Received: 14 August 2017 | Revised: 7 February 2018 | Accepted: 20 April 2018

DOI: 10.1111/gec3.12374

ARTICLE

WILEY

Challenges to representing the population from new forms of consumer data

Guy Lansley | James Cheshire

Department of Geography, University College London, United Kingdom

Correspondence
Guy Lansley, Department of Geography, University College London, United Kingdom.
Email: g.lansley@ucl.ac.uk

Funding information
UK ESRC Consumer Data Research Centre (CDRC), Grant/Award Number: ES/L011840/1

Abstract

A large share of human activity is now routinely captured and stored by commercial organisations as consumer data. These new forms of data have provided many new and exciting opportunities for understanding population trends at fine spatial scales. In many settings, they have reduced our dependence on theory and traditional modelling approaches and fundamentally changed how geographers approach producing representations of the population. However, consumer data are typically created for commercial benefit and do not have academic standard data quality controls, and so, in repurposing the data for social research, we encounter inherent issues of veracity. Without an understanding of uncertainty, social scientists can risk overstretching consumer datasets beyond the specific populations or phenomena they directly pertain to. Moreover, the technical characteristics of large and complex datasets also make it challenging to generate valid information efficiently. Therefore, this paper reviews the major challenges to harnessing consumer data to produce valid spatial representations of the population at large.

1 | INTRODUCTION

A substantial—and increasing—range of human activity is now being captured and stored digitally by commercial organisations as consumer data (OECD, 2013). Much of these data contain spatial attributes, which include references to locations (e.g., addresses) that either assist in the identification of customers or delivery of products, or, through new technologies that enable the precise detection of coordinates (e.g., using georeferenced social media services on GPS-enabled devices). All benefit from technological innovations that facilitate their collection, storage, and analysis, and they contribute to a wealth of new data available encompassed by the term "Big Data."

"Consumer data" is defined by the UK Government's Competition and Markets Authority (CMA) as "any information firms might collect from and about consumers that is used, or intended to be used, to support commercial activities" (CMA, 2015). This includes data that

- consumers offer voluntarily ("declared data")—for instance, when transacting or registering for a service;
- consumers generate and supply passively ("observed data")—for instance, on social media or when their online browsing activity is tracked; and
- are generated by first and third parties as a result of analysis or in combination with other data (CMA, 2015).

Consumer datasets have unlocked the possibility of conducting large-scale population research in new topics and settings. Moreover, they could supplement or even replace traditional long-form censuses in the future due to their coverage and the breadth of the phenomena they represent collectively (Dugmore, 2010). Efforts to gain access to commercial data for research purposes are also motivated by an increasingly acknowledged necessity to broaden the base of possible data sources for social research since traditional sources of data—such as surveys—are costly to administer and in many cases suffering from decreasing response rates. In the UK, for example, this process was catalysed by the "Beyond 2011" Census initiative (Ralphs & Tutton, 2011). Unfortunately, many datasets remain difficult to access to the extent that only a very small proportion of large datasets on the population are released in the public domain due to their strategic importance, commercial value, and potentially disclosive attributes.

What is more, there has been a relatively limited appraisal of the validity of consumer datasets as indicators of population characteristics (Folch, Spielman, & Manduca, 2017; Lee & Kang, 2015). Therefore, whilst we remain optimistic about the value of such data, this paper discusses some of the challenges to producing representative population research with spatial data routinely collected by commercial organisations. These challenges have been identified through our experiences assembling, processing, and analysing a wide range of datasets as part of the UK's Consumer Data Research Centre (CDRC). They fall into three key categories:

1. Epistemological. Human geography as a discipline has had to adapt to a new era of knowledge discovery and is still deeply divided.
2. Veracity. Data quality issues are paramount and can be difficult to comprehend.
3. Technical. The volume, velocity, and the variety of Big Data can make them technically challenging to handle.

2 | LARGE CONSUMER DATASETS AND THEORY IN GEOGRAPHIC RESEARCH

All geographic data are models of the world or its processes in some form; they can only ever be partial representations of reality (Longley, Goodchild, Maguire, & Rhind, 2015). This is particularly true of consumer data. Geography as a discipline has long been concerned with devising representations of the population and their spatial characteristics, often utilising imperfect data (Trewartha, 1953). Much of the fundamentals of the analysis of geospatial population data were first devised in the quantitative revolution in human geography of the 1960s and 1970s. The revolution integrated scientific theory and rigour into human geography following concerns that the discipline had become unobjective and lacked theoretical reasoning (Johnston, 2008). It reintroduced older spatial theories that were not previously taught due to the inability to test them efficiently (Gould, 1979). For instance, Brian Berry and William Garrison (1958) tested the principles of the central place theorem using data from Snohomish County. Their research was the first real critical attempt to identify general trends from spatial data. The paradigm shift was enabled by new technologies and the storage and accessibility of large geographically referenced population datasets. New opportunities for Geographic Information Systems (GIS) to develop arose and techniques for spatial analysis were devised; to a large extent, the principles of spatial analysis have barely changed since (Openshaw, 1991). In response to the perception that GIS was simply software—rather than a field of critical enquiry—Geographic Information Science (GIScience) was established as a discipline, and it sought to a reach a compromise between idiographic (description-seeking) and nomothetic (law-seeking) approaches through methodological techniques that account for variations across space (Fotheringham, Brunsdon, & Charlton, 2002).

