Zum Abschluss werden die drei Schätzer (direkter, synthetischer und zusammengesetzter) anhand des mittleren quadratischen Fehlers verglichen und bewertet.[8]

Ein wichtiger Vorteil des gewählten Ansatzes liegt in der sofortigen Verfügbarkeit der Daten sowie in der Art der verfügbaren Daten (sowohl Stichproben- als auch Totalmaterial für das Jahr 2007). Für die Anwendung einer modellbasierten Methode wäre ein beträchtlicher Mehraufwand für Modellentwicklung, Gewinnung und Aufbereitung der Daten notwendig, ohne dass dies zwingend zu besseren Schätzergebnissen führen würde.

Aufgrund methodischer Veränderungen der Schweinebestandserhebung ab dem Jahr 2010 ist eine Übertragung der Ergebnisse dieses Beitrags auf künftige Erhebungen allerdings unsicher. Zwischen 1999 und 2009 wurde die Erhebung der Viehbestände im Mai zusammen mit der Bodennutzungshaupterhebung durchgeführt.[9] Dabei gab es eine gemeinsame Stichprobe für die beiden Erhebungen, was durch eine Schichtung nach Größenklassen der landwirtschaftlich genutzten Fläche einerseits und nach Produktionsrichtungen der Betriebe (zum Beispiel große Tierbestände) andererseits ermöglicht wurde. Seit dem Jahr 2010 wird eine separate Stichprobe für die Schweinebestandserhebung gezogen, in der nur Schweine haltende Betriebe enthalten sind. Zudem wurden die Erfassungsgrenzen erhöht (10 Zuchtsauen beziehungsweise 50 Schweine). Dieser Beitrag ist deshalb als Grundlage für eine erste Einschätzung der Methode und deren eventuelle Weiterverwendung und -entwicklung zu verstehen.

3 Anwendung von Small Area-Schätzverfahren für die Schätzung von Kreisdaten

3.1 Direkter Schätzer

Am 3. Mai 2007 erreichten 80 453 Schweine haltende Betriebe in 420 Landkreisen die Erfassungsgrenze der Agrar-

4 Siehe Münnich, R./Schmidt, K.: „Small Area Estimation in der Bevölkerungsstatistik", Statistisches Landesamt Baden-Württemberg (Herausgeber): Baden-Württemberg in Wort und Zahl, Heft 3/2002, Seite 139 ff.

5 Siehe Rao, J. N. K. (Fußnote 2), hier: Seite 75 ff.

6 Siehe Australian Bureau of Statistics (Herausgeber): "A Guide to Small Area Estimation – Version 1.1." im Internet unter *www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocument, Complete Dokument, May 2006* (abgerufen am 8. November 2011).

7 Für das Jahr 2007 liegt für die Viehbestände sowohl Stichproben- als auch Totalmaterial vor.

8 Alle Auswertungen wurden mit dem Programm SAS durchgeführt.

9 Siehe Statistisches Bundesamt (Herausgeber): „Methodische Grundlagen der Strukturerhebung in landwirtschaftlichen Betrieben 2007", im Internet unter *www.destatis.de*, im Bereich Publikationen → Fachveröffentlichungen → Land- und Forstwirtschaft → Landwirtschaftszählung.

strukturerhebung[10]. Davon waren 20 924 Betriebe in 399 Kreisen im Stichprobenmaterial enthalten. Für 21 Landkreise ohne Stichprobeneinheiten ist somit keine direkte Schätzung der Schweinebestände für diesen Zeitpunkt möglich (siehe zum Beispiel Landkreis J in Tabelle 1). Zumindest für diese Landkreise ist also ein alternatives Schätzverfahren notwendig.

Die landwirtschaftlichen Betriebe in der Stichprobe der Agrarstrukturerhebung 2007 sind nach Bundesländern und weiteren 26 Merkmalen geschichtet. Als Schichtungsmerkmale dienen unter anderem Größenklassen der landwirtschaftlich genutzten Fläche und die Produktionsschwerpunkte der Betriebe. Zudem wurde je Bundesland eine Schicht für Neuzugänge eingerichtet. Insgesamt gibt es somit 432 Schichten. Der direkte Schätzer (D) für die Schweinebestände auf Kreisebene ist daher[11]:

$$\hat{X}^{d}_{k2007} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in k} X_{hi}$$

mit

$k = 1, 2, ..., 399$: Landkreise

$h = 1, 2, ..., 432$: Schichten

N_h : Schichtumfang der Schicht h

n_h : Stichprobenumfang in Schicht h

x_{hi} (mit $i = 1, 2, ..., n_h$) : Schweinebestand der Stichprobeneinheit i in Schicht h und Landkreis k

$\frac{N_h}{n_h}$ entspricht dem Hochrechnungsfaktor in Schicht h.

