Editorial

Registerforschung: Chancen, Risiken und Herausforderungen

Wie alle empirischen Wissenschaften sind die Sozialwissenschaften auf Beobachtungen ihres Untersuchungsgegenstandes angewiesen. Für viele wirtschafts- und gesellschaftspolitische Fragestellungen und Herausforderungen sind empirische Daten Voraussetzung für eine evidenzbasierte Evaluierung und Planung von Maßnahmen. Die öffentliche Hand ist dabei eine wichtige Datenquelle, viele Informationen aus Wirtschaft und Gesellschaft laufen in ihren Datenbanken zusammen. Um den Zugang zu diesen Register-, Verwaltungs- und für die amtliche Statistik erhobenen Daten für die Wissenschaft ist in jüngster Zeit eine rege Debatte entstanden.

Die Positionen reichen von der Forderung nach einem allgemeinen Recht auf diese Daten, über einen einfachen und unbürokratischen Zugang zumindest für die Wissenschaft bis hin zur völligen Ablehnung jeglicher über den *Status quo* hinausgehender Einsichtnahme und Verwendung. Aus den antagonistischen Positionen wird ersichtlich, dass die Frage der Sammlung, Bereitstellung und Nutzung von Daten von unterschiedlichen Interessen geleitet und von Machtverhältnissen geprägt ist.

ÖkonomInnen, Parteien und zivilgesellschaftliche Organisationen erhoffen in den Datenbeständen der öffentlichen Verwaltung einen Wissensschatz zur Beantwortung vieler Fragen. Andererseits ist die Skepsis gegenüber dem offenen Zugang zu Individualdaten in der Bevölkerung gestiegen, auch angesichts der stets wachsenden Möglichkeiten der Aus- und Verwertung großer Datenbestände.

Da beide Argumente ihre Berechtigung haben und beide Interessen nachvollziehbar sind, wollen wir die Chancen und Risiken beleuchten, die eine Nutzung dieser Datensammlungen mit sich bringt.

Ursprünge staatlicher Datenerhebung

Die älteste Form staatlicher Datenerhebung sind Volkszählungen, deren Ursprünge bis ins alte Ägypten zurückreichen. Das bekannteste historische Beispiel einer solch staatlich organisierten Großregistrierung ist der von Kaiser Augustus zu Christi Geburt beauftragte Zensus. Begründet und motiviert wurde diese Erhebung mit der Identifizierung

von wehrfähigen – und im Bedarfsfall auch wehrpflichtigen – Männern, sowie mit der Feststellung des potenziellen Steueraufkommens und der Organisation der Steuereintreibung. Nicht ohne Grund entzogen sich jene, die mächtig genug waren, über lange Phasen diesen Erhebungen.

Dass Volkszählungen historisch zur Aufrechterhaltung von Herrschaftsverhältnissen eingesetzt wurden, zeigt auch Thomas Piketty in seinem neuen Buch „Kapital und Ideologie". Die ersten Volkszählungen der britischen Kolonialmacht in Indien diente den Besatzern, sich ein umfassendes Bild über zur Zwangsarbeit einsetzbare Bevölkerungsgruppen zu verschaffen. Diese Beispiele verdeutlichen, weshalb umfassende Erhebungen der Zentralmacht häufig mit staatlicher Zwangsgewalt durchgeführt werden mussten.

Register und Kataster schufen aber zugleich auch Rechtssicherheit für die Bevölkerung. In diesem Kontext ist der Franziszeische Kataster von Kaiser Franz I. berühmt geworden. Als erstes umfassendes Grundbuch wurde damit eine wesentliche Grundlage für den Schutz des Privateigentums an Grund und Boden gelegt. Und so trägt auch die staatliche Datenerhebung – wie viele Errungenschaften der Aufklärung – das Janusgesicht von Kontrolle aber auch Schutz durch die Zentralmacht.

Entscheidungen darüber, welche Daten (zentral) verfügbar sind und welche nicht, sind deshalb stets auch Ausdruck von Macht- und Herrschaftsverhältnissen. So ist es ist kein Zufall, dass in Österreich versicherte Arbeitsverhältnisse inzwischen lückenlos dokumentiert und auch wissenschaftlich auswertbar sind, es aber keine verpflichtende Erhebung der Privatvermögen gibt.

Der Umgang der Demokratie mit Datensammlungen

Der Umgang mit staatlichen Datensammlungen erfordert eine Abwägung von Kontrolle und Freiheit, von Fürsorge und Eigenverantwortung. Das müssen keineswegs unauflösbare Dichotomien sein. Vielmehr fordern diese Gegenüberstellungen auf zu berücksichtigen, wer wen kontrolliert, wessen Freiheit geschützt wird, wer für wen sorgen muss, und wer sich um sich selbst kümmern muss oder darf.

Manche argumentieren, es wäre Verschwendung, bereits vorhandene Daten nicht zur Gewinnung neuer Erkenntnisse zu nutzen. Sie ignorieren dabei schlichtweg die demokratiepolitischen Gefahren. Denn bei genauer Betrachtung kann dies einen Freibrief für die Überwachung der BürgerInnen, die Kontrolle politischer GegnerInnen oder die Kommerzialisierung öffentlicher Daten bedeuten.

