Klingwort, Jonas

Article

Die Verwendung von Straßensensoren und Capture-recapture-Techniken zur Messfehlerkorrektur in Surveys

WISTA – Wirtschaft und Statistik

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

Suggested Citation: Klingwort, Jonas (2021) : Die Verwendung von Straßensensoren und Capture-recapture-Techniken zur Messfehlerkorrektur in Surveys, WISTA – Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 73, Iss. 1, pp. 49-58

This Version is available at:
https://hdl.handle.net/10419/230954

DIE VERWENDUNG VON STRASSENSENSOREN UND CAPTURE-RECAPTURE-TECHNIKEN ZUR MESS-FEHLERKORREKTUR IN SURVEYS

Dr. Jonas Klingwort

↘ **Schlüsselwörter:** Big data – Registerdaten – Record Linkage – Unterberichterstattung – Datenvalidierung

ZUSAMMENFASSUNG

Dieser Artikel stellt eine Methode vor, welche Survey-, Sensor- und administrative Daten verknüpft und unter Anwendung von Capture-recapture-Techniken (CRC) die Korrektur von Messfehlern in Surveys ermöglicht. Dazu werden die Antworten des niederländischen Surveys zu Güterverkehr mit externen Sensormessungen verknüpft. Mittels CRC-Techniken wird die Unterberichterstattung geschätzt, wobei administrative Daten Informationen über Fahrzeuge und Halter liefern, die zur Modellierung der Heterogenität in den Beobachtungen verwendet werden. Die Ergebnisse zeigen, dass die Surveyschätzungen aufgrund von Unterberichterstattung negativ verzerrt sind.

↘ ***Keywords:*** *big data – register data – record linkage – underreporting – data validation*

ABSTRACT

This article presents a method that links survey, sensor, and administrative data and, by using capture-recapture techniques (CRC), allows to correct measurement errors in surveys. To this end, the responses to the Dutch Road Freight Transport survey are linked to records from a road sensor network. Using CRC techniques, underreporting is estimated, with administrative data providing information on vehicles and owners which is used to model the heterogeneity in the observations. The results show that the survey estimates are downward biased due to underreporting.

Dr. Jonas Klingwort

war in seiner Zeit als Doktorand als wissenschaftlicher Mitarbeiter in der Research Methodology Group der Universität Duisburg-Essen und als Statistiker bei Statistics Netherlands (CBS) tätig, unter anderem im Center for Big Data Statistics. Gegenwärtig arbeitet er als Methodiker beim CBS in der Abteilung „Methodology Research & Development“. Seine Doktorarbeit mit dem Titel „Correcting Survey Measurement Error With Big Data from Road Sensors Through Capture-recapture“ wurde unter der Betreuung von Prof. Dr. Rainer Schnell, Dr. Bart Buelens und Dr. Joep Burger in einem internationalen Kooperationsprojekt zwischen den beiden Institutionen angefertigt. Für seine Dissertation, aus der er Auszüge im vorliegenden Artikel vorstellt, wurde er 2020 mit dem Gerhard-Fürst-Preis des Statistischen Bundesamtes ausgezeichnet.

1

Einleitung

Die Verwendung von Daten, die nicht auf Zufallsstichproben basieren, wird in der amtlichen Statistik immer relevanter. Ein Grund dafür ist, dass Statistikbehörden mit der Nachfrage nach günstigeren, detaillierten und schnelleren Statistiken konfrontiert sind, und die Belastung der Befragten möglichst reduziert werden soll (Daas und andere, 2015). Big Data wird das Potenzial zugesprochen, diesen Ansprüchen gerecht werden zu können. Diese Daten werden in der Regel von automatisierten Systemen oder Sensoren aufgezeichnet, wobei es sich um beliebige Geräte handeln kann, die Informationen über physikalische Elemente und menschliches Verhalten erfassen (Shekhar, 2009). Dabei sind die datengenerierenden Mechanismen meist unbekannt, sodass die Daten nicht uneingeschränkt zur Produktion amtlicher Statistiken verwendet werden können. Genauer gesagt ist aufgrund des fehlenden Stichprobendesigns kein design-basierter Inferenzschluss möglich. Weiterhin enthalten die Datenbestände häufig nur undefinierte Subpopulationen oder nur wenige Ziel- oder Hilfsvariablen (Buelens und andere, 2014; Schnell, 2019). Ein vielversprechender Ansatz, Big Data in der amtlichen Statistik zu verwenden, ist die Kombination dieser Daten mit bestehenden Datenbeständen (Citro, 2014; Lohr/Raghunathan, 2017). Dabei ist zur Verknüpfung mehrerer Datensätze Record-Linkage auf Mikroebene ein zentrales Verfahren (Schnell, 2016). Wenn ein Sensor und ein Survey unabhängig voneinander eine identische Zielvariable messen, durch einen eindeutigen Identifikator verknüpft und mit Registerdaten angereichert werden können, wird ein maximaler Informationsgewinn erzielt (Japec und andere, 2015). In diesem Fall kann der Begriff Big Data zu Identifiable Big Data erweitert werden (Shlomo/Goldstein, 2015). Allerdings sind empirische Studien notwendig, um den Nutzen der Verknüpfung von Surveydaten mit Big Data und Registerdaten zu evaluieren.

