Fachstatistische Anwendungen

Datenschutzkonzept zur Nutzung von SOEPgeo im Forschungsdatenzentrum SOEP am DIW Berlin

von **Jan Goebel** und **Bernd Pauer**

In den empirischen Sozial- und Wirtschaftswissenschaften werden Forschungsdaten immer wichtiger, aber auch komplexer. In die Erstellung von Daten fließt nicht zuletzt ein großer zeitlicher und/oder personeller Aufwand. Daher sollten Daten für die Forschung möglichst breit nutzbar sein, um ihren maximalen Nutzen entfalten zu können. Empirische Daten tragen dazu bei, dass wissenschaftliche Aussagen auf einer transparenten Grundlage stehen. Sie anderen Wissenschaftlerinnen und Wissenschaftlern zur Verfügung zu stellen, gehört zur guten wissenschaftlichen Praxis. Aus diesen Gründen sollten empirische Daten, welche die Grundlage von wissenschaftlichen Analysen bilden, immer auch als Forschungsdaten angemessen archiviert und allen Forschenden für eine potenzielle Nachnutzung oder Reanalyse zur Verfügung stehen.

1. Einleitung

Bereits zu Beginn des vergangenen Jahrhunderts wurden die Beziehungen zwischen dem Menschen und dem Raum festgestellt (Werlen 2008). In diesem Zusammenhang wurden in der sozialökologischen Forschung der Chicagoer Schule die Interaktionen zwischen der räumlichen Stadtumwelt und dem Handeln des Individuums untersucht (Burgess 1925). Die auf diesen Ansätzen aufbauenden bzw. weiterführenden Kategorisierungen von Städten und Stadtregionen bestehen für Deutschland zunächst als Modell der Stadtregionen (Boustedt 1953). Weitere spezifisch für Deutschland entwickelte Typisierungen, wie die BIK-Regionen oder die laufende Raumbeobachtung des Bundesinstituts für Bau-, Stadt- und Raumforschung (BBSR), beschreiben den Raum mangels detaillierter räumlicher Daten auf administrativen Raumeinheiten.

Dabei kommt der Bedeutung des Raumes für die Sozialforschung eine immer größer werdende Rolle zu, da durch raumbezogene Informationen die Beziehungen zwischen dem Menschen und dem Raum, also die Interaktion zwischen dem Individuum und dem räumlichen Kontext, besser untersucht werden können. Einerseits ist der Bedeutungszuwachs des Raumes auf die gestiegene Verfügbarkeit von räumlichen Daten zurückzuführen, was letztlich auch auf technische Entwicklungen in den Informationstechnologien wie den geographischen Informationssystemen (GIS), globalen Positionierungssystemen (GPS) und vor allem satellitengestützter Erdbeobachtung zurückzuführen ist. Die Relevanz raumbezogener Daten für die Sozialforschung wird als spatial turn für die Sozialwissenschaften bezeichnet (Goodchild 2007) und die Georeferenzierung von sozialwissenschaftlichen Daten als eine der großen positiven Herausforderungen der letzten Jahre interpretiert (RatSWD 2010).

Ein gänzlich neuer Zugang zur raumbezogenen Auswertung der Daten des SOEP ist über die vom Forschungsdatenzentrum des SOEP (FDZ SOEP) eingerichtete Infrastruktur „SOEPgeo" möglich. An Gastarbeitsplätzen wird im FDZ SOEP die Möglichkeit bereitgestellt, Geo-Koordinaten für die einzelnen Haushalte zu nutzen. Das Konzept ermöglicht es, dass unter streng kontrollierten Datenschutzbedingungen Mikrodaten des SOEP zusammen mit geokodierten externen Daten für wissenschaftliche Zwecke ausgewertet werden können. Die folgenden Ausführungen beschreiben im Detail, wie hoch sensitive geokodierte Adressinformationen von befragten Haushalten Wissenschaftlerinnen und Wissenschaftlern zugänglich gemacht werden können, ohne dass es zu datenschutzrechtlichen Abstrichen kommt.

1.1 Das SOEP

Das Sozio-oekonomische Panel (SOEP) ist eine repräsentative Wiederholungsbefragung, die bereits seit nahezu 30 Jahren läuft (Wagner et al. 2007). Im Auftrag des DIW Berlin werden jedes Jahr in Deutschland über 20 000 Personen aus rund 11 000 Haushalten von TNS Infratest Sozialforschung befragt. Die Daten geben Auskunft zu Fragen über Einkommen, Erwerbstätigkeit, Bildung oder Gesundheit. Weil jedes Jahr die gleichen Personen befragt werden, können langfristige soziale und gesellschaftliche Trends besonders gut verfolgt werden. Eine detaillierte Beschreibung und weiterführende Dokumentation des Surveys findet sich auf der Webseite des FDZ SOEP (http://www.diw.de/soepfdz) oder auf dem Metadatenportal des SOEP (http://data.soep.de/studies/2).