Der Standardfehler des direkten Schätzers für die Schweinebestände auf Kreisebene lässt sich folgendermaßen schätzen[12]:

$$s_{\hat{X}^{d}_{k2007}} = \sqrt{\sum_{h=1}^{h} \frac{N_h}{n_h}\left(\frac{N_h}{n_h} - 1\right) s^2_{hx_{k2007}}}$$

$$\text{mit } s^2_{hx_{k2007}} = \frac{1}{n_h - 1}\left(\sum_{i \in k2007} x^2_{hi} - \frac{1}{n_h}\left(\sum_{i \in k2007} x_{hi}\right)^2\right)$$

$$\text{und } x_{hi} = \begin{cases} x_{hi} \text{für } i \in k2007 \\ 0 \qquad \text{sonst} \end{cases}$$

Der relative Standardfehler des direkten Schätzers wird wie folgt berechnet:

$$\text{RSTF(D)} = \frac{s_{\hat{X}^{d}_{k2007}}}{\hat{X}^{d}_{k2007}}$$

Tabelle 1 zeigt eine Auswahl der direkten Schätzer für die Schweinebestände auf Kreisebene, aufsteigend sortiert nach Höhe des relativen Standardfehlers. Dargestellt sind die tatsächliche Anzahl der Schweine im Kreis nach dem Totalmaterial der Agrarstrukturerhebung 2007, die Ergebnisse des direkten Schätzers aus dem Stichprobenmaterial der Agrarstrukturerhebung 2007, die relativen Standardfehler des direkten Schätzers sowie die Anzahl der im Totalmaterial und im Stichprobenmaterial enthaltenen Betriebe.

Tabelle 1 Ausgewählte Ergebnisse der direkten Schätzer des Schweinebestands 2007 auf Kreisebene

Landkreis	Schweinebestand 2007		Relativer Standardfehler des direkten Schätzers	Anzahl der Betriebe in der	
	Totalerhebung[1]	Direkter Schätzer[2]		Totalerhebung	Stichprobe
A	238	4	0,0	5	1
B	94 414	94 680	0,2	81	47
C	830 303	850 517	5,1	1 530	382
D	7 416	6 190	10,0	11	2
E	44 445	36 411	15,2	215	38
F	12 153	15 522	22,0	91	21
G	4 216	4 864	36,4	72	17
H	2 457	3 147	50,1	47	10
I	9	6	70,8	2	1
J	714	.	.	4	0

1 Ergebnis der Agrarstrukturerhebung 2007.
2 Anhand des Stichprobenmaterials der Agrarstrukturerhebung 2007.

Der relative Standardfehler des direkten Schätzers schwankt zwischen 0 % und 97 %, der Mittelwert liegt bei 20 %. In elf Kreisen beträgt der relative Standardfehler des direkten Schätzers null, was nur bei einem Hochrechnungsfaktor von eins in allen am Kreis beteiligten Schichten möglich ist, das heißt wenn alle Schichten Totalschichten[13] sind. Das kommt einerseits in Stadtstaaten vor, welche aufgrund ihrer Größe meist nur Totalschichten beinhalten. Aus diesem Grund ist das in diesem Artikel vorgestellte Schätzverfahren für Stadtstaaten uninteressant. Bei anderen Landkreisen kommt es aufgrund der gemeinsamen Stichprobe jedoch vor, dass landwirtschaftliche Betriebe aus Totalschichten enthalten sind, deren Hauptproduktionsrichtung nicht die Schweinehaltung ist (zum Beispiel Gartenbaubetriebe). In diesen Landkreisen wird der Standardfehler für Merkmale zu Schweinebeständen dadurch unterschätzt (siehe zum Beispiel Landkreis A in Tabelle 1). Beim zusammengesetzten Schätzer muss dies später berücksichtigt werden, sonst würde der direkte Schätzer für die Schweinebestände in diesen Landkreisen mit einem zu starken Gewicht in den zusammengesetzten Schätzer eingehen und die Ergebnisse verfälschen (siehe auch Abschnitt 3.3). Dieses Problem fällt allerdings ab 2010 mit der neuen Stichprobe für die Schweinebestandserhebung weg (siehe Kapitel 2).

10 Betriebe mit 2 Hektar und mehr landwirtschaftlich genutzter Fläche oder einer Mindestzahl an Nutztieren (zum Beispiel acht Rinder oder acht Schweine) oder einer Mindestfläche an Sonderkulturen, die zu Erwerbszwecken genutzt wird (zum Beispiel 30 Ar Obstfläche oder 30 Ar bestockte Rebfläche).