In diesem Artikel werden das niederländische Straßengüterverkehr Survey, Weigh-in-Motion-Sensordaten (WiM), das niederländische Fahrzeugregister und das niederländische Unternehmensregister auf Mikroebene verknüpft. Alle Datenbestände stammen aus dem Jahr 2015. Aufgrund von Unterberichterstattung ist zu erwarten, dass die Surveyschätzungen der betrachteten Zielvariablen negativ verzerrt sind, das heißt, es findet eine Unterschätzung der Zielgröße statt. Unter Verwendung der WiM-Daten wird das Ausmaß der Verzerrung quantifiziert und korrigiert. Die Korrektur basiert auf einer Anwendung von Capture-recapture-Techniken (CRC). Diese Techniken wurden ursprünglich in der Ökologie und Biologie entwickelt, um (unbekannte) Populationsgrößen zu schätzen. Die Survey- und Sensorbeobachtungen werden dabei als zwei unabhängige Erfassungen betrachtet. Die Register liefern Kovariaten zur Modellierung der Erfassungswahrscheinlichkeiten in Survey und Sensoren. Diese Arbeit ist eine aktuelle empirische Studie und ein neues Beispiel für multisource statistics (de Waal und andere, 2017) in der amtlichen Statistik.[1]

2

Hintergrund

Zeitbasierte Tagebuchbefragungen zu Verkehr und Mobilität sind mit einem hohen Bearbeitungs- beziehungsweise Beantwortungsaufwand verbunden. Um den Aufwand zu reduzieren, antworten die Befragten ungenau oder gar nicht. Folglich erzielen diese Surveys oft niedrige Rücklaufquoten und unterschätzen die interessierende(n) Variable(n) (Richardson und andere, 1996; Meyburg/Rahman, 2003; Krishnamurty, 2008). Bisher wurden diese Surveys unter Verwendung von GPS-Empfängern oder Mobiltelefonen validiert, um das Ausmaß der Unterberichterstattung abzuschätzen (Wang und andere, 2018).

Pearson (2001) berichtete für die erste GPS-basierte Mobilitätsbefragung (1997 in den Vereinigten Staaten durchgeführt) eine Unterberichterstattung von bis zu 31 %. Für das California Statewide Household Travel Survey berichteten Wolf und andere (2003) eine Unterberichterstattung von bis zu 42 %. Bricka/Bhat (2006) dokumentierten bei Vergleichen mehrerer GPS-basierter Mobilitätsbefragungen in den Vereinigten Staaten Untererfassungen von Fahrten zwischen 11 und 81 %. Stopher/Greaves (2007) berichteten für das Sydney Household Travel Survey (2004) eine Unterberichterstattung von 7 %.

1 In diesem Beitrag werden nur notwendige technische Details erläutert. Sämtliche Details finden sich bei Klingwort und andere (2019), Klingwort (2020) sowie Klingwort und andere (erwartet 2021).

Die Ergebnisse früherer Studien im Bereich von Mobilitätsbefragungen zeigen zum einen, dass es Hinweise auf Unterberichterstattung gibt, und zum anderen, dass das Ausmaß sowohl innerhalb als auch zwischen den berichteten Ergebnissen variiert. Zu den häufig auftretenden Problemen in diesen Studien zählen technische Probleme mit GPS-Geräten und unterschiedliche Datenqualität zwischen den GPS-Gerätetypen (Sun und andere, 2017). Weiterhin berichten Bricka und andere (2012) und Shen/Stopher (2014) von Schwierigkeiten beim Abgleich von aufgezeichneten und berichteten Daten.

In dieser Arbeit werden anstelle von mobilen GPS-Empfängern fest installierte Straßensensoren verwendet, um das Ausmaß der Unterberichterstattung in den Surveyschätzungen zu quantifizieren und zu korrigieren.