1.2 Regionaldaten im SOEP

Das SOEP bietet vielfältige Möglichkeiten für regionalisierte Analysen. Die enthaltenen regionalen Informationen lassen sich in zwei Kategorien unterteilen:

1. Beschreibung des regionalen oder räumlichen Kontexts
   a) Informationen gewonnen aus der Befragungssituation zur Nachbarschaft bzw. Wohnumgebung der Befragten,
   b) Informationen zur Beschreibung der näheren Umgebung an Hand von verknüpften Indikatoren der microm GmbH,
   c) Informationen zu Regionstypisierungen, in denen der befragte Haushalt lebt.
2. Offizieller Schlüssel der regionalen Einheiten zum jeweiligen Erhebungsjahr (administrative Einheiten, Postleitzahlen oder Koordinaten), um eine Verknüpfbarkeit mit externen Daten (z.B. aus der amtlichen Statistik) zu gewährleisten.

Bis auf die Informationen unter Punkt 1 a) sind alle diese Informationen datenschutzrechtlich sensibel und daher nicht in der Datenlieferung des Standard Scientific Use Files enthalten. Alle Informationen können aber für die unabhängige wissenschaftliche Forschung genutzt werden, allerdings unter zum Teil stark erhöhten Datenschutzanforderungen, insbesondere im Zugang zu diesen Daten. Eine Übersicht über die im SOEP vorhandenen Regionalinformationen findet sich in Goebel 2013.

1.3 Koordinaten

Im Rahmen einer Kooperation mit der microm Micromarketing-Systeme und Consult GmbH war es seit 2004 möglich, den SOEP-Haushalten Indikatoren zu ihrer konkreten kleinräumigen Umgebung zuzuspielen. microm entwickelt seit 1992 mikrogeographische Daten, ursprünglich mit dem Ziel, Unternehmen bei der räumlichen Verortung ihrer Kunden- oder Zielgruppen zu unterstützen. Teile dieser Informationen sind auch für sozialwissenschaftliche Fragestellungen von Bedeutung, wie beispielsweise die Variable „microm Status“ oder der Anteil von Personen mit Migrationshintergrund. Die microm Datenbasis ermöglicht es flächendeckend, die ca. 41 Mill. bundesdeutschen Haushalte hinsichtlich verschiedenster Kriterien räumlich abzubilden. Die räumliche Verortung der SOEP-Haushalte wird direkt beim Erhebungsinstitut (TNS Infratest Sozialforschung), das als Einzige Namen und Klartext-Adresse vorhält, vorgenommen. Die verknüpften Daten werden von Infratest (ohne Klartext-Adressen und Namen) an das FDZ SOEP geliefert.

Die dabei zugespielten Daten enthalten auch die Koordinaten, die zur Verknüpfung der Datenquellen notwendig sind. Am FDZ SOEP werden diese Daten wiederum aufgespalten, sodass die inhaltlichen Indikatoren (z. B. microm status oder Anteil der Migrantinnen und Migranten) an einem normalen Gastarbeitsplatz Wissenschaftlerinnen und Wissenschaftlern zugänglich sind. Diese Zugangsmöglichkeiten beinhalten jedoch keinen Zugriff auf die Koordinaten. Damit Wissenschaftlerinnen und Wissenschaftlern auch das große Potenzial von geografisch referenzierten Koordinaten nutzen können, musste erst eine zusätzliche Infrastruktur aufgebaut werden. Einer der großen Vorteile bei der Integration der räumlichen Komponente über die Verortung der Befragungshaushalte ist, dass herkömmliche SOEP-Auswertungen immer auf administrative Raumeinheiten beschränkt sind, während über die Gebietsauswahl durch Geo-Koordinaten jeder beliebige Raum definiert werden und eine Unabhängigkeit von Gebietsstandsveränderungen in den Verwaltungseinheiten erzielt werden kann.

Zentraler Teil des entwickelten Datenschutzkonzeptes ist dabei, dass die Geo-Koordinaten der SOEP-Haushalte von den Erhebungsinformationen grundsätzlich getrennt gehalten werden. Keine Wissenschaftlerin und kein Wissenschaftler hat zu irgendeinem Zeitpunkt Zugriff auf Koordinate und Befragungsinformation. Die Erzeugung von inhaltlichen Indikatoren ist nur innerhalb eines speziell geschützten und abgeschotteten Systems möglich. Der Datennutzer hat daher keinen gleichzeitigen Zugriff auf die SOEP-Erhebungsdaten und die Geo-Koordinaten der SOEP-Haushalte. Die Ergebnisse werden nur streng anonymisiert präsentiert.