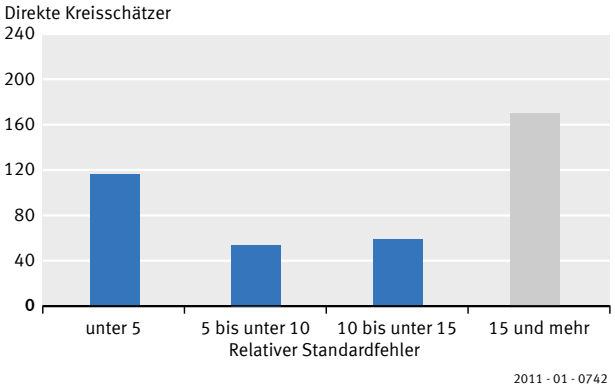
11 **Siehe Krug, W./Nourney, M./Schmidt, J.: „Wirtschafts- und Sozialstatistik: Gewinnung von Daten“, Seite 116, München 2001.**

12 Siehe Krug, W. und andere (Fußnote 11), hier: Seite 116 f.

13 Totalschichten sind Schichten, in denen alle Betriebe erfasst werden (Stichprobenumfang von 100 %). Es handelt sich dabei um Schichten mit wenigen Erhebungseinheiten, aus denen nur schwer eine repräsentative Stichprobe gezogen werden kann (Beispiel: Stadtstaaten), und/oder solche Schichten, die einen großen Einfluss auf das Ergebnis haben (zum Beispiel Betriebe mit großen Tierbeständen). Auch die Zugangsschicht, der neu erfasste Betriebe zugeordnet werden, ist als Totalschicht konzipiert (siehe auch Fußnote 9).

In Schaubild 1 ist die Häufigkeitsverteilung für den relativen Standardfehler der direkten Schätzer für die Schweinebestände auf Kreisebene zu sehen. In 116 Landkreisen liegt der relative Standardfehler unter 5 %, in 170 Landkreisen unter 10 % und in 229 Landkreisen unter 15 %. Dies zeigt, dass der direkte Schätzer durchaus gute bis sehr gute Ergebnisse für einen Teil der Landkreise liefert. In 170 Landkreisen betrug der relative Standardfehler allerdings 15 % oder mehr, sodass für diese Kreise keine Ergebnisse veröffentlicht werden können. Für 21 Landkreise ist wie schon erwähnt keine direkte Schätzung möglich. Das bedeutet, dass für 45 % der 420 Landkreise die Werte für den Schweinebestand unbekannt oder sehr unsicher sind.

Schaubild 1 Histogramm der relativen Standardfehler der direkten Kreisschätzer



3.2 Synthetischer Schätzer

In einem zweiten Schritt wird der Anteil der Schweine in den einzelnen Kreisen im Jahr 2003 als Hilfsinformation für die Schätzung der Schweinebestände auf Kreisebene im Jahr 2007 verwendet. Für Kreise mit Gebietsstandsänderungen wurden die Ergebnisse des Jahres 2003 mithilfe eines flächenbezogenen Umrechnungsfaktors des Bundesinstituts für Bau-, Stadt- und Raumforschung auf die Kreisgrenzen des Jahres 2007 umgerechnet. Den 434 Kreisen mit Schweine haltenden Betrieben im Jahr 2003 entsprachen umgerechnet 426 Landkreise in den Kreisgrenzen von 2007. Im Vergleich zu den 420 Landkreisen mit Schweine haltenden Betrieben nach dem Ergebnis der Agrarstrukturerhebung 2007 sind dies sechs Landkreise mehr – in diesen Landkreisen wurde die Schweinehaltung zwischen den Jahren 2003 und 2007 anscheinend aufgegeben.

Der Anteil der Schweine in einem Kreis an allen Schweinen im jeweiligen Bundesland im Jahr 2003 in den Kreisgrenzen von 2007 lässt sich ausdrücken als:

$$a_{k2003} = \frac{X_{k2003}}{X_{l2003}}$$

mit

$X_{k2003} = \sum_{i \in k2003} X_i$: Anzahl der Schweine am 3. Mai 2003 im Landkreis k, umgerechnet auf die Kreisgrenzen des Jahres 2007, mit $k = 1, \ldots, 426$,

und

$X_{l2003} = \sum_{i \in l2003} X_i$: Anzahl der Schweine am 3. Mai 2003 im Bundesland l, wobei $l = 1, \ldots, 16$

Der synthetische Schätzer (S) für die Schweinebestände auf Kreisebene ist wie folgt definiert:

$$\hat{X}^s_{k2007} = a_{k2003} \cdot \hat{X}^d_{l2007},$$

wobei $\hat{X}^d_{l2007} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i \in l2007} x_{hi}$ den direkten Schätzer der Schweinebestände auf Bundeslandebene aus dem Stichprobenmaterial der Agrarstrukturerhebung 2007 darstellt.