3

Daten

3.1 Surveydaten

Der niederländische Survey zu Straßengüterverkehr wird von Statistics Netherlands durchgeführt. Ein zentrales Ziel des Surveys ist die Schätzung des gesamten transportierten Sendungsgewichts (*W*), das von niederländischen Nutzfahrzeugen transportiert wird. Weiterhin wird die Gesamtzahl der Fahrzeugtage (*D*) in dieser Studie als zusätzliche Zielvariable analysiert (keine Messfehlerkorrekturen erforderlich, siehe Abschnitt 3.2). Ein Fahrzeugtag *D* ist definiert als ein Tag, an dem ein Fahrzeug in den Niederlanden auf der Straße war.

Die Zielpopulation ist die niederländische Nutzfahrzeugflotte, wobei militärische und landwirtschaftliche Fahrzeuge sowie Fahrzeuge, die älter als 25 Jahre sind, ausgeschlossen werden. Außerdem werden nur Fahrzeuge mit einem Gewicht von mindestens 3,5 t (Fahrzeugleergewicht + Zuladung) berücksichtigt (Centraal Bureau voor de Statistiek, 2017). Die Population besteht aus etwa 135 000 Kennzeichen und wird vierteljährlich aktualisiert.

In jedem Quartal wurde eine geschichtete Zufallsstichprobe gezogen, die im Jahr 2015 insgesamt 33 817 eindeutige Fahrzeug-Wochen-Kombinationen umfasste, das heißt im Durchschnitt etwa 650 Fahrzeuge pro Woche. Die Fahrzeughalter müssen für eine zugewiesene Woche angeben, an welchen Tagen das Fahrzeug genutzt und wie viel Sendungsgewicht transportiert wurde. Keine Angabe ist erforderlich, wenn das Fahrzeug nicht zu Transportzwecken verwendet wurde.

Insgesamt wurden 22 454 Fahrzeuge (66 %) als an mindestens einem Tag in der zugewiesenen Woche genutzt gemeldet. In der zugewiesenen Woche nicht genutzt wurden 5 304 Fahrzeuge (16 %), 2 462 Fahrzeuge (7 %) waren nicht mehr im Besitz des Halters und in 3 597 Fällen (11 %) erfolgte keine Meldung (Nonresponse).

Die Option anzugeben, dass das Fahrzeug nicht genutzt wurde, reduziert die Belastung des Befragten erheblich, da in diesem Fall nur wenige Teile des Fragebogens beantwortet werden müssen. Dies ist die erwartete Hauptursache für die Unterberichterstattung. Eine weitere Möglichkeit, um mit minimaler Belastung zu antworten, ist, nur einen einzigen Tag zu melden. Der CRC-Ansatz ermöglicht es, den Effekt dieser Antwortstrategien auf die Surveyschätzungen zu quantifizieren und zu korrigieren.

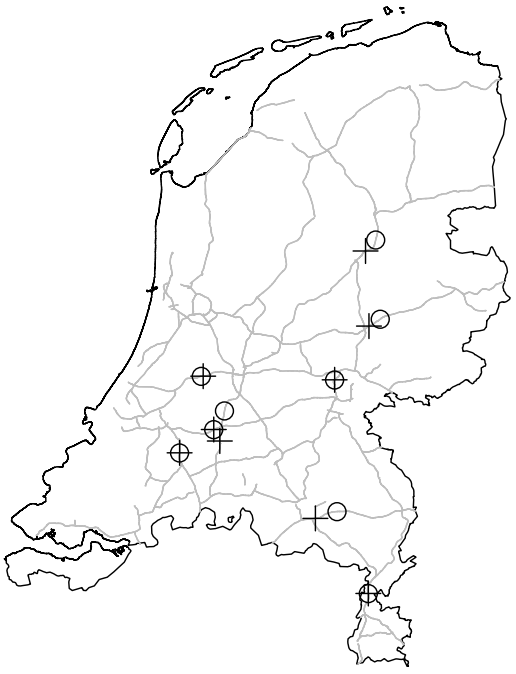
3.2 Sensordaten

Das niederländische Weigh-in-Motion-Straßensensornetzwerk dient der Erkennung von strafbar überladenen Transport- und anderen Schwertransportfahrzeugen sowie der Durchsetzung von Strafen (Federal Highway Administration, 2007). Im Jahr 2015 waren neun Sensorsysteme, in beide Richtungen messend, auf niederländischen Autobahnen im Einsatz, daraus resultierten 18 Messpunkte. ↘ Grafik 1 Wenn ein Fahrzeug eine Station passiert, wird es gewogen, klassifiziert, es werden Fotos der Kennzeichen gemacht und ein Zeitstempel wird gespeichert. Die letzten beiden Merkmale ermöglichen die Verknüpfung mit den Surveybeobachtungen und zusätzlichen Registerinformationen. Die einzelnen Achsgewichte eines Fahrzeugs werden gemessen und zum Gesamtgewicht aufaddiert.

