Datenqualität und Selektivitäten digitaler Daten

Alte und neue digitale und analoge Datensorten im Vergleich

Nina Baur und Peter Graeff

Beitrag zur Veranstaltung »Digitale Daten und neue Methoden: Chancen und Herausforderungen für die Soziologie« der Sektion »Wissenschafts- und Technikforschung«

Der sozialwissenschaftliche Diskurs um digitale Daten

Bevor wir verschiedene digitale und analoge Datensorten vergleichen, möchten wir eine Präzisierung des Diskurses um digitale Daten vornehmen: Es lassen sich derzeit im Rahmen der methodologischen Debatte um digitale Daten zwei große Forschungsfelder identifizieren, die sich mit diesem Thema befassen: die Methoden der empirischen Sozialforschung und die Computational Social Sciences. Während sich die Computational Social Sciences hauptsächlich mit dem Thema der Datenaufbereitung und Datenanalyse befassen, haben die Methoden der empirischen Sozialforschung eine sehr lange Tradition, sich mit der Frage der Datenqualität, Stichprobenqualität und Selektivitäten zu befassen. Anders als bisweilen nahelegt wird (z.B. Diekmann 2016), ist es nicht so, dass die Methodenforschung sich erst seit einigen Jahren mit dem Thema befasst, sondern sogar schon sehr lange und sehr ausführlich. In der Tat reicht die Debatte um digitale Daten zurück bis in die 1960er Jahre (Thaller 2017 [1990]). Wir fokussieren im Folgenden auf die Fragen der Daten- und Stichprobenqualität aus Sicht der Methoden der empirischen Sozialforschung.

Was sind „digitale Daten“?

Die erste Frage, die sich im Hinblick auf Datenqualität und Stichprobenqualität stellt, ist, was eigentlich „digitale Daten“ und – in vielen wissenschaftlichen Debatten oft fast synonym verwendet – sogenannte „Big Data“ sind: Digitale Daten werden oft gleichgesetzt mit Big Data und mit dem Web 2.0 assoziiert. Wenn Sozialforschende über „digitale Daten“ sprechen, meinen sie häufig Social-Media-Daten – wie Twitter-Daten (Einspänner et al. 2014; Mayerl, Faas 2019), Facebook-Daten (Schneider, Harknett 2019; Schrape, Siri 2019) oder YouTube-Videos (Traue, Schünzel 2019) –, Geodaten (Lakes 2019; Kandt 2019) und mobile Daten (Bähr et al. 2020), an Daten, die aus neuen Technologien generiert worden sind – seien es Aufnahmen von Videokameras (Knoblauch, Tuma 2020; Tuma 2017), von Smart Devices (Koch 2019), Smart Homes, oder Smart Citys, das heißt insgesamt: Daten, die ein Nebenprodukt von digitaler

Kommunikation sind. Bei genauerem Hinsehen ist dieser Begriff aber in doppelter Hinsicht sehr unpräzise.

Digitale Daten und Big Data (Massendaten)

In vielen Debatten werden „digitale Daten" fälschlicherweise mit „Big Data" („Massendaten") gleichgesetzt: Abgesehen davon, dass sehr viele digitalen Daten keine Massendaten sind, sind Big Data methodologisch gesehen eine Variante der sogenannten prozessproduzierten Daten (Baur 2009), die als Nebenprodukt sozialer Prozesse entstehen und eine der ältesten Datenformen überhaupt sind – man denke zum Beispiel an staatliche administrative Daten, die etwa im Rahmen von Volkszählungen, von Einwohnermeldeämtern oder von den Sozialversicherungen produziert werden (Wallgren, Wallgren 2014). Aber auch Nachrichten (Iglesias et al. 2016) in Zeitungen oder im Fernsehen sind eine klassische Variante von Big Data, ebenso wie Daten, die von wirtschaftlichen und zivilgesellschaftlichen Akteuren produziert werden – seien es Personalabteilungen, die Personaldaten ihrer Mitarbeiter verwalten, oder Firmen, die Kundendaten über Kontaktformulare administrieren. Alle diese Datensorten existieren schon sehr, sehr lange. Insbesondere in Mitteleuropa wurden sie spätestens zu Beginn des 19. Jahrhunderts systematisch als Herrschaftstechnik eingesetzt – die methodologische Kritik an ihrer vermeintlichen „Objektivität" ist zu Beginn des 20. Jahrhunderts eine der Wiegen der deutschen Soziologie (Baur et al. 2020, S. 213–216) – und zwar in Form der Kritik und der Auseinandersetzung mit den Verzerrungen dieser Daten, die sowohl ein wesentlicher Impuls für die Entwicklung der Deutschen Gesellschaft für Soziologie (DGS), als auch Anlass der damaligen fachlichen Debatten war, die in Deutschland letztendlich zur Institutionalisierung von Theorie und Methoden sowie international zur Entwicklung von forschungsinduzierten Methoden in der Soziologie führten (Lepsius 1979). Forschungsinduzierte Daten – wie etwa die Befragung oder Beobachtung – erlauben Forschenden (im Gegensatz zu prozessproduzierten Daten), den Prozess der Stichprobenziehung und Datengenerierung und damit verbundene etwaige Fehler zu kontrollieren.

