Anonymisierung von Betriebsdaten durch Erzeugung synthetischer Datensätze

Jörg Drechsler, Kompetenzzentrum Empirische Methoden
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit

Vortrag im Rahmen des Institutskolloquiums
4. Mai 2011, 16:15 Uhr
Seminarraum, Ludwigstraße 33 I

Statistische Ämter wie das Statistische Bundesamt und andere Forschungseinrichtungen tragen anhand von Registern und Befragungen viele wichtige statistische Informationen über Wirtschaft und Gesellschaft zusammen. Um eine effiziente Nutzung der erhobenen Daten zu fördern und eine breitgefächerte wissenschaftliche Forschung zu ermöglichen, wäre es wünschenswert, diese Informationen der interessierten Fachöffentlichkeit uneingeschränkt zur Verfügung stellen zu können. Ein uneingeschränkter Datenzugang könnte die wissenschaftliche Forschung stimulieren, Forschungsaufträge zu anstehenden politischen Entscheidungen auf eine solide Datengrundlage stellen, und Studierenden die Möglichkeit bieten, schon während ihres Studiums empirische Analysemethoden für komplexe Datensätze praktisch anzuwenden. Dies ist allerdings aus datenschutzrechtlichen Gründen nur in seltenen Fällen möglich. Ein einfaches Entfernen identifizierender Merkmale wie Name und Anschrift ist oft nicht ausreichend, um die Anonymität der Befragungsteilnehmer zu gewährleisten. Gerade bei Betriebsdaten genügen oft wenige Variablen wie Branche und Umsatz, um einen Betrieb eindeutig zu identifizieren. Deshalb ist der Zugang zu Betriebsdaten für externe Wissenschaftler stark reglementiert.

Ein sehr innovativer und in Europa noch relativ unbekannter Ansatz, der dazu beitragen kann, die Datenweitergabe zu vereinfachen, ist die Erzeugung synthetischer Datensätze. Bei diesem Verfahren, das auf den Ideen der multiplen Imputation beruht, werden die Original-werte durch mehrere künstliche Werte ersetzt, die möglichst ähnliche Verteilungs-eigenschaften wie die Originaldaten aufweisen. Diese künstlichen oder synthetischen Daten werden dann der Allgemeinheit zur Verfügung gestellt. Potenzielle Nutzer können unter Verwendung einfacher Formeln für eine Vielzahl von Auswertungen valide Ergebnisse erzielen. Da es sich um rein fiktive Werte handelt, ist das Re-Identifikationsrisiko in der Regel zu vernachlässigen.

Der Vortrag soll eine Einführung in die Erzeugung synthetischer Daten bieten. Es werden verschiedene Varianten dieses Verfahrens vorgestellt und ihre Vor- und Nachteile diskutiert. Anhand des IAB Betriebspanels wird gezeigt, dass das Konzept eine Datenweitergabe ermöglicht, bei der einerseits der zugesicherte Datenschutz eingehalten wird, andererseits aber auch das Analysepotenzial der anonymisierten Daten erhalten bleibt.