Die notwendigen Datenbereinigungen wurden anhand der Richtlinien von Enright/OBrien (2011) und Experteninformationen der niederländischen Straßenverwaltung durchgeführt. Das transportierte Sendungsgewicht be-

Grafik 1
Sensorsysteme auf niederländischen Autobahnen

Sensornetzwerk aus 9 verschiedenen Systemen mit 18 Messstationen



Beispiel der WiM-Software, die einen vorbeifahrenden Lkw aufzeichnet

2020 - 0048

rechnet sich durch Subtraktion des Leergewichts vom gemessenen Gesamtgewicht. Informationen zum Anhängergewicht konnten nicht in allen Fällen verknüpft werden. Das hintere Nummernschild wurde von der OCR-Software (Optical Character Recognition) in 11 340 Fällen nicht erkannt, oder der Anhänger war in 5 980 Fällen nicht im Fahrzeugregister eingetragen. Dabei handelte es sich um nicht in den Niederlanden gemeldete Anhänger. Waren das vordere und das hintere Kennzeichen identisch, wurde kein Anhänger gezogen und das Anhängergewicht wurde auf 0 t gesetzt. Um die verbleibenden fehlenden Werte zu ersetzen, wurde eine Mittelwert-Imputation angewendet.

3.3 Verknüpfung von Survey- und Sensordaten

Die Verknüpfung der Survey- und Sensordaten erfolgte über einen eindeutigen Identifikator, der aus Nummernschild und Datum besteht. Die resultierende Kontingenztabelle zeigt ↘ Tabelle 1 für die Variable *D* (Tag) und ↘ Tabelle 2 für Variable *W* (Gewicht). Die Anzahl für *D* und *W*, die weder im Survey gemeldet noch von den Sensoren erfasst wurde, ist unbekannt und wird jeweils mit dem CRC-Schätzer geschätzt. Die Zeilensummen der Angaben im Survey entsprechen den ungewichteten Surveyschätzungen.

Tabelle 1
Erfassungen von Fahrzeugtagen[1] in Survey- und Sensordaten

Surveydaten	Sensordaten		Summe
	erfasst	nicht erfasst	
gemeldet	34 284	60 522	94 806
nicht gemeldet	9 727	?	?
Summe	44 011	?	?

1 Fahrzeugtage: Variable D.

Tabelle 2
Erfassungen von transportiertem Gewicht | 1 in Survey- und Sensordaten

Surveydaten	Sensordaten		Summe
	erfasst	nicht erfasst	
gemeldet	591	879	1 470
nicht gemeldet	139	?	?
Summe	730	?	?

1 Transportiertes Gewicht: Variable W in Kilotonnen (kt).

3.4 Capture-recapture-Annahmen

Ursprünglich wurde Capture-recapture entwickelt, um die (unbekannte) Größe einer Tierpopulation zu schätzen (International Working Group for Disease Monitoring and Forecasting, 1995). Erste Anwendungen finden sich bei Graunt im 16. Jahrhundert und bei Laplace 1802 zur Schätzung der Populationsgröße in England beziehungsweise Frankreich (Hald, 1990; Stigler, 1986; Cochran, 1978). Damit CRC-Schätzungen unverzerrt sind, müssen die im Folgenden erläuterten Annahmen erfüllt sein.

› Die Annahme der Unabhängigkeit zwischen den Datensätzen: Diese Annahme betrifft die Erfassungswahrscheinlichkeit von Elementen in den verschiedenen Datenquellen, welche unabhängig voneinander sein müssen. Bei den verwendeten Survey- und Sensordatensätzen ist diese Annahme erfüllt. Dass ein Fahrzeug in die Stichprobe gelangt, ist unabhängig von einer möglichen Erfassung in den Sensoren.

› Die Annahme einer offenen oder geschlossenen Population: In der vorliegenden Studie wird die Population als geschlossen betrachtet. Das heißt, es gibt keine Elemente, welche die Population verlassen und es werden keine weiteren Elemente hinzugefügt. Dadurch, dass die Elemente im Survey als Studienpopulation definiert sind, können keine neu zugelassenen Fahrzeuge in die gezogene Stichprobe eintreten. Es ist jedoch möglich, dass Fahrzeuge während eines Stichprobenquartals aus der Population ausscheiden. Dies ist bei Fahrzeugen, die zum Ende eines Quartals berichten müssen, wahrscheinlicher, da seit der Ziehung der Stichprobe mehr Zeit verstrichen ist. Dieser Fall dürfte sehr selten eintreten, sodass die Annahme einer geschlossenen Population als erfüllt betrachtet wird.