Digitale Daten und analoge Daten

Der Begriff der „digitalen Daten" ist auch deshalb unpräzise, weil er suggeriert, dass es sich um ein neuartiges Phänomen des Web 2.0 handelt, obwohl in den Sozialwissenschaften bereits seit den 1960ern viele Daten entweder rein digital erhoben oder im Zuge der Datenaufbereitung digitalisiert wurden. Es existieren auch zahlreiche Hybridformen, und zwar unabhängig von der Datensorte. Diese Hybridformen können sich zunächst auf der Ebene der Datenerhebung manifestieren. So können etwa standardisierte Befragungen als Mixed-Mode-Befragungen aus (offline) schriftlich-postalischer Befragung (Reuband 2019) und Online-Befragung (Wagner-Schelewsky, Hering 2019) durchgeführt werden. Bei CATI („Computer-Aided Telephone Interviewing") und CAPI („Computer-Aided Personal Interviewing") findet die Datenerhebung zwar telefonisch bzw. persönlich-mündlich statt, die Daten werden aber bereits während der Datenerhebung digital erfasst (Fuchs 1994). Dies gilt aber auch für die Bereitstellung der Daten sowohl für eine breite Öffentlichkeit, als auch für wissenschaftliche Analysen. So sind etwa viele staatlich administrativen Daten mittlerweile auch online verfügbar, etwa über Destatis oder den RatSWD. Die meisten Zeitungsnachrichten sind heute sowohl in einer Online-, als auch in einer Printversion verfügbar, ebenso wie die Daten vieler zivilgesellschaftlichen Akteure.

Begriffliche Präzisierung des Begriffs der „digitalen Daten“

Das Problem an dieser mangelnden begrifflichen Präzision ist, dass sie den methodologischen Diskurs um Datenqualität erschwert, weshalb wir eine andere begriffliche Fassung vorschlagen, die präzisere methodologische Debatten erlaubt. Konkret schlagen wir vor, zunächst zwischen *forschungsinduzierten Daten* – in ihren klassischen Formen der Befragung bzw. des Interviews und die Beobachtung bzw. der Ethnografie – und *prozessproduzierten Daten* zu unterscheiden. Weiterhin schlagen wir vor, bei prozessproduzierte Daten „*klassische prozessproduzierten Daten*“ – wie die Daten der amtlichen Statistik und anderer administrative Daten staatlicher Akteure bzw. zivilgesellschaftlichen Akteuren oder auch Zeitungsdaten – von „*neuartigen prozessproduzierte Daten*“ zu unterscheiden, die als Nebenprodukte digitaler Kommunikation entstehen (Graeff, Baur 2020). Übernimmt man diese Terminologie, dann lassen sich eine Reihe von Querdimensionen identifizieren:

1. Fast alle diese Daten können digital oder analog oder hybrid sein, mit einer Ausnahme: Die neuartigen prozessproduzierten Daten sind fast ausnahmslos rein digital.
2. Die Fallzahlen können klein sein oder eben groß – „big“. Mit anderen Worten: Alle Daten (nicht nur die digitalen) können auch theoretisch „big“ werden.
3. Daten können offen („qualitativ“) oder standardisiert („quantitativ“) – bzw. schwach oder stark strukturiert – sein, oder es kann sich um gemischte Daten („mixed“) handeln. Prozessproduzierten Massendaten („big data“) – unabhängig davon, ob es sich um klassische oder neuartige prozessproduzierte Daten handelt – sind fast immer gemischte Daten. So enthält fast jedes Verwaltungsdokument neben standardisiert erfassten Informationen wie Alter und Geschlecht auch offene Felder (ähnlich wie bei offenen Fragen in standardisierten Befragungen). Webseiten (Arora et al. 2016; Schünzel, Traue 2019) enthalten neben Log-Dateien (Schmitz, Yanenko 2019) meistens auch Bilder und Text (Williams et al. 2019) usw. Daher ist die Forschung mit prozessproduzierten Massendaten („big data“) fast immer automatisch Mixed-Methods-Forschung (Baur et al. 2020, S. 221) – und zwar nicht, weil sich dies die Forschenden (wie bei forschungsinduzierten Daten) bewusst aussuchen, sondern weil die Daten eben diese Form annehmen.