As data became abundant, new opportunities for representing populations across a range of scales emerged. For example, the use of aggregate data is no longer a necessity; it is optional since contemporary data and the technology to process it enable research to focus on the roles of individuals within complex systems (Batty, 2013b). Analysis can be both general and specific. Such advances have been used to support the idea that we are experiencing a second paradigm shift in quantitative research (Hey, Tansley, & Tolle, 2009; Johnston et al., 2014). Previously research on population geography was grounded in theory that could not be empirically tested due to the dearth of data. For example, in the context of retail geography, researchers have previously estimated where customers live using spatial interaction models, which assume customers are drawn to retail outlets based on their size (or attractiveness) and relative proximity alone (Huff, 1964). However, today, customer loyalty databases routinely record where large numbers of customers live and shop and therefore provide a viable alternative to models based on aggregate population data.

Research increasingly grounded in empirical data counters one of the core antipositivist critiques of models: that they assume all individuals act rationally. Using data containing near-complete populations, for example, or, at the very least, much larger samples, researchers can also represent those that seemingly act irrationally. To that end, we can observe previously unobservable exceptions occur where human behaviour is not constrained by the laws determined for models. As such, it has been argued that large enough datasets eliminate the need for theory as it offers this empiricist knowledge on populations (Anderson, 2008). Indeed, Anderson (2008) has famously argued that we could see an end of theory as "with enough data, the numbers speak for themselves." He noted that the volume, velocity, and detail of some datasets meant that with data mining techniques "correlation is enough."

This view has been widely tempered with frequent reminders, and examples, of a standard tenet of statistics: Correlation does not imply causation. A notable example is Google Flu Trends, which aimed to estimate the spread of influenza in near real time by tracking search terms on Google. Although their results originally correlated with findings from the Centers for Disease Control and Prevention, after some time, the methodology greatly overpredicted flu. Researchers had identified patterns in their data but did not have a complete understanding of the correlations (Harford, 2014). It is likely that many terms were overfitted as identifiers, and many of these may be seasonal in frequency. In addition, the researchers did not anticipate the role of the media in fuelling public panic over the flu (Lazer, Kennedy, King, & Vespignani, 2014). By letting data speak for themselves, researchers neglect key geographic and sociological concepts that may underpin trends and limit their research to the ideographic knowledge that the pioneers of quantitative geography sought to move beyond.

Whilst the advent of Big Data has resurfaced epistemological debates about knowledge discovery (Cresswell, 2013, 2014), we suggest that a balance between idiographic and nomothetic approaches is still required to fully understand the world's processes. Inductive modelling techniques rarely offer explanations (Miller & Goodchild, 2015), although there is no doubt that the extent of data available now has reduced our dependence on traditional theory-driven techniques. However, theory is still integral to human geography. As geographers seek to extrapolate data to represent the broader population they must take into account both spatial dependence and spatial heterogeneity (Anselin, 1990). Therefore, fields such as GIScience have had to adopt both lenses. Theory and domain-specific knowledge at the very least are still required to select appropriate techniques and to interpret trends (Miller & Goodchild, 2015). Whilst previous scientific research was based on identifying a problem, many geographers have opted to explore Big Data first, then to identify an appropriate research question afterwards as they may offer new insights into underlying phenomena that were previously unobserved and overlooked (Harris et al., 2017). Often, the hypotheses are tailored to fit the data and be conclusive rather than addressing the original critical questions (Barnes, 2013).

In addition to the divisions between quantitative paradigms in geography, most of the criticisms of broader positivistic research that surfaced following the quantitative revolution have not been addressed (Christensen, 1982). This is because data may be larger, but they are most commonly still numerical indicators of particular actions or characteristics sampled from a broader population. Research utilising it must, therefore, be equally reductionist, and as researchers, we should be cognisant of this (Kitchin, 2015). What is more, positivistic approaches typically neglect metaphysical concepts, which may be fundamental to behaviour and actions (Tuan, 1976). Indeed, there are still some key geographic issues where positivistic research can offer little contribution, such as understanding the formulation of geopolitical divisions (Harvey, 1973). Whilst positivists attempt to avoid philosophical paradigms and remain scientifically objective, they usually remain external observers that are restricted to a top-down view, which may be subject to bias (Gregory, 1978). Consequently, critical geographers have distanced themselves from positivism, arguing that it lacks fundamental ideological and ontological bases that are crucial for understanding real-world problems.

That said, there is a growing sense that antipathy towards data is no longer a sustainable position for those who have historically had an aversion to it. There remain many who will need convincing, for example, Cresswell (2014) argues that is important for geographers to know enough to question “the stream of numbers that gets pumped out by corporations, governments and the media,” whilst suggesting the well-founded calls for continuous training in quantitative methods are inhabiting “some of the same ground that … big data for the production of profit or the manipulation of populations does” (p. 58). Such training, however, is essential if geographers (and social scientists more broadly) are to meaningfully engage with a data-driven world. As researchers, we would much rather be part of the process of creating social good from new forms of data rather than shouting from the sidelines. To that end, a way forward has been proposed by Wilson (2017) who suggests “New Lines” of enquiry that cut through the historic disciplinary divides that perhaps no longer apply in the Big Data era. This constructive approach advocates that dismissal should be replaced with dialogue and lays the foundations for a much richer treatment of Big Data within geography.