2. Allgemeine Beschreibung von SOEPgeo

SOEPgeo ermöglicht es den Nutzerinnen und Nutzern des SOEP, geokodierte Daten für wissenschaftliche Zwecke innerhalb des DIW Berlin auszuwerten. Bisher konnten Wissenschaftlerinnen und Wissenschaftler diese Informationen nicht nutzen. Mit SOEPgeo ist es nunmehr möglich, dies in einer speziellen Arbeitsumgebung zu realisieren. Wissenschaftlerinnen und Wissenschaftler müssen eine Datenschutzverpflichtung unterzeichnen und der Zugriff auf die Daten wird vollständig protokolliert.

Grundlage des Konzepts ist es, dass während allen von Nutzerinnen und Nutzern ausgeführten Analysen die Geo-Koordinate der SOEP-Haushalte von den eigentlichen Erhebungsinformationen getrennt gehalten wird. Zur Erzeugung von inhaltlichen Indikatoren innerhalb eines Geo-Informations Systems (GIS) oder innerhalb des Statistik Pakets R sind nur die Koordinaten ohne weitere Informationen über den Haushalt oder Personen in diesem Haushalt notwendig. Ein Zuspielen der in einem GIS erzeugten Indikatoren erfolgt durch Beschäftigte des FDZ ohne die Möglichkeit eines Zugriffs durch die Nutzerinnen und Nutzer. Im Folgenden wird das Konzept im Detail dargestellt.

3. Grundlagen des Datenschutzkonzepts

3.1 Juristische Grundlagen

Das Datenschutzkonzept beruht auf dem Konzept der „faktischen Anonymisierung“ von Daten, das auch von der amtlichen Statistik angewandt wird. Zugänglich gemacht werden „Scientific Use Files“ (SUF). Dabei sind Daten theoretisch deanonymisierbar (im Gegensatz zu absolut anonymisierten Mikrodaten, die als Public Use Files für jedermann verfügbar sind). Die Deanonymisierung würde einen unverhältnismäßigen Aufwand bedeuten und

ist mit juristischen Sanktionen belegt. Das heißt, dass die Auswertung nur auf Basis eines speziellen Datennutzungsvertrages, der ausschließlich mit vertrauenswürdigen Forschungseinrichtungen abgeschlossen wird, möglich ist und die erhobenen Daten faktisch anonymisiert sind (insbesondere keine Namen oder Klartexte enthalten). Das Deanonymisieren durch indirekte Verfahren wird deutlich erschwert und ist mit juristischen Sanktionen belegt.

Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (§ 3 Abs. 1 BDSG[1]). Keine personenbezogenen Daten sind zusammenfassende Angaben oder aggregierte Daten, da sie keiner Einzelperson zugeordnet werden können. Bestimmt sind die Daten, wenn sie sich direkt auf eine bestimmte Person beziehen, d. h. einen unmittelbaren Rückschluss auf die Identität einer Person, in erster Linie anhand ihres Namens und Geburtsdatums, zulassen. Bestimmbar sind sie, wenn sie eine Person nicht als solche identifiziert, jedoch mithilfe anderer Informationen und Zusatzwissen identifizierbar machen, d. h. die Möglichkeit besteht, die Identität einer Person festzustellen.

Wie der Interministerielle Ausschusses für Geoinformationswesen treffend feststellt, besteht das Problem bei georeferenzierten Daten darin, „dass sich die Abgrenzung, wann es sich um ein – datenschutzrechtlich nicht geschütztes – bloßes Sachdatum handelt und in welchen Fällen ein Personenbezug vorliegt, für den Einzelfall als sehr schwierig erweisen kann. Damit ein Geodatum überhaupt Personenbezug aufweisen kann, muss es sich auf eine bestimmte oder bestimmbare Person beziehen.“ [2]

Da es sich bei den Geo-Koordinaten der Haushalte um sensitive Daten handelt, die in Verbindung mit den Befragungsdaten eine Deanonymisierung von teilnehmenden Personen enorm erleichtern können, werden die üblichen Maßnahmen für Scientific Use Files durch SOEPgeo deutlich verstärkt. Durch SOEPgeo wird ausgeschlossen, dass ein Deanonymisierungsversuch der Befragungsdaten auf Basis der Geo-Koordinaten durchgeführt werden kann.