Vorteile der begrifflichen Präzisierung

Was sind nun die Vorteile, wenn man die von uns vorgeschlagene Terminologie statt der bisher üblichen Unterscheidung in digitale und nicht digitale Daten verwendet? Der erste Vorteil ist, dass eine Entemotionalisierung des Diskurses möglich ist. In vielen Methodendebatten begegnen wir entweder vehementen Verfechtern der klassischen Forschung mit forschungsinduzierten Daten oder der Forschung mit neuartigen prozessproduzierten Daten. In beiden Fällen gewinnt die Argumentation häufig fast ideologische Züge mit der Verteufelung der jeweils anderen Seite. Verwendet man die von uns vorgeschlagene Systematisierung, dann sind digitale Daten – insbesondere neuartige prozessproduzierte Daten – eine Datensorte unter vielen Daten. Dies macht die Datensorten untereinander im Hinblick auf das zu untersuchende soziale Phänomen einordbar – welche Datensorte dann für eine konkrete Forschungsthematik zu bevorzugen ist, wird zu einer empirischen Frage. Die bisherige Methodenforschung deutet darauf hin, dass diese Frage nicht grundsätzlich beantwortet werden kann, sondern von der konkreten Fragestellung abhängt. Ein zweiter und noch größerer Vorteil ist – wie wir im Folgenden zeigen werden –, dass die Debatte um digitale Daten – und insbesondere die um als „neuartige prozessproduzierte Daten“ – durch die Einordnung in dieses Schema in den bereits existierenden Methoden Diskurs integriert werden kann.

Abschätzung der Datenqualität verschiedener Datensorten

Die Methodenforschung verwendet oft Modelle des Forschungsprozesses, um Stärken, Schwächen sowie mögliche Fehlerquellen im Forschungsprozess aufzudecken. Diese reichen von der Konzeptualisierung über die Datensammlung – also Stichprobenziehung oder gesammelte Datenerhebung – über die Archivierung und Datenzugang, Datenaufbereitung, Datenanalyse bis zur Generalisierung der Ergebnisse.

Forschungsinduzierte Daten: Fehlerkunde

So existieren für die Survey-Forschung Ablaufmodelle zum „Survey Life Cycle" (Biemer 2010; Hill et al. 2019), die aufzeigen, was in diesem Prozess wann passiert und welche Fehler dann auftreten können. Auf dieser Basis hat die Survey-Forschung eine „*Fehlerkunde*" entwickelt, die bestimmte Fehler benennt und erlaubt, vergleichend für konkrete Forschungsprozesse abzuschätzen, ob und in welchem Maße diese Fehler aufgetreten sind, ob und wie sie vermeidbar gewesen wären – der „Total Survey Error" (TSE) (Groves, Lyberg 2010) versucht, den Gesamtfehler zu bestimmen. Dieses Ablaufmodell wird sowohl prospektiv bei der Forschungsplanung verwendet, um den Forschungsprozess zu verbessern und Fehler zu vermeiden, als auch im Nachhinein, um Fehler und Güte der Daten und Analyse abschätzen zu können.

Prozessproduzierte Daten: Datenkunde

Ein äquivalentes Modell gibt es seit langem für standardisierte klassische prozessproduzierten Daten: Wolfgang Bick und Peter J. Müller (1980, 1984) erläuterten in den 1980ern, dass verschiedene soziale Prozesse bei der Entstehung und Aufbewahrung prozessproduzierter Daten diese verzerren können. Bick und Müller (1980, 1984) sprechen bewusst nicht von einer „Fehlerkunde" (wie in der Survey-Forschung) sondern von einer „*Datenkunde*". Sie argumentieren, dass Verzerrungen prozessproduzierter Daten aus Perspektive sozialwissenschaftlicher Datennutzenden keine Verzerrung der Daten aus der Perspektive der Institutionen ist, die diese Daten produziert haben. Vielmehr liegen hinter diesen „Verzerrungen" meist bewusste Entscheidungen in den administrativen Entstehungs- und Aufbewahrungsprozessen. Diese Prozesse der Datenproduktion führen allerdings dazu, dass Big Data eben nicht – wie in manchen jüngeren Debatten oft suggeriert wird – ein Abbild sozialer Wirklichkeit sind. Vielmehr führen mehrere sich überlagernde soziale Prozesse dazu, dass die in den Daten enthaltenen Informationen von der sozialen Wirklichkeit abweichen:

- Zunächst verfügen die datenproduzierenden Institutionen über eine sogenannte „*Verwaltungstheorie*" (Bick, Müller 1984), die bestimmt, warum sie welche Daten nach welchen Regeln überhaupt erheben – und welche eben nicht. So erhebt vielleicht eine Firma Kundendaten, um etwa Bestellungen abzuwickeln – und schließt damit automatisch alle Nichtkunden aus. Dies ist dann aber kein Fehler, sondern eine spezifische Eigenheit der Daten – zur Kundenverwaltung benötigt die Firma keine Informationen über ihre Nichtkunden. Warum sollte sie diese also erheben?
- Weiterhin lässt sich auf der Ebene des die Daten erfassenden und bearbeitenden Personals eine sogenannte „*Sachbearbeiter-Logik*" (Bick, Müller 1984) beobachten, die in organisationale Regelungen und Alltagszwänge eingebettet ist, so dass sich Konventionen zur Datenbearbeitung entwickeln. Hierzu gehören etwa Abkürzungsregeln, Sonderregeln und Ausführungsbestimmungen. Hierzu gehören aber auch schnöde Eingabefehler, die umso wahrscheinlicher sind, je weniger dringlich der Vorgang ist. Das Personal kann aber bei der Dateneingabe auch Abkürzungen vornehmen, wenn es bestimmte Regeln nicht verstanden hat, usw.

- Hinzu kommt die sogenannte „*Klientenlogik*“ bzw. „*Kundenlogik*“ (Bick, Müller 1984): Auch die Personen, deren Daten erfasst werden, antizipieren den Datenerhebungsprozess und passen das, was sie preisgeben, entsprechend an. So wäre es etwa denkbar, dass ein und dieselbe Person gegenüber dem Finanzamt versucht, das Einkommen herunterzurechnen, während sie es zum Beispiel in anderen Kontexten wünschenswert ist, dass das Einkommen relativ hoch erscheint, damit man zum Beispiel bei einem Visaantrag als einkommensstarker Reisender oder bei der Bank als besonders kreditwürdig eingeordnet wird.
- Schließlich ist auch eine „*Plattform-Logik*“ zu beobachten, d.h. die Technologie der Datenerhebung bestimmt, welche Daten wie erfasst werden können – und welche nicht. So lassen viele klassischen prozessproduzierten Daten nicht zu, dass visuelle Informationen mit erfasst werden.

Bick und Müller (1984) argumentieren, dass für prozessproduzierte Daten – äquivalent zu forschungsinduzierten Daten – auch ein Ablaufmodell erstellt werden kann, das ermöglicht abzuschätzen, in welchen Schritten des Forschungsprozesses welche soziale Prozesse ablaufen und wie die Daten infolge welcher Selektivitäten von sozialer Wirklichkeit abweichen. Wie bereits erwähnt, ist das Argument nicht, dass hier ein Fehler entsteht, sondern dass ganz bewusst über den Verwaltungsprozess blinde Flecken erzeugt werden, die so lange kein Problem sind, solange sie später in der Forschung nicht unreflektiert übernommen werden. Ein Beispiel für eine unreflektierte Übernahme von blinden Flecken sind arbeits- und organisationssoziologische Untersuchungen auf Basis der iab-Daten. Da diese auf dem Datenbestand der Bundesagentur für Arbeit beruhen, enthalten sie vor allem Daten zu abhängigen sozialversicherungspflichtigen Beschäftigten und viele Lücken. So fehlen etwa viele Selbständige und Freiberufler im Datenbestand komplett. Dies ist so lange kein Problem, solange die abhängigen sozialversicherungspflichtigen Beschäftigten auch die intendierte Grundgesamtheit darstellen. Dies wird allerdings dann zum Problem, sobald Forschende annehmen, dass sie mit den iab-Daten eine Vollerhebung *aller* Beschäftigten analysieren. Das Bick-Müller-Modell erlaubt es, genau solche blinden Flecken aufzudecken.

Datenqualität von neuartigen prozessproduzierten Daten (digitale Daten aus dem Web 2.0)

Wie sich neuartige prozessproduzierte Daten – wie etwa digitale Daten aus dem Web 2.0 – von anderen Daten hinsichtlich der Datenqualität unterscheiden, ist eine große offene Forschungsfrage. Allerdings deutet unsere aktuell laufende gemeinsame Arbeit (Baur et al. 2020; Graeff, Baur 2020) darauf hin, dass das Bick-Müller-Modell zur Erfassung der Datenqualität von prozessproduzierten Daten ohne Probleme auch auf neuartige prozessproduzierte Daten übertragen werden kann. Auf Basis unserer eigenen Arbeit und des aktuellen Stands der Forschung kann man zum Vergleich der Datenqualität von neuartigen prozessproduzierten Daten und anderen Daten mehrere Aussagen treffen.