3 | THE VERACITY OF CONSUMER DATA

As discussed above, the shift from data-poor to data-rich does not mean we have reached a point where we can accurately depict the full complexities of the human condition (Amin & Thrift, 2002). Two fundamental issues of representation remain. Firstly, they are limited by demographic biases since no samples are objectively random. Secondly, as the data are often by-products of transactions of some kind, there are frequently data quality issues when they are repurposed. For example, in Figure 1, several thousand journeys taken by runners during a typical week in London are mapped. These have been recorded by a fitness app and reveal fascinating patterns about desirable running routes, turning small paths across parks into busy routes, whilst the major roads disappear. These tracks can provide useful information for city planners looking to encourage jogging, but they also contain biases. Fewer paths are seen in poorer areas to the east and north-east; there are also errors from the GPS or when a jogger boards public transport without deactivating the app. This section details how an awareness of such issues is integral to successful analysis.

3.1 | Demographic biases

It is increasingly argued that Big Data has changed how we approach sampling in geography and the social sciences more broadly. We have shifted from utilising data built from carefully constructed sampling frames to harnessing

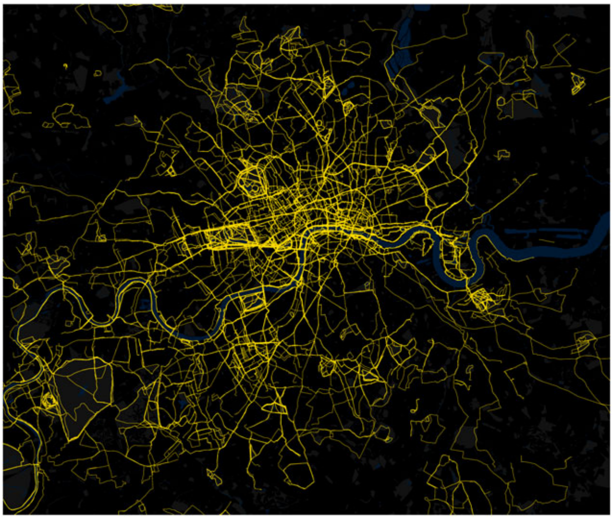

FIGURE 1 Running routes recorded by a fitness app in London

bulks of Big Data which represent potentially very biased populations with little knowledge of their provenance. This can be especially challenging for consumer data given that a single source rarely contains a sufficient breadth of variables to measure geodemographic heterogeneity (Lee et al., 2016).

A historical example of demographic biases in big datasets describes when the Literary Digest once ran a postal poll to predict the outcome of the 1936 presidential election in the USA. Despite achieving 2.4 million returns, the poll incorrectly indicated that Landon would beat Roosevelt. This was in part due to their sampling; the magazine had utilised vehicle registration lists and telephone directories to acquire names and addresses for their ballots. Both data sources were biased to consumers of higher social status at the time (Squire, 1988).

Large consumer datasets are prone to biases, and it is not unknown for a minority of adults to contribute the majority of data where data contributions are by-products of services (Longley, Adnan, & Lansley, 2015). This issue is particularly challenging for positivist social scientists who are more familiar with normal distributions and using linear models to extrapolate their data in order to represent the population at large. Many consumer datasets are self-selecting samples in one form or another. Action is required for data to be generated, and actions are almost never evenly distributed across the population. For instance, customer loyalty databases will over-represent the geodemographic characteristics of individuals who are more likely to visit the given store chain, in addition to the characteristics of those who are near their stores (Wright & Sparks, 1999). This makes it challenging to represent broader consumption patterns across a population or link it to independent variables as the sample is not neutral. Building representative samples has always been a key consideration in social research, but in the era of Big Data, these issues can be easily missed when dealing with millions, if not billions, of data points that were previously unobtainable. Data biases exist across all scales. In the case of Twitter (which has become synonymous with Big Data in geographic research due to its large size and widespread availability, in comparison to data from the likes of Facebook), those who generate data via the social network represent a small part of society. For example, within the UK, it has been observed that Twitter users who share their location are skewed towards young adults and the White British ethnicity (Longley et al., 2015). It is therefore unsurprising that Twitter data provide a very cumbersome tool for predicting real-world phenomena which relate to the wider society, such as elections results (Gayo-Avello, 2012). In terms of the built environment, social media can show significant clustering, as Folch et al. (2017) demonstrate in Phoenix. They compared the geography of restaurants identified by a social media source (yelp.com) and those from an administrative source. Only about one third of restaurants in each dataset were common between them, with Yelp offering much better coverage in downtown areas compared to the greater consistency of the administrative data across the study region. The datasets combined offered a more detailed picture, but independently, they could not be considered equivalent.

