Big Data und Repräsentativität

von Cornelius Puschmann

Spiegel oder Schlagschatten?

In den bisherigen Beiträgen der Big-Data-Reihe haben Jochen Mayerl, Jan-Felix Schrape und Simon Munzert ganz unterschiedliche Aspekte dessen beleuchtet, was Big Data aus Sicht der Sozialwissenschaften ausmacht oder ausmachen könnte. Konkret wurden Beobachtungen zur wachsenden Bedeutung interdisziplinärer Kooperationen (Mayerl), zur Notwendigkeit des Erwerbs neuer Kompetenzen innerhalb der Sozialwissenschaften (Munzert), und zu Big Data als Utopie und „Erwartungsraum" in Abgrenzung zu einem klar umgrenzten wissenschaftlichen Konzept (Schrape) vorgestellt. Darauf aufbauend möchte ich nun einen Aspekt der Big-Data-Thematik hervorheben, für dessen Erforschung die Kommunikations- und Medienwissenschaft besonderes Rüstzeug mitbringt, nämlich die zahlreichen Verzerrungen, die sich bei der Betrachtung der Gesellschaft durch die Medien – gerade auch die digitalen – zwangsläufig ergeben[1]. Es geht um die Interpretation digitaler Spuren ("digital trace data"), welche oft die Grundlage von Big-Data-Analysen darstellen, und speziell die mangelnde Repräsentativität dieser Spuren.

Digitale Spuren als "low hanging fruit"

Bei weitem nicht alles, was mit den Begriffen "Big Data" oder "Computational Social Science" in Verbindung gebracht werden kann, hat mit digitalen Spuren zu tun. Auch digitalisierte Inhalte (Mitschriften von Parlamentsdebatten, Parteiprogramme), Erhebungsdaten (Freitextantworten aus Fragebögen, Interviewtranskripte) und eine Vielzahl von weiteren Informationen aus öffentlich zugänglichen Datenbanken lassen sich sowohl manuell als auch computergestützt und mit einer Kombination unterschiedlicher Verfahren analysieren. Umgekehrt gehört auch die Anwendung von Simulationstechniken, die ganz ohne empirische Daten auskommen, etwa der agentenbasierten Modellierung, zum Repertoire der Computational Social Science[2]. Die Größe des Datensatzes allein ist ebenso wenig entscheidend wie der Einsatz bestimmter Verfahren – "in some cases, small is best", wie Mayerl betont.[3]

Trotz der potenziellen Vielfalt der Thematik ist das Interesse an Daten aus sozialen Medien wie Facebook und Twitter besonders groß. Vielleicht gilt das gerade für diejenigen Wissenschaftler, die erst über die Daten und Methoden zu den Sozialwissenschaften gekommen sind, etwa aus der Informatik und der Physik, für die die Zugänglichkeit und der Umfang der Daten eine inhärent wichtige Rolle spielen.[4] Zeynep Tufekci spricht in diesem Zusammenhang von Twitter als "model organism" dieser Forschungsrichtung[5], ähnlich wie die Drosophila (die Schwarzbäuchige Fruchtfliege oder Taufliege) auch aus forschungsökonomischen

Gründen als Modellorganismus für die Biologie dient. Die Ursachen für diese Vorliebe liegen auf der Hand: Einerseits nutzen immer mehr Menschen in immer mehr Ländern die sozialen Medien intensiv, nicht nur zur Unterhaltung und Kontaktpflege, sondern auch, um sich über Politik oder Gesundheitsthemen zu informieren. Andererseits sind die Daten speziell bei Twitter vergleichsweise leicht und in großem Umfang zugänglich. Digitale Spuren geben inzwischen nicht mehr nur Aufschluss über die Präferenzen einiger *early adopters,* sondern bilden einen nicht unerheblichen Teil der alltäglichen Mediennutzung von immerhin 320 Millionen Menschen (Twitter) bzw. knapp 1,8 Milliarden Menschen (Facebook) ab.

Soziale Medien als Spiegel oder Schlagschatten der Gesellschaft?