Der relative Standardfehler des synthetischen Schätzers für die Schweinebestände auf Kreisebene ist:

$$\mathrm{RSTF(S)} = \frac{s_{\hat{X}^s_{k2007}}}{\hat{X}^s_{k2007}}$$

Er entspricht dem relativen Standardfehler des direkten Schätzers der Schweinebestände in dem Bundesland, in welchem sich der Landkreis befindet (siehe Tabelle 2). Dies kann aus der Varianz des synthetischen Schätzers abgeleitet werden:

$$\mathrm{Var}\left(\hat{X}^s_{k2007}\right) = \mathrm{Var}\left(\hat{X}^d_{l2007} \cdot a_{k2003}\right) = {a_{k2003}}^2 \cdot \mathrm{Var}\left(\hat{X}^d_{l2007}\right)$$

$$\mathrm{RSTF(S)} = \frac{s_{\hat{X}^s_{k2007}}}{\hat{X}^s_{k2007}} = \frac{\sqrt{Var\left(\hat{X}^s_{k2007}\right)}}{\hat{X}^s_{k2007}} = \frac{\sqrt{{a_{k2003}}^2 \cdot Var\left(\hat{X}^d_{l2007}\right)}}{\hat{X}^s_{k2007}}$$

$$= \frac{a_{k2003} \cdot s_{\hat{X}^d_{l2007}}}{a_{k2003} \cdot \hat{X}^d_{l2007}} = \frac{s_{\hat{X}^d_{l2007}}}{\hat{X}^d_{l2007}}$$

Tabelle 2 Relative Standardfehler der direkten Schätzer des Schweinebestands 2007 auf Länderebene

	Relativer Standardfehler
Schleswig-Holstein	1,4
Hamburg	0,3
Niedersachsen	1,2
Bremen	0,0
Nordrhein-Westfalen	1,2
Hessen	1,0
Rheinland-Pfalz	2,8
Baden-Württemberg	1,5
Bayern	1,2
Saarland	5,7
Berlin	0,0
Brandenburg	0,2
Mecklenburg-Vorpommern	1,1
Sachsen	0,2
Sachsen-Anhalt	0,4
Thüringen	0,1

Die Varianz des synthetischen Schätzers für die Schweinebestände auf Kreisebene ist gering, dafür wird allerdings eine Verzerrung in Kauf genommen, die durch die Verwendung von Hilfsinformationen aus dem Jahr 2003 (Anteile der Schweinebestände je Kreis) entsteht.

Die Verzerrung ist definiert als die Abweichung des Erwartungswerts des Schätzers vom wahren Wert:

$$\text{Verzerrung}\left(\hat{X}^{s}_{k2007}\right) = \mathrm{E}\left(\hat{X}^{s}_{k2007}\right) - X_{k2007}$$
$$= \mathrm{E}\left(a_{k2003}\hat{X}^{d}_{l2007}\right) - X_{k2007} = a_{k2003}\,\mathrm{E}\left(\hat{X}^{d}_{l2007}\right) - X_{k2007}$$
$$= a_{k2003}X_{l2007} - X_{k2007} = a_{k2003}X_{l2007} - a_{k2007}X_{l2007}$$
$$= \left(a_{k2003} - a_{k2007}\right)X_{l2007}$$

Die Verzerrung entsteht hauptsächlich durch die Veränderung der Anteile (Anzahl der Schweine im Landkreis geteilt durch die Anzahl der Schweine im zugehörigen Bundesland) zwischen den Jahren 2003 und 2007, die im synthetischen Schätzer nicht berücksichtigt wird. Für die sechs Landkreise, in denen im Jahr 2003 zum Beispiel noch Schweine gehalten wurden, vier Jahre später nach dem Ergebnis der Agrarstrukturerhebung 2007 aber nicht mehr, weist der synthetische Schätzer auch für das Jahr 2007 noch Schweine aus. In den Landkreisen mit Gebietsstandsänderungen zwischen den Jahren 2003 und 2007 kommt eine zusätzliche Verzerrung durch die flächenproportionale Umrechnung der Schweinezahlen auf die Landkreisgrenzen des Jahres 2007 hinzu.