Eine reflektiertere Argumentation lautet, dass es angesichts der per

Zwang gesammelten Daten wichtig ist, diese auch für ÖkonomInnen, BürgerInnen und zivilgesellschaftlichen Organisationen nutzbar zu machen, um ein Gegengewicht (oder zumindest eine Kontrolle) der Staatsgewalt sicherzustellen. Dieser Argumentation entsprechen die inzwischen verbreitete Politik der Open Government Data bzw. das Prinzip des allgemeinen Zugangs zu Verwaltungsdaten, solange nichts für die Geheimhaltung spricht. Doch auch in diesem Fall sind es staatliche AkteurInnen, die über die Auswahl der bereitgestellten und der zurückgehaltenen Datensätze entscheiden.

Die Sammlung, Bereitstellung und Auswertung von Daten gehen als aktive Handlungen immer auf die Interessen der jeweiligen AkteurInnen zurück. Staatliche Organisationen, Unternehmen, Parteien, zivilgesellschaftliche Organisationen und auch ÖkonomInnen fordern Veränderungen oder Erhalt des Bestehenden nicht zuletzt auf Basis von Eigen- und Gruppeninteressen. Die Frage, wer Zugriff zu welchen Daten und in welcher Form bekommt, lässt sich also nicht pauschal beantworten. Es braucht Mechanismen der Abwägung sowie Checks und Balances. Im folgenden Abschnitt werden wir auf einige aus unserer Sicht relevante Abwägungen eingehen und die verschiedenen Interessen beleuchten.

Risiken bei der Verwendung von Registerdaten

Ein zentraler Bereich, in dem die Risiken der Verwendung individueller Daten schon lange diskutiert werden, ist der Schutz persönlicher Daten. Darunter fällt etwa die Wohnadresse. Durch Verfahren wie den „differenziellen Datenschutz“ („differential privacy“) wird versucht, einen Kompromiss zwischen Datenschutz und Erkenntnissinteressen von ÖkonomInnen zu finden. Es gibt aber auch weitere Gründe, mit Registerdaten sehr vorsichtig umzugehen. Gerade diese werden auch mangels technischer Lösung weniger oft behandelt.

So stellt sich bei jeder auf Individualdaten beruhenden Untersuchung das Problem, dass häufig aus dem Vorhandensein individueller Merkmale allgemeine Schlüsse auf das Verhalten von zumindest in diesen Merkmalen vergleichbaren Personen gezogen werden. Ein derartiger Vergleich verletzt – auch wenn er zutrifft – das Recht der einzelnen Menschen auf eine vorurteilsfreie Beurteilung.

a. Schutz persönlicher Daten

Bei den meisten Personen wird sich ein unangenehmes Gefühl einschleichen, wenn sie erkennen wie detailliert ihr Leben heutzutage in

verschiedenen Datenbeständen dokumentiert ist. Allen, die daran zweifeln, sei ein Blick in auf ihr Google-Profil und die von Google erstellten Monatsberichte empfohlen. Jedenfalls ist es im digitalen Zeitalter für private Konzerne möglich, umfassende Datenerfassungen und Bewegungsprofile ihrer NutzerInnen anzulegen. Diese teils offene, teils verborgene Datensammlung durch große, unkontrollierte Technologiekonzerne bedeutet für viele Menschen einen ungewollten Eingriff in ihre Privatsphäre und eine kommerzielle Verwertung persönlicher Informationen.

Die Hoffnung, es handle sich „nur" um private Sammlungen, bei denen niemand gezwungen ist mitzumachen, ist trügerisch. Es wird etwa immer üblicher, dass bei Bewerbungen der eigene *Facebook-Account* angegeben werden muss, und bei immer mehr Jobs gehört der „private" *Twitterfeed* zur Tätigkeitsbeschreibung. Und die in der Corona-Pandemie entwickelte *Handy-App* hat einen Konflikt zwischen unsolidarischer Nicht-Teilnahme am *Contact Tracing* und der Freigabe der Standortdaten an Google aufgemacht. Das sind also Entscheidungen, die nur mehr bei sehr weiter Interpretation als frei gelten. Dazu kommt, dass angesichts der weiter steigenden NutzerInnenzahlen ein Ausschluss von diesen Plattformen einem Rückzug aus der gesellschaftlichen Teilhabe nahekommt. Ein bloßes *„love it or leave it"* wird der Bedeutung dieser Plattformen nicht gerecht.

Bei öffentlichen Datensätzen, insbesondere bei Registerdaten, ist diese Freiheit (also die Möglichkeit zum Rausoptieren) üblicherweise gar nicht gegeben. So werden etwa Sozialversicherungsdaten gesammelt und gespeichert, um die Pensionsansprüche der Beschäftigten berechnen zu können. Bei der Debatte zur e-card fällt auf, dass die Datenfreigabe für Forschungszwecke nicht mit dem sonstigen Entzug wichtiger Funktionalitäten, wie der e-Medikation, erzwungen werden darf[1].