› Die Elemente in den Datenquellen müssen zur interessierenden Population gehören. Diese Annahme ist erfüllt, da die in der Stichprobe erfassten Fahrzeuge per Definition zur Grundgesamtheit gehören und die von den Sensoren erkannten Fahrzeuge durch die Verknüpfung ihres Kennzeichens mit dem Register selektiert werden.

› Die Antwort- oder Erfassungswahrscheinlichkeiten für die Elemente sollten in mindestens einer Datenquelle homogen sein (Zwane/van der Heijden, 2004; van der Heijden und andere, 2017). Diese Annahme wird durch die Modellierung von Erfassungswahrscheinlichkeiten als Funktion von Hilfsinformationen erfüllt (siehe Kapitel 4).

› Die Annahme der perfekten beziehungsweise eindeutigen Verknüpfung der Elemente: Diese Annahme kann in beiden Datenquellen aus unterschiedlichen Gründen gelegentlich verletzt werden. Die Fahrzeugbesitzer melden möglicherweise zu wenige, zu viele oder die falschen Daten. Weiterhin müssen die Halter den Tag der Beladung melden, während die Sensoren den Tag der Fahrt erfassen. Die Sensoren erfassen Fahrzeuge, auch dann, wenn diese nicht zu Transportzwecken genutzt werden. Weiterhin erfassen die Kameras nicht immer ein Kennzeichen, sodass eine Verknüpfung nicht möglich ist. Diese potenziellen Verletzungen der fünften Annahme werden in diesem Beitrag nicht behandelt. Details dazu finden sich bei Klingwort und andere (erwartet 2021).

3.5 Registerdaten

Zur Modellierung der Erfassungswahrscheinlichkeiten werden Variablen aus dem Fahrzeug- und Unternehmensregister verwendet. Das Fahrzeugregister liefert sowohl technische als auch nicht technische Fahrzeugmerkmale. Das Unternehmensregister liefert Merkmale über den Fahrzeughalter. Die Registerdaten werden auf Mikroebene verknüpft, unter Verwendung eines eindeutigen Identifikators, bestehend aus Kennzeichen und Quartal.

4

Methodik

Die Unterberichterstattung wird geschätzt, indem die für selektive Ausfälle korrigierten Surveyschätzungen mit CRC-Schätzungen verglichen werden. Die CRC-Schätzungen korrigieren für selektive Ausfälle und Messfehler. Die Unterberichterstattung im Survey wird geschätzt als die relative Differenz RD zwischen der Surveyschätzung ($\hat{Y}^{SVY}$) und der CRC-Schätzung ($\hat{Y}^{CRC}$), wobei die CRC-Schätzung als Vergleichswert verwendet wird:

$$RD = \frac{\hat{Y}^{SVY}}{\hat{Y}^{CRC}} - 1 \quad (1)$$

In diesem Artikel wird ein CRC-Schätzer, der auf einem log-linearen Modell basiert, verwendet. Die Verwendung log-linearer Modelle zur Schätzung der Populationsgröße in geschlossenen Populationen geht auf Fienberg (1972) zurück. Das log-lineare Modell wird mit Hilfsinformationen aus dem Register erweitert, um die Heterogenität in den Erfassungswahrscheinlichkeiten in beiden Datensätzen zu modellieren. Weitere CRC-Schätzer für diese Anwendung, wie der Lincoln-Petersen Schätzer (Petersen, 1894; Lincoln, 1935) oder der von Huggins (1989) und Alho (1990) vorgeschlagene Schätzer finden sich bei Klingwort (2020).