Digitale Daten als Plattformdaten

Zunächst muss man unterscheiden zwischen der Nutzung des Internets „an sich“ und der Nutzung spezifischer Plattformen. Die Datenerhebung im Web 2.0 erfolgt nicht über das gesamte Netz, sondern es werden immer Daten aus spezifischen Plattformen (wie Twitter, Facebook, Spiegel Online etc.)

erhoben – d.h. methodologisch wird mit konkreten Plattformdaten gearbeitet. Da die meisten Nutzer nur wenige spezifische Plattformen nutzen und nicht das gesamte Internet, ist es wichtig zu wissen, wer welche Plattform wie und wie häufig nutzt.

Digitale Spaltungen und Datenqualität

Weiterhin muss beachtet werden, dass es unterschiedliche Gründe für Nichtnutzung gibt, die dann auch möglicherweise beeinflussen, wie die Daten verzerrt sind (Baur et al. 2020, S. 225–233): Nutzungshindernisse können einerseits die Form des *Nicht-Nutzen-Könnens* annehmen und reichen hier von Zugangsbeschränkungen zum Internet oder zu bestimmten Endgeräten über mangelnde Fremdsprachenkenntnisse und Analphabetentum, staatliche Einschränkungen der Mediennutzung, hohe Kosten bis hin zu mangelndem Wissen und Nutzungskompetenzen. Nutzungshindernisse können aber auch die Form des *Nicht-Nutzen-Wollens* annehmen. In methodologischen Debatten wird oft vergessen, dass – wie die Innovations- und Techniksoziologie vielfach gezeigt hat – neue Technologien (und dazu gehört auch das Web 2.0) nur dann adaptiert werden, wenn sie für Menschen gegenüber althergebrachten Technologien eine Verbesserung erbringen – und sehr viele Menschen sehen im Web 2.0 keinen Mehrwert.

Die Nichtnutzung des Internets führt zu den sogenannten *digitalen Spaltungen* („digital divides“) – und diese sind erheblich. So lag 2019 die globale Nichtnutzungsrate bei 46 Prozent (ITU 2019, S. 2), das heißt, dass fast die Hälfte der Weltbevölkerung noch nicht einmal einen Zugang zum Internet hatte, mit der Folge, dass diese Personen keinerlei digitale Spuren hinterlassen und damit auch nicht in den Daten auftauchen. Methodologisch relevant ist, dass diese Muster der (Nicht-)Nutzung des Internets nicht zufällig verteilt sind, sondern (a) im globalen Maßstab weitgehend (wenn auch nicht ausschließlich) Mustern globaler Ungleichheit folgen und mehr oder weniger das postkoloniale Machtgefüge reproduzieren sowie (b) innerhalb von spezifischen Gesellschaften weitgehend bestehenden bekannten Mustern sozialer Ungleichheit folgen – so sind Alter, Bildung und Geschlecht sehr relevante Kategorien für die Nutzungsraten. Konkret ist die Person, die mit der geringsten Wahrscheinlichkeit im Internet Spuren hinterlässt, eine arme ältere Frau aus dem ländlichen Raum in Afrika (Baur et al. 2020, S. 223-225). Dieser Umstand hat für sozialwissenschaftliche Forschung in unterschiedlichen Bereichen Relevanz.

Digitale Spaltungen sind auf der einen Seite methodologisch relevant, weil sich die neuartigen prozessproduzierten Daten eben nicht ohne Weiteres oder gar uneingeschränkt als Ersatz für klassische Sozialforschung eignen. Fruchtbarer erscheint die Anwendung von Auswertungstechniken der Computational Social Sciences auf klassische Datentypen.

Digitale Spaltungen sind aber auf der anderen Seite auch gesellschaftspolitisch relevant, weil – wenn Sozialforschende diese Daten unreflektiert verwenden – sie für viele Themen systematisch blinde Flecken generieren. Dadurch, dass diese digitalen Ausschlüsse so umfangreich sind, impliziert die Verwendung neuartiger prozessproduzierter Daten den Ausschluss der Weltperspektive großer Teile der Weltbevölkerung. Damit können wir als Soziologinnen und Soziologen Gefahr laufen, soziale Ungleichheit nicht nur zu reproduzieren, sondern – wenn die auf diesen Daten basierenden Ergebnisse dann später für gesellschaftspolitische Entscheidungen verwendet werden – sogar zu verstärken, weil in der Regel die ohnehin schon benachteiligten sozialen Gruppen in der Analyse fehlen.

