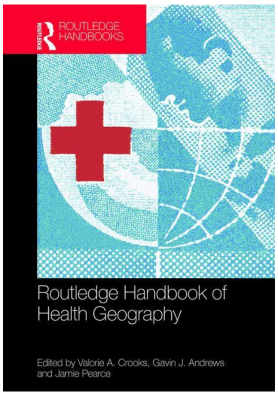

Routledge Handbook of Health Geography

Valorie A. Crooks, Gavin J. Andrews, Jamie Pearce

Health geography and the big data revolution

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315104584-46
Alec Davies, Mark A. Green
Published online on: 11 Jun 2018

46

HEALTH GEOGRAPHY AND THE BIG DATA REVOLUTION

Alec Davies and Mark A. Green

The era of big data is upon us. We have more data than ever before, greater computing power to process and analyze such data, and access to new forms of data opening up novel research avenues. The promise offered through big data has brought great potential for researchers to answer many long-standing questions and provide more accurate solutions to problems (Boyd and Crawford, 2012), although significant challenges remain (Khoury and Ioannidis, 2014). In this chapter, we outline the varying definitions of "big data," the applications across health and health geography and the challenges and opportunities associated with using big data.

Defining "big data"

The definition of "big data" is not universally agreed upon (Herland, Khoshgoftaar and Wald, 2014). "Big data" traditionally applies to data of size beyond the capabilities of commonly used software within a tolerable time frame, with size constantly increasing with computing capacity (Manovich, 2011). While size inherently matters (for example, data contained within the US health-care system exceeded 150 exabytes [or 150 billion gigabytes] in 2010, Andreu-Perez et al., 2015), a simpler means for thinking about what is considered as *big* might be data that cannot be opened, stored and processed in the memory of a desktop or laptop computer.

Kitchin (2014) identifies the three Vs of big data research. *Volume*, referring to the size of data, is often over-emphasized in big data (Birkin, Clarke and Clarke, 2017). Data size is making way to the capacity to search, aggregate and cross-reference large datasets as the defining feature of big data (Boyd and Crawford, 2012). *Velocity* refers to data that is generated (and received) quickly, often in or near to real-time. For example, Twitter data is generated frequently as individuals tweet throughout their days to provide a continuous stream of data. The high velocity of data received creates practical issues in the processing and analyzing of such data. Data are often received in a *variety* of heterogeneous types and forms, and often completely unstructured, thereby placing emphasis on manipulating such data (Andreu-Perez et al., 2015). *Variability* has also been touted as an additional *V*, in that what data is received may change in type or quality over time due to updated processes or practice.

Boyd and Crawford (2012) define "big data" as an interplay of technology, analysis and mythology. Technology enables computation, analysis allows interpretation and mythology is the belief that big data offer a higher form of intelligence. Such a holistic definition helps us move away from a technical definition to one that is more akin to the philosophy of *doing* big data research. In this vein, Andreu-Perez et al. (2015) recommend including *value* in any definition. There are many smaller datasets available that can help answer most research questions. Big data should complement or extend these analyses.

Alongside the growth in popularity of big data, we have also seen increasing usage of an interrelated phrase – *new forms of data*. The phrase refers to the increasing availability of data sources not traditionally used for research (e.g., internet search data, social media, loyalty-card records, mobile devices or wearable technology). Connelly et al. (2016) distinguish between *made* (collected for a specific research purpose) and *found* (which may be valuable for researchers but was collected for a non-research purpose – alternatively titled *repurposed*) data. They are *new* in the sense that these data have only recently started becoming available to researchers. For example, social-media organizations (e.g., Facebook, Twitter) did not exist 10–15 years ago, yet the data that they generate have increasingly been used by researchers to understand health-related behaviors (Khoury and Ioannidis, 2014). The *internet of things* (the interconnectedness of everyday objects through the inclusion of internet-based technologies) and *smart cities* (the interconnectedness of communication and technology used to manage a city) offer exciting directions for new forms of data to better understand urban dynamics.

These paradigms of *big data* and *new forms of data* have been enabled through the *datafication* of society. Data are generated in all aspects of our daily routines, whether it is purchasing food using our loyalty and credit cards at a supermarket, visiting a doctor or using public transport. Many of these data were being generated in the past at similar levels that would be considered big data today. The difference is that in the past many of these data were simply not being stored or analyzed. Costs and difficulties of creating, storing and analyzing data have caused limited scope and temporality in large datasets. As such, we have seen national datasets such as census compilation being confined to coarse temporal (e.g., every 10 years) and spatial (through aggregation to small geographical zones) resolutions (Kitchin, 2014). The continual improvements in computing power, alongside the development of cloud computing, have rendered these issues obsolete. The production of data through many actions of everyday life (and subsequent ability to exploit such data) have seen big data termed the *new oil* (Andreu-Perez et al., 2015).