4.1 Modellselektion

Um ein optimales Modell für den CRC-Schätzer zu wählen, wurde ein schrittweises Auswahlverfahren auf Basis des BIC (Bayesian information criterion – ein Kriterium zur Modellauswahl) verwendet (Burnham/Anderson, 2004). Es wurden nur Haupteffekte der Variablen berücksichtigt. Die Modellauswahl für das log-lineare Modell basiert auf Logit-Modellen, da in log-linearen Modellen nur kategoriale Variablen verwendet werden können. Unter Verwendung des von Huggins (1989) und Alho (1990) vorgeschlagenen Ansatzes wurden zwei unabhängige Logit-Modelle berechnet, mit welchen die Erfassungswahrscheinlichkeiten der Fahrzeuge im Survey und in den Sensoren berechnet wurden. Für das log-lineare Modell wurden die fünf Variablen mit der größten Erklärungskraft in den beiden Logit-Modellen verwendet. Die gewählten Variablen sind die maximale Masse des gezogenen Anhängers, die Provinz des Fahrzeugbesitzers, die Klassifizierung der wirtschaftlichen Aktivität, das Baujahr des Fahrzeugs, der Fahrzeugtyp, die Größe der Fahrzeugflotte und die Leistung (kW). Zusätzlich wurden die Anzahl der verwendeten Sensorsysteme und ein Wochenendindikator in das Modell aufgenommen.

4.2 Varianzschätzung

Um die Genauigkeit von $\hat{Y}^{SVY}$ und $\hat{Y}^{CRC}$ zu schätzen, wurde Perzentil-Bootstrapping angewendet. Es wurden B = 3 000 Stichproben mit Zurücklegen gezogen, jede mit der Größe des ursprünglichen Datensatzes. Die 2,5- und 97,5-Perzentile der Bootstrap-Schätzungen wurden verwendet, um das 95-%-Konfidenzintervall zu schätzen.

5

Ergebnisse

Die relative Differenz zwischen der design-basierten Surveyschätzung und der CRC-Schätzung beträgt sowohl für *D* als auch für *W* etwa 18 % (siehe ↘ Tabelle 3). Die Differenz zwischen dem ungewichteten und gewichteten Surveyschätzer beträgt für *D* etwa 8 % (94 806 in Tabelle 1) und etwa 2 % für W (1 470 Kilotonnen [kt] in Tabelle 2). Da die CRC-Schätzer für Nonresponse und Messfehler korrigieren, ist die wahrscheinlichste Erklä-

Tabelle 3

Survey- und CRC-Schätzungen für die Variablen D und W (in kt), Bootstrapped Mittelwert, Standardfehler, Konfidenzintervall und Unterberichterstattung

Schätzer	Punktschätzung	Bootstrap-Mittelwert	Bootstrap-Standardfehler	Bootstrap-95 % Konfidenzintervall	Geschätzte Unterberichterstattung in %	Bootstrap 95 % Konfidenzintervall
$\hat{D}^{SVY}$	102 273	102 266	408	[101 474, 103 059]	– 18,4	[– 19,2, – 17,6]
$\hat{D}^{CRC}$	125 327	125 350	619	[124 125, 126 572]		
$\hat{W}^{SVY}$	1 499	1 499	9,4	[1 481, 1 518]	– 18,3	[– 19,3, – 17,4]
$\hat{W}^{CRC}$	1 835	1 835	11	[1 815, 1 857]		

rung für die Differenz, dass die Unterschätzung im Survey auf Unterberichterstattung zurückzuführen ist. Alternative Erklärungen untersuchen Klingwort und andere (erwartet 2021).

6

Diskussion und Fazit

Die hier vorgestellte Studie ist die erste empirische Studie, die CRC-Techniken zur Korrektur von Messfehlern in Umfragen verwendet und dabei Umfrage-, Register- und Sensordaten kombiniert. Für zwei Zielvariablen wurde die Unterberichterstattung im niederländischen Survey zu Güterverkehr geschätzt. Basierend auf einem loglinearen CRC-Schätzer liegt die Unterberichterstattung für die Gesamtzahl der Fahrzeugtage und das transportierte Sendungsgewicht bei 18 %. Die hier vorgestellte Methode ist auf jede Validierungsstudie anwendbar, bei der Erhebungs-, Register- und Sensordaten (oder jede andere externe Big-Data-Quelle) auf Mikroebene mittels eines eindeutigen Identifikators verknüpft werden können.

Eine zentrale Limitation dieser Studie ist, dass es nicht möglich war, die Anzahl der falsch-positiven Verknüpfungen zu schätzen. Diese Fehlerquelle ist zentral, da die CRC-Schätzungen sensitiv auf solche Fehler reagieren. Diese Limitation behandeln Klingwort und andere (erwartet 2021).

Diese Studie zeigt, dass Sensordaten in Kombination mit dem CRC-Schätzer ein valides Werkzeug zur Schätzung von Unterberichterstattung in Umfragen darstellen. Diese Studie ist eine nützliche Referenz für Statistikerinnen und Statistiker im Bereich der Verkehrsforschung oder in der amtlichen Statistik, wenn Umfrage-, Sensor- und Registerdaten miteinander verknüpft werden können und CRC angewendet werden kann.