In der aktuellen Diskussion geht es den WissenschaftlerInnen, die sich öffentlich äußern, in erster Linie um den Zugang zu Mikrodaten der öffentlichen Verwaltung und von Statistik Austria. Von privaten Anbietern, etwa von Google, gesammelte Daten werden von diesen Forderungen nicht direkt berührt. Aber die Verknüpfung der beiden Datenquellen passiert heute schon öfter als allgemein bekannt. Preiserhebungen der Statistik (z. B. Verbraucherpreisindex) basieren häufig auf Webscraping, dabei werden Internetseiten automatisiert nach relevanten Informationen durchsucht. Und während der Corona-Krise hat etwa das Statistische Bundesamt Deutschland experimentelle Daten zum Nachzeichnen von Mobilitätsveränderungen während der Corona-Krise verwendet. Auch Statistik Austria hat angekündigt, sich in Zukunft verstärkt der experimentellen Statistik zu widmen und damit auch den privaten Raum „Internet" für die Datenproduktion nutzen zu wollen.

In all diesen Bereichen müssen selbstverständlich die Grundsätze des Datenschutzes beachtet werden, die auch in der europäischen Datenschutzgrundverordnung (DSGVO) festgehalten sind. Diese führt sieben Prinzipien an: 1. Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz; 2. Zweckbindung; 3. Datenminimierung, Datensparsamkeit; 4. Richtigkeit; 5. Speicher(-zeit)begrenzung; 6. Integrität und Vertraulichkeit sowie 7. Rechenschaftspflicht.

Diese Pflichten gelten zwar nur für die Verarbeitung personenbezogener Daten, der Begriff ist aber durchaus weit gefasst, womit viele der bereits heute von ÖkonomInnen genutzten Daten als personenbezogene Informationen gelten. Auch wenn die Einhaltung des Datenschutzes schon als Innovations- und Forschungsbremse beklagt wurde, so wichtig ist es gerade für die ÖkonomInnen, dass bereits in der Organisation des Datenzugangs Vorkehrungen getroffen werden, die Probleme mit dem Datenschutz ausschließen. Gerade junge ForscherInnen an kleineren Instituten haben in der Regel nicht die juristische und technische Unterstützung, die notwendig ist, um sie vor Datentücken, rechtlichen Fallen und unbeabsichtigten Datenschutzverletzungen zu bewahren. Im Sinne einer freien Forschung sollte ihnen ein Zugang gewährt werden, der die Verletzung des Datenschutzes ausschließt.

b. Schutz vor Diskriminierung

Im Gegensatz zu den Gefahren für den Datenschutz werden die potenziellen Auswirkungen der Nutzung individueller Daten auch für nicht unmittelbar erfasste Personen oft übersehen. Dies sei an einem fiktiven Beispiel illustriert: Eine Studie könnte feststellen, dass Männer, die am 5. Oktober in Salzburg geboren wurden, ein im Vergleich zum Durchschnitt der SalzburgerInnen 80% höheres Risiko haben, an einer Krankheit zu erkranken, deren Behandlung extrem teuer ist.

ÖkonomInnen werden nun verstehen, dass es für die AnbieterInnen von Versicherungen unprofitabel ist am 5. Oktober geborene Personen zum selben Tarif wie allen anderen SalzburgerInnen zu versichern. Für diese Gruppe würde es dann wegen verschiedener Versicherungsprämien (zu) teuer sich gegen dieses hohe Gesundheitsrisiko abzusichern. Aus neoklassischer Sicht wäre das gar keine Benachteiligung, sondern – betrachtet man den Erwartungswert – eine faire Lösung. Allerdings können alle aus dieser Gruppe, die nicht an dieser Krankheit erkrankten, zurecht argumentieren, dass sie aufgrund eines unverschuldeten Umstandes, daran gehindert wurden, sich gegen dieses Risiko zu schützen, sie also diskriminiert wurden.

Das Beispiel ist zwar ein Extremes, aber keineswegs vollkommen fiktiv. Es gibt ausreichend Berichte, insbesondere über Auswertungen

von Gesundheitsdaten mit Hilfe von neuronalen Netzen, die derartige Ungleichbehandlungen nahelegen. Ein jüngstes Beispiel bot der Versuch der britischen Regierung, die Abschlussnoten von SchülerInnen einerseits zu schätzen und dann andererseits entsprechend zu korrigieren, und zwar nach schulspezifischen Erfahrungswerten.

Ein solcher Algorithmus hätte wohl auch die universitären Erfolgsaussichten von Gabriele Possanner von Ehrenthald, der ersten Medizinerin, Cäcilie Wendt, der ersten Mathematikerin, oder Elise Richter, der ersten Habilitandin an der Uni Wien, sehr gering eingeschätzt und große Karrieren verhindern können. Die Wahrscheinlichkeitsrechnung ist für die Einzelperson ein schwacher Trost: Kaum jemand bekommt genügend Lebenschancen um seine mittlere Lebensqualität am Erwartungswert auszurichten.

