Zusammenfassung der Dissertation

'Correcting Survey Measurement Error With Big Data from Road Sensors Through Capture-recapture'

Jonas Klingwort

Die Verwendung von Sensordaten die nicht auf Zufallsstichproben basieren, wird in den Sozialwissenschaften und der amtlichen Statistik immer relevanter. Diese Entwicklung lässt sich durch sinkende Rücklaufquoten bei Surveys, steigende Datenerhebungskosten, die Nachfrage nach regelmäßigeren Statistiken und durch eine allgemeine Diskussion über die Qualität von Survey-Daten erklären. Allerdings werden Sensordaten aufgrund ihres unbekannten Datengenerierungsprozesses derzeit nur selten in der amtlichen Statistik verwendet. Die Integration von Sensordaten könnte besonders nützlich sein, wenn diese mit Erhebungs- und Registerdaten verknüpft werden können. Um den Nutzen der Verknüpfung von Survey-Daten mit Register- und Sensordaten zu evaluieren, mit dem Ziel die Genauigkeit von Survey-Punktschätzern zu verbessern, ist empirische Forschung zu verknüpfbaren Datensätzen erforderlich.

Insbesondere Surveys in Form von zeit-basierten Tagebuchbefragungen verursachen einen hohen Beantwortungsaufwand und führen zu niedrigen Rücklaufquoten. In der Vergangenheit wurden Surveys zu Mobilität und Verkehr mit mobilen GPS-Geräten validiert und korrigiert. Es zeigte sich, dass diese Erhebungen, aufgrund von Unterberichtung von z.B. Fahrten, negativ verzerrte Schätzungen aufweisen. Übertragen auf das Total Survey Error Framework kann Unterberichtung als eine Funktion von Mess- und Nonresponse-Fehlern betrachtet werden. Messfehler sind in der Regel schwer zu quantifizieren, da externe Quellen für die Validierung nur selten zur Verfügung stehen. Nonresponse-Fehler, der andere Teil des Frameworks, können durch den Vergleich der Verteilung von Hilfsinformationen zwischen Stichprobe und Grundgesamtheit korrigiert werden. In dieser Arbeit werden anstelle von mobilen GPS-Geräten externe Daten von fest-installierten Straßensensoren zur Schätzung der Unterberichtung in der niederländischen Güterverkehrserhebung verwendet.

In der Arbeit werden die niederländische Güterverkehrserhebung und die Straßensensordaten des Weigh-in-Motion-Straßensensornetzes der niederländischen Straßenverwaltung von 2015 verwendet. Das Survey basiert auf einer Zufallsstichprobe von Lkw-/Transportfahrzeugbesitzern, die Fahrten und transportiertes Gewicht der Güter für das Fahrzeug in der Stichprobe in einer bestimmten Woche angeben müssen. 18 Sensorstationen auf den niederländischen Autobahnen wiegen kontinuierlich jedes vorbeifahrende Lkw-/Transportfahrzeug und verwenden ein Kamerasystem, das die Nummernschilder erfasst, um die Fahrzeuge zu identifizieren. Jedes Fahrzeug des Surveys kann mit den entsprechenden Beobachtungen in den Sensordaten und mit amtlichen Registern verknüpft werden, wobei die Kombination aus Kennzeichen und Zeitstempel als eindeutige Kennung verwendet wird. Da das nationale Fahrzeugregister das Leergewicht jedes Fahrzeugs und Anhängers beinhaltet, kann das Gewicht der transportierten Güter berechnet werden. Folglich messen die Sensoren und die Erhebung unabhängig voneinander die gleichen Variablen: das Vorkommen von Fahrten und das entsprechende Gewicht der transportierten Güter. Weitere Variablen die in den Registern enthalten sind, sind z.B. technische Spezifikationen der Fahrzeuge und administrative Angaben der Fahrzeughalter.