LITERATURVERZEICHNIS

Alho, Juha M. *Logistic Regression in Capture-Recapture Models.* In: Biometrics. Volume 46. Ausgabe 3/1990, Seite 623 ff.

Bricka, Stacey/Bhat, Chandra R. *Comparative Analysis of Global Positioning System-Based and Travel Survey-Based Data.* In: Transportation Research Record: Journal of the Transportation Research Board. Volume 1972. Ausgabe 1/2006, Seite 9 ff.

Bricka, Stacey/Sen, Sudeshna/Paleti, Rajesh/Bhat, Chandra R. *An Analysis of the Factors Influencing Differences in Survey-Reported and GPS-Recorded Trips.* In: Transportation Research Part C. Volume 21. Ausgabe 1/2012, Seite 67 ff.

Buelens, Bart/Daas, Piet/Burger, Joep/Puts, Marco/van den Brakel, Jan. *Selectivity of Big data. Statistics Netherlands Discussion Paper.* The Hague/Heerlen 2014.

Burnham, Kenneth P./Anderson, David R. *Multimodel Inference: Understanding AIC and BIC in Model Selection.* In: Sociological Methods & Research. Volume 33. Ausgabe 2/2004, Seite 261 ff.

Centraal Bureau voor de Statistiek. *Basisbestanden Goederenwegvervoer 2015.* The Hague/Heerlen 2017.

Citro, Constance F. *From Multiple Modes for Surveys to Multiple Data Sources for Estimates.* In: Survey Methodology. Volume 40. Ausgabe 2/2014, Seite 137 ff.

Cochran, William G. *Laplace's Ratio Estimator.* In: David, H. A. (Herausgeber). Contributions to Survey Sampling and Applied Statistics. New York 1978, Seite 3 ff.

Daas, Piet J. H./Puts, Marco J./Buelens, Bart/van den Hurk, Paul A. M. *Big Data as a Source for Official Statistics.* In: Journal of Official Statistics. Volume 31. Ausgabe 2/2015, Seite 249 ff.

de Waal, Ton/van Delden, Arnout/Scholtus, Sander. *Multi-source Statistics: Basic Situations and Methods.* Statistics Netherlands Discussion Paper. The Hague/Heerlen 2017.

Enright, Bernard/OBrien, Eugene J. *Cleaning Weigh-in-motion Data: Techniques and Recommendations.* Dublin Institute of Technology & University College Dublin 2011.

Federal Highway Administration. *Effective Use of Weigh-in-motion Data: The Netherlands Case Study.* 2007. Office of International Programs. FHWA/US DOT (HPIP). Publication No. FHWA-PL-07-028 HPIP/10-07(3.5)EW.

Fienberg, Stephen E. *The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables.* In: Biometrika. Volume 59. Ausgabe 3/1972, Seite 591 ff.

Hald, A. *A History of Probability and Statistics and Their Applications Before 1750.* New York 1990.

LITERATURVERZEICHNIS

Huggins, R. M. *On the Statistical Analysis of Capture Experiments*. In: Biometrika. Volume 76. Ausgabe 1/1989, Seite 133 ff.

International Working Group for Disease Monitoring and Forecasting. *Capture-recapture and Multiple Record Systems Estimation II: Applications in Human Diseases*. In: American Journal of Epidemiology. Volume 142. Ausgabe 10/1995, Seite 1059 ff.

Japec, Lilli/Kreuter, Frauke/Berg, Marcus/Biemer, Paul/Decker, Paul/Lampe, Cliff/Lane, Julia/O'Neil, Cathy/Usher, Abe. *Big Data in Survey Research: AAPOR Task Force Report*. In: Public Opinion Quarterly. Volume 79. Ausgabe 4/2015, Seite 839 ff.

Klingwort, Jonas. *Correcting Survey Measurement Error With Big Data from Road Sensors Through Capture-recapture*. Dissertation. Universität Duisburg-Essen 2020. doi: 10.17185/duepublico/72081.

Klingwort, Jonas/Buelens, Bart /Schnell, Rainer. *Capture-Recapture Techniques for Transport Survey Estimate Adjustment Using Permanently Installed Highway-Sensors*. In: Social Science Computer Review. Online first. 2019. doi: 10.1177/0894439319874684.

Klingwort, Jonas/Burger, Joep/Buelens, Bart/Schnell, Rainer. *Understanding the difference in freight transport estimates with and without road sensor data*. In: Travel Behaviour & Society (im Review-Prozess, erwartet 2021).