Die Genauigkeit eines verzerrten Schätzers wird durch den mittleren quadratischen Fehler (*mean squared error*, MSE) gemessen, der sich aus der Varianz (stichprobenbedingter Fehler) und der Verzerrung (systematischer Fehler) zusammensetzt[14]:

$$\text{MSE} = \text{Varianz} + \text{Verzerrung}^2$$

In Analogie zum relativen Standardfehler für unverzerrte Schätzer gibt es für verzerrte Schätzer die relative Wurzel aus dem mittleren quadratischen Fehler [*relative root mean squared error* (RRMSE)], welche folgendermaßen berechnet wird[15]:

$$\text{RRMSE}(S) = \frac{\sqrt{MSE_{\hat{X}^{s}_{k2007}}}}{\hat{X}^{s}_{k2007}}$$

Tabelle 3 ergänzt Tabelle 1 um die Ergebnisse des synthetischen Schätzers und deren relative Wurzel aus dem mittleren quadratischen Fehler. Landkreis K stellt einen der sechs Landkreise dar, in welchen zwar im Jahr 2003, nicht jedoch im Jahr 2007 Schweine haltende Betriebe existierten. Deshalb liefert der direkte Schätzer für diesen Landkreis kein Ergebnis, der synthetische Schätzer aufgrund seiner Verzerrung einen positiven Wert. Die relative Wurzel des mittleren quadratischen Fehlers des synthetischen Schätzers ist entsprechend hoch (100 %).

Die Ergebnisse zeigen beispielhaft, dass der synthetische Schätzer teilweise zu einer Verschlechterung und teilweise zu einer Verbesserung der Ergebnisse führt. Insbesondere bei hohen relativen Standardfehlern der direkten Schätzer scheint der synthetische Schätzer überlegen zu sein (zum Beispiel in den Landkreisen H und I). Allerdings gibt es auch hier starke Ausreißer: Der Maximalwert der relativen Wurzel aus dem mittleren quadratischen Fehler des synthetischen Schätzers liegt bei 191 %, der Durchschnitt ist allerdings niedriger als beim direkten Schätzer (18 %). Die relative Wurzel aus dem mittleren quadratischen Fehler des synthetischen Schätzers ist insbesondere in jenen Landkreisen hoch, in welchen keine direkten Schätzer zur Verfügung stehen (durchschnittlich 71 %).

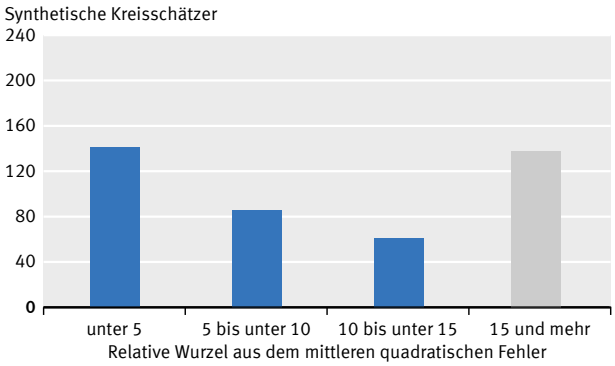
Tabelle 3 Ausgewählte Ergebnisse der direkten und synthetischen Schätzer des Schweinebestands 2007 auf Kreisebene

Landkreis	Schweinebestand 2007			Relativer Standardfehler des direkten Schätzers	Relative Wurzel aus dem mittleren quadratischen Fehler des synthetischen Schätzers
	Totalerhebung[1]	Direkter Schätzer[2]	Synthetischer Schätzer[3]		
A	238	4	250	0,0	2,8
B	94 414	94 680	95 669	0,2	1,4
C	830 303	850 518	803 528	5,1	3,7
D	7 416	6 190	8 260	10,0	11,4
E	44 445	36 410	38 758	15,2	12,1
F	12 153	15 521	13 790	22,0	14,6
G	4 216	4 863	5 565	36,4	24,2
H	2 457	3 147	2 605	50,1	8,4
I	9	6	11	70,8	19,4
J	714	.	1 269	.	43,2
K	0	.	2	.	100,2

1 Ergebnis der Agrarstrukturerhebung 2007.
2 Anhand des Stichprobenmaterials der Agrarstrukturerhebung 2007.
3 Unter Verwendung des Anteils der Schweine in einem Kreis an allen Schweinen im jeweiligen Bundesland 2003 als Hilfsinformation.

