

Bedarf der empirischen Sozialforschung an Geodaten

Jan Goebel

August 15, 2011

Vorbemerkung

Die Bedeutung von raumbezogenen Daten in der empirischen Sozialforschung ist in den letzten Jahren erheblich gestiegen. Tickamyer (2000) spricht davon, “that the natural and built environment the design of space and place, shapes social relation”.¹ Goodchild und Janelle (2004) sehen im Raumbezug eine der fundamentalen Kategorien in der Menschen die Welt erfahren.² Die zunehmende Bedeutung des Raums in den Sozialwissenschaften wird daher auch als “spatial turn” beschrieben.³ In der empirischen Sozialforschung heißt das natürlich, dass der Bedarf an Daten zur Beschreibung des menschliche Umfelds ebenfalls enorm gestiegen ist, dies trifft insbesondere auf georeferenzierte Daten zu.

Die im folgenden ausgeführte Problemanalyse und Einschätzung der derzeitigen Datenlage entsprechen der Sicht eines empirischen Sozialforschers und nicht der eines Geodatenexperten. Eine tiefere Einbindung georeferenzierter Daten in sozialwissenschaftliche Problemstellungen zeigen den deutlichen Bedarf einer engeren Kooperation zwischen den beteiligten Disziplinen.

Implementierung des Raumbezug in der empirischen Sozialforschung

Der klassische Einbezug des Raums in empirischen sozialwissenschaftlichen Studien folgte bisher über charakterisierende Merkmale der Verwaltungseinheit, in der die beobachtete Einheit (z. Bsp. Haushalte oder Personen) verortet ist. So wird zum Beispiel der Einfluss des regionalen Arbeitsmarktes über die Arbeitslosenquote des Kreises in ein Regressionsmodell mit aufgenommen, oder in einem Mehrebenenmodell geschätzt. Die

¹ Ann R. Tickamyer, Space Matters! Spatial Inequality in Future Sociology, Contemporary Sociology Vol. 29, No. 6 , 2000, pp. 805-813.

² Goodchild, M.F. and Janelle, D.G. (2004) Spatially Integrated Social Science. New York: Oxford University Press.

³ Goodchild, M. F. The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science, Journal of Official Statistics, Vol.23, No.3, 2007. pp. 269-283.

Auswahl dieses Vorgehens hat üblicherweise nur einen Grund: die Verfügbarkeit der Daten. Dies beinhaltet allerdings wiederum zwei Aspekte, zum einen die Verfügbarkeit der raumbeschreibenden Daten auf einer bestimmten Regionsebene und zum anderen die Möglichkeit an die genutzten Mikrodaten nur Daten auf bestimmten Regionsebenen zuspiesen zu können (meist ist in Haushaltsbefragungen eine Verortung der Einheiten bis auf die Kreisebene zum Teil auch bis auf die Postleitzahlenebene möglich).

Diese bisher übliche Vorgehensweise greift an mehreren Stellen zu kurz. Zwar liegen dank der amtlichen Statistik und dem BBSR eine Vielzahl an regionalen Indikatoren vor, diese basieren jedoch auf administrativen und damit für die lebensweltliche Erfahrung künstlichen Grenzen (üblicherweise NUTS1 bis LAU2, z.B. Kreise oder Gemeinden). Hinzu kommt, dass die Grenzen über die Zeit Änderungen ausgesetzt sind und das umso stärker, je kleinräumiger die Ebene (insbesondere in Ostdeutschland und durch die Bildung von Stadtregionen wie in Hannover oder Aachen) ist. Auch die Lage des Haushalts / der Person innerhalb der administrativen Einheit wird ausgeblendet. So kann es je nach Fragestellung sehr unterschiedliche Konsequenzen haben, ob der beobachtete Haushalt in der Mitte eines Kreises liegt oder sehr nah an der Grenze. Bei denen an der Grenze liegenden Haushalten ist die Information über die Situation im Nachbarkreis ein entscheidender Faktor. Auch die Zugänglichkeit der umliegenden Regionen durch z. Bsp. ÖPNV oder Bundesstraßen kann wichtig sein.

Wissenschaftliche Fragen bei der Beschreibung des räumlichen Umfelds sind normalerweise unabhängig von administrativen Einheiten und greifen meist auf das unscharfe Konstrukt "Nachbarschaft" zurück. Nachbarschaft liegt daher "irgendwo" zwischen der bisherigen Einbeziehung von Regionalisierung von (kleinräumigem) Haushalt und (großräumiger) Region (z. Bsp. Arbeitsmarkt). Eine tragfähige Operationalisierung ist bisher kaum oder gar nicht gelungen. Meist erfolgt die Operationalisierung über Kreisinformationen auf Grund der Datenverfügbarkeit, oder neuerdings auch über Postleitzahlen.

