DOI: 10.2478/wd-2023-0207

sciendo

Wirtschaftsdienst, 2023, 103(11), 746-749

JEL: R00, H00, C80

Thomas K. Bauer, Christina Gathmann, Ines Helm, Christian Raffer, Sebastian Siegloch, Sandra Schaffner, Andreas Lichter*

Regionaldaten: Verknüpfungen ermöglichen

Regionale Ungleichheiten, etwa bei dem Einkommen oder dem Beschäftigungsniveau, sind ausgeprägt und persistent. Um ein besseres Verständnis über die Ursprünge und Auswirkungen regionaler Disparitäten sowie der Wirksamkeit regionalpolitischer Maßnahmen zu erlangen, sind detaillierte empirische Analysen verschiedener, politisch höchst relevanter Aspekte notwendig. Welche Ursachen hat regionale Ungleichheit? Welche Faktoren behindern eine Angleichung regionaler Lebensverhältnisse und welche Faktoren können eine Angleichung befördern? Welche politischen Maßnahmen können regionale Wirtschaftsräume und Arbeitsmärkte stärken? Die Grundlage derartiger empirischer Analysen sind qualitativ hochwertige Regionaldaten. Eine Umfrage unter ca. 500 Mitgliedern des Vereins für Socialpolitik (VfS) zu den Problemen des Datenzugangs für die Wissenschaft offenbarte jedoch insbesondere bezüglich der Qualität, der Zugangsmöglichkeiten zu und der Verknüpfung von Regionaldaten erhebliche Defizite. Knapp 80 % der befragten Mitglieder mit einem regionalökonomischen Forschungsinteresse sind mit dem Zugang zu Regionaldaten eher oder sehr unzufrieden, und 73 % gaben an, dass sie Projekte aufgrund von Defiziten beim Zugang zu Regionaldaten nicht durchführen konnten.

Problemlage bei existierenden Regionaldaten

Im Allgemeinen existiert eine Vielzahl an Datenquellen, die potenziell für empirische regionalökonomische Analysen verwendet werden können. Neben den amtlichen Daten des Statistischen Bundesamts, den INKAR-Daten des Bundesinstituts für Bau-, Stadt- und Raumforschung (BBSR) oder den Daten der Urban-Audit-Initiative der statistischen Ämter der Städte[1] stehen Daten mit Regionalbezug bei Forschungseinrichtungen und privaten Anbietern zur Verfügung oder können über die Verwendung geeigneter Algorithmen aus Internetangeboten oder Satellitendaten gewonnen werden.

So erfreulich der Umstand eines potenziell sehr umfangreichen Angebots an Regionaldatenquellen sein mag, besteht für Wissenschaftler:innen das Problem einer weitgehenden Intransparenz hinsichtlich der tatsächlich zur Verfügung stehenden Regionaldaten. Insbesondere bei Befragungsdaten und administrativen Daten ist häufig nicht bekannt, ob und über welchen Zeitraum sowie auf welcher Aggregationsebene Regionalindikatoren zur Verfügung stehen. Da Metainformationen über Regionaldaten vielfach fehlen und kaum standardisiert sind, erfolgt die Identifikation geeigneter Regionalinformationen für eine spezifische Forschungsfrage üblicherweise dezentral und händisch. Die Entwicklung eines geeigneten Metadatensystems und – darauf aufbauend – eines zentralen Repositoriums für Regionaldaten könnte diese Intransparenz erheblich verringern.

Ein ungleich größeres Problem liegt jedoch darin, dass die zur Verfügung stehenden Regionaldaten häufig auf einer zu hohen regionalen Aggregationsebene bereitgestellt werden. Regionale Daten werden vorwiegend auf Basis der Grenzen administrativer Verwaltungseinheiten (Gemeinden, Kreise, Regierungsbezirke, Bundesländer) oder unterschiedlich abgegrenzter Gebietseinheiten (Arbeitsamtsbezirke, Mikrozensuskreisregionen oder Raumordnungsregionen) zur Verfügung gestellt, wobei letztere nicht in allen Fällen mit den Grenzen administrativer Verwaltungseinheiten kompatibel sind. Darüber hinaus können sich verschiedene regionale Abgrenzungen über die Zeit ändern, beispielsweise aufgrund von Eingemeindungen oder Kreisreformen. Damit wird die Generierung von Paneldaten, die gerade für regionalökonomische Analysen notwendig sind, erheblich erschwert.[2]