Im Zentrum des technischen Schutzes steht, dass im Rahmen von SOEPgeo keinerlei Mikrodaten in Verbindung mit Geo-Koordinaten genutzt werden können. Eine Verknüpfung und damit missbräuchliche Nutzung der Mikrodaten, um die einzelnen Koordinaten einer bestimmbaren Person zuzuordnen, wird durch software- und hardware-technische Vorkehrungen für Benutzer ausgeschlossen. Dass heißt, die Deanonymisierung ist nur mit unverhältnismäßigem Aufwand möglich. Die Deanonymisierung von ausgewählten Einzelfällen bringt auch keinen erkennbaren Nutzen und Forscherinnen und Forscher würden ihre Karriere ruinieren, wenn durch die vollständige Protokollierung des Datenzugriffs Deanonymisierung bzw. Deanonymisierungsversuche aufgedeckt werden.

Alle Zugriffe werden vollständig protokolliert. Das heißt – im Gegensatz zur Distribution von faktisch anonymisierten Scientific Use Files – ist eine vollständige Kontrolle der Analysen jederzeit möglich, und es werden keine Mikrodaten weitergegeben. Ein weiterer wesentlicher Bestandteil des Datenschutzkonzeptes von SOEPgeo ist, dass der Zugriff nur während eines Aufenthalts am DIW Berlin möglich ist.

3.2 Technische Grundlagen

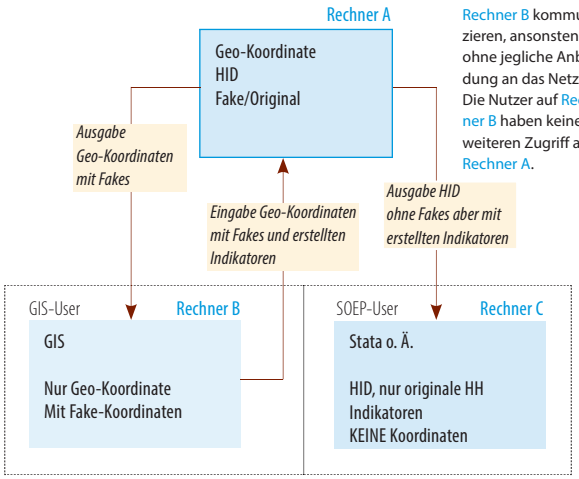
Zur Umsetzung des Datenschutzkonzepts von SOEPgeo wurden im DIW Berlin für das FDZ SOEP drei virtuelle Server[3] speziell konfiguriert (ein Schema des Setups findet sich in Abbildung 1). Auf dem ersten Rechner (Rechner A) ist als Einziges die Zuordnung der SOEP-Haushalts-ID zu den jeweiligen Geo-Koordinaten gespeichert. Dieser Rechner ist dafür verantwortlich, entweder die Geo-Koordinaten (ohne Haushalts-ID Zuordnung) oder die SOEP-Haushalts-IDs mit den neu erstellten Indikatoren (ohne die entsprechenden Geo-Koordinaten) über vordefinierte Schnittstellen auszugeben. Der zweite Rechner (Rechner B) wird zur Analyse der Geo-Koordinaten genutzt und der dritte Rechner (Rechner C) zur Analyse der SOEP-Befragungsdaten inklusive der aus dem Raumbezug gewonnenen zusätzlichen Indikatoren (ohne Geo-Koordinaten). Die Analyse der Geo-Koordinaten und die Analyse der eigentlichen SOEP-Daten erfolgt dabei auf den getrennten Systemen im jeweiligen Datenkontext in den Rollen als GIS-Nutzer oder SOEP-Nutzer. Beide Rollen können sich nicht überschneiden und es ist auch nicht möglich, Daten direkt zu übermitteln.

1 Bundesdatenschutzgesetz (BDSG) in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), zuletzt geändert durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814).

2 Siehe Behördenleitfaden zum Datenschutz bei Geodaten und -diensten, http://www.imagi.de/SharedDocs/Downloads/IMAGI/DE/Imagi/behoerdenleitfaden.pdf?__blob=publicationFile, Zugriff 30.07.2014.

3 Auf dem Hostsystem laufen nur diese drei virtuellen Maschinen.

1 | Schematische Darstellung von SOEPgeo



Rechner A kann nur über klar definierte Schnittstellen mit zwei Nutzern auf Rechner B kommunizieren, ansonsten ohne jegliche Anbindung an das Netz. Die Nutzer auf Rechner B haben keinen weiteren Zugriff auf Rechner A.