It is also clear that many new forms of data are dependent on reliable access to the Internet. At the global scale, connectivity is improving, but developing nations and those with less information equality still lag behind (Graham & Foster, 2014; Sui, Goodchild, & Elwood, 2013). Even within developed countries, there can be surprising differences in access. For example, Figure 2 shows the different download speeds experienced across London; it follows that those with slower speeds will have more limited online behaviours and this can impact on the kinds of data they generate.

It is therefore important to remain aware of over-representation and under-representation in consumer data (and Big Data more broadly) and acknowledge that both will vary depending on the context. For example, it is common for data to only be inclusive of persons who meet particular criteria, such as being above a particular age or being a resident. Consequently, teenagers are under-represented in the commercial context despite their engagement with consumer services (Spero & Stone, 2004).

3.2 | Considerations for repurposing consumer data for academic research

One of the merits of consumer datasets is that they are generated passively from everyday activities such as store transactions or posting on social media. This enables the reliable stream of data on a range of attributes that are otherwise difficult and costly to obtain. However, whilst social surveys objectively curate the data collection process so that data directly measure phenomena for social research, consumer data are generated by third parties as a by-product. Researchers should, therefore, consider the connotations of repurposing Big Data to represent people and their actions. In addition to demographic biases, consumer data also have restricted spatial and temporal coverage which are vary depending on the nature of data collection. For instance, whilst London's travel smart card system (Oyster) is an excellent source of spatiotemporal data on population mobility, it is primarily a ticketing system, not a system for travel surveillance (Reades, Zhong, Manley, Milton, & Batty, 2016). Whilst the data can be informative of where and when Oyster cards were used, they cannot hone the precise origin and destinations of journeys. Likewise, retail data generally record who purchases products, not who actually consumes them. Data, therefore, are fundamentally linked to the generating actions, most of which are controlled by commercial organisations; therefore, they should not be considered neutral (Kitchin, 2014a).

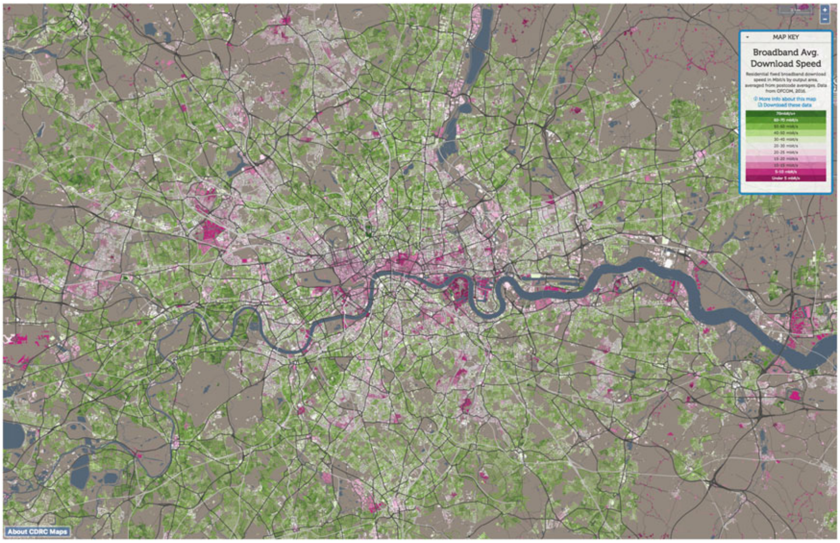

FIGURE 2 Average broadband download speeds across London. Source: maps.cdrc.ac.uk

The previous section identified that whilst consumer data may inherently exclude certain social groups, they may also exclude or misrepresent particular phenomena too, since they are "technically, economically, ethically, temporally, spatially and philosophically" framed (Kitchin, 2014b). Therefore, where possible, researchers should attempt to contextualise the data through the exploration of existing variables, data linkage, or through known trends and laws in order to better understand the manifestation of characteristics and activities (Crampton et al., 2013).

It should also be acknowledged that many new forms of data have not been subject to rigorous quality controls, particularly since the commodification of geographic data has meant some organisations risk overvaluing volume at the expense of quality (Dalton & Thatcher, 2015). Data quality issues typically originate from how the data were collected, either due to technological deficiencies or human error. It may be difficult to maintain the quality of geographic datasets over time as there are often little motivations for individuals to update their records. Government and consumer databases frequently retain incorrect residential records after adults change address (The Electoral Commission, 2016). This limitation is understandable since the costs associated with regularly validating data are prohibitive in the commercial context. We are therefore left with datasets that can consist of an unknown volume of noise that can obscure otherwise notable trends (Batty, 2013a) and that are prone to erroneous data points (Wilson, 2015). Indeed, incentives (such as the allocation of funding or increased revenue) may lead to individuals or organisations intentionally influencing or even falsifying records (Connelly, Playford, Gayle, & Dibben, 2016). Yet in many cases, consumer data can contribute information where traditional datasets cannot and so they have significant research value (Mayer-Schönberger & Cukier, 2013), provided the extent of noise and data quality issues are understood and accounted for. Through the careful development of heuristics, problematic records can be filtered, as Figure 3 shows. It displays the relative proportion of customers from a major loyalty card database that were estimated to have changed address by age.