Big data and health

If big data was around in 1854 when cholera swept through London, John Snow could have used Global Positioning System (GPS) data, mortality records and environmental data to identify the disease vectors and implement a solution within hours (Drexler, 2014; Khoury and Ioannidis, 2014).

Big data stand to improve health by providing insights into the causes and outcomes of disease, better drug targets for precision and personalized medicine, and enhanced disease prediction and prevention (Khoury and Ioannidis, 2014; Raghupathi and Raghupathi, 2014). Although traditional health-data sources are rarely the largest in terms of size, the potential offered through data linkage to newer forms of data (e.g., linking survey data to purchasing behaviors through loyalty-card records) could offer novel and exciting insights into health-related behaviors (Herland, Khoshgoftaar and Wald, 2014).

Evidence-based medicine has long adopted a practice of generating evidence from trials and experiments, using data to inform decisions. The evidence-based movement was founded on the belief that scientific inquiry is superior to expert opinion and testimonials, meaning the industry was ahead of many others in recognizing the value of data-guided decision-making (Murdoch and Detsky, 2013). Despite this long-standing support for evidence-informed decisions, health-care delivery has been slow to utilize the rich information routinely collected within its own data frameworks (Safran et al., 2007). There are significant opportunities for health-care researchers and policy-makers to learn from other data-driven revolutions that have been successful at applying large datasets (Kayyali, Knott and Van Kuiken, 2013; Weber and Kohane, 2014), such as astronomic telescopic information, retail-sales data or internet search engine data (Murdoch and Detsky, 2013).

That being said, big data have been used for some facets of health science. The management of pandemics such an influenza has seen opportunities for heterogeneous information from managed and unmanaged sources processed, mined and turned into decision actions to control outbreaks (Andreu-Perez et al., 2015). An example of such application is Google Flu Trends, which predicts outbreaks of flu from flu-related internet searches (Pentland, Reid and Heibeck, 2013; Herland, Khoshgoftaar and Wald, 2014). Similar studies have used Twitter data to estimate the prevalence of flu (Lampos, De Bie and Cristianini, 2010), as well as the likelihood of whence an individual developed a case of food poisoning (Sadilek et al., 2013). Cell-phone data were recently used in West Africa to track travel patterns and predict where Ebola might spread to after an outbreak in a town or village, allowing public-health officials to set up early preventative barriers for containing the spread of the disease (Wesolowski et al., 2014). Sensor information has accelerated in usage in recent years, bringing forward the notion of the internet of things, with wearables being at the forefront (Andreu-Perez et al., 2015). Such wearable technologies offer opportunities for improved prediction of health outcomes and are being combined with interventions to evaluate their success or monitor progress.

It is not the case, though, that big data will improve health applications overnight. In 2013, Google Flu Trends overestimated doctor visits for influenza-associated conditions by twice what validated data suggested it should be (Butler, 2013). In the past, the same service has also underestimated the H1N1 pandemic due to how people were searching for the condition (Cook et al., 2011). While these issues are partly due to the novelty of their algorithms (in Google Flu Trends' case, due to overfitting of data), issues that can be overcome, Lazer et al. (2014) acknowledge that *big data hubris* has seen many individuals overestimate the potential of big data. Big data should only supplement, rather than replace, traditional approaches for data collection and analysis. For these reasons, it is important to stress the *value* part of any big data application.

Applications in health geography

While the possibilities are wide-ranging for the applications of big data in health geography, we focus on a few examples to illustrate feasible opportunities within the field.

Developing new metrics

Much of the hype surrounding big data has been about the perceived endless analytical applications, but many of the actual applications have been somewhat basic in statistical and analytical terms. The most common application has been the aggregation and development of new indicators to aid the measurement of health phenomena. The opening up of big data (particularly new forms of data) creates opportunities for health geographers to better understand the contribution of environmental features on health. One example can be seen in Daras et al. (2017), who use consumer data on the location of retail outlets to calculate network distances for all postcodes (and also aggregated to small geographical zones) in Great Britain to their nearest amenities such as fast-food outlets, pubs or off-licenses. Developing national-level metrics from traditional and non-traditional sources and making these openly available (data are available here: http://maps.cdrc.ac.uk/) helps researchers and policy-makers better understand the role of neighborhood features on health and health-related behaviors without having to undertake major data manipulation.