3.2.1 Hostsystem der SOEPgeo Maschinen

Die Hardware und der darauf installierte Hypervisor für die drei virtuellen Maschinen der SOEPgeo Installation steht hinter der allgemeinen Perimeter Firewall des DIW Berlin und hat nur die zum Betrieb notwendigen Dienste aktiviert, in diesem Fall die Management Software zur Verwaltung der drei Maschinen.

Der Zutritt ist durch die Installation im Rechenzentrum des DIW Berlin in einem abgeschlossenen Datenschutzschrank gesichert.

Die drei SOEPgeo Systeme haben namentlich bekannte Systemadministratoren, die nach einer Anmeldung explizit in die Adminstratorenrolle wechseln müssen, was, wie der Zugang, ebenfalls protokolliert wird.

3.2.2 Rechner A: Kombination von Haushalts-ID und Geo-Koordinate

Der eingesetzte Rechner für die Kombination der Geo-Koordinaten und der SOEP-Haushalts-ID ist nur von Rechner B oder Rechner C über eine interne Netzwerkschnittstelle, die auch nicht aus dem DIW-internen Netz angesprochen werden kann, erreichbar. Er kennt keine sogenannte „Default Route“ und ist somit nicht in der Lage, mit anderen Rechnern zu kommunizieren.

Zugriff auf den Rechner A ist nur für die Systemadministratoren mit dem Secure Shell Public-Key Verfahren möglich. Da auf diesem Rechner die Verbindung der Koordinate zu den Haushalts-IDs liegt, gelten für diesen Rechner besondere Schutzmaßnahmen.

Die dortige Datenbank beinhaltet den Umsteigeschlüssel von der Geo-Koordinate zur SOEP-Haushalts-ID und die Information, ob für diese Geo-Koordinate überhaupt ein SOEP-Haushalt im SOEP befragt wurde, es sich also um eine „wahre“ Koordinate oder um eine gefälschte Koordinate handelt. Um die eindeutige Zuordnung der Koordinate zu einem Haus mit einem SOEP-Haushalt aufzulösen, werden zusätzliche Geo-Koordinaten in die Datenbank mit aufgenommen (im Verhältnis 1:1). Da das Ziehungsdesign des SOEP zu einer Klumpung von Geo-Koordinaten führt, wurden die gefälschten Geo-Koordinaten nur im Umkreis der im SOEP bereits vertretenen Haushalte hinzugefügt, um auch auf diesem Weg eine Unterscheidung von gefälschter und richtiger Koordinate nicht zu ermöglichen.

Werden von einem GIS-Nutzer Indikatoren generiert, so werden anschließend durch Beschäftigte des FDZ SOEP auf Rechner A die Geo-Koordinaten entfernt und die SOEP-Haushalts-ID angespielt. Die Indikatoren mit Haushaltsnummern werden dann auf den jeweiligen Nutzerbereich des Rechners C geschrieben. Auf diesen Bereich können die jeweils definierten SOEP-Nutzer des Rechners C zugreifen.

3.2.3 Rechner B + C: Analyserechner

Zugriff auf die Rechner B + C ist nur mit dem Secure Shell Public-Key Verfahren innerhalb des DIW Berlin möglich. Auf den Analyserechnern werden (neben dem Systemadministrator) zwei getrennte Nutzerkreise eingerichtet: die GIS-Nutzer auf Rechner B und die SOEP-Nutzer auf Rechner C.

Der GIS-Nutzer darf nur auf die Datenbestände der Geo-Koordinaten zugreifen und kann nicht unterscheiden, hinter welcher der Koordinaten sich eine reale SOEP-Adresse befindet und hinter welcher nicht. Die Geo-Koordinaten werden ohne Verknüpfung zur SOEP-Haushalts-ID auf diesen Rechner durch die Systemadministration abgespeichert und einmal jährlich aktualisiert. Als einziges Attribut enthalten die Geo-Koordinaten ein Erhebungsjahr, um eine zeitliche Zuordnung zu ermöglichen. Die gefälschten Koordinatenpunkte bilden dabei ebenso reale Mobilitätsmuster ab wie die tatsächlich beobachteten Haushalte, das heißt, auch durch den Einbezug der temporalen Information können Wissenschaftlerinnen und Wissenschaftler daher eine echte und eine gefälschte Koordinate nicht unterscheiden.

Der GIS-Nutzer kann nicht auf Rechner A zugreifen und daher nicht die Verknüpfung von Koordinate und Haushalts-ID lesen. Der Datenfluss geht hier ausschließlich über das Konto der Systemadministration in Richtung Rechner A und beinhaltet die innerhalb des GIS-Systems erstellten Indikatoren auf Geo-Koordinaten-Ebene. Als GIS-Software kommen die folgenden Open Source Software Produkte zum Einsatz:

- R (**http://www.r-project.org/**),
- QGIS (**http://www.qgis.org/**),
- GRASS (**http://grass.osgeo.org/**).