Auf der Ebene der Postleitzahlen sind allerdings deutlich weniger Indikatoren verfügbar, da diese im amtlichen Berichtswesen keine Rolle spielen. Unter anderem daher gibt es eine Kooperation zwischen dem SOEP, dem RWI und dem IAB, um Indikatoren auf der Ebene der Postleitzahlen dem SOEP zuzuspielen.

Aber auch die weitere regionale Verfeinerung auf Postleitzahlenebene löst das grundlegende Problem nicht, dass der Zuschnitt der Regionen nicht erfolgte um Nachbarschaften zu bestimmen und dass unklar ist, was die richtige Abgrenzung von Nachbarschaft überhaupt bedeutet. Benötigt werden zur Bestimmung von Nachbarschaften aber nicht nur möglichst kleinräumige Informationen, sondern auch die Möglichkeit diese räumlich zueinander in Bezug zu setzen und/oder zu aggregieren. Ebenso wichtig ist die Möglichkeit Nachbarschaften für jede Beobachtung individuell differenzieren zu können. Eine Annäherung an diese Problematik ist mit Hilfe einer räumlichen Projektion der bereits verfügbaren amtlichen Daten auf die entsprechenden Grenzen der Kreise oder Gemeinden möglich. Bei Vorliegen von georeferenzierten Befragungsdaten (z. Bsp. auf der Ebene der Straßenabschnitte um den Datenschutz der Befragten zu wahren) kann z. Bsp. durch die Berechnung eines Umkreises um jeden Beobachtungspunkt ein

(auch gewichteter) Mittelwert des jeweiligen Indikators berechnet werden.⁴ Idealerweise ist dem Forscher aber ein georeferenzierter Datensatz des externen Indikators bereits zugänglich, so dass die Rohdaten zum Beispiel beim Statistischen Bundesamt genutzt wurden ohne auf administrative Grenzen zu aggregieren, dieses digitale Kartenmaterial wäre dann als Vektor- und Rasterdaten dem Forscher zur Verfügung zu stellen. Detaillierte Daten von Bebauung (Liegenschaftskataster), Grundstückspreisen, sowie Straßen und Schienen würden es erlauben für jeden Haushalt individuell eine Nachbarschaft zu definieren, z. Bsp. abgetrennt durch größere Straßen, Flüsse oder Parks und mit ähnlichen Grundstückspreisen.

Ein Beispiel hierfür wäre die Ausweisung der Siedlungsstruktur, die sich bisher immer entlang der Gemeindegrenzen ändert. Zwar werden umliegende Informationen genutzt, der veröffentlichte Indikator ist aber der siedlungsstrukturelle Gemeindetyp auf der Ebene der Gemeinden. Um hochverdichtete Großstädte herum gibt es daher eine scharfe Grenze mit meist starken Unterschieden, die sich in der Realität mit den Ausläufern einer Großstadt meist nicht wiederfinden.⁵

Hier konnte mit Hilfe von Geobasis- und Geofachdaten oder den Corinne Daten Zusatzinformationen genutzt werden, um die Siedlungsstruktur flächendeckend ohne administrative Grenzen zu bestimmen und das Resultat als georeferenzierte Dateien der Wissenschaft zur Verfügung gestellt werden, ohne dass bestehende Datenschutzbestimmungen verletzt werden oder geändert werden müssten.⁶ Bei Indikatoren, die datenschutzrechtlich problematisch sind, könnte durch Rasterdaten (z.B. Dichten von Arbeitslosenzahlen, siehe auch Vortrag Schraepler) eine Nutzungsmöglichkeit geschaffen werden.

Ein weiteres Problem in der Nutzung von Geodaten in der empirischen Sozialforschung liegt darin, dass fast immer kleinräumige Daten für ganz Deutschland benötigt werden. Das findet seinen Grund in der Auswahl und der Größe von Zufallsstichproben. Um auf der Ebene der Mikrodaten noch genügend große Fallzahlen für die Analysen zu haben,⁷ müssen daher große Analyseregionen ausgewählt werden (z. Bsp. West- oder Ostdeutschland oder Großstädte mit mehr als 100.000 Einwohnern). Dies hat zur

⁴ Hierbei ist anzumerken, dass es bei diesem Verfahren zu keinem Zeitpunkt notwendig ist, die Straßenabschnittskoordinate mit den Befragungsdaten zu verbinden. Die Berechnung der Nachbarschaftsindikatoren braucht keinerlei Befragungsinformation und der Indikator kann nach der Berechnung über einen Haushaltsschlüssel an die Befragungsdaten zugespielt werden. Dass dies nur innerhalb eines FDZ bzw. mit Hilfe einer kontrollierten Datenfernverarbeitung möglich ist, ist selbstverständlich.