Finally, the sourcing of large datasets from beyond the rigours of academic research can present some important ethical and legal constraints. The ethical concerns raised about the use of geographic datasets in John Pickles' Ground Truth are still present today (Pickles, 1995) and have in fact been amplified as more data are collected. Breaches of privacy, for example, can impact public opinions of businesses so there are considerable efforts to safeguard big datasets, and these have impeded their integration into the academic community (Duckham & Kulik, 2006). Indeed, access to these remains a significant barrier to research. A large share of research on data pertaining to individuals is undertaken within the commercial sector where the focus is on efficiency rather than scientific robustness (Lazer et al., 2014). But once access has been granted the researcher needs to ensure they are complicit with data protection procedures and also have a clear ethical steer. In many ways, the rapid developments in the collection and application of new forms of data have outpaced the thinking on what is and is not an acceptable use of data.

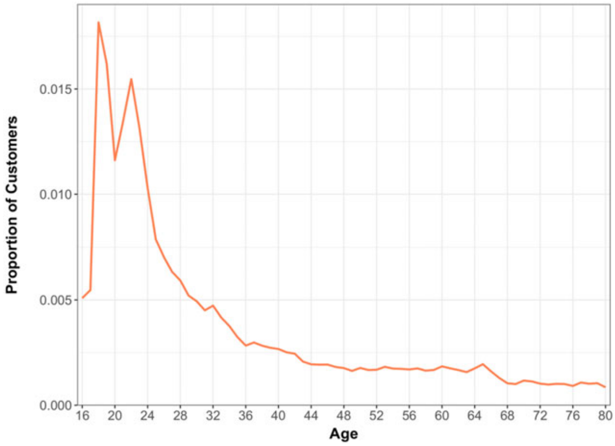

FIGURE 3 Ages recorded at the time of estimated change of address from a major high street retailer loyalty card dataset, normalised by total customers per year of age. Source: Lloyd and Cheshire (2018)

The British Academy and Royal Society's (2017) report entitled *Data Management and Use: Governance in the 21st Century* offers a valuable primer on the current issues and how they may be best addressed as individuals continue to generate ever greater volumes of personal data.

4 | HANDLING CONSUMER DATASETS IN SOCIAL SCIENCE RESEARCH

Laney's (2001) core descriptive characteristics that define Big Data, volume, velocity, and variety, can themselves pose particular challenges to geospatial population research with consumer data. Particularly since geography—and the social sciences more broadly—have been slow to adapt to new techniques for analysing large datasets, there remains a shortage of people who can work and engage with the new data landscape (see ESRC/RGS/AHRC, 2013; Ruppert, 2013).

Firstly, the size and scale of some datasets are a noteworthy obstacle. Many are released at the individual level with precise spatial referencing information (such as coordinates). Whilst this has freed research from the constraints of forming representations based on spatial aggregations in many settings, scale is still integral to spatial analysis and visualisation. Sometimes, zooming in too far can make localised patterns undetectable (González-Bailón, 2013). Furthermore, atomistic fallacies arise when incorrect inferences on a population are deduced from individual-level data (Lee et al., 2016). The researcher, therefore, maintains an integral role in producing selective (yet objective) representations to avoid issues of information overload (Zhang et al., 2012). In addition, spatial processes are computationally intensive as they typically need to also control for additional factors such as spatial autocorrelation and spatial nonstationarity so data reduction can be a necessity (Shekhar, Evans, Kang, & Mohan, 2011). With inputs at increasingly granular spatiotemporal scales, data reduction approaches can be more tailored to specific applications—or geographic areas—and remain an essential step in the transformation of raw data into information. In addition, consumer data often need to be aggregated to match traditional data sources in order to estimate their coverage and extrapolate trends. Consequently, data reduction, aggregation, and descriptive techniques are still paramount to understanding some geographic phenomena.

More data on human activities are being released in real time in the form of social media postings or travel card usage in major cities (Batty, 2012). However, it is challenging to keep up with the velocity of these big datasets since most research has developed from analysis on stationary databases. Indeed, our theories about social flows in cities, for example, have considered them as either static or timeless entities (Batty & Cheshire, 2011). Whilst there was urban research on the long-term dynamics of cities, consumer data have enabled researchers to gather data about dynamics at a more granular temporal scale. However, to date, no platforms offered by academics have been able to harness the full potential of real-time data generated by smart cities (Batty, 2017), and techniques for trend detection in real time are still fairly rudimentary.

It is clear that new techniques from computer sciences and statistics have enabled researchers to generate impressive insights (such as those from machine learning and artificial intelligence) from a variety of data types. For example, text mining techniques have empowered geographers to quantify previously intangible social media documents in order to detect trends in activities and opinions across space (Lansley & Longley, 2016). However, there are dangers of simply recycling these methods for social research as they are inherently reductionist and lack sociological theoretical reasoning (Kitchin, 2014b). Consequently, the importance of geocomputation has re-emerged 25 years since its introduction by Stan Openshaw and colleagues, due to the requirement to analyse large and complex geographic data whilst retaining a basic understanding of the principles of spatial analysis (Harris et al., 2017).