Veränderung der Machtbalancen vom Staat zur Wirtschaft

Dieses grundsätzliche Dilemma wird dadurch verschärft, dass mit der Veränderung der Dateneigentümerschaft auch eine Verschiebung der Machtbalancen (Elias 1969, S. 123) vom Staat und der Bevöl-

kerung hin zu multinationalen Konzernen erfolgt (Baur et al. 2020, S. 221–222) – auch dies ist ein Thema, dass einer weiteren Diskussion würdig wäre.

Verstärkung der Rekursivität von Daten und Gesellschaft

Auch wenn es nicht neu ist, dass Gesellschaft selbstreflexiv ist und sozialwissenschaftliche Analysen zu selbsterfüllenden Prophezeiungen werden können, so ist es doch sehr wahrscheinlich, dass sich infolge der größeren Geschwindigkeit der Datenproduktion sowie der Zunahme von undurchschaubaren Algorithmen die Rückkopplungsschleifen (Traue 2020) und die Rekursivität zwischen Daten und Gesellschaft erhöhen oder zumindest beschleunigen – wie sich dies auf die Datenqualität auswirkt, ist unklar. Schon seit geraumer Zeit greifen die großen Datenproduzenten von Makrodaten (wie die Weltbank) zur Datengenerierung von sozialen Phänomenen wie Korruption oder Lebenszufriedenheit auf Quellen anderer Datenproduzenten zurück (Arndt, Oman 2006; Rohwer 2009). Damit wird nicht nur die Unabhängigkeit der Datenerhebungen in Zweifel gezogen, eine empirische Analyse mit Hilfe dieser Daten verliert eine ihrer wichtigsten Grundannahmen mit unklaren Konsequenzen für die messtheoretischen Desiderate wie Validität und Reliabilität. Die digitalen Möglichkeiten erhöhen zwar die Verfügbarkeit solcher Daten und meist für jeden, suggerieren aber oft durch die Reputation der datenproduzierenden Institutionen eine valide Erfassung des quantifizierten sozialen Phänomens. Damit werden Anreize gesetzt, die Daten (auch mangels eines intransparenten und schwierig nachzuvollziehenden Konstruktionsvorganges) in nicht reflexiver Weise zu verwenden.

Es ist durchaus wahrscheinlich, dass mangelnde Reflexion dieser Rekursivität zu gesellschaftspolitisch unerwünschten Nebenfolgen führen. Das lässt sich auch abseits sozialwissenschaftlicher Makrodaten am lebensweltlichen Beispiel der Corona-App zeigen: Wenn Nutzende die App installiert haben und sie regelmäßig nutzen, ist es naheliegend, dass – wenn die App anzeigt, dass die Nutzenden ein „geringes Infektionsrisiko“ haben – diese auch davon ausgehen, dass sie kein oder ein geringes Krankheitsrisiko haben und daher gesund sind. Als Konsequenz ist es dann naheliegend, auf Masken, Abstandsregeln und andere Maßnahmen zur Seuchenprävention zu verzichten – oder gar die Existenz bzw. Gefahr durch die Pandemie grundsätzlich infrage zu stellen. Solche und ähnliche Schlüsse können aber aus der Corona-App aus verschiedenen Gründen *nicht* gezogen werden:

Erstens können infolge der digitalen Ausschlüsse große Teile der Bevölkerung die Corona-App gar nutzen, selbst wenn sie wollten. So gelten 2019 in Deutschland 18 Prozent der Bevölkerung als digital Abseitsstehende, die kaum oder sogar keinerlei digitale Spuren hinterlassen (Initiative D21 e. V. 2020). Es hat folglich nur ein Teil der Bevölkerung ein Smartphone oder ist fähig, die App zu installieren – und kann dann auch folglich keine Spuren auf der Corona-App hinterlassen.

Zweitens können Menschen sich bewusst entscheiden, die App nicht zu installieren – etwa aus Datenschutzbedenken, weil sie nicht an die Existenz oder Gefährlichkeit der Pandemie glauben, oder weil sie die App für sinnlos halten (Dehmel et al. 2020).

Doch selbst, wenn sie ein Smartphone haben und die App installiert haben, so bedeutet dies, drittens, nicht notwendig, dass die Corona-App korrekt ihre Bewegungsmuster erfasst. Man denke etwa an eine ältere Person, die zum Telefonieren hauptsächlich das Festnetztelefon und ihr Smartphone ausschließlich wegen der Zusatzfunktionen nutzt: Sudoku spielen, Nachrichten lesen, ein Foto vom Garten im Schnee machen. Solche Personen tragen das Smartphone vermutlich nicht immer bei sich, sondern lassen es oft zu Hause, wenn sie unterwegs sind. In diesem Fall verzeichnet die Corona-App dann auch keinerlei Corona-Kontakte und Risiko, weil sich das Smartphone nicht bewegt, sondern nur sein Nutzer. Aber auch eine Person, die ihr Smartphone immer bei sich trägt, trägt es vermutlich gerade in Hochrisiko-Situationen nicht bei sich – etwa beim Sport, beim Baden am Strand im Urlaub oder

beim Sex, also in Situationen, in denen man anderen Menschen potenziell sehr nahe ist und wegen der körperlichen Bewegung einen erhöhten Virenausstoß über die Atmung hat, sollte man infiziert sein.