In Einzelfällen werden Daten in Form von Shapefiles zur Verfügung gestellt, wobei jedoch für die meisten Gemeinden in Deutschland keine Shapefiles einzelner Stadtteile existieren. Daten privater Anbieter werden häufig in Straßenabschnitten gruppiert, bei landwirtschaftlichen Daten und dem Zensus kommen regionale Raster zum Einsatz,

1 Siehe https://www.staedtestatistik.de/arbeitsgemeinschaften/kosis/urban-audit (6. November 2023).

© Der/die Autor:in 2023. Open Access: Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht (creativecommons.org/licenses/by/4.0/deed.de).

Open Access wird durch die ZBW – Leibniz-Informationszentrum Wirtschaft gefördert.

* Dies ist eine Kurzversion der Stellungnahme der Unterarbeitsgruppe Regionaldaten. Die Langfassung findet sich hier: www.socialpolitik.de/vfs_stellungnahmen_daten_regionaldaten.

2 Zwar existieren Routinen, die versuchen, veränderte Raumabgrenzungen über die Zeit zu harmonisieren. Diese stehen aber lediglich für einen begrenzten Zeitraum (beispielsweise seit 1990) und nicht für alle in verschiedenen Daten verwendeten Raumabgrenzungen zur Verfügung.

wobei die einzelnen Beobachtungseinheiten häufig in Raster der Größe 500x500 Meter aggregiert werden. Diese Raster sind europaweit verfügbar, da sie der europäischen INSPIRE-Richtlinie folgen. Schließlich stehen für eine steigende Zahl von Regionalinformationen exakte Geopositionen (mit Angabe des Breiten- und Längengrads) zur Verfügung. Diese Daten werden üblicherweise aus Satellitendaten, Kartenangeboten aus dem Internet oder durch Georeferenzierung von Adressen gewonnen.

Prinzipiell stehen in vielen Befragungsdaten und administrativen Daten kleinräumige Regionalinformationen zur Verfügung, werden jedoch der Wissenschaft nicht zur Verfügung gestellt. So existieren für jeden sozialversicherungspflichtigen Beschäftigten kleinräumige Informationen zum Wohn- und Arbeitsort. Der Zugriff auf diese Daten und damit die Analyse regionaler Arbeitsmärkte, der regionalen Mobilität von Beschäftigten oder der Determinanten des Pendlerverhaltens ist jedoch nur in sehr spezifischen Fällen möglich. Der AfiD-Unternehmensdatensatz wird der Wissenschaft nicht von allen Bundesländern zur Verfügung gestellt und beim Taxpayer-Panel, der wichtigsten Mikro-Steuerstatistik des Landes, sind Gemeindeinformationen erst ab dem Jahr 2012 verfügbar, für Bayern allerdings nur in pseudonomisierter Form.[3] Weitere Beispiele umfassen Daten der statistischen Ämter größerer Gemeinden, die teilweise Informationen auf Stadtteilbezirken erfassen, Grundsteuerdaten, Daten der Bundesnetzagentur, die Transaktionsdaten der Gutachterausschüsse von Immobilien und eine Vielzahl von Befragungsdaten, wie beispielsweise das Nationale Bildungspanel oder der ALLBUS. Wissenschaftler:innen haben zu diesen Daten zumeist keinen oder nur einen eingeschränkten Zugang. Vielmehr werden kleinräumige Regionalinformationen vor allem aus datenschutzrechtlichen Gründen häufig nicht oder nur auf einer relativ stark aggregierten regionalen Ebene zur Verfügung gestellt, da eine Verknüpfung selbst faktisch anonymisierter Befragungsdaten mit Regionalinformationen aus anderen Quellen die Identifikation einzelner Individuen oder Unternehmen vereinfachen könnte.