Gegen die Verwendung von noch mehr Individualdaten sprechen die, gerade in der Ökonomie, gut erforschten negativen Konsequenzen statistischer Diskriminierung auf Basis von Unterschieden zwischen Gruppen. Insbesondere bei starken Resultaten besteht die Gefahr, dass Menschen ihr Verhalten anpassen und eventuell besonders talentierte Personen mit sehr "schlechten" Merkmalen gewisse Ausbildungen, Berufe oder Positionen gar nicht mehr anstreben. In anderen Worten, Diskriminierung wirkt selbstbestätigend, wie auch im einflussreichen Papier von Spences (1973) gezeigt wird. Gleichzeitig haben neuere Modelle zur statistischen Diskriminierung gezeigt, dass Kenntnis über detailliertere Charakteristika von Individuen und nicht nur deren Gruppenzugehörigkeit der statistischen Diskriminierung entgegenwirken können (Lang/Kahn-Lang Spitzer 2020).

c. Schutz der Selbstbestimmung

Neben dem Schutz vor Diskriminierung ist auch der Schutz der informationellen Selbstbestimmung relevant. ÖkonomInnen tendieren dazu, den in dieser Frage Betroffenen rasch strategische Motive zu unterstellen. Diese würden dazu führen, die Effizienz öffentlicher Maßnahmensteuerung zu beeinträchtigen. Ebenso plausibel ist die Vermutung, dass Personen ihre Beteiligung verweigern könnten, weil sie die untersuchten Zusammenhänge in ihrem speziellen Fall für nicht untersuchbar halten.

Zur Illustration sei die konkrete Frage angeführt, wie sich das Ende von Arbeitslosengeldzahlungen auf die Wiederbeschäftigung auswirkt. Betrachtet man hier den in den Verwaltungsdaten kodierten Status Arbeitslosigkeit, so sieht man eine signifikante Häufung der Abgänge zum Ende des Arbeitslosengeldbezuges. In anderen Worten, es scheint als würden sich Betroffene so lange arbeitslos melden wie sie bezugsbe-

rechtigt sind; im Umkehrschluss würde eine Verkürzung der Anspruchsberechtigung die Arbeitslosigkeit verkürzen. Das ist allerdings ein Fehlschluss: Eine exzellente Studie mithilfe von Registerdaten des Sozialversicherungssystems (Card et al. 2007) zeigt, dass dieser Anstieg weitgehend verschwindet, wenn anstelle des Verlassens der Arbeitslosenversicherung (die ja auch in Richtung Sozialhilfe oder Ausstieg aus dem Arbeitsmarkt gehen kann) die Aufnahme einer neuen Beschäftigung betrachtet wird.

Nur ein breiteres, auch die Erfahrungswelt der Betroffenen einbeziehendes Forschungsdesign, kann hier zu realistischen Schlussfolgerungen kommen. Doch selbst wenn den Betroffenen klar ist, dass es diesen Zusammenhang in ihrem Fall gab und sie nur befürchten, dass die Studie ihren eigenen Interessen schaden würde, ist es fraglich, unter welchen Bedingungen man sie direkt oder indirekt zwingen kann, ihre Daten zur Verfügung zu stellen. Im Fall von Steuern beispielsweise wird diesbezüglich ein sehr strenger Maßstab angelegt, eine Konsolidierung von Einkommensteuerdaten und Daten aus der Kapitalertragsbesteuerung ist nach wie vor nicht möglich. Die Begründung, dass dies angesichts des Endbesteuerungscharakters der KESt auch logisch sei, ist nicht plausibel, da gerade die Nicht-Konsolidierbarkeit der wesentliche politische Grund für die Einführung der Endbesteuerung war.

Es würde der Sache der Wissenschaft schaden, sich über legitime Forderungen der BürgerInnen nach informationeller Selbstbestimmung hinwegzusetzen. Beim Auftreten von Datenlecks, Hackerattacken oder groben Fehlern im Datenschutz würde die Verletzung dieses Grundrechts schnell zu einem Verlust der öffentlichen Unterstützung und dem Ausschluss der Wissenschaft aus der Nutzung solcher Daten führen.

Die Potenziale von Mikrodatenforschung

Viele aggregierte und gruppierte Wirtschaftsdaten sind für ForscherInnen und Interessierte einfach zugänglich. Doch viele gesellschafts- und wirtschaftspolitische Fragestellungen lassen sich nur schwer anhand dieser Zahlen beantworten. Der Rückschluss aus Summen und Durchschnittswerten auf individuelle Dynamiken ist nicht möglich.

Es droht ein „ökologischer Fehlschluss“, wenn auf Basis aggregierter Daten, welche die Merkmale von Gruppen darstellen, fälschlicherweise auf Individuen geschlossen wird. Ein häufig zitiertes Beispiel kommt von W. S. Robinson (1950): Für die USA in den 1930er-Jahre stellte er eine starke positive Korrelation (0.77) zwischen dem Anteil von Schwarzen an der Gesamtbevölkerung und dem Anteil der AnalphabetInnen auf Ebene der Bundestaaten fest. Dieser starke Zusammenhang steht

aber in deutlichem Widerspruch zur viel schwächeren Korrelation zwischen diesen Merkmalen auf Individualebene (0.20). Robinson verwies anhand dieses Beispiels auf die mit der Verwendung aggregierter und gruppierter Daten einhergehenden Gefahren.