In einem Vortrag an der Stanford University vor drei Jahren hat der Physiker Duncan Watts die Sozialwissenschaften mit der Astronomie verglichen. Entsprechend argumentierte er, durch soziale Medien und das Internet lasse sich die Gesellschaft so bequem beobachten, wie wir Himmelskörper durch ein Teleskop betrachten können.[6] Mit dieser Meinung ist er nicht allein. Inzwischen haben Physiker, Mathematiker und Informatiker ein wachsendes Interesse daran, die großen Mengen an Daten systematisch auszuwerten, die digitale Plattformen wie Facebook und Twitter generieren, um aus ihnen Rückschlüsse über menschliches Verhalten zu ziehen. Derartige Forschung hat sowohl eine akademische als auch eine praktische Dimension und lässt die Grenzen zwischen Wissenschaft und Marktforschung häufig verschwimmen. Unter Sozialwissenschaftlern gilt sie als kontrovers, nicht nur, da sie mitunter ethische Fragen aufwirft. Sie lässt auch auf eine Vorstellung von Verhalten und Kommunikation schließen, die mit den prägenden sozialwissenschaftlichen Theorieströmungen der letzten Jahrzehnte nicht immer in Einklang gebracht werden kann.
Im Übrigen liefert Watts eine griffige Metapher, der er möglicherweise inzwischen selbst nicht mehr uneingeschränkt zustimmen würde: Soziale Medien erlauben die Beobachtung der Gesellschaft *durch* das Internet, nicht etwa nur die Beschreibung des Internets und seiner Communities als etwas Partikuläres und Eigenes. Damit setzt er sich von den „Internet Studies“ der vergangenen Jahrzehnte deutlich ab, die genau diese Eigenheiten des Netzes in den Vordergrund gestellt haben. Gleichwohl setzt Watts’ Vergleich mit dem Teleskop unweigerlich voraus, dass die sozialen Medien uns ein adäquates Bild der Gesellschaft liefern, es sich also um ein geeignetes wissenschaftliches Instrument handelt – und das dies wirklich der Fall ist, erscheint durchaus fragwürdig. Genau auf diese Grundspannung verweist auch der Titel des Beitrags von Jochen Mayerl. Wenn wir so viele Daten über das Verhalten von Menschen haben, warum brauchen wir dann noch Umfragen?
Zwei Probleme ergeben sich für die Big-Data-Forschung, auf die Mayerl bereits unter den Stichworten *Indikatoren-Problem* und *(mangelnde) Repräsentativität* hingewiesen hat und die ich noch weiter ausdifferenzieren möchte. Zum einen wird in Analysen digitaler Spuren häufig auf Indikatoren zurückgegriffen, die sich kaum standardisieren lassen – jedenfalls nicht im sozialwissenschaftlichen Verständnis. Ein gutes Beispiel ist die sogenannte Sentimentanalyse, also die Messung der

emotionalen Valenz eines Textes. Mit einer Reihe von methodischen Ansätzen kann man Sentimentanalysen durchführen, aber die Validität und intersubjektive Nachprüfbarkeit der Ergebnisse lassen oftmals zu wünschen übrig. Das hält die Industrieforschung nicht davon ab, trotzdem Sentimentanalysen durchzuführen; es beschränkt aber deren Nutzen für die sozialwissenschaftliche Forschung nachhaltig. Selbst wenn manuelle und computergestützte Inhaltsanalyse kombiniert werden, lassen sich anhand von Äußerungen schwerlich zulässige Annahmen über das Verhalten und die Motive von Nutzern formulieren. Das macht solche Verfahren für eine Reihe von Fragestellungen ungeeignet.[7]