Die heterogenen Aggregationsebenen der zur Verfügung stehenden Regionaldaten führen bei wissenschaftlichen Analysen zu erheblichen Herausforderungen. Stehen unterschiedliche Datensätze oder verschiedene relevante Variablen nur auf unterschiedlichen regionalen Ebenen zur Verfügung, müssen die Daten auf die jeweils höchste Ebene aggregiert werden. Will man beispielsweise Befragungsdaten, in denen für die einzelnen Beobachtungseinheiten lediglich Regionalidentifikatoren auf der Bundeslandebene zur Verfügung stehen, mit kleinräumigen Regionalinformationen aus anderen Datenquellen verknüpfen, müssen letztere auf die Bundeslandebene aggregiert werden. Darüber hinaus ist dies aufgrund der fehlenden Kompatibilität verschiedener Raumabgrenzungen häufig nicht eindeutig möglich. Somit werden aufgrund der Notwendigkeit der Aggregation der Regionalinformationen auf die jeweils höchste zur Verfügung stehende regionale Abgrenzung die adäquat zu analysie-

3 Siehe https://www.forschungsdatenzentrum.de/de/steuern/tpp (6. November 2023).

Prof. Dr. Thomas K. Bauer ist Mitglied im Vorstand und Vizepräsident des RWI – Leibniz-Institut für Wirtschaftsforschung in Essen sowie Lehrstuhlinhaber für Empirische Wirtschaftsforschung an der Ruhr-Universität Bochum.

Prof. Christina Gathmann, Ph.D., ist Leiterin der Arbeitsmarktabteilung am Luxembourg Institute of Socio-Economic Research (LISER) und Professorin an der Universität Luxemburg.

Prof. Ines Helm, Ph.D., ist Ass.-Professorin am Department of Economics an der LMU München und Research Affiliate am Forschungsinstitut zur Zukunft der Arbeit (IZA).

Dr. Christian Raffer ist wissenschaftlicher Mitarbeiter am Forschungsbereichs Infrastruktur, Wirtschaft und Finanzen am Deutschen Institut für Urbanistik gGmbH in Berlin.

Prof. Dr. Sebastian Siegloch ist Professor für Volkswirtschaft an der Universität Köln and Mitglied der DFG Excellence Cluster ECONtribute der Universitäten Köln und Bonn.

Dr. Sandra Schaffner ist Leiterin des Forschungsdatenzentrums (FDZ Ruhr) am RWI in Essen und Mitglied des Kompetenzbereichs Arbeitsmärkte, Bildung, Bevölkerung.

Prof. Dr. Andreas Lichter ist Juniorprofessor für VWL, insbesondere Angewandte Mikroökonomie am Düsseldorf Institute for Competition Economics (DICE) der Heinrich-Heine-Universität Düsseldorf.

renden regionalökonomischen Fragestellungen erheblich eingeschränkt. So ist beispielsweise die für die Wissenschaft zur Verfügung stehende niedrigste Aggregationsebene im Regionalfile des Mikrozensus die sogenannte Mikrozensuskreisregion (MZKR), eine konstruierte Raumeinheit mit mindestens 100.000 Einwohner:innen. Zwar kann man dem Mikrozensus prinzipiell auf Gemeindeebene Information zuspielen, jedoch keine Auswertungen auf dieser durchführen. Regionalökonomische Analysen schließen sich damit de facto aus.

Neben den Problemen einer (zu) hohen Aggregationsebene und des fehlenden Zugriffs auf existierende kleinräumige Informationen ist der Bezug von regionalen amtlichen Daten aus mehr als einem Bundesland, die nicht bereits über www.regionalstatistik.de angeboten werden, sehr aufwendig und mit erheblichen Kosten verbunden. In der Regel erfolgt die Bereitstellung derartiger Daten über eine „koordinierte Abfrage", die vom Statistischen Landesamt des Bundeslandes der bzw. des jeweiligen Forschenden durchgeführt wird. Dieses fragt dann die Möglichkeit und die Kosten der Bereitstellung der Daten bei allen anderen Landesämtern an und erstellt ein Angebot. Dieses Prozedere kann mehrere Wochen dauern.[4] Ebenso problematisch sind die Kosten der Bereitstellung dieser Daten. Insbesondere bei Paneldaten können die entstehenden Kosten sehr hoch sein, da die Rechnungsgrundlage der Datenbereitstellung nicht nur die Zahl der Einheiten umfasst, sondern für jedes bezogene Jahr gesondert gezahlt werden muss. Insbesondere für Nachwuchsforschende stellen diese Kosten der Datenbereitstellung eine prohibitive Hürde dar.