Krishnamurty, Parvati. *Diary*. In: Lavrakas, Paul J. (Herausgeber). Encyclopedia of Survey Research Methods. Band 1. Thousand Oaks 2008, Seite 197 ff.

Lincoln, Frederick C. *The Waterfowl Flyways of North America*. Washington 1935.

Lohr, Sharon L./ Raghunathan, Trivellore E. *Combining Survey Data with Other Data Sources*. In: Statistical Science. Volume 32. Ausgabe 2/2017, Seite 293 ff.

Meyburg, Arnim H./Rahman, Shams. *The Challenges of Freight and Commercial Transport Surveys*. In: Stopher, Peter/Jones, Peter (Herausgeber). Transport Survey Quality and Innovation. Bingley 2003, Seite 443 ff.

Pearson, D. *Global Positioning System (GPS) and Travel Surveys: Results from the 1997 Austin Household Survey*. Paper presented at the Eighth Conference on the Application of Transportation Planning Methods. Corpus Christi 2001.

Petersen, Carl Georg Johannes. *On the Biology of Our Flat-fishes*. Kopenhagen 1894.

Richardson, Antony J./Ampt, Elizabeth S./Meyburg, Arnim H. *Nonresponse Issues in Household Travel Surveys*. In: Conference on Household Travel Surveys: New Concepts and Research Needs. Washington 1996, Seite 79 ff.

Schnell, Rainer. *Record Linkage*. In: Wolf, Christof/Joye, Dominique/Smith, Tom W./Fu, Yang-chih (Herausgeber). The SAGE Handbook of Survey Methodology. Thousand Oaks 2016, Seite 662 ff.

LITERATURVERZEICHNIS

Schnell, Rainer. *Big Data aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt.* In: /Arránz Becker, Oliver/ Lois, Daniel (Herausgeber). Soziologie und soziale Praxis. Wiesbaden 2019, Seite 101 ff.

Shekhar, Shashi. *Foreword.* In: Ganguly, Auroop R./Gama, João/Omitaomu, Olufemi A./ Gaber, Mohamed Medhat/Vatsavai, Ranga Raju (Herausgeber). Knowledge Discovery From Sensor Data. Boca Raton 2009, Seite ix f.

Shen, Li/Stopher, Peter. *Review of GPS Travel Survey and GPS Data-processing Methods.* In: Transport Reviews. Volume 34. Ausgabe 3/2014, Seite 316 ff.

Shlomo, Natalie/Goldstein, Harvey. *Editorial: Big Data in Social Research.* In: Journal of the Royal Statistical Society. Series A. Volume 178. Ausgabe 4/2015, Seite 787 ff.

Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge 1986.

Stopher, Peter/Greaves, Stephen P. *Household Travel Surveys: Where Are We Going?* In: Transportation Research Part A. Volume 41. Ausgabe 5/2007, Seite 367 ff.

Sun, Qian (Chayn)/Odolinski, Robert/Xia, Jianhong (Cecilia)/Foster, Jonathan/ Falkmer, Torbjörn/Lee, Hoe. *Validating the Efficacy of GPS Tracking Vehicle Movement for Driving Behaviour Assessment.* In: Travel Behaviour and Society. Volume 6. Januar 2017, Seite 32 ff.

van der Heijden, Peter G. M./Cruyff, Maarten/Whittaker, Joe/Bakker, Bart F. M. /Smith, Paul A. *Dual and Multiple System Estimation: Fully Observed and Incomplete Covariates.* In: Böhning, Dankmar/van der Heijden, Peter G. M./Bunge, John (Herausgeber). Capture-recapture Methods for the Social and Medical Sciences. Boca Raton 2017, Seite 213 ff.

Wang, Zhenzhen/He Sylvia Y./Leung, Yee. *Applying Mobile Phone Data to Travel Behaviour Research: A Literature Review.* In: Travel Behaviour and Society. Volume 11. April 2018. Seite 141 ff.

Wolf, Jean/Oliveira, Marcelo/Thompson, Miriam. *Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-enhanced Household Travel Survey.* In: Journal of the Transportation Research Board. Volume 1854, Seite 189 ff.

Zwane, Eugene N./van der Heijden, Peter G. M. *Semiparametric Models for Capture-recapture Studies with Covariates.* In: Computational Statistics & Data Analysis. Volume 47. Ausgabe 4/2004, Seite 729 ff.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktionsleitung: Juliane Gude
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2021
Das Archiv älterer Ausgaben finden Sie unter www.destatis.de

Artikelnummer: 1010200-21001-4, ISSN 1619-2907