Zweierlei Unrepräsentativität: Variierende Nutzungsmuster und „Lautstärke“

Das Problem der Repräsentativität möchte ich in größerem Detail beschreiben, als das in diesem Zusammenhang oftmals getan wird, weil es sich teilweise aus den besonderen Voraussetzungen der Onlinekommunikation ergibt. Es herrscht allgemeiner Konsens darüber, dass Daten aus den sozialen Medien nicht bevölkerungsrepräsentativ sind, da nicht alle sozialen Gruppen gleichermaßen im Internet aktiv sind. Allerdings wird selten genau herausgearbeitet, in welcher Hinsicht dies der Fall ist. In einer aktuellen Studie ermittelt Grant Blank[8] entsprechende Daten für Twitter anhand von bevölkerungsrepräsentativen Umfrageergebnissen. Er kommt zu dem Schluss, Twitter eigne sich für eine Vielzahl soziologischer und politikwissenschaftlicher Forschungsfragen keinesfalls, weil britische Twitternutzer jünger, wohlhabender und gebildeter sind als durchschnittliche Internetnutzer. Letztere sind ihrerseits jünger, wohlhabender und gebildeter als der Bevölkerungsdurchschnitt.
Obschon derartige Daten nur begrenztes Erkenntnispotenzial aufweisen (für Deutschland gilt mit hoher Wahrscheinlichkeit das Gleiche), werden regelmäßig Versuche unternommen, Wahlentscheidungen oder die öffentliche Meinung anhand von Twitter zu analysieren, als schaue man durch ein Teleskop auf die eigene Nachbarschaft.[9] Bei Facebook sieht das Bild aufgrund der höheren Verbreitung und einer anderen Zusammensetzung der Nutzerschaft zwar etwas weniger verzerrt aus, das Problem besteht aber prinzipiell im gleichen Umfang. Es reicht zudem nicht aus, davon zu sprechen, dass die Nutzergemeinschaften dieser Plattformen weltweit nicht repräsentativ sind. Vielmehr gilt es unter anderem, signifikante nationale Unterschiede herauszuarbeiten. In Deutschland wird Twitter beispielsweise von anderen Bevölkerungsgruppen genutzt als in den USA oder in Saudi-Arabien; Gleiches gilt für andere Plattformen.
Ein weiterer, bisher nahezu vollkommen vernachlässigter Aspekt ergibt sich aus dem Verhältnis *aller* User zu den Nutzern, die den Großteil der Aktivität erzeugen – sei es in Form von Äußerungen, Klicks oder der Nutzungsdauer. In einer Studie zur Nutzung der E-Petitionsplattform des Deutschen Bundestags habe ich gemeinsam mit meinen Kollegen Jan H. Schmidt und Marco Bastos sehr aktive mit weniger aktiven und sporadisch aktiven Nutzern verglichen, also danach gefragt, wie sich quantitative Unterschiede im Aktivitätsniveau mit qualitativen Unterschieden etwa bei den bevorzugten Themenfeldern vergleichen lassen.[10] Dabei stellten wir fest,

dass der Anteil der männlichen Nutzer, der insgesamt bei circa 50 Prozent lag, in der Gruppe der hochaktiven Nutzer auf 75 Prozent anstieg. Ebenso konnten wir Unterschiede bei den Themenpräferenzen zwischen hochaktiven und sporadischen Nutzern feststellen.
All dies mag erwartbar sein; die Befunde deuten aber auf ein Wahrnehmungsproblem hin, dem wir uns hinsichtlich der ‚Gleichwertigkeit' von Nutzerstimmen stellen müssen. Oftmals wird implizit davon ausgegangen, dass Repräsentativität erreicht wäre, wenn die soziodemographischen Eigenschaften der Facebook- und Twitternutzer denen der Gesamtbevölkerung entsprächen. Da aber eine kleine Gruppe von Nutzern innerhalb dieser Plattformen für einen solch hohen Anteil der Gesamtaktivität verantwortlich ist, erfahren wir schlussendlich primär etwas über die Präferenzen dieser Teilgruppe. Vieles deutet darauf hin, dass wir damit meilenweit von dem entfernt sind, was Watts sich unter einem Teleskop für die Gesellschaft vorstellt.