Das Vermeiden von „ökologischen Fehlschlüssen" ist eines von vielen Problemen, die das Verlangen der Wissenschaft nach möglichst umfassenden Einzelfalldaten – etwa auf Ebene von Personen oder Unternehmen – begründet.

Fest steht, dass Daten in aggregierter Form zur Beantwortung vieler Fragestellungen nicht ausreichen. Denn häufig stehen ja die Unterschiede innerhalb von Gruppen im Zentrum des Interesses. Etwa in der Ungleichheitsforschung, wenn die Forschungsfrage eine Betrachtung der Unterschiede innerhalb der Gruppe des obersten Prozentes der Einkommensverteilung bedingt. Besonders die Untersuchung heterogener Effekte – zum Beispiel der Arbeitslosenversicherung oder von Arbeitsmarktpolitiken entlang der Einkommensverteilung – hat die Arbeitsmarktökonomie in den letzten Jahren vorangebracht.

Entscheidend ist, dass ForscherInnen Daten zur Verfügung haben, die eine adäquate Beantwortung der Forschungsfrage erlauben. Nicht jede wissenschaftliche Untersuchung in Sozialwissenschaften oder Medizin benötigt Einzelfalldaten, manche Fragen sind ohne diese detaillierte Grundlage aber nicht behandelbar.

Laut einem viel diskutierten Papier von Angrist/Pischke (2010) ist die empirische (Mikro-)Ökonomie – beginnend Mitte der 1990er-Jahre – von einer „Revolution der Glaubwürdigkeit" geprägt. Angetrieben wurde diese auch von Entwicklungen im Bereich des *Designs* von Forschungsarbeiten. Schlussendlich ist es das *Design* der Forschung, das für immer mehr ÖkonomInnen darüber entscheidet, ob eine Studie als glaubwürdig gilt oder nicht. Detailliertere und weniger fehlergeplagte Daten – insbesondere in Form von Individualdaten aus Registern – können dazu einen Beitrag leisten, sind aber kein Garant. Jedenfalls werden Datenanalysen in weiten Bereichen der Ökonomie heute ernster genommen und Annahmen transparenter offengelegt, als dies noch im letzten Jahrhundert der Fall war. Die bessere Nachvollziehbarkeit und gestärkte Glaubwürdigkeit hat insbesondere in Arbeitsmarktökonomie, Finanzwissenschaft und Entwicklungsökonomie die politische Relevanz empirischer Forschungsergebnisse gestärkt, auch wenn die externe Validität häufig auf Kosten der internen Validität geht.

Auf Registerdaten basierende Forschungsarbeiten erfordern auch eine ernsthafte Auseinandersetzung mit einer lang andauernden Fehlentwicklung in der ökomischen Forschung. McCloskey und Ziliak (2007) nannten diese treffend den „Kult der statistischen Signifikanz". Bei großen Fallzahlen ist es wenig verwunderlich, dass fast jeder Effekt

statistisch signifikant ist. Anstatt „der Suche nach niedrigen p-Werten“ müssten sich ÖkonomInnen nun eigentlich mit Fragen der sozialen oder ökonomischen Bedeutsamkeit von Effekten auseinandersetzen. Andere Disziplinen, insbesondere die quantitativ orientierten Felder in der Politikwissenschaft, sind hier schon deutlich weiter und lehnen die Publikation von und Orientierung an „p-Werten“ zunehmend ab. Schlussendlich müssen Ergebnisse danach beurteilt werden, ob sie ökonomisch, politisch oder sozial relevante Größen erreichen und nicht nur, ob sie statistisch nachweisbar sind.

In Summe ist die Forderung nach einem transparenten und möglichst kostenfreien Zugang zu Statistik- und Registerdaten für uns als ForscherInnen nachvollziehbar und unter Berücksichtigung der hier aufgeworfenen Risiken und Beschränkungen auch zu unterstützen. Wir tappen in vielen Bereich unseres Lebens im Dunkeln, ein verbesserter Datenzugang könnte hier durchaus ein neues Licht auf viele soziale und ökonomische Probleme werfen und eine wissenschaftliche Basis für deren Lösung bieten.

Erste Ansätze: Das Angebot der Statistik Austria

Die lauten Forderungen nach einem möglichst breiten Zugang zu Mikrodaten sollen nicht darüber hinwegtäuschen, dass auch Statistik Austria als Amt der Republik in den letzten Jahren bereits deutliche Fortschritte im Bereich des Datenzugangs für ForscherInnen unternommen hat. Damit ÖkonomInnen ihren Forschungsinteressen nachgehen können, bietet Statistik Austria auf Einzelfallebene derzeit zwei Mikrodaten-Formate an:

- standardisierte Datensätze und
- aufgabenspezifische Datensätze.