Insgesamt erfasst die Corona-App nur einen Bruchteil aller Kontakte zu anderen Personen (Dehmel et al. 2020), und gerade in Hochrisiko-Situationen in Bezug auf die Übertragung wird diese Situation vermutlich eher seltener erfasst. Damit kann die Corona-App aber gar nicht das tatsächliche Übertragungsrisiko anzeigen – dies wird genau dann zum Problem, wenn auf Basis dieser Daten Entscheidungen getroffen werden. Dies – und hiermit wollen wir schließen – ist genau einer der Gründe, warum es unserer Meinung nach sehr wichtig ist, der Datenqualität digitaler Daten künftig mehr Aufmerksamkeit zu schenken.

Literatur

Arndt, Christiane, und Charles Oman. 2006. *Uses and Abuses of Governance Indicators*. Paris: OECD Development Centre Studies.

Arora, Sanjay K., Yin Li, Jan Youtie, und Philip Shapira. 2016. Using the wayback machine to mine websites in the social sciences: a methodological resource. *Journal of the Association for Information Science and Technology*, 67(8):1904–1915.

Bähr, Sebastian, Georg-Christoph Haas, Florian Keusch, Frauke Kreuter, und Mark Trappmann. 2020. Missing data and other measurement quality issues in mobile geolocation sensor data. *Social Science Computer Review* 38(1):10–24.

Baur, Nina, Peter Graeff, Lilli Braunisch, und Malte Schweia. 2020. The Quality of Big Data. Development, Problems, and Possibilities of Use of Process-Generated Data in the Digital Age. *Historical Social Research* 45(3):209–243.

Baur, Nina. 2009. Measurement and Selection Bias in Longitudinal Data. A Framework for Re-Opening the Discussion on Data Quality and Generalizability of Social Bookkeeping Data. *Historical Social Research* 34 (3):9–50. doi: 10.12759/hsr.34.2009.3.9-50.

Bick, Wolfgang, und Peter J. Müller. 1980. The nature of process-produced data – towards a social scientific source criticism. In *Historical Social Research. The Use of Historical and Process-Produced Data. Historisch-Sozialwissenschaftliche Forschungen Band 6*, Hrsg. Jerome M. Clubb und Erwin K. Scheuch, 369–413. Stuttgart: Klett-Cotta.

Bick, Wolfgang, und Peter J. Müller. 1984. Sozialwissenschaftliche Datenkunde für prozeßproduzierte Daten. Entstehungsbedingungen und Indikatorenqualität. In *Sozialforschung und Verwaltungsdaten. Historisch-Sozialwissenschaftliche Forschungen Band 17,* Hrsg. Wolfgang Bick, Reinhard Mann und Peter J. Müller, 123–159. Stuttgart: Klett-Cotta.

Biemer, Paul P. 2010. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly* 74(5):817–848.

Dehmel, Susanne, Peter Kenning, Gert G. Wagner, Christa Liedtke, Hans W. Micklitz, und Lousa Specht-Riemenschneider. 2020. *Die Wirksamkeit der Corona-Warn-App wird sich nur im Praxistest zeigen. Der Datenschutz ist nur eine von vielen Herausforderungen*. Veröffentlichungen des Sachverständigenrats für Verbraucherfragen. Berlin: Sachverständigenrat für Verbraucherfragen.

Diekmann, Andreas. 2016. Gesellschaft der Daten. Die Soziologie muss sich neu erfinden. *Süddeutsche Zeitung*. 25.09.2016. www.sueddeutsche.de/kultur/geisteswissenschaften-die-gesellschaft-der-daten-1.3178096 (Zugegriffen: 15. Jan. 2021)

Einspänner, Jessica, Mark Dang-Anh, und Caja Thimm. 2014. *Twitter and Society*, Hrsg. Katrin Weller, Axel Bruns, Jean Burgess, Meria Mahnt und Cornelius Puschmann, 97–108. Bern: Peter Lang.

Elias, Norbert. 1969. *Die höfische Gesellschaft*. Frankfurt am Main: Suhrkamp.

Fuchs, Marek. 1994. *Umfrageforschung mit Telefon und Computer. Einführung in die computergestützte telefonische Befragung*. Weinheim: Psychologie Verlags Union.