Fehlende Regionaldaten

Neben den aufgezeigten Problemen existieren darüber hinaus bedeutende Datenlücken. Für viele regionalökonomische Fragestellungen, wie beispielsweise die Analyse regionaler Ungleichheiten, wären Informationen über räumliche Preisniveauunterschiede zentral, nicht zuletzt da regionale Unterschiede in den nominalen Einkommen nur unzureichende Aussagekraft über die regionale reale Kaufkraft zulassen. Deshalb wird sowohl aus Kreisen der Wissenschaft als auch der Politik seit langem die Erhebung regionaler Preisindizes gefordert. Konzeptionen zur Machbarkeit der Erhebung regionaler Preisindizes liegen vor.[5] Das BBSR und das Institut der deutschen Wirtschaft (IW) haben im Oktober 2023 einen regionalen Preisindex vorgestellt. Wir sehen in diesem Bereich die dringende Notwendigkeit, die jährliche Erhebung regionaler Preisindizes und deren Bereitstellung für wissenschaftliche Analysen nachhaltig sicherzustellen. Zudem wäre eine flächendeckende Bereitstellung der Daten der regionalen Gutachterausschüsse wünschenswert, um das erhebliche Informationsdefizit über den deutschen Immobilienmarkt zumindest teilweise zu beheben. Auch wären Daten zum Handel von Gütern und Dienstleistungen zwischen Regionen innerhalb Deutschlands bedeutsam, um regionale wirtschaftliche Verflechtungen abbilden und analysieren zu können.

Die Entwicklung regionaler Disparitäten ist zudem ein langfristiger Prozess. Daher erfordert die empirische Analyse der Determinanten und der Dynamik regionaler Ungleichheiten die Verfügbarkeit sehr langer Beobachtungszeiträume. Nicht zuletzt deshalb versuchen Wissenschaftler:innen verstärkt, historische Bestände an regionalen Indikatoren zu identifizieren und maschinenlesbar zu digitalisieren. Trotz erheblicher technischer Fortschritte in der automatisierten Erkennung und Verarbeitung von Texten (Textmining), ist die Übertragung und Georeferenzierung digitalisierter Bibliotheksbestände in ein Format, das von statistischen Softwarepaketen gelesen werden kann, nach wie vor personal- und zeitintensiv. Dies liegt neben den bereits erwähnten Veränderungen von Gemeinde- oder Kreisgrenzen über die Zeit beispielsweise auch an zeitlichen Veränderungen der Schreibweisen von Gemeindenamen. Darüber hinaus wurde der Bestand der in Bibliotheken zur Verfügung stehenden historischen Veröffentlichungen der Statistischen Ämter (beispielsweise durch das Statistische Amt des Deutschen Reiches oder der Weimarer Republik) bisher noch nicht vollständig digitalisiert. Hier liegt aus unserer Sicht ein weitgehend ungehobener Schatz an Regionaldaten, deren Erfassung die Forschung zu den Determinanten und der Entwicklung regionaler Disparitäten erheblich befördern würde.

Handlungsvorschläge

Vor dem Hintergrund der beschriebenen Datenprobleme für empirische regionalpolitische Analysen ergeben sich folgende Handlungsvorschläge:

1. Eine qualitative Verbesserung des Angebots an Regionaldaten in Deutschland erfordert die Entwicklung eines geeigneten Metadatensystems und darauf aufbauend die systematische Erfassung, Speicherung und Bereitstellung bereits existierender amtlicher sowie nichtamtlicher Regionalinformationen in einem einheitlichen Repositorium, um die derzeit bestehende Intransparenz des Bestands an Regionalinformationen aufzulösen. Das Repositorium könnte in einer zentralen Informationsstelle für Forschungsdaten angesie-

4 In Frankreich kann man einen Antrag auf Datenbereitstellung digital auf einer zentralen Plattform einreichen. Die verantwortlichen Datenbereitsteller prüfen den Antrag und stellen gegebenenfalls Nachfragen. Üblicherweise wird der Antrag innerhalb von wenigen Tagen genehmigt.