Unfortunately, despite the growth in data science outside of academia, there is a skills deficit in geocomputation within social sciences (Rey, 2009). The fundamentals taught in popular textbooks in quantitative analysis have barely changed since the 1990s (Kitchin, 2013). This raises doubts about the future quality of education in data sciences, particularly because students also have very little exposure to commercial big datasets. As demonstrated by the variety of data produced, researchers often need to blend techniques from different disciplines in order to harness information from consumer datasets. It is therefore essential to encourage greater cooperation between academic and commercial sectors to ensure that more research and training can utilise consumer datasets. We also need to promote the development of skills such as geocoding and address matching which are essential for maximising the

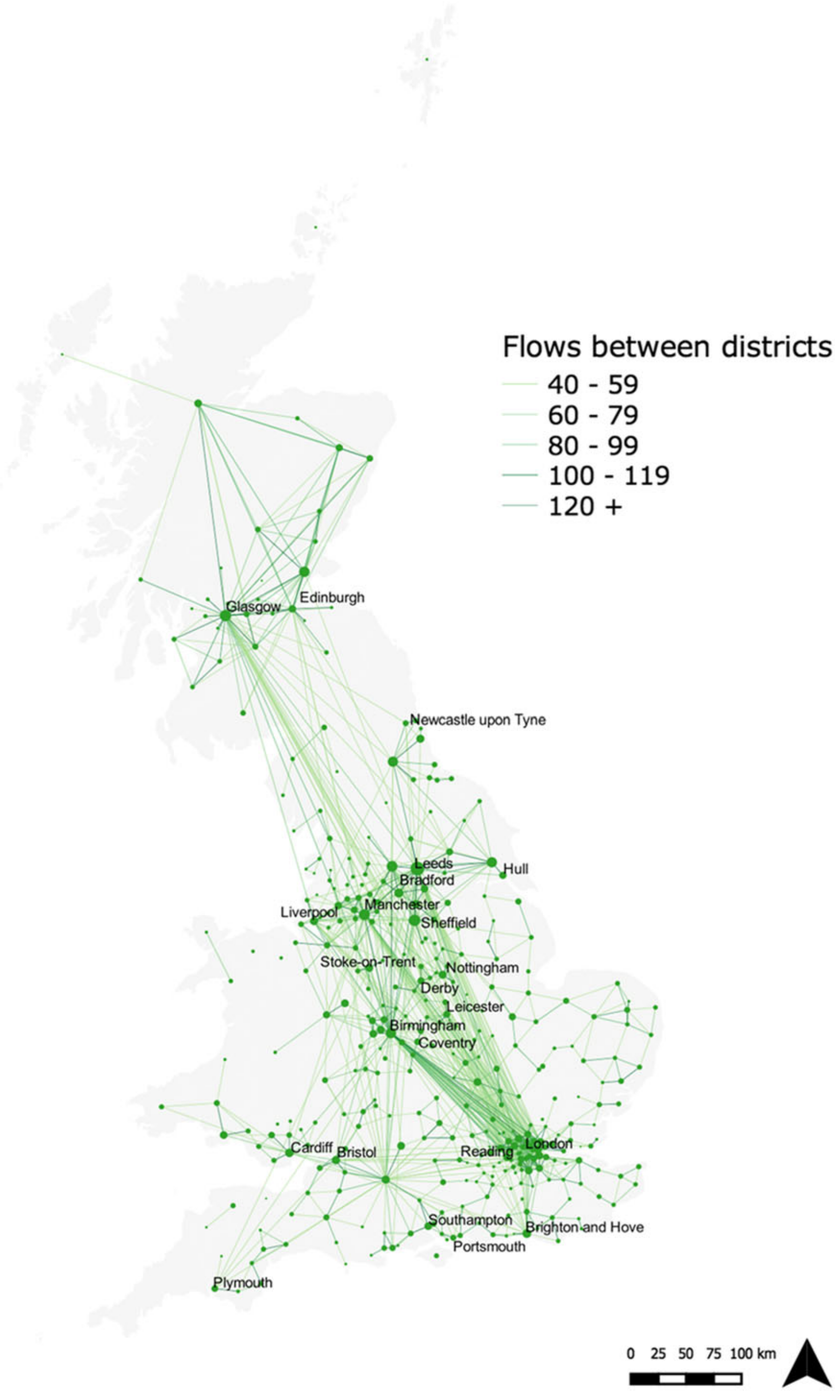

FIGURE 4 The estimated migration flows between local authority districts in Great Britain derived from novel data linkage techniques applied on the 2013 and 2014 consumer registers. Only flows of more than 40 people are shown

value of said data for population research. For example, Figure 4 presents research which sought to estimate migration flows through novel data linkage approaches on two population registers that comprise Great Britain's electoral roll and various additional consumer sources.

5 | CONCLUSIONS

In order to utilise consumer data as indicators of the population at large, we must first consider data as models of real-world phenomena. Every data point is a static digital representation of a particular characteristic or process. Therefore, it is important to understand how the data are structured and what they represent as greater understandings of data can assist hypothesis generation and improve our understanding of uncertainty. For instance, sometimes, data may measure seemingly intangible concepts and it may be possible that if they were recorded differently, entirely different patterns would emerge. We must also remain considerate of scale and how space and time are represented when devising a methodology. Data may be aggregated into units that are not natural and may not align with other records. In addition, the phenomena being studied may be spatially dependent in some sense. Likewise, the data may be released at regular temporal intervals or as a singular snapshot.