Schaubild 2 zeigt die Verteilung der relativen Wurzel aus dem mittleren quadratischen Fehler für den synthetischen Schätzer. In 141 Landkreisen liegt die relative Wurzel aus dem mittleren quadratischen Fehler unter 5 %, in 227 Landkreisen unter 10 % und in 288 Landkreisen unter 15 %. Dies stellt im Vergleich zum direkten Schätzer eine leichte Verdichtung der „besseren" Ergebnisse dar; der Anteil veröffentlichungswürdiger Ergebnisse, gemessen an einer relativen Wurzel aus dem mittleren quadratischen Fehler

Schaubild 2 Histogramm der relativen Wurzel aus dem mittleren quadratischen Fehler der synthetischen Kreisschätzer



14 Siehe Krug, W. und andere (Fußnote 11), hier: Seite 25. Der mittlere quadratische Fehler eines unverzerrten Schätzers entspricht also seiner Varianz.
15 Bei einem unverzerrten Schätzer sind der RRMSE und der relative Standardfehler gleich.

unter 15 %, ist auf 69 % (im Vergleich zu 55 % beim direkten Schätzer) gestiegen[16].

3.3 Zusammengesetzter Schätzer

Der zusammengesetzte Schätzer (Z) stellt eine Kombination aus dem direkten und dem synthetischen Schätzer dar[17]:

$$\hat{X}^z_{k2007} = w_k \cdot \hat{X}^d_{k2007} + (1-w_k)\cdot \hat{X}^s_{k2007}$$

Das Gewicht w_k legt dabei den Anteil des direkten Schätzers, das Gewicht $(1-w_k)$ den Anteil des synthetischen Schätzers am zusammengesetzten Schätzer fest. Das optimale Gewicht w^*_k kann durch die Minimierung des mittleren quadratischen Fehlers bestimmt werden[18] und entspricht ungefähr dem Anteil des mittleren quadratischen Fehlers des synthetischen Schätzers am Gesamtfehler der zwei Schätzer:

$$W^*_k = \frac{MSE\left(\hat{X}^s_{k2007}\right)}{MSE\left(\hat{X}^d_{k2007}\right) + MSE\left(\hat{X}^s_{k2007}\right)}$$

Je größer also der mittlere quadratische Fehler des synthetischen Schätzers im Vergleich zu jenem des direkten Schätzers ist, desto stärker geht der direkte Schätzer in das Schätzergebnis ein.

Der mittlere quadratische Fehler des zusammengesetzten Schätzers mit dem ungefähren optimalen Gewicht kann auch folgendermaßen geschrieben werden:

$$MSE\left(\hat{X}^z_{k2007}\right) = w^*_k MSE\left(\hat{X}^d_{k2007}\right) = \left(1-w^*_k\right) MSE\left(\hat{X}^s_{k2007}\right)$$

Bestenfalls kann der mittlere quadratische Fehler des direkten beziehungsweise des synthetischen Schätzers also halbiert werden.[19] Bei einem Gewicht von null oder eins stimmt der zusammengesetzte Schätzer mit dem jeweils präziseren Schätzer überein und der Informationsgewinn des zusammengesetzten Schätzers ist null.

Die Verzerrung des zusammengesetzten Schätzers kann wie folgt berechnet werden:

$$\begin{aligned}
&\text{Verzerrung} = \mathrm{E}\left(\hat{X}^z_{k2007}\right) - X_{k2007} \\
&= \mathrm{E}\left[w_k \cdot \hat{X}^d_{k2007} + (1-w_k)\cdot \hat{X}^s_{k2007}\right] - X_{k2007} \\
&= w_k \cdot X_{k2007} + (1-w_k)\cdot \mathrm{E}\left(\hat{X}^s_{k2007}\right) - X_{k2007} \\
&= w_k \cdot X_{k2007} + (1-w_k)\cdot a_{2003}\mathrm{E}\left(\hat{X}^d_{l2007}\right) - X_{k2007} \\
&= w_k \cdot X_{k2007} + (1-w_k)\cdot a_{2003} X_{l2007} - X_{k2007} \\
&= X_{l2007}\left[a_{2007}(w_k - 1) + (1-w_k)\cdot a_{2003}\right]
\end{aligned}$$

Tabelle 4 entspricht Tabelle 2 erweitert um die Ergebnisse des zusammengesetzten Schätzers und deren relative Wurzel aus dem mittleren quadratischen Fehler. Für Stadtstaaten wurde nur der direkte, für Landkreise ohne Stichprobeneinheiten nur der synthetische Schätzer verwendet. In Landkreisen mit einem relativen Standardfehler des direkten Schätzers von null, die nicht zu den Stadtstaaten gehören, wurde der synthetische Schätzer angewendet, um eine künstliche Unterschätzung des relativen Fehlers zu verhindern (siehe auch Abschnitt 3.1).

Die Ergebnisse zeigen beispielhaft, dass der zusammengesetzte Schätzer wie erwartet die größten Vorteile bringt, wenn keiner der zwei enthaltenen Schätzer stark dominiert (zum Beispiel im Landkreis E oder im Landkreis F). Wenn das Gewicht nahe bei 0 oder 1 liegt, kann der zusammengesetzte Schätzer keine signifikante Verbesserung des Ergebnisses herbeiführen.