Im ersten Fall reicht das Angebot von Mikrozensusdaten über Daten der Konsumerhebung, Erwachsenenbildung, Gesundheitsbefragung, Registerzählung, diverse Steuerdaten bis zu den Reisegewohnheiten der Bevölkerung. Im zweiten Fall bedarf es einer Spezifizierung des Datenwunsches seitens der ForscherInnen, die Erfüllung der Datenanforderung durch Statistik Austria ist hierbei auch mit (hohen) Kosten verbunden.

Statistik Austria ist die größte offizielle Datenproduzentin und darum bemüht, wirtschafts- und gesellschaftspolitische Statistik-Informationen in hoher Qualität zur Verfügung zu stellen. Zu ihren Aufgaben gehört dabei auch die Dokumentation der Datengewinnung, die unabhängige und mehrstufige Überprüfung der Qualität von Statistiken sowie die weitgehende europäische und internationale Kohärenz der Daten.

Diese Aufgaben wurden allerdings immer wieder durch budgetäre Beschränkungen und politische Interventionen zur großen Herausforderung. Da eine breitere Öffnung für die Wissenschaft auch mit einem höheren Aufwand einhergeht, müssten zunächst die benötigten Rahmenbedingungen und Ressourcen für Statistik Austria und ihre MitarbeiterInnen gesichert werden.

Während die Verfügbarkeit von Daten zu natürlichen Personen bereits relativ weit fortgeschritten ist, lässt die Transparenz und der Zugang zu Daten von juristischen Personen noch viele Verbesserungen zu. Dies ist umso erstaunlicher, als Daten juristischer Personen nicht unter den Anwendungsbereich des Datenschutzes fallen und daher diesbezüglich deutlich weniger schwierig freizugeben wären. Nicht zuletzt aufgrund dieser deutlich geringeren Zugänglichkeit konzentrieren sich auch viele Forderungen aus den Wirtschaftswissenschaften auf die Bereitstellung von mehr Unternehmensdaten.

Die meisten aktuell zugänglichen Mikrodaten haben Stichprobencharakter. Die Forderung der Wissenschaft nach Vollerhebungsdaten wie etwa Unternehmensregister, Melderegister oder Sozialversicherungsdaten, birgt zusätzlichen Aufwand bezüglich Infrastruktur, MitarbeiterInnen und, nicht zu vergessen, heikle Änderungen im Statistikgesetz.

Registerdaten bergen ein großes Potenzial für die Forschung. Das zeigt sich nicht zuletzt darin, dass viele entscheidende Erkenntnisgewinne in den Sozialwissenschaften des letzten Jahrzehnts erst durch die Nutzung von Einzelfalldaten möglich wurden. Besonders die Verknüpfung von Befragungsdaten, wie der Einkommenserhebung EU-SILC, mit Register- und Administrativdaten bietet solche Möglichkeiten. So kann der Vorteil von Befragungen – sie enthalten oft Informationen über theoretisch gut definierte Konzepte – mit den Vorteilen der Register verknüpft werden. Umgekehrt stehen am Anfang jeder repräsentativen Stichprobenerhebung Informationen aus Vollerhebungen und Registern, zum Beispiel die Berechnung der Beobachtungsgewichte, die eine Stichprobe erst repräsentativ machen.

Der Schutz von privaten Daten ist schon jetzt eine wichtige Voraussetzung für die Arbeit von Statistik Austria. Daten werden nur anonymisiert bzw. pseudoanonymisiert weitergegeben, womit die individuellen ForscherInnen diesbezüglich von einer umfassenden datenschutzrechtlichen Verantwortung befreit sind.

Mit der Öffnung der Register für Forschungszwecke gilt es, international etablierte Methoden, die den Schutz personenbezogener Daten gewährleisten und die missbräuchliche Verwendung der gewonnenen Erkenntnisse und Daten zu verhindern, umzusetzen und gegebenenfalls zu verbessern.

Ein Vorschlag für eine Synthese

In diesem Beitrag versuchen wir, unterschiedliche Interessen hinter den Positionen für und gegen einen breiteren Zugang zu amtlichen Individualdaten zu beleuchten. Es besteht ein Konflikt zwischen dem Recht auf informationelle Selbstbestimmung – Menschen sollen selbst entscheiden, wer unter welchen Bedingungen Daten in Bezug auf ihre Person verwenden darf – und dem Recht auf umfassende Kenntnis des eigenen Lebensumfelds, der Welt, in der wir leben. Hinter der Frage, welche Daten von wem und warum erhoben werden, stehen Interessen und Machtverhältnisse.

Auch wenn nie alle Interessen gleichzeitig bedient werden können, scheint es aussichtsreich, Rahmenbedingungen für die jeweiligen Wünsche, Rechte und Ansprüche zu schaffen. Im diesem abschließenden Abschnitt bieten wir einen Vorschlag für eine Synthese der unterschiedlichen Interessen, und legen Überlegungen für den Aufbau eines geplanten „Austria Micro Data Center" dar.