We must also be cognisant of how data were collected. Given that most consumer datasets used in social research are by-products of activities, it is important to be realistic about the extent of which the data can be repurposed for valid population research. Where possible, we should contextualise data and account for all possible fallacies that will undoubtedly arise from the data collection procedures. In addition, we should also evaluate the spatial and temporal coverage of such procedures. For instance, social media users may be more likely to post georeferenced content in places and times where social activities are occurring.

With an understanding of how the data are collected, it is then crucial to pay attention to demographic biases. Most consumer datasets contain outliers and are inherently unrepresentative of particular social groups who do not engage in the data collection process. Such biases may be experienced differently at various different scales. Through data linkage and comparisons to official statistics, it is possible to measure data biases to an extent. Whilst it is very difficult to fully account for who may not be present in the data, at the very least, estimates of uncertainty can be produced.

It is clear that new forms of geospatial consumer data can contribute to new and useful representations of the population and their activities. In many settings, they have the potential to reduce our dependence on theories, small samples, and official surveys and reveal otherwise unobservable trends about the population. Whilst this changes the way many researchers can approach hypothesis formulation, the fundamental challenges to positivistic geographic research do not evaporate with increased volumes of data. Deprived of tailored questions, social scientists can risk overstretching these new forms of data beyond the specific populations they tend to pertain to. Without such sensitivity, size can be seen to trump validity to the detriment of the certainty and representativeness of research findings.

ACKNOWLEDGEMENTS

This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1.

ORCID

Guy Lansley http://orcid.org/0000-0002-3406-178X
James Cheshire http://orcid.org/0000-0003-4552-5989

REFERENCES

Amin, A., & Thrift, N. (2002). *Cities: Reimagining the urban*. London: Polity.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*, *16*(7). http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 16 July 2017).

Anselin, L. (1990). What is special about spatial data? Alternative perspectives on spatial data analysis. In D. A. Griffith (Ed.), *Spatial statistics, past, present and future* (pp. 63–77). Ann Arbor: MI.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302.

Batty, M. (2012). Smart cities, big data. *Environment and Planning B: Planning and Design*, *39*, 191–193.

Batty, M. (2013a). Big data, big issues. *Geographical*, *85*(1), 75.

Batty, M. (2013b). *The new science of cities*. Cambridge, MA: MIT Press.

Batty, M. (2017). Producing smart cities. In R. Kitchin, T. Lauriault, & M. Wilson (Eds.), *Understanding spatial media*. London: Sage Publications. In Press

Batty, M., & Cheshire, J. (2011). Cities as flows, cities of flows. *Environment and Planning B*, *38*(2), 195–196.

Berry, B., & Garrison, W. L. (1958). The functional bases of the central place hierarchy. *Economic Geography*, *34*, 145–154.

Christensen, K. (1982). Geography as a human science: A philosophic critique of the positivist-humanist split. In P. R. Gould, & G. Olsson (Eds.), *A search for common ground* (pp. 37–57). London: Pion.

CMA. (2015). The commercial use of consumer data. Report on the CMA's call for information. Online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/435817/The_commercial_use_of_consumer_data.pdf (Accessed: 01.12.17)

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1–12.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139.

Cresswell, T. (2013). *Geographic thought: A critical introduction*. New York: Wiley-Blackwell.

Cresswell, T. (2014). Déjà vu all over again: Spatial science, quantitative revolutions and the culture of numbers. *Dialogues in Human Geography*, *4*(1), 54–58.

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial "big data" and geodemographics. *Big Data & Society*, *2*(2), 1–15.

Duckham, M., & Kulik, L. (2006). Location privacy and location-aware computing. In J. Drummond, R. Billen, D. Forrest, & E. João (Eds.), *Dynamic and mobile GIS: Investigating change in space and time* (Vol. 3) (pp. 35–51). Boca Raton: CRC Press.

Dugmore, K. (2010). *Information collected by Commercial Companies: What might be of value to Official Statistics? The case of the UK Office for National Statistics*. London: Demographic Decisions Ltd.

ESRC/RGS/AHRC (2013). *International benchmarking review of UK human geography*. London, UK: ESRC/RGS/AHRC.

Folch, D. C., Spielman, S. E., & Manduca, R. (2017). Fast food data: Where user-generated content works and where it does not. *Geographical Analysis* https://doi.org/10.1111/gean.12149, *50*, 125–140.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: Wiley.

Gayo-Avello, D. (2012). "I wanted to predict elections with Twitter and all I got was this lousy paper"—A balanced survey on election prediction using Twitter data. arXiv Preprint arXiv:1204.6441.

González-Bailón, S. (2013). 'Big data' and the capillaries of human geography. *Dialogues in Human Geography*, *3*(3), 292–296.

Gould, P. (1979). Geography 1957–1977: The Augean period. *Annals of the Association of American Geographers*, *69*(1), 139–151.