5 Siehe beispielsweise Möller et al. (Hrsg.) (2010) und BBSR (2009).

delt werden. Die Erfassung sollte dabei interdisziplinär erfolgen (beispielsweise unter Einbeziehung existierender Daten der Raumforschung) und die Ämter der Städtestatistik sowie historische Datenbestände einbeziehen. Darüber hinaus sollte das regionale Metadatensystem die potenziell niedrigste Ebene administrativer Daten und mit öffentlichen Mitteln finanzierter Befragungsdaten erfassen.

2. Die Löschung von Regionalidentifikatoren in Datensätzen sollte mittelfristig ausgeschlossen werden. Regelungen sind so anzupassen, dass die möglichst niedrigste räumliche Ebene einer Beobachtungseinheit nachhaltig als exakter Geocode gespeichert wird. Bei Individualdaten wäre dies die Wohnadresse, bei Unternehmen und Betrieben deren Standortadresse. Damit könnte man eine flexible Aggregation der Daten sowie eine adäquate Verknüpfung verschiedener Regionaldaten sicherstellen.

3. Die Verknüpfung von Mikrodaten mit Regionalinformationen sollte auf der niedrigsten regionalen Aggregationsebene angestrebt werden. Die datenschutzkonforme Bereitstellung derartiger Daten könnte über einen Datentreuhänder in einer zentralen Informationsstelle für Forschungsdaten erfolgen und Wissenschaftler:innen über einen Remote Access mit Outputprüfung oder Gastwissenschaftleraufenthalt zur Verfügung gestellt werden.

4. Amtliche Daten mit Regionalinformationen, die nicht bereits über www.regionalstatistik.de für alle Bundesländer angeboten werden, sollten standardmäßig beim Statistischen Bundesamt vorgehalten und der Wissenschaft kostenlos zur Verfügung gestellt werden, damit die derzeit für alle Beteiligten kosten- und zeitintensiven koordinierten Abfragen über die Statistischen Landesämter entfallen können.

5. Existierende Datenbestände mit kleinräumigen Informationen sollten der Wissenschaft (unter erleichtertem Voraussetzungen) zur Verfügung stehen. Schließlich wäre eine stärkere Förderung der Identifikation, Digitalisierung und maschinenlesbaren Bereitstellung von statistischen Archivmaterialien wünschenswert.

6. Bedeutende Lücken in der Bereitstellung von Regionalinformationen sollten schnellstmöglich behoben werden. Dies beinhaltet insbesondere die Erhebung jährlicher regionaler Preisindizes, aber auch die Erfassung des Handels von Gütern und Dienstleistungen zwischen Regionen innerhalb Deutschlands.

Die Erfassung bzw. Beibehaltung kleinräumiger Regionalinformationen, die Abkehr von regionaler Aggregation, die Bereitstellung von Informationen in einem Metadatensystem und die datenschutzadäquate Bereitstellung der Daten mit Regionalidentifikatoren auf der jeweils kleinstmöglichen regionalen Ebene für die Wissenschaft könnte unsere Erkenntnisse über die Ursachen und Konsequenzen regionaler Disparitäten und damit die Grundlage evidenzbasierter Politikberatung im Bereich der Regionalökonomik erheblich verbessern.

Literatur

BBSR – Bundesinstitut für Bau-, Stadt- und Raumforschung (Hrsg.) (2009), Regionaler Preisindex, Berichte, Bd. 30.

Möller, J., E. Hohmann und D. Huschka (Hrsg.) (2010), *Der weiße Fleck – zur Konzeption und Machbarkeit regionaler Preisindizes*, IAB-Bibliothek (Gutachten), 324, W. Bertelsmann Verlag, https://doi.org/10.3278/300713w.

Title: ***Regional Data: Enabling Linkages***

Abstract: *Regional disparities are pronounced and persistent. Regional data of high quality are needed to improve the understanding of the origins and consequences of regional disparities and the effectiveness of regional policies through empirical analysis. Against the background of existing problems with regard to both the access to and the provision of high-quality regional data, this paper derives recommendations for action to improve the supply of regional data for the purpose of scientific analyses.*