Allerdings gibt es auch beim zusammengesetzten Schätzer starke Ausreißer, die maximale relative Wurzel aus dem mitt-

16 Der Anteil bezieht sich auf die 420 Landkreise mit tatsächlicher Schweinehaltung im Mai 2007.

17 Siehe Rao, J. N. K. (Fußnote 2), hier: Seite 57.

18 Siehe Rao, J. N. K. (Fußnote 2), hier: Seite 58.

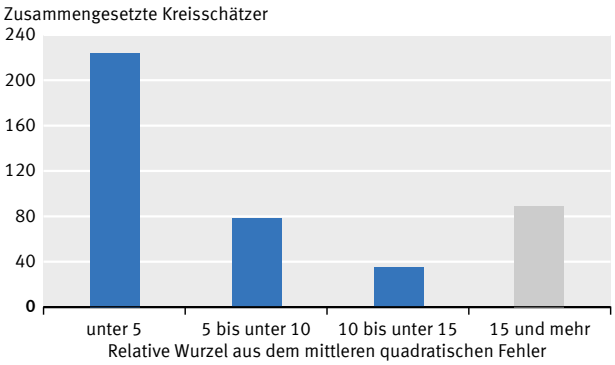
19 Siehe Rao, J. N. K. (Fußnote 2), hier: Seite 58.

Tabelle 4 Ausgewählte Ergebnisse der direkten, synthetischen und zusammengesetzten Schätzer des Schweinebestands 2007 auf Kreisebene

Landkreis	Schweinebestand 2007				Relativer Standardfehler des direkten Schätzers	Relative Wurzel aus dem mittleren quadratischen Fehler[5]		Anteiliges Gewicht des direkten Schätzers[6]
	Totalerhebung[1]	Direkter Schätzer[2]	Synthetischer Schätzer[3]	Zusammengesetzter Schätzer[4]		des synthetischen Schätzers	des zusammengesetzten Schätzers	
A	238	4	250,3	4	0,0	2,8	0,0	1,00
B	94 414	94 680	95 669	94 702	0,2	1,4	0,2	0,98
C	830 303	850 517	803 528	818 415	5,1	3,7	3,0	0,32
D	7 416	6 190	8 260	6 816	10,0	11,4	7,6	0,70
E	44 445	36 411	38 758	37 775	15,2	12,1	9,5	0,42
F	12 153	15 522	13 790	14 236	22,0	14,6	12,2	0,26
G	4 216	4 864	5 565	5 309	36,4	24,2	20,2	0,37
H	2 457	3 147	2 605	2 615	50,1	8,4	8,3	0,02
I	9	6	11	10	70,8	19,4	19,1	0,20
J	714	.	1 269	1 269	.	43,2	43,2	0,00
K	0	.	2	2	.	100,2	100,2	0,00

1 Ergebnis der Agrarstrukturerhebung 2007.
2 Anhand des Stichprobenmaterials der Agrarstrukturerhebung 2007.
3 Unter Verwendung des Anteils der Schweine in einem Kreis an allen Schweinen im jeweiligen Bundesland 2003 als Hilfsinformation.
4 Kombination aus dem direkten und dem synthetischen Schätzer.
5 Beim direkten Schätzer entspricht diese dem relativen Standardfehler.
6 Am zusammengesetzten Schätzer.

Schaubild 3 Histogramm der relativen Wurzel aus dem mittleren quadratischen Fehler der zusammengesetzten Kreisschätzer



leren quadratischen Fehler beträgt rund 184 %, der Durchschnitt liegt allerdings deutlich niedriger als beim direkten und synthetischen Schätzer, nämlich bei 12 %. Die Verteilung der relativen Wurzel aus dem mittleren quadratischen Fehler des zusammengesetzten Schätzers ist in Schaubild 3 zu sehen.

In 224 Landkreisen liegt die relative Wurzel aus dem mittleren quadratischen Fehler unter 5 %, in 302 Landkreisen unter 10 % und in 337 Landkreisen unter 15 %. Dies stellt eine deutliche Verbesserung im Vergleich zu den Ergebnissen des direkten und synthetischen Schätzers dar, da nun – gemessen an der 15 %-Grenze für die relative Wurzel aus dem mittleren quadratischen Fehler – für 80 % der 420 Landkreise mit Schweinehaltung im Mai 2007 Schätzwerte veröffentlicht werden könnten. Durch die Anwendung des zusammengesetzten Schätzers wurde somit die Anzahl veröffentlichungswürdiger Werte signifikant gesteigert. Die Zusammenfassung dieser Ergebnisse ist in Tabelle 5 zu sehen.