In der SOEP-Nutzer-Rolle darf nur auf die Datenbestände des SOEP, ohne Geo-Koordinaten, aber mit den im GIS gebildeten Indikatoren, zugegriffen werden. Dabei kann ebenfalls nicht auf die Datenbank des Rechners A zugegriffen werden. Die Rechte beschränken sich auf das Lesen der Indikatoren in Verbindung mit den SOEP-Haushalts-IDs, wie sie von Rechner A rausgeschrieben wurden. Als Statistik-Software kommt zum einen Stata (**http://www.stata.com**) und zum anderen die Open Source Software R (**http://www.r-project.org/**) zum Einsatz.

3.3 Zugangskontrolle

Jedes Einloggen auf den Rechner A und B und jeder Zugriff auf die Postgres Datenbank des Rechners A wird protokolliert und mit Datum und Nutzerkennung von der SOEP-Gruppe elektronisch gespeichert. Dadurch kann jegliche Fehlverwendung auch im Nachhinein erkannt und individuell zugeordnet werden.

Möchten Wissenschaftlerinnen und Wissenschaftler für ihre SOEP-basierte Forschung auf Koordinateninformationen zugreifen und dafür SOEPgeo nutzen, so muss der Vertragsnehmer neben einem gültigen SOEP-Standardweitergabevertrag auch einen Sondernutzungsvertrag unterzeichnen. Erst aufgrund dieser geprüften Informationen können Wissenschaftlerinnen und Wissenschaftler als SOEPgeo-Datennutzer am DIW Berlin arbeiten.

3.4 Zugriffskontrolle

Die Zugriffskontrolle auf Rechner B+C erlaubt den jeweiligen Nutzerinnen und Nutzern den Lesezugriff auf die bereitgestellten Daten. Gleichzeitig sind die Ergebnisse und Verfahren anderer Nutzerinnen und Nutzer nicht von vornherein zugreifbar.

Arbeits- oder Projektgruppen mit mehreren Nutzerinnen und Nutzern können unter einer gemeinsamen Gruppe zusammengefasst werden. Feiner granulierte Zugriffsrechte ließen sich noch darüber hinaus abbilden, wurden aber bislang nicht benötigt.

4. Nutzungsbeispiele

Im Folgenden soll anhand ausgewählter Beispiele dargestellt werden, wie Forscherinnen und Forscher bisher den neuen Infrastruktur Service SOEPgeo nutzten und welche Forschungsfragen dadurch analysiert werden konnten.

Bauernschuster, Falck und Woessmann nutzten den zeitlich und räumlich ungleichmäßigen Ausbau und damit den Zugang zu Breitband-Internet in Deutschland, um den Zusammenhang von Internet und sozialem Kapital zu untersuchen. Dazu nutzten sie die Daten der Telekom, um über die Distanz zum nächsten OPAL-fähigen Zugangspunkt festzustellen, wann für welche Haushalte Breitband-Internet zur Verfügung stand. Ihre Ergebnisse zeigen, dass kein negativer, sondern eher ein positiver Effekt vom Internetzugang auf soziales Kapital ausgeht (Bauernschuster, Falck und Woessmann 2014).

Verschiedene Forscherinnen und Forscher nutzen die Geo-Koordinaten, um den Einfluss von Naturkatastrophen auf das Verhalten oder Befinden der Menschen zu untersuchen. Goebel, Kerkel u. a. (2013) untersuchten den Einfluss der Katastrophe von Fukushima auf die Umweltsorgen in Deutschland und kontrollierten dabei die Distanz zum nächsten Atomkraftwerk in Deutschland, Frankreich oder Tschechien. Berlemann, Steinhardt und Tutt 2014 stellten auf einer Präsentation anlässlich des „Spring Meeting of Young Economists“ (2014) ihre Arbeiten zum Einfluss des Elbe-Hochwassers 2002 auf das Spar- und Umzugsverhalten vor. Ebenfalls die Daten aus Satellitenaufnahmen zu einer Hochwasser-Katastrophe (allerdings die des Jahres 2013) nutzten Avdeenko, Krekel und Martinoty 2014, um den Zusammenhang mit Umweltsorgen und Einstellungen zum Klimawandel zu untersuchen.

Die allgemeineren Auswirkungen der unmittelbaren Umgebung zeigen unter anderem Kolbe, Krekel und Wüstemann (2014). Sie nutzten flächendeckende Daten aus dem European Urban Atlas für deutsche Großstädte, um relevante Grünflächen zu identifizieren. Ihre Analyse beleuchtet damit die quantitative Bedeutung von Grünflächen für das Wohlbefinden der Stadtbewohner.

