Data analytics: on the cusp of using new sources?

Paul Norman, Alan Marshall & Nik Lomax

Abstract

Many data analytics settings are moving from conventional formal data sources like censuses and large scale social surveys via the increasing use of administrative data, towards using less formal 'Big Data' sources. As society changes, researchers seek to combine sources of information to capture social, economic and demographic processes in response to new ways of living. Big Data has the potential to clarify our understanding of the social world yet may confuse if not used with a clear understanding of its limitations with a focus on the ways it can complement rather than supplant existing data.

Introduction

In many data analytics settings there are ambitions and pressures to move away from using more conventional formal data sources like censuses and large scale planned social surveys via the increasing use of administrative data (e.g. benefits) and towards the use of less formal 'Big Data' sources. Big Data has emerged as a major theme for social science researchers over the past decade. For example, the UK's Economic and Social Research Council has committed major resources to a 'Big Data network' aimed at exploring the links between administrative data, information held by business and Government and data from the Third Sector and social media.

In this paper we point to some of the characteristics and the strengths and weaknesses of different types of data. We welcome the exciting opportunities provided by new sources of information under the umbrella term Big Data although we argue that more traditional sources remain essential for new sources to be compared with to check whether a 'reliable' picture is emerging about the aspect of society we are investigating. As society changes researchers must seek to combine sources of information to reliably capture social, economic and demographic processes in response to new ways of living and new data issues. In this context Big Data, like other sources, has the potential to clarify our understanding of the social world around us yet may confuse if not used with a clear understanding of its limitations with a focus on the ways it can complement rather than supplant existing data.

For good governance and informed policies and for the provision of public and private / commercial goods and services, the research community (e.g. government officers, academics, consultants, journalists) and the general public need data on how many people there are (by age, sex and demographic components) along with data on the population characteristics (their attributes such as qualifications, occupation, ethnic group, marital status, health) and their living arrangements (household group, dwelling types and ownership). We refer to these collectively as 'socio-demographic' data.

Such data might be used to tell us how health inequalities are changing across generations, the extent to which education determines later outcomes or the extent of discrimination faced by ethnic minorities over time, generation and place. Without data, answers to such questions become reliant on anecdotal evidence which may lead to inaccurate conclusions. The 'Perceptions are not reality project' (Ipsos MORI, 2014) demonstrates that the public overestimate the proportion of teenagers who become pregnant, who are elderly, who are immigrants or from particular ethnic minority groups. Quantitative data collected in a systematic way are essential to inform the public and policymakers on specific issues.

Subsections below will appraise the advantages and limitations of three styles of data sources differentiated loosely by date of development and their coverage of the population. First, there are the national censuses. The aim of a census is to capture information about everybody living in a defined geographic area (usually a country). Second, there are large scale social surveys which are planned to capture information about a sufficient number of people, who are also sufficiently representative of the general population, so that approximate conclusions can be drawn about the general population. Third, there is administrative data. This information derives from organisations (often government departments) whose day to day activities involve formal interactions with the general public and a realisation that the public good may be served by making the data available for third party usage in research. Finally, we consider the transition being made from more conventional to Big Data sources.

The types of data sources discussed below are those which are readily available to most researchers most of the time. Sources where access is restricted for the general research community due to financial costs or confidentiality constraints requiring access behind a firewall are not detailed here.

Census data

Censuses have been taken since ancient times with most countries in the contemporary era conducting a census (Holdsworth et al., 2013). The aim of a census is to collect data on 100% of the population across a whole nation at the same point in time and this process is usually repeated at regular intervals (every five or ten years). The topics and questions may be repeated from one census to the next so that change over time can be determined but the responsible organisation will consult users before the next census to obtain views on potential questions. Respondents usually complete a 'tick box' questionnaire with the questions aiming to elicit factual information.