Nur selten verschaffen sich ÖkonomInnen noch vor der Forschungstätigkeit ein Bewusstsein über den Ursprung und die damit verbundenen Eigenheiten der verwendeten Daten. Gerade bei sensiblen, administrativen Individualdaten scheint dies aber unumgänglich. Denn die Sammlung von Daten durch den Staat findet in einem politischen und historischen Kontext statt. Wessen Daten erhoben und geschützt werden ist somit auch eine Frage der gesellschaftlichen und ökonomischen Machtverhältnisse, die beim Verwenden der Daten berücksichtigt werden müssen. Auch bei der Nutzung muss der Vorteil für die Allgemeinheit klar erkennbar sein und Nachteile für die Betroffenen minimiert aber auch transparent kommuniziert werden.

Ein besonders brennendes Beispiel für potenzielle Gefahren sind Gesundheitsdaten. Die Krankenkassen verfügen in Österreich über eine fast lückenlose Kranken-, Berufs-, Einkommens- und Wohnortsgeschichte aller Versicherten, von den Zähnen bis zu den Zehen wird fast jede Erkrankung oder besser Behandlung abgerechnet und dokumentiert. Wie interessant solche Daten aus kommerzieller Sicht sind, zeigte sich erst jüngst wieder bei den umstrittenen Datensammelaktionen durch Technologiekonzerne wie Google (Singer/Wakabayashi 2019), bei denen solche Gesundheitsdaten als wichtige zukünftige Geldquelle gesehen werden. Zugleich ermöglichen genau diese Daten aber auch viele für die Öffentlichkeit nützliche Erkenntnisse. Angesichts der monetären Versuchungen steht außer Zweifel, dass ausreichende Maßnahmen getroffen werden müssen, um Missbrauch, Diebstahl oder illegitime Nutzung dieser Daten zu verhindern.

Ähnlich wie bei Statistik Austria gilt es aber auch hier zuerst Struktu-

ren zu schaffen, in denen der Ausbau der Datenkompetenzen auf sicheren datenschutzrechtlichen Füßen steht. Diese Struktur muss in der Lage sein, die Sinnhaftigkeit von Datenanfragen aus der Forschung ebenso zu beurteilen wie ihre Vereinbarkeit mit dem Datenschutz. Sie braucht andererseits auch die Kompetenz, ForscherInnen bei der adäquaten Interpretation der Daten zu unterstützen. Gerade bei Daten, die nicht gezielt für wissenschaftliche Fragen erhoben und entsprechen operationalisiert wurden, ist die korrekte Verwendung der Daten wichtig.

Auch im Rahmen von wissenschaftlichen Studien an Universitäten müssen adäquate Vorkehrungen getroffen werden. Der Drittmittelanteil aus privaten Quellen an der universitären Forschung wirft die Frage auf, ob die Beschränkung des Datenzugangs auf Universitäten keine Beschränkung für die Datennutzung durch zahlungskräftige InteressentInnen bedeuten würde (laut Unidata kamen 2018 fast 200 Mio. €, die für universitäre Forschung zur Verfügung standen, von Privatpersonen, Unternehmen, Privatstiftungen und Vereinen[2]). Oft ist keine ausreichende Transparenz dieser Drittmittelforschungen gegeben.

Hier sind Vorsichtsmaßnahmen notwendig, weil es kaum ein größeres Risiko für die sinnvolle Verwendung von Daten gibt, als das Auftreten von Missbrauch. Schon ein verhältnismäßig kleiner Skandal könnte die gesamte Registerdatenforschung auf Jahre verhindern.

Hinsichtlich der Etablierung eines „Austrian Micro Data Center" kann das „Austrian Center for Labor Economics and the Analysis of the Welfare State" in Linz ein Vorbild sein, wo die Aufbereitung von Sozialversicherungsdaten geleistet wurde. Dieses Zentrum ermöglichte vielen ForscherInnen beachtliche wissenschaftliche Erfolge. Aufbauend auf diese Erfahrungen könnte ein geplantes „Austrian Micro Data Center" tatsächlich ein großer Fortschritt für die wissenschaftliche Gemeinschaft bedeuten. Gerade beim Aufbau einer solchen Institution müssten die in diesem Beitrag diskutierten Risiken beachtet und die notwendigen Rahmenbedingungen garantiert werden.

So könnte etwa die demokratische Kontrolle, die rechtliche Kontrolle durch Datenschutzräte und Gerichte wie auch die wissenschaftliche Kontrolle durch passend besetzte Gremien mit entsprechenden Ressourcen sichergestellt werden.