Mapping services, such as Google Street View, have also been mined as a source of information about neighborhoods, such as aesthetics, location of food outlets and land use (Bethlehem et al., 2014). Fully automating the process has proved difficult but represents a useful research avenue, particularly with new developments in image processing and machine learning. Google has also incorporated air-quality measures onto their street-view cars to collect data on levels of nitric oxide, nitrogen dioxide and black carbon (Tuxen-Bethman, 2017), and the data can be requested freely from Google for research purposes. There is also growing interest in using remote-sensing data to develop desk-based audits of features of the physical and built environment (Charreire et al., 2016).

The increasing use of social media across the population has also offered a novel source of information on human relationships and social interactions. Twitter data is the most commonly used social-media data source, as unlike other platforms it makes a portion of its data openly available to researchers. The complexity of processing such data has been a limitation. Strategies employed have ranged from taking random samples of tweets, to selecting phrases, keywords or hashtags (e.g., Eichstaedt et al. 2015), to using machine learning to classify the content of posts (e.g., Nguyen et al., 2016). Many of these approaches can be combined with the geotagged information provided in tweets to explicitly incorporate the spatial dimension. For example, Nguyen et al. (2016) used machine-learning techniques to classify whether individuals were tweeting about fast food or high-calorie/energy-dense foods. Through linking these data to zip codes, they found that where there were higher rates of such tweets, there was a positive correlation to the number of fast-food outlets and the state-level prevalence of obesity. Similar studies have been undertaken using Twitter to measure geographical patterns in happiness (Gore, Diallo and Padilla, 2015), physical activity (Nguyen et al., 2016), diet (Nguyen et al., 2016; Widener and Li, 2014) and psychological distress (Eichstaedt et al. 2015).

Data linkage

One of the early promises of big data was the ability to connect multiple sources of information and combine their separate sources to create a comprehensive data resource. While there has been good progress in linking administrative sources (e.g., health records), there have been few examples using commercial data sources. This is partly a result of the difficulty of linking several datasets with no common identifier between them, but it is also due to the trepidation of data custodians who are wary of sharing or linking data to consumer sources that may reveal insights to competitors. Many organizations are understandably cautious, although initiatives such as the Consumer Data Research Centre (CDRC) have started a process of brokering conversations between organizations and researchers to help facilitate collaborations. Ethical issues around security, data ownership and privacy are important and require careful consideration before progress will be made. As boyd and Crawford (2012) argue, there is a fine line between accessibility and ethics. Some of these issues are being alleviated through examples of good practice, but progress remains slow.

One data source that has received much attention for data linkage has been the potential offered by loyalty-card records. Consumer loyalty cards were originally introduced as a marketing tool to provide customers with incentives to remain brand loyal (Mauri, 2003; Sharp and Sharp, 1997). Many organizations, particularly supermarkets, soon realized their potential for understanding consumer behaviors, helping them tailor experiences (e.g., through person-specific price coupons for items) and deciding which products to stock in stores. While they are an imperfect data source, since not all individuals have loyalty cards for the places where they shop (or some individuals have multiple cards), they provide objective data on behaviors that is important given the under-reporting of dietary behaviors common in traditional self-reported sources of data (Flegal, 1999).

A recent example demonstrating the potential of supermarket data can be seen in a study by Silver et al. (2017), who used point-of-sale data on purchasing behaviors in supermarkets to evaluate the impact of the introduction of a tax on sugar-sweetened beverages in the American city of Berkeley. Several countries and local governments have proposed introducing a similar tax as an approach to promote healthier diets (and therefore reduce the risk of developing conditions such as type 2 diabetes). The quality of evidence as to whether such a policy would be effective is limited and often relies on modeled estimates. Silver and colleagues' results suggested that one year after the introduction of the tax, sales of sugar-sweetened beverages declined by 9.6%, compared to an increase of 6.9% in supermarkets where the tax was not introduced. These findings were novel and important in understanding how successful a similar tax could be elsewhere. While the data were not loyalty-card records (loyalty-card data would have strengthened the quality of the study through providing repeated-measures data allowing for a better understanding of changes in behaviors post-intervention), they still demonstrate the usefulness and potential of consumer data.

Challenges of big data

While big data offer many possibilities for health geography, there are several important limitations. Ethics is important, and we have touched on security issues previously from the point of researchers and data custodians, but privacy concerns also matter to the public. The public may not have signed up to take part in external research knowingly, nor agreed for their data to be linked to other sources.