The dissemination of census outputs enables the analysis of the size and key characteristics of the population for a range of geographies from national and down to local levels. Despite the strengths, there are issues with census data. The periodic nature of census taking and the time taken to release outputs means that data are out of date by the time of release. Respondents make mistakes when answering questions and some people do not fill out a census form. The impact of 'underenumeration' (for example, due to the uncounted emigration of young adult males or protesters to the 'poll tax') can be alleviated (Norman et al., 2008) and where outputs do not have the socio-demographic detail needed for a purpose, this can be estimated (Lomax & Norman, 2016)

In many countries, the future of census taking is being questioned with arguments linked to spiralling costs and to topic content with the concept of a snapshot of the population at one time point becoming less relevant, especially in the face of increasingly mobile populations (Dugmore et al., 2011; Yacyshyn & Swanson, 2011). Alternative approaches to census design are presented by the United Nations (UN, 2016). When the UK Census came under threat, a campaign was launched to ensure that users demonstrated the value of having small area statistics and for people to assess what would *not* be possible without a census (Elias et al., 2013). The census is the source which informs the delineation of small area geographies used for the dissemination of other data sources, some of which are noted below. Without a census, defining small area geographies will be challenging (Norman, 2013).

Survey data

We refer here to large scale surveys, often undertaken by government departments. Data collection is usually motivated around the need to provide information about a particular topic including health, labour force participation, crime, elections, social attitudes, personal expenditure, etc. The framework of the study may be cross-sectional whereby data are collected over a few months by professional interviewers and released, often on an annual basis. The same questions may be asked in successive years, but different people will be interviewed. Survey development and data collection is a rigorous process with, for example, focus groups informing question development to ensure validity and careful training of interviewers to avoid bias in the collection of data. Institutional and other non-household populations tend to be outside the coverage of these surveys. Carr-Hill (2015) usefully reviews population sub-groups about whom it is hard to access data.

An advantage of these surveys compared with census data is that a much greater variety of questions can be asked and in a timely manner to fit with demand. Survey data collection is not necessarily independent of the census since sampling strategies often use census geographies to ensure sufficient coverage of population types. Similarly, the weighting of general population surveys to ensure representativeness is invariably calculated using census data (Crockett et al., 2011; Dawes et al., 2014). Where a survey is of population sub-groups not captured adequately in the census then comparisons or weightings cannot be achieved. Disadvantages of survey data include the increasing challenges of non-response (Buckley & King-Hele, 2014) and, like the census, large costs of data collection and processing and threats of discontinuation if funding was withdrawn (Pearson, 2016).

Administrative data

Administrative data refers to information formally collected primarily for an organisation's activities. This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service. Government departments are the main (although not exclusive) collectors of large administrative databases, including welfare, tax, health and educational record systems. These datasets have for many years been used to produce official statistics to inform policy-making. When disseminated for use by others, administrative data are invariably released using census / electoral geographies.

Administrative datasets are typically very large, covering counts or samples of individuals and time periods (at least annually) which are not normally financially or logistically achievable through traditional census and survey methods. Alongside cost savings, the wide ranging scope and national coverages of administrative data are main advantages for research purposes. Other advantages include: relieving the burden on census or survey respondents and providing data on individuals who would not normally respond to surveys e.g. welfare / benefits. Census and survey non-response occurs because people are not willing to fill out a questionnaire and other people may also be hard to reach (i.e. they don't get a questionnaire delivered) because they are in temporary accommodation, etc. This may be important because if you are trying to identify areas in need of investment to address unemployment, if the unemployed are less likely to respond to a census / survey then measures will be biased. There is an incentive (financial!) for the unemployed to get their benefit so they do fill out their claim forms.

Criticisms levelled at administrative resources relate to the lack of control the researcher has on data collection. This can affect what you can do with the data in terms of data types, variable definition and geography. To some extent this is true of other secondary sources (but as noted above, there are consultations on census content) though with administrative sources there is more risk of a lack of utility. Administrative data do not count certain groups who do not use particular services and are subject to inaccuracy for analysis over time when rules around eligibility or the social acceptability of using particular services change. Used carefully though, time-series analysis is possible (Norman & Bambra, 2007).

With the future of censuses coming under pressure, the use of administrative data to provide socio-demographic information is being encouraged (ONS, 2014). Good progress has been made to demonstrate that administrative sources have utility to aid in counting the population (Harper & Mayhew, 2012; UN, 2016) and in using administrative sources to investigate the relationship between health and deprivation (Ajebon & Norman, 2006). Combining administrative data and data from social surveys and censuses can provide excellent insights (Marshall et al., 2013).