Plattformreaktive Daten

Eine weitere Komplikation besteht darin, dass digitale Umgebungen sich selbst auf die Spuren auswirken, welche die Nutzer erzeugen. Jürgen Pfeffer hat dies eindrucksvoll mit einer Zeitreihenanalyse von Freundschaftsbeziehungen auf Facebook nach Einführung des "People You May Know"-Features demonstriert.[11] Besagte Facebook-Funktion macht Nutzern aktiv Vorschläge, mit wem sie sich anfreunden könnten, um deren Vernetzung zu fördern – mit großem Erfolg. Die resultierende plötzliche Zunahme der Freundschaftsbeziehungen, die man in einer traditionellen Erhebung nach Ausschluss von Messfehlern oder irregulären Ereignissen wahrscheinlich mit einem veränderten Nutzerverhalten erklären würde, wird aber längst nicht immer so transparent, wie es im Falle einer neuen Facebook-Funktion der Fall ist. Sehr oft bleiben technische Veränderungen, welche letztendlich beeinflussen, was Nutzer sehen oder anklicken können, nicht nur den Nutzern selbst, sondern auch den Wissenschaftlern, die deren Verhalten untersuchen, verborgen, was zu offensichtlichen methodischen Problemen führt.
Damit sind die verkomplizierenden Faktoren immer noch nicht erschöpft: Nutzer reagieren auch auf andere Nutzer, und zwar nicht nur auf die ‚vorgesehene' Art und Weise, sondern auch mit subversiven Taktiken. Ein konkretes Beispiel ist das Meldeverhalten auf Facebook. Die Möglichkeit, anstößige Beiträge dem Betreiber zu melden, wird zum Teil ganz bewusst gegen Inhalte eingesetzt, die keineswegs den von Facebook entwickelten Vorschriften widersprechen, sondern die manchen Nutzern schlicht nicht gefallen oder gegen die sie politisch opponieren. Organisieren sich ausreichend viele Nutzer auf diese Art und Weise, können bestimmte Inhalte gesperrt (Facebook, Twitter) oder Suchergebnisse mit rassistischen Inhalten verknüpft werden (Google). Strategisches Nutzerverhalten kann also das Ziel haben, andere Nutzer oder die Erfahrung anderer Nutzer auf der Plattform zu beeinflussen, statt ‚authentisches' Verhalten abzubilden. Von nicht-reaktiven Daten zu sprechen, ist also insofern durchaus irreführend, als dass die Wissenschaftler zwar keinen Einfluss ausüben, die Plattform und ihre Nutzer aber sehr wohl. Gerade der Unwille, die Plattform selbst als Kommunikationsraum und

als Akteur zu begreifen, schränkt das Sichtfeld der Computational Social Science bislang nachhaltig ein.

Fazit: Medieneffekte überall

Man könnte aus alledem den falschen Schluss ziehen, dass computergestützte Analyseverfahren für die Sozialwissenschaften keine Rolle spielen. Damit würde man allerdings verkennen, dass einerseits nicht alle sozialwissenschaftlichen Forschungsdesigns die exakt gleichen Anforderungen an Repräsentativität stellen, und andererseits die ‚Einmischung' digitaler Plattformen ins Nutzerverhalten nicht automatisch ein Störfaktor sein muss, wenn man ihr nur ausreichend Rechnung trägt. Zum ersten Punkt ist anzumerken: Big-Data-Verfahren eignen sich zum Teil für Analyseschritte, die zwar selektieren, filtern oder priorisieren, aber nicht automatisch quantitative sozialwissenschaftliche Methoden ersetzen. Dies gilt besonders für das Text Mining[12], das oftmals eher explorativ eingesetzt wird. Es spricht nichts dagegen, Meinungen oder Standpunkte aus den sozialen Medien zur Grundlage weitergehender Analyseschritte zu machen, die dann auf standardisierten Methoden beruhen. Wichtig ist: Nicht jede Analyse, in der computergestützte Verfahren zum Einsatz kommen, ist automatisch quantitative Forschung im traditionellen Sinn oder muss es sein. Zugleich folgt daraus, dass alternative Datenquellen die traditionellen nicht ersetzen können.
Zweitens sind es *gerade* die Effekte der Medien (auf Social Media bezogen: die Effekte des Plattformdesigns und der in Social Media-Plattformen integrierten Selektions- und Filtermechanismen), die mich als Kommunikationswissenschaftler interessieren. Vielleicht sind sie für die Forschung gar ergiebiger als die „unverfälschten" Informationen, die Duncan Watts sich erhoffte. Die Utopie der „Wahrhaftigkeit" von digitalen Spuren – hier unbedingt zu unterscheiden von deren Objektivität – nimmt so zwar Schaden, aber das dürfte nur diejenigen enttäuschen, die (wie Jochen Mayerl treffend feststellt) von einer Substitution etablierter Verfahren durch Big Data ausgehen statt von einer sinnvollen Ergänzung. Big Data liefert oftmals ein sehr verzerrtes Bild, aber durchaus eines, welches sich zu betrachten lohnt – vor allem dann, wenn man es vor dem Hintergrund anderer Datenquellen tut, wie etwa Andreas Jungherr[13] oder Grant Blank.

Lernt programmieren!