Krekel und Zerrahn (2014) untersuchten den Einfluss von Windenergieanlagen auf die Wohnzufriedenheit, indem sie die Distanz zum nächstgelegenen Windrad im Zeitraum 2000 bis 2012 errechneten. Die Kombination von Paneldaten und flächendeckender Geoinformationen machte es hier möglich, auch kausale Erklärungsmodelle zu schätzen. Ebenfalls flächendeckende Daten, allerdings zur Luftverschmutzung, nutzten Voigtländer u. a. 2011. Mit Hilfe von Daten zur Luftbelastung aus EURAD-IM untersuchten sie den Einfluss auf die im SOEP erfragten subjektiven Gesundheitsmaße.

Eine wichtige Anwendung von Geo-Koordinaten in der sozialwissenschaftlichen Forschung ist die Möglichkeit, Varianz unterhalb der Gemeindeebene zu nutzen, also den Effekt von Stadt je nach Lage innerhalb der Stadt differenziert schätzen zu können. Goebel und Wurm (2010) nutzten den Abstand zu einer verallgemeinerten Stadtgrenze, die sich nicht nach der offiziellen Gemeindegrenze richtet, sondern nach der faktischen Bebauungsstruktur. Dabei verschnitten sie Daten von CORINE Land Cover mit der Lage der SOEP-Haushalte und berechneten den Abstand zum Zentrum und zur Stadtgrenze. Sie zeigten dabei, dass das Armutsrisiko am Stadtrand am niedrigsten ist und in beide Richtungen (zum Stadtzentrum und in Richtung ländlicher Raum) zunimmt.

5. Fazit

Am FDZ SOEP ist seit Beginn der Pilotphase im Jahr 2010 die Infrastruktur zur Nutzung von GEO-Koordinaten in Verbindung mit sozialwissenschaftlichen Befragungsdaten für Wissenschaftlerinnen und Wissenschaftler zugänglich. Die bisherigen Erfahrungen sind durchweg positiv.

Insgesamt nimmt die wissenschaftliche Community den Service sehr gut an, obwohl die Arbeitsbedingungen schwieriger sind als am eigenen Arbeitsplatzrechner und teilweise das Arbeiten durch die technische Trennung mehr Zeit beansprucht. Die Einschätzung, dass der Zugang zu den Originaldaten für wissenschaftliche Analysen wichtiger ist als der Komfort bei den Analysen, hat sich als richtig herausgestellt. Trotz der erhöhten Anforderungen an die Wissenschaftlerinnen und Wissenschaftler wird der Service von SOEP-Nutzern honoriert.

Insgesamt gesehen hält sich auch der Arbeitsaufwand für die Prüfung der Ergebnisse im Rahmen der normalen Outputprüfungen an einem Gastarbeitsplatz in einem Forschungsdatenzentrum in Grenzen.

Ein Problem, das sich in Zukunft stellt, ist damit verbunden, dass immer mehr geokodierte flächendeckende Daten und auch neue Schätzmethoden verfügbar werden. Diese an sich sehr erfreuliche Entwicklung hat jedoch den Nachteil, dass die Datenmengen, die die Server verarbeiten müssen, steigen. Die Einbindung von Abstandsmatrizen, also berechnete Distanzen von jeder beobachteten Einheit zu jeder anderen, stellt im Moment noch sehr hohe Anforderungen an die Rechner und ist auch im gegenwärtigen Setting nur schwer umzusetzen.

Eine weitere Herausforderung wird es sein, den Zugang zu SOEPgeo auch von Gastarbeitsplätzen in anderen Forschungsdatenzentren per Remote-Zugriff zu ermöglichen, wie es bereits für Regionalinformationen bis zu den Kreisen möglich ist (das sogenannte FDZ-im-FDZ-Modell, siehe auch Heining und Bender 2012).

Dr. Jan Goebel ist stellvertretender Leiter des Sozio-oekonomischen Panel (SOEP). Sein Arbeitsbereich: Data-Operation und Forschungsdatenzentrum (SOEP FDZ), die Arbeitsschwerpunkte sind: Datenweitergabe SOEP/SOEPinfo, Regionaldaten, Armuts- und Ungleichheitsforschung, Einkommensverteilung und -dynamik, Item-Non-Response und Imputation.

Bernd Pauer ist Mitarbeiter der Abteilung *Forschungsinfrastruktur* im DIW Berlin, IT-Sicherheit, er ist stellvertretender Datenschutzbeauftragter des DIW Berlin.