Aufgabe eines derartigen Zentrums müsste es auch sein, technische Möglichkeiten zur Datenverarbeitung und zur Wahrung des Datenschutzes weiterzuentwickeln. Hinsichtlich des Datenschutzes gilt das Prinzip der „differenziellen Privatheit". Es bietet eine technische Antwort auf Datenschutzbedenken und kommt etwa in den USA im Rahmen des Zensus 2020 zur Anwendung. Dieses Prinzip erlaubt einen an die Fragestellung angepassten Grad der Datenanonymisierung. Das Eichhörnchen-Prinzip vieler ForscherInnen, im Zweifel besser einen

größeren als einen kleineren Variablenbestand anzufordern, schafft unnötige Mehrarbeit bei der Bereitstellung und Anonymisierung von Datenbeständen. Auch hier können technische Datenschutzlösungen Abhilfe verschaffen, so dass nie mehr Daten freigegeben werden, als für eine spezifische Berechnung notwendig sind. Dänemark, Schweden oder auch die Niederlande gelten hinsichtlich der Bereitstellung von Register- und Administrativdaten als besonders liberal. Österreich ist weniger offen mit den Daten seiner Wohn- und Arbeitsbevölkerung. Aus den Erfahrungen dieser Länder in Bezug auf die Entwicklung einer Forschungsdateninfrastruktur kann gelernt werden.

Es wird vor allem an den ForscherInnen liegen, die Öffentlichkeit zu überzeugen, dass ihre Fragestellungen für die breite Bevölkerung interessant sind oder sogar zur Verbesserung der Lage der Bevölkerung beitragen und nicht ein Projekt zur Vergrößerung des Herrschaftswissens oder ausschließlich zur Beschleunigung der eigenen wissenschaftlichen Karriere dienen.

Anmerkungen

[1] Zur Debatte der Datenfreigabe siehe https://www.derstandard.at/story/2000116442277/coronavirus-krankenkassen-geben-daten-fuer-forschung-frei sowie https://www.derstandard.at/story/2000116383998/kampf-gegen-coronavirus-patientenanwalt-will-freigabe-von-elga-daten, und zu den Möglichkeiten des *Opt Out* siehe https://www.chipkarte.at/cdscontent/?contentid=10007.678580. Dabei soll niemand die Möglichkeit der e-Medikation verlieren, nur weil man nicht bereit ist, persönliche Daten pauschal für Forschungsprojekte zur Verfügung zu stellen.

[2] https://unidata.gv.at/Pages/auswertungen.aspx Tab.7.6.

Literatur

Angrist, Joshua D. und Jörn-Steffen Pischke (2010): „The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." Journal of Economic Perspectives 24 (2) 3-30. https://doi.org/10.1257/jep.24.2.3.

Britton, Jack; Waltmann, Ben (2020): The government's A Level failure leaves universities in the lurch. Institute for Fiscal Studies (IFS). Online verfügbar unter https://www.ifs.org.uk/publications/14978.

Card, David, Raj Chetty, and Andrea Weber (2007): „The Spike at Benefit Exhaustion: Leaving the Unemployment System or Starting a New Job?" The American Economic Review 97 (2) 113-18. http://www.jstor.org/stable/30034431.

Carrell, Severin (2020): Over 120,000 Scottish exam grades to be reinstated after row. Online verfügbar unter https://www.theguardian.com/politics/2020/aug/11/over-100000-scottish-exam-grades-to-be-reinstated-after-row.

Lang, Kevin und Ariella Kahn-Lang Spitzer (2020): „Race Discrimination: An Economic Perspective." Journal of Economic Perspectives 34 (2) 68-89. https://doi.org/10.1257/jep.34.2.68.

Robinson, W. S. (1950): „Ecological Correlations and the Behavior of Individuals." American Sociological Review 15 (3): 351. https://doi.org/10.2307/2087176.

Singer, Natasha; Wakabayashi, Daisuke (2019): Google to Store and Analyze Millions of Health Records – The New York Times, 11.11.2019, zuletzt geprüft am 8.10.2020.661Z.

Spence, Michael. (1973): „Job Market Signaling." The Quarterly Journal of Economics 87 (3) (1973) 355-74. http://www.jstor.org/stable/1882010.

Verordnung (EU) 2016/679 (14.10.2020.000Z): Verordnung (EU) 2016/679 des Europäischen Parlament und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung). DSGVO. Fundstelle: Artikel 5. In: Amtsblatt der Europäischen Union (1977-0642) 59 (L 119), S. 1-89. Online verfügbar unter http://data.europa.eu/eli/reg/2016/679/oj, zuletzt geprüft am 14.10.2020.967Z.

Ziegler, Elke (2016): Wir bezahlen, Sie forschen? – science.ORF.at. Hg. v. science.ORF.at. ORF. Online verfügbar unter https://science.orf.at/v2/stories/2808343/, zuletzt aktualisiert am 8.10.2020.000Z, zuletzt geprüft am 8.10.2020.891Z.

Ziegler, Elke (2019): Transparenzstudie zu Drittmitteln gescheitert – science.ORF.at. Hg. v. science.ORF.at. ORF. Online verfügbar unter https://science.orf.at/v2/stories/2958831/, zuletzt aktualisiert am 8.10.2020.000Z, zuletzt geprüft am 8.10.2020.795Z.

Ziliak, Stephen T.; McCloskey, Deirdre N. (2007): The cult of statistical significance. How the standard error costs us jobs, justice, and lives (Economics, cognition, and society series).

Zuboff, Shoshana (2018): Das Zeitalter des Überwachungskapitalismus. Frankfurt, New York: Campus Verlag. Online verfügbar unter https://www.content-select.com/index.php?id=bib%5Fview&ean=9783593439433.