Wie also mit diesen Herausforderungen umgehen? Simon Munzert weist auf einen entscheidenden Punkt hin: Programmierkenntnisse sind ausgesprochen wichtig, um eine genuin kritische sozialwissenschaftliche Forschung anhand von Big Data betreiben zu können. Programmierkurse sollten mittelfristig Einzug in unsere Curricula halten, und dabei sollte Open-Source-Lösungen wie R oder Python unbedingt der Vorzug gegenüber proprietären Softwarepaketen gegeben werden. Nicht zuletzt können wir so eine gemeinsame Arbeitsbasis mit ‚den' Informatikern schaffen, die die Zusammenarbeit über Disziplingrenzen hinweg entscheidend erleichtern wird.
Wer argumentiert, ‚die' Informatiker seien den Sozialwissenschaften uneinholbar

weit voraus, wenn es um die Anwendung computergestützter Analyseverfahren geht, verkennt, dass die Informatik eine dynamische und vielfältige Disziplin mit hohem Spezialisierungsgrad ist. Häufig werden Probleme erforscht, die mit einer (wie auch immer genau ausgestalteten) „computergestützten Sozialwissenschaft“ wenig zu tun haben. Hinzu kommt, dass sozialwissenschaftliche Grundlagen von Wissenschaftlern aus anderen Bereichen (auch der Physik, Mathematik, Computerlinguistik etc.) ebenfalls Stück für Stück erarbeitet werden müssen. Die Idee, dass Soziologen, Politikwissenschaftler und Kommunikationswissenschaftler programmieren lernen, scheint keineswegs ambitionierter, als der Wunsch, dass Informatiker sich sozialwissenschaftlich weiterbilden. Beides ist zu begrüßen und wird nicht das Ende ‚unserer‘ Methoden oder ‚unseres‘ Faches bedeuten. Vielmehr kann es dazu beitragen, die erhitzte Debatte um Big Data als Heilsbringer oder Menetekel der Sozialwissenschaften wieder auf den Boden der Tatsachen zurückzubringen.
Dieser Beitrag ist Teil eines Themenschwerpunkts zu Big Data. Weitere Texte finden Sie hier.

Fußnoten

1 Für eine umfassende Diskussion von Social Media-Daten als Zerrspiegel und den Konsequenzen für die Sozialwissenschaften, vgl. Andreas Jungherr, Analyzing political communication with digital trace data, Heidelberg 2015.

2 Annie Waldherr, Emergence of news waves. A social simulation approach, in: Journal of Communication 64 (2014), S. 852–873.

3 Jochen Mayerl, Bedeutet 'Big Data' das Ende der sozialwissenschaftlichen Methodenforschung?, in: Soziopolis, 29. Dezember 2015.

4 Claudio Cioffi-Revilla, Computational social science, in: Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010), 3, S. 259–271.

5 Zeynep Tufekci, Big questions for social media big data: Representativeness, validity and other methodological pitfalls, in: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, Palo Alto, CA, 2014, S. 505–514.

6 Duncan Watts, Computational Social Science. Exciting Progress and Grand Challenges, Vortrag an der Stanford University, 19. September 2013.

7 Vgl. Andreas Jungherr / Pascal Jürgens / Harald Schoen, Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. “Predicting elections with Twitter: What 140 characters reveal about political sentiment.”, in: Social Science Computer Review 30 (2012), 2, S. 229–234.

8 Grant Blank, The Digital Divide Among Twitter Users and Its Implications for Social Research, in: Social Science Computer Review 2016, S. 1–19.

9 Vgl. Andrak Tumasjan et al., Predicting elections with Twitter: What 140 characters reveal about political sentiment, in: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010), S. 178–185.

10 Cornelius Puschmann / Marco T. Bastos / Jan-Hinrik Schmidt, Birds of a feather petition together? Characterizing e-petitioning through the lens of platform data, in:

Information, Communication & Society 20 (2017), 2, S. 203–220 [im Erscheinen].

11 Jürgen Pfeffer, Vortrag beim Computational Social Science Winter Symposium, 2.–3. Dezember 2015, GESIS, Köln.

12 Vgl. Matthias Lemke / Gregor Wiedemann (Hrsg.), Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse, Heidelberg 2016.

13 Andreas Jungherr, Analyzing Political Communication with Digital Trace Data, Heidelberg 2015.