Graeff, Peter, und Nina Baur. 2020. Digital Data, Administrative Data, and Survey Compared: Updating the Classical Toolbox for Assessing Data Quality of Big Data, Exemplified by the Generation of Corruption Data. *Historical Social Research* 45(3):244–269.

Groves, Robert M., und Lars Lyberg. 2010. Total survey error. Past, present, and future. *Public Opinion Quarterly* 74(5);849–879.

Hill, Craig A., Paul Biemer, Trent Buskirk, Mario Callegaro, Anna Lucia Cordova Cazar, Adam Eck, Lili Japec, Antja Kirchner, Stas Kolenikov, Lars Lyberg, und Patric Sturgis. 2019. Exploring new statistical frontiers at the intersection of survey science and big data: Convergence at „BigSurv 18". *Survey Research Methods* 13(1):123–135.

Iglesias, José Antonio, Alexandra Tiemblo, Agapito Ledezma, und Araceli Sanchis. 2016. Web news mining in an evolving framework. *Information Fusion* 28:90–98.

Initiative D21 e. V. 2020. *D21-Digital-Index 2019/2020. Jährliches Lagebild zur Digitalen Gesellschaft*. https://initiatived21.de/app/uploads/2020/02/d21_index2019_2020.pdf (Zugegriffen: 15. Jan. 2021)

ITU (International Telecommunication Union). 2019. *Measuring Digital Development. Facts and Figures 2019*. Geneva: ITU. www.itu.int/en/ITUD/Statistics/Documents/facts/FactsFigures2019.pdf (Zugegriffen: 02. Feb. 2020)

Kandt, Jens. 2019. Geotracking. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1353–1360. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_100.

Knoblauch, Hubert, und Rene Tuma. 2020. Videography. An Interpretive Approach to Video-Recorded Micro-Social Interaction. In *The Sage Handbook of Visual Methods*, Hrsg. Eric Margolis und Luc Pauwels, 129–142. Los Angeles: Dage.

Koch, Gertraud. 2019. Digitale Selbstvermessung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1089–1102. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_77.

Lakes, Tobia. 2019. Geodaten. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1345–1352. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_99.

Lepsius, Rainer M. 1979. Die Entwicklung der Soziologie nach dem Zweiten Weltkrieg 1945 bis 1967. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (Sonderheft) 21:25–70.

Mayerl, Jochen, und Thorsten Faas. 2019. Quantitative Analyse von Twitter und anderer usergenerierter Kommunikation. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1027–1040. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_73.

Reuband, Karl-Heinz. 2019. Schriftlich-postalische Befragung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 769–786. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_53.

Rohwer, Anja. 2009. Measuring corruption: A comparison between the Transparency International´s Corruption Perception Index and the World Bank's Worldwide Governance Indicators. *CESifo DICE Report 7.3*, 42–52.

Schmitz, Andreas, und Olga Yanenko. 2019. Web Server Logs und Logfiles. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 991–999. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_70.

Schneider, Daniel, und Kristen Harknett. 2019. What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, 1–33.

Schrape, Jan-Felix, und Jasmin Siri. 2019. Facebook und andere soziale Medien. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1053–1064. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_75.

Schünzel, Anja, und Boris Traue. 2019. Websites. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1001–1013. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_71.

Thaller, Manfred 2017 [1990]. Entzauberungen: Die Entwicklung einer fachspezifischen historischen Datenverarbeitung in der Bundesrepublik. *Historical Social Research*, Supplement 29, 178–192. doi:10.12759/hsr.suppl.29.2017.178-192.

Traue, Boris, und Anja Schünzel. 2019. YouTube und andere Webvideos. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 1065–1077. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_76.

Traue, Boris. 2020. *Selbstautorisierungen. Die Transformation des Wissens in der Kommunikationsgesellschaft*. Habilitationsschrift. TU Berlin.

Tuma, Rene. 2017. *Videoprofis im Alltag – Die kommunikative Vielfalt der Videoanalyse*. Wiesbaden: Springer VS.

Wagner-Schelewsky, Pia, und Linda Hering. 2019. Online-Befragung. In *Handbuch Methoden der empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius, 787–800. Wiesbaden: Springer VS. doi: 10.1007/978-3-658-21308-4_54.

Wallgren, Anders, und Britt Wallgren. 2014. *Register-based Statistics. Statistical Methods for Administrative Data. Second Edition.* Chichester: John Wiley & Sons.

Williams, Nora Webb, Andreu Casas, und John D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification (Elements in Quantitative and Computational Methods for the Social Sciences).* Cambridge: Cambridge University Press. doi:10.1017/978110886074.