Datasets are considerably bigger, but not necessarily better. Moving away from thinking about the *big* in big data and incorporating boyd and Crawford's philosophy is more appropriate. We should not view a difference between *big* and *small* data; they are both data. Data should be judged on what they can add to a research question, as well as the quality that they offer (more eloquently put by the equation *garbage in equals garbage out*). Rather than a *big data revolution*, we should view future research within the *all-data revolution* paradigm whereby the importance is placed on innovative analytics to better understand the world (Lazer et al., 2014). In this regard, viewing the field as (geographic) data science seems a better fit, since it avoids many of these issues while preserving the approaches toward big data analytics that are important and sought after.

We need to continue placing emphasis on understanding what a sample offers, rather than using it purely because it ticks the *big data* box (a large amount of the studies using Twitter data fall into this category, with their primary interest in using *novel* data rather than necessarily improving on past approaches). This is important given that data are often not collected for research purposes, and, therefore, using repurposed data requires an understanding of the context and meaning of any data (boyd and Crawford, 2012) – which can be difficult to ascertain, as metadata is typically lacking. Classic research is hypothesis-driven, and data collection is therefore focused to answering a specific question, improving its *value*. Repurposed data move us away from this, and so data should not be taken at face value alone.

The issue of value in large datasets is also statistical: we need to understand how representative any data are. For example, Twitter data are inherently biased – Twitter is popular with younger populations, and only a small proportion (<1%) of tweets contain any geotagged information (and indeed those who geotag tweets are demographically different from those who do not: Sloan and Morgan, 2015). Understanding the appropriateness and validity of any data source is simply good statistical practice (hence our prior argument about the importance of discussing big data within a data-science or all-data context). Accuracy and precision are very different statistical issues, and the greater precision offered through larger sample sizes will not make up for a lack of accuracy.

Big data suffers gigantic noise, and big error can plague big data, where spurious correlations and ecological fallacies can multiply and mislead findings (Khoury and Ioannidis, 2014). Large sample sizes can distort many common statistical techniques and require us to approach analyses using newer methods (e.g., machine learning). This is one of the reasons why many studies opt for simple aggregation of data, since this is usually where the main patterns can be observed. Removing noise from the vast amounts of data also needs to be tackled for big data to be translated into gains in societal well-being (Khoury and Ioannidis, 2014).

Many big data analytics focus on leveraging large sample sizes to make better predictions; however, often in health geography we are more interested in causal inference. Health geographers have an important role to play here in ensuring that correct study design is applied to improve our (causal) understanding of the world. There are important and exciting opportunities available for natural experiments involving such data. For instance, one could envisage analyzing how changes in the environment are associated with changes in purchasing behavior. Although big data offer endless possibilities, only through carefully designed studies can we understand how best to utilize such data.

Conclusion

There has been a lot of interest around the opportunities big data pose, both across and within many disciplines. While many challenges remain, there is still a lot of promise, and we should not forget that such systems and technology are still fairly new. Indeed, many companies are only just realizing themselves what potential their data offers for improving their own insights! The opportunities opened up through the big-data revolution within health geography are due not only to new data avenues, but also to the analytical and modeling possibilities of larger datasets. Many of these possibilities have yet to be realized, as typical applications involve simple aggregations of data using non-complex methods. This should not be viewed as a criticism, though. Results may be initially unclear in large data, and only through the expertise of careful interpretation will we discover valuable insights (Van Dijck, 2014). As such, health geographers should construct their studies carefully to ensure that future developments are effective at helping us better understand the world.