Literatur

- Avdeenko, Alexandra, Christian Krekel und Laurine Martinoty (2014). "Natural Disasters and Environmental Concerns: The Case of the 2013 Flood in Germany". Unveröffentlichter Vortrag SOEP User Conference 2014. Berlin. url: http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.465838.de/107_soep2014abstract_avdeenko_etal.pdf (besucht am 06. 08. 2014).
- Bauernschuster, Stefan, Oliver Falck und Ludger Woessmann (2014). "Surfing alone? The internet and social capital: Evidence from an unforeseeable technological mistake". In: Journal of Public Economics 117, S. 73–89. issn: 0047-2727. doi: http://dx.doi.org/10.1016/j.jpubeco.2014. 05.007. url: http://www.sciencedirect.com/science/article/pii/ S0047272714001145
- Berlemann, Michael, Max Steinhardt und Jascha Tutt (2014). Natural disasters and the savings rate. Some micro-evidence from a natural experiment. Unveröffentlichter Vortrag "Spring Meeting of Young Economists".
- Boustedt, Olaf (1953). „Die Stadtregion. Ein Beitrag zur Abgrenzung städtischer Agglomerationen". In: Allgemeines Statistisches Archiv 37, S. 13–26.
- Burgess, Ernest W. (1925). "The growth of the city: an introduction to a research project". In: The City. Hrsg. von Robert E. Park, Ernest W. Burgess und Roderick D. McKenzie. Univ. of Chicago Press.
- Goebel, Jan (2013). „Regionalisierungsmöglichkeiten des Sozio-oekonomischen Panels (SOEP)". In: Regionale Standards, Hrsg. von der Arbeitsgruppe Regionale Standards. 2. vollständig überarbeitete und erweiterte Auflage. GESIS.
- Goebel, Jan, Christian Krekel u. a. (2013). Natural Disaster, Policy Action, and Mental WellBeing: The Case of Fukushima. SOEPpaper 599. Berlin: DIW/SOEP. url: http://www.diw.de/documents/publikationen/73/diw_01.c.430617.de/diw_sp0599.pdf
- Goebel, Jan und Michael Wurm (2010). „Räumliche Unterschiede im Armutsrisiko in Ost- und Westdeutschland". In: Leben in Ost- und Westdeutschland: Eine sozialwissenschaftliche Bilanz der deutschen Einheit 1990-2010. Hrsg. von Peter Krause und Ilona Ostner. Frankfurt/Main: Campus. url: http://www.diw.de/documents/publikationen/73/diw_01.c.362278.de/diw_sp0321.pdf
- Goodchild, Michael F. (2007). "The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science". In: Journal of Official Statistics 23.3, S. 269–283. url: http://www.geog.ucsb.edu/~good/papers/436.pdf
- Heining, Jörg und Stefan Bender (2012): „Technische und organisatorische Maßnahmen für den Fernzugriff auf die Mikrodaten des Forschungsdatenzentrums der Bundesagentur für Arbeit". FDZ-Methodenreport, 08/2012, Nürnberg.
- Kolbe, Jens, Christian Krekel und Henry Wüstemann (5.–9. November 2014). "The Greener, the Happier? The Effects of Urban Green and Abandoned Areas on Residential Well-Being". Unveröffentlichter Vortrag, 67th Annual Meeting of the Gerontological Society of America. Washington, USA.
- Krekel, Christian und Alexander Zerrahn (5.–6. Juni 2014). "Sowing the Wind and Reaping the Whirlwind? The Effect of Wind Turbines on Residential Well-Being". Unveröffentlichter Vortrag, Energy, Environment and Well-Being : Workshop. Universität Oldenburg. Delmenhorst.
- (RatSWD), German Data Forum, Hrsg. (2010). Building on Progress. Expanding the Research Infrastructure for the Social, Economic and Behavioral Sciences. Budrich UniPress. url: http://www.ratswd.de/publikationen/building-on-progress
- Voigtländer, Sven u. a. (2011). Using geographically referenced data on environmental exposures for public health research: a feasibility study based on the German Socio-Economic Panel Study (SOEP). SOEPpaper 386. Berlin: DIW Berlin. url: http://www.diw.de/documents/publikationen/73/diw_01.c.375907.de/diw_sp0386.pdf
- Wagner, Gert G., Joachim R. Frick und Jürgen Schupp (2007). "The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements". In: Schmollers Jahrbuch (Journal of Applied Social Science Studies) 127.1, S. 139–169.
- Werlen, Benno (2008). Sozialgeographie. 3. Aufl. Bern: HauptVerlag.