Im Rahmen dieser Arbeit wurde eine objektive Methode zur Schätzung des Ausmaßes der Unterberichtung und der Korrektur von Punktschätzern entwickelt, die auf der Anwendung von Capture-Recapture Techniken basiert. Insgesamt werden sechs verschiedene Schätzer in der Arbeit verwendet. Genauer gesagt werden ein post-stratifizierter Survey-Schätzer, eine naive Erweiterung des Survey-Schätzers, zwei Conditional-Likelihood Capture-Recapture Schätzer und zwei Unconditional-Likelihood Capture-Recapture Schätzer angewendet, verglichen und diskutiert. Die Capture-Recapture Schätzer korrigieren sowohl den Nonresponse- als auch den Messfehler. Der Survey-Schätzer wird nur für selektiven Nonresponse korrigiert. Daher kann eine mögliche Differenz zwischen Capture-Recapture- und Survey-Schätzer auf einen Messfehler zurückgeführt werden. Da die Capture-Recapture Annahme homogener Erfassungswahrscheinlichkeiten verletzt ist, wird die Heterogenität der Erfassungswahrscheinlichkeiten mittels logistischer Regression und log-linearer Modelle modelliert. Die Auswirkungen vereinzelter Verletzungen der Perfect Linkage-Annahme werden im Rahmen von Sensitivitätsanalysen bewertet. Die Flexibilität sowie die Grenzen der verwendeten Schätzer werden in einer stratifizierten Capture-Recapture Analyse bewertet.

Alle Capture-Recapture Schätzer liefern größere Schätzungen für die betrachteten Zielvariablen als der Survey-Schätzer. Entsprechend dem empfohlenen log-linearen Schätzer beträgt die Unterberichterstattung für das Vorkommen von Fahrten bei ca. 18% und bei 23% für das transportierte Gewicht der Güter. Unter Berücksichtigung der Ergebnisse in der Literatur zu Unterberichtung bei Mobilitäts- und Verkehrserhebungen liefert die vorgeschlagene Kombination von Datenquellen und Methoden plausible Ergebnisse. Die stratifizierten Schätzungen zeigen zum Teil größere Anteile an Unterberichtung in einzelnen Strata, z.B. bei kleineren Unternehmen und Fahrzeugen, die nicht für kommerzielle Zwecke verwendet werden. Die Stratifizierung zeigt aber auch, dass Capture-Recapture Schätzer bei z.B. kleinen Strata ungeeignet sind. Die Sensitivitätsanalysen zeigen, dass die Unconditional-Likelihood Capture-Recapture Schätzer robust gegenüber Fehlern in den Survey-Antworten sind. Bei Fehlern in den Sensorbeobachtungen sind die Unconditional-Likelihood Schätzer sensibel gegenüber falsch-positiven Verknüpfungen, aber robust gegenüber OCR-Fehlern.

Die Dissertation hat aber auch zu weiteren Fragen geführt und gezeigt, dass weitergehende Forschung notwendig ist. Hier werden jedoch nur einige Punkte genannt. Zunächst muss die Wahrscheinlichkeit von falsch-positiven Verknüpfungen geschätzt werden, da dieser Anteil die Ergebnisse beeinflusst. Zweitens können die entwickelten Capture-Recapture Modelle verbessert werden, z.B. durch die Verwendung von Interaktionstermen. Drittens sollte die Analyse hinsichtlich der Verallgemeinerbarkeit auf weitere Jahre ausgedehnt werden. Viertens kann der Prozess der Sensordatenaufbereitung verbessert werden, da in einigen Fällen negative transportierte Gewichte berechnet wurden.

Diese Arbeit demonstriert eine spezifische Verwendung von Big Data in der amtlichen Statistik zur Abschätzung des Bias durch Unterberichtung in Surveys. Die Methode ist jedoch nicht auf die amtliche Statistik beschränkt, sondern kann auch in anderen Disziplinen wie z.B. den Sozialwissenschaften verwendet werden. Weiterhin ist diese Methode für jede Art von Validierungsstudie geeignet, bei der Erhebungs-, Register- und Sensordaten (oder jede andere Big-Data Datenquelle) auf Mikroebene mit einem eindeutigen Identifikator verknüpft werden können. Die Arbeit ist ein neues Beispiel für Multi-Source Statistiken, ein vielversprechender Ansatz, um den Nutzen von Sensordaten im Bereich der amtlichen Statistik zu verbessern.