References

Andreu-Perez, J., Poon, C., Merrifiedl, R., Wong, S.T. and Yang, G. Z. (2015). Big data for health. *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp. 1193–1208.
Bethlehem, J., Mackenbach, J., Ben-Rebah, M., Compernolle, S. Glonti, K., Bárdos, H., Rutter, H. R., Charreire, H., Oppert, J.-M., Brug, J. Lakerveld, J. (2014). The SPOTLIGHT virtual audit tool: a valid and reliable tool to assess obesogenic characteristics of the built environment. *International Journal of Health Geographics*, 13(52), pp. 2–8.
Birkin, M., Clarke, G. and Clarke, M. (2017). *Retail location planning in an era of multi-channel growth*. London: Routledge.
boyd, d. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5), pp. 662–679.
Butler, D. (2013). When Google got flu wrong. *Nature*, 494, pp. 155–156.
Charreire, H., Mackenbach, J., Ouasti, M., et al. (2016). Using remote sensing to define environmental characteristics related to physical activity and dietary behaviours: a systematic review (the SPOTLIGHT project). *Health & Place*, 25, pp. 1–9.
Connelly, R., Playford, C., Gayle, V. and Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, pp. 1–12.
Cook, S., Conrad, C., Fowlkes, A. and Mohebbi, M. (2011). Assessing Google Flu Trends performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS One*, 6(8), pp. 1–8.
Daras, K., Davies, A., Green, M. and Singleton, A. (2017). Developing indicators for measuring health-related features of neighbourhoods. In: P. Longley, J. Cheshire and A. Singleton, eds., *Consumer data analytics*. London: UCL Press.
Drexler, M. (2014). Big data's big visionary. [online] *Harvard Public Health Magazine*. Available at: www.hsph.harvard.edu/magazine/magazine_article/big-datas-big-visionary/. [Accessed 20 June 2017].
Eichstaedt, J., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H. and Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2), pp. 159–169.
Flegal, K. (1999). Evaluating epidemiologic evidence of the effects of food and nutrient exposures. *American Journal of Clinical Nutrition*, 69, pp. 1339–1344.
Gore, R. J., Diallo, S. and Padilla, J. (2015). You are what you tweet: connecting the geographic variation in America's obesity rate to Twitter content. *PloS One*, 10(9), pp. 1–16.
Herland, M., Khoshgoftaar, T. and Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, 1(1), pp. 1–35.
Kayyali, B., Knott, D. and Van Kuiken, S. (2013). The big-data revolution in US health care: accelerating value and innovation. *McKinsey & Company*, pp. 1–13.
Khoury, M. and Ioannidis, J. (2014). Big data meets public health. *Science*, 346(6213), pp. 1054–1055.
Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), pp. 1–12.
Lampos, V., De Bie, T. and Cristianini, N. (2010). Flu detector – tracking epidemics on Twitter. In: J. L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, eds., *Machine learning and knowledge discovery in databases*. ECML PKDD 2010. Lecture Notes in Computer Science, 6323. Berlin, Heidelberg: Springer, pp. 599–602.
Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176), pp. 1203–1205.
Manovich, L. (2011). Trending: the promises and the challenges of big social data. *Debates in the Digital Humanities*, 2, pp. 460–475.
Mauri, C. (2003). Card loyalty. A new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services*, 10(1), pp. 13–25.
Murdoch, T. and Detsky, A. (2013). The inevitable application of big data to health care. *Journal of the American Medical Association*, 309(13), pp. 1351–1352.

Nguyen, Q., Li, D., Meng, H., Nsoesie, E., Li, F. and Wen, M. (2016). Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR Public Health and Surveillance*, 2(2), pp. 1–16.

Pentland, A., Reid, T. and Heibeck, T. (2013). *Big data and health: revolutionizing medicine and public health*. Report for the Big Data and Health Working Group. N.p.: World Innovation Summit for Health (WISH).

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1), pp. 1–10.

Sadilek, A., Brennan, S., Kautz, H. and Silenzio, V. (2013). nEmesis: which restaurants should you avoid today? *AAAI Publications, First AAAI Conference on Human Computation and Crowdsourcing*, pp. 138–146.

Safran, C., Bloomrosen, M., Hammond, W. E., et al. (2007). Toward a national framework for the secondary use of health data: an American medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1), pp. 1–9.

Sharp, B. and Sharp, A. (1997). Loyalty programs and their impact on repeat-purchase loyalty patterns. *International Journal of Research in Marketing*, 14(5), pp. 473–486.

Silver, L. D., Ng, S., Ryan-Ibarra, S., et al. (2017). Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: a before-and-after study. *PLoS Medicine*, 14(4), pp. 1–19.

Sloan, L. and Morgan, J. (2015). Who Tweets with their location? Understanding the relationship between demographic characteristics and the use of Geoservices and Geotagging on Twitter. *PLoS One*, 10(11), pp. 1–15.

Tuxen-Bethman, K. (2017). *Let's clear the air: mapping our environment for our health*. [online] The Keyword. Available at: https://blog.google/products/maps/lets-clear-air-mapping-our-environment-our-health/. [Accessed 20 June 2017].

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), pp. 197–208.

Weber, G. and Kohane, I. (2014). Finding the missing link for big biomedical data. *Journal of the American Medical Association*, 311(24), pp. 2479–2480.

Wesolowski, A., Buckee, C., Bengtsson, L., et al. (2014). Commentary: containing the ebola outbreak – the potential and challenge of mobile network data. *PLoS Currents Outbreaks*, 29(1), pp. 1–20.

Widener, M. and Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, 54, pp. 189–197.