Tabelle 5 Vergleich der Schätzergebnisse nach der Höhe der relativen Wurzel aus dem mittleren quadratischen Fehler

Relative Wurzel aus dem mittleren quadratischen Fehler[1] kleiner als	Zahl der Kreise, für die das zutrifft, und Anteil an allen Kreisen[2]		
	beim direkten Schätzer	beim synthetischen Schätzer	beim zusammengesetzten Schätzer
5 %	116 (28 %)	141 (34 %)	224 (53 %)
10 %	170 (40 %)	227 (54 %)	302 (72 %)
15 %	229 (55 %)	288 (69 %)	337 (80 %)

1 Beim direkten Schätzer entspricht diese dem relativen Standardfehler.
2 Landkreise mit Schweinehaltung (420) im Mai 2007.

4 Fazit

Der Vergleich der direkten, synthetischen und zusammengesetzten Schätzerergebnisse zeigt, dass die Anzahl „guter" Ergebnisse (gemessen an einer relativen Wurzel aus dem mittleren quadratischen Fehler unter 15 %) durch die Anwendung des zusammengesetzten Schätzers signifikant gesteigert werden konnte. Allerdings bleibt auch beim zusammengesetzten Schätzer ein Teil der Resultate unbefriedigend. Für die Verwendung im Emissionsmodell könnten die Werte dennoch ausreichend genau sein, da das Modell selbst mit großen Unsicherheiten behaftet ist und sich die Ungenauigkeiten eines kleinen Teils der Kreiszahlen nicht signifikant auf die berechneten Emissionen auf Bundeslandebene auswirken dürften. Die Summe der Schweinebestände in den betroffenen Kreisen entspricht zudem nur ein bis zwei Prozent des Schweinebestands insgesamt in Deutschland.

Um die Methode weiterentwickeln zu können, müssen insbesondere folgende Fragen untersucht werden: Lassen sich die Ergebnisse auf die Schweinebestandserhebung ab dem Jahr 2010 (neue Erfassungsgrenze, neuer Stichprobenplan) übertragen? Wie kann der mittlere quadratische Fehler des synthetischen Schätzers in Zukunft ohne die Informationen aus einer Totalerhebung ermittelt werden? Sind die Schätzergebnisse zeitlich konsistent und kohärent? Kann die Genauigkeit der Schätzmethode durch das Heranziehen von Daten aus Registern, zum Beispiel der HIT-Datenbank[20], erhöht werden? Dies wäre der Fall, wenn die Veränderung der Anteile der Schweine je Kreis mit solchen Daten näherungsweise abgebildet und die Verzerrung des synthetischen und zusammengesetzten Schätzers dadurch deutlich reduziert werden könnte.

20 In dieser Datenbank werden seit dem Jahr 2008 alle Meldungen zur Zahl der Schweine erfasst.

Auszug aus Wirtschaft und Statistik

Herausgeber
Statistisches Bundesamt, Wiesbaden

www.destatis.de

Schriftleitung
Roderich Egeler, Präsident des Statistischen Bundesamtes

Brigitte Reimann (verantwortlich für den Inhalt)
Telefon: +49 (0) 6 11 / 75 20 86

Ihr Kontakt zu uns
www.destatis.de/kontakt

Statistischer Informationsservice
Telefon: +49 (0) 6 11 / 75 24 05
Telefax: +49 (0) 6 11 / 75 33 30

Abkürzungen

WiSta	=	Wirtschaft und Statistik
MD	=	Monatsdurchschnitt
VjD	=	Vierteljahresdurchschnitt
HjD	=	Halbjahresdurchschnitt
JD	=	Jahresdurchschnitt
D	=	Durchschnitt (bei nicht addierfähigen Größen)
Vj	=	Vierteljahr
Hj	=	Halbjahr
a. n. g.	=	anderweitig nicht genannt
o. a. S.	=	ohne ausgeprägten Schwerpunkt
St	=	Stück
Mill.	=	Million
Mrd.	=	Milliarde

Zeichenerklärung

p	=	vorläufige Zahl
r	=	berichtigte Zahl
s	=	geschätzte Zahl
–	=	nichts vorhanden
0	=	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
.	=	Zahlenwert unbekannt oder geheim zu halten
. . .	=	Angabe fällt später an
X	=	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
I oder —	=	grundsätzliche Änderung innerhalb einer Reihe, die den zeitlichen Vergleich beeinträchtigt
/	=	keine Angaben, da Zahlenwert nicht sicher genug
( )	=	Aussagewert eingeschränkt, da der Zahlenwert statistisch relativ unsicher ist

Abweichungen in den Summen ergeben sich durch Runden der Zahlen.