⁵ Eine Kartedesiedlungsstrukturellen Gemeindetyps findet sich unter:
http://www.bbsr.bund.de/cln_032/nn_103086/BBSR/DE/Raumbeobachtung/Werkzeuge/Raumabgrenzungen/SiedlungsstrukturelleGebietstypen/Gemeindetypen/download_karte08.pdf,templateId=raw,property=publicationFile.pdf/download_karte08_pdf.pdf.

⁶ Eine weiteres Beispiel wäre eine georeferenzierte Liste aller Schulen und Kitas in Deutschland (möglichst inkl. Anzahl an Plätzen und Schulform) um für jeden befragten Haushalt die Versorgungssituation im näheren Umkreis zu beschreiben. Diese Informationen liegen im Grunde bereits vor.

⁷ Und nicht zu Letzt, um Aussagen über die räumlichen Effekte in Deutschland machen zu können und nicht über die Situation in einer spezifische Kommune. Die Wichtigkeit regional spezifischer Analysen möchte ich dabei nicht in Zweifel ziehen. Diese werden aber meist in Kooperation mit der Kommune vor Ort erarbeitet und haben dadurch weniger Probleme mit dem Zugang von kleinräumigen Daten.

Folge, dass Geodaten im Rahmen der normalen Forschungspraxis in Universitäten und Forschungsinstituten nicht genutzt werden können; auf Grund der hohen Kosten, wie bereits Müller und Meinel in Ihrem Vortrag ausführten.

Der föderale Aufbau der Bundesrepublik spiegelt sich auch in der Haltung der Geodaten wider und erschwert den Zugang zu solchen Daten enorm. Ein Beispiel sind Informationen zum Umgebungslärm. Seit dem 30. Juni 2007 sind auf Basis der EU-Richtlinie “2002/49/EG Umgebungslärm” Städte und Gemeinden mit mehr als 250.000 Einwohnern verpflichtet, der interessierten Öffentlichkeit ein örtliches Lärmkataster zur Verfügung zu stellen. Diese Daten bleiben allerdings im Besitz der Kommune und die jeweilige Darstellung für die Öffentlichkeit ist je nach Kommune sehr unterschiedlich, zum Teil geschieht dies nur über pdf-Dateien,⁸ zum Teil über sehr ausgefeilte GIS-Viewer in sogenannten Geo-Portalen. Ein Versuch diese Daten für Auswertungen mit dem SOEP zu nutzen kam zu folgendem Ergebnis:

- Der download der eigentlichen digitalen Karten war in keiner der Kommunen möglich
- Auf eine Anfrage wurde zum Teil sehr langsam oder gar nicht reagiert. Die übliche Reaktion war Unklarheit darüber, ob diese Daten für wissenschaftliche Zwecke genutzt werden dürfen.
- In nur einem Fall (Duisburg) war es möglich, kostenfrei und schnell die digitale Karte zu erhalten.

Vergessen werden sollte jedoch nicht die notwendige Anpassung der Ausbildung von Sozialwissenschaftlern, um diese Daten auch effektiv nutzen zu können. Ebenfalls wichtig ist die Entwicklung von auf die Belange von fachfremden Wissenschaftlern angepasste (GIS) Software, damit solche Wissenschaftler mit den räumlichen Daten umgehen können.⁹ Als Beispiel bietet sich hier das Center for Spatially Integrated Social Science an (<http://www.csiss.org/>) oder das Advanced Spatial Analysis Program¹⁰.

⁸ Zum Teil wurden wohl auch die darunterliegenden georeferenzierten Rohdaten gelöscht, da die Erstellung der pdf Dateien von externen Anbietern durchgeführt wurde.

⁹ Für die Statistik Software R bestehen bereits Pakete die räumliche Daten und deren Analyse in den normalen Analyseablauf leicht integrieren lassen.

¹⁰ Gefördert von The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) <http://www.csiss.org/GISPopSci/>.