Graham, M., & Foster, C. (2014). Geographies of information inequality in sub-Saharan Africa. Oxford Internet Institute, University of Oxford, U.K. http://cii.oii.ox.ac.uk/geographies-of-information-inequality-in-sub-saharan-africa/

Gregory, D. (1978). *Ideology, science and human geography*. London: Hutchinson.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Harris, R., O'Sullivan, D., Gahegan, M., Charlton, M., Comber, L., Longley, P., … Evans, A. (2017). More bark than bytes? Reflections on 21+ years of geocomputation. *Environment and Planning B: Urban Analytics and City Science*, *44*(4), 598–617.

Harvey, D. (1973). *Social justice and the city*. London: Edward Arnold.

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond: Microsoft Research.

Huff, D. L. (1964). Defining and estimating a trading area. *The Journal of Marketing*, *28*, 34–38.

Johnston, R. (2008). Geography and the social science tradition. In S. L. Holloway, S. P. Price, & G. Valentine (Eds.), *Key concepts in geography* (2nd ed.) (pp. 46–65). London: Sage.

Johnston, R., Harris, R., Jones, K., Manley, D., Sabel, C., & Wang, W. W. (2014). Mutual misunderstanding and avoidance, misrepresentations and disciplinary politics: Spatial science and quantitative analysis in (United Kingdom) geographical curricula. *Dialogues in Human Geography*, *4*(1), 3–25.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267.

Kitchin, R. (2014a). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12.

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Kitchin, R. (2015). Positivistic geography. In S. Aitken, & G. Valentine (Eds.), *Approaches in human geography*. London: Sage.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. Online: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (Accessed 26/06/2017).

Lansley, G., & Longley, P. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Lee, E. C., Asher, J. M., Goldlust, S., Kraemer, J. D., Lawson, A. B., & Bansal, S. (2016). Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. *The Journal of Infectious Diseases*, *214*(4), S409–S413.

Lee, J. G., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research*, *2*(2), 74–81.

Lloyd, A., & Cheshire, J. (2018). Detecting uncertainty in loyalty card data. *Applied Spatial Analysis and Policy*. https://doi.org/10.1007/s12061-018-9250-1

Longley, P. A., Adnan, M., & Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A*, *47*(2), 465–484.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems*. Hoboken, NJ: John Wiley & Sons.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461.

OECD (2013). *New data for understanding the human condition: International perspectives*. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. Paris: OECD Publishing.

Openshaw, S. (1991). A view on the GIS crisis in geography, or, using GIS to put Humpty-Dumpty back together again. *Environment and Planning a*, *23*(5), 621–628.

Pickles, J. (Ed.) (1995). *Ground truth: The social implications of geographic information systems*. New York: Guilford Press.

Ralphs, M., & Tutton, P. (2011). Beyond 2011: International models for census taking: Current processes and future developments. Beyond 2011 Project, Office for National Statistics. Online: http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes%E2%80%94projects/beyond-2011/news/reports-andpublications/early-reports-and-research-papers/international-models-for-census-taking.pdf (Accessed 01/04/17)

Reades, J., Zhong, C., Manley, E. D., Milton, R., & Batty, M. (2016). Finding pearls in London's oysters. *Built Environment*, *42*(3), 365–381.

Rey, S. J. (2009). Show me the code: Spatial analysis and open source. *Journal of Geographical Systems*, *11*, 191–207.

Royal Society. (2017). Data management and use: Governance in the 21st century—A British Academy and Royal Society project. Source: https://royalsociety.org/topics-policy/projects/data-governance/ (accessed August 2017).

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography*, *3*(3), 268–273.

Shekhar, S., Evans, M. R., Kang, J. M., & Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 193–214.

Spero, I., & Stone, M. (2004). Agents of change: How young consumers are changing the world of marketing. *Qualitative Market Research: An International Journal*, *7*(2), 153–159.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, *52*(1), 125–133.

Sui, D., Goodchild, M., & Elwood, S. (2013). Volunteered geographic information, the exaflood, and the growing digital divide. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice* (pp. 1–12). Berlin: Springer.

The Electoral Commission (2016). The December 2015 electoral registers in Great Britain, accuracy and completeness of the registers in Great Britain and the transition to Individual Electoral Registration. The Electoral Commission Report, July 2016.

Trewartha, G. T. (1953). A case for population geography. *Annals of the Association of American Geographers*, *43*(2), 71–97.

Tuan, Y.-F. (1976). Humanistic geography. *Annals of the Association of American Geographers*, *66*, 266–276.

Wilson, M. W. (2015). Morgan Freeman is dead and other big data stories. *Cultural Geographies*, *22*(2), 345–349.

Wilson, M. W. (2017). *New lines: Critical GIS and the trouble of the map*. University of Minnesota Press.

Wright, C., & Sparks, L. (1999). Loyalty saturation in retailing: Exploring the end of retail loyalty cards. *International Journal of Retail and Distribution Management*, *27*(10), 429–439.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., & Keim, D. (2012). Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*: 173–182

Guy Lansley is a Research Associate at the Consumer Data Research Centre and the Department of Geography, University College London (UCL).

James Cheshire is a Senior Lecturer in Quantitative Geography at UCL and Deputy Director of the Consumer Data Research Centre.

How to cite this article: Lansley G, Cheshire J. Challenges to representing the population from new forms of consumer data. *Geography Compass*. 2018;e12374. https://doi.org/10.1111/gec3.12374