Although the purposes of administrative data collection was not for general research, the formality of the processes and that organisations are largely 'official' places administrative data closer in definition to the census and survey sources noted above but is a stepping stone towards Big Data.

On the cusp of using new sources

Kitchin (2015) points to various features of Big Data including its size, creation in (or close to) real time, its aim for total population coverage, its great detail on data entities, its relational nature enabling the joining of Big Data sources and its flexibility in terms of adding fields. In broad terms we see Big Data as informal sources whereby the data have been created as part of a process of everyday life. Subsequent knowledge about the existence of sources and an intrigue to plug a particular gap leads to their usage. 'Big' implies that issues which researchers may have about the quality of the data will be offset sufficiently by the quantity such that conclusions can be drawn (though Garbage In may well be Garbage Out). Big Data offers incredibly detailed information at near real time from a variety of sources such as financial transaction data, internet and mobile phone use with the potential to deliver valuable new understanding of society. What 'Big' can mean to researchers is that existing tools they are comfortable with no longer work because of the volume (and sometimes velocity) of the data. Often what is required is a bigger and faster computer, or a re-engineering of the methods for them to work on the data within a reasonable time frame. There may also be the need for the development or innovation of new analyses or visualisation methods.

Potential advantages of Big Data can be considered both spatially and temporally. One contribution of the census discussed earlier is its role in defining administrative or statistical geographies for research purposes. Big Data offers alternatives to these geographies. Since record level information in Big Data sources is often reported for individuals, for specific households or events, and because there is (generally) less aggregation before data are delivered, there is an opportunity to conduct analysis for more flexible geographies based on point data, or the use of postcodes as an alternative to more traditional administrative geographies. For example, the detailed analysis of migration distance using postcode to postcode moves is used to compare patterns for a hierarchy of administrative geographies in the UK by Stillwell and Thomas (2016) who find that results vary substantially by the geography used. The timeliness (or even real time nature) of Big Data sources offers considerable advantages for research into dynamic processes. Where survey or administrative data are released periodically to a timetable set by the data holders, Big Data are often available more regularly. For example by using Application Programming Interfaces (APIs), researchers can get immediate access to the latest data from a provider without the need to request, agree access and download data (as is the case with more traditional forms of data). This also presents opportunities for streamlining research and making the process more efficient: if the data download and extraction can be automated then this time can be better devoted to analysing those data. Through Big Data there has been an explosion in the amount and variety of data available for research and this opens up new opportunities to study things in novel and interesting ways and indeed allow us to study things that have evaded study before now.

As the volume of Big Data increases, infrastructure is being developed which is able to deal with its storage, access and security. One example is the Consumer Data Research Centre, part of the ESRC funded Big Data network, which is forging academic partnerships with commercial data organisations to procure data for research purposes. There has also been a recent focus by national statistics agencies to develop in-house Big Data teams (e.g. the Office for National Statistics Big Data Project), with an eye on utilising data which in the past have been the sole domain of commercial organisations. With this legitimisation of Big Data infrastructure comes the research expertise to deal with data effectively and ethically. Big Data analysis is quickly moving from a niche to a mainstream practice and is becoming better regulated and more transparent in the process. As this continues, Big Data is likely to take on a more 'formal' role in the data landscape of socio-demographic analysis. There is, however, a need for further work to be undertaken to ensure the quality and consistency of Big Data sources is adequate for the purposes of socio-demographic research as it is likely that quantity does not offset issues of data quality.

Smith (2010: 4) tracks the use of opinion polls through to social media analyses and warns that "new alternatives to survey research in general and public opinion research in particular have not established their superiority either scientifically or empirically." Similarly, Ye and He (2016: 813) "caution against shifting towards a big-data-driven research too hastily." We do not believe that Big Data can or should be used without the existence of formally planned data collection sources. This is for a variety of reasons.

The first reason relates to reliability and issues of bias. Whilst a new source may well have utility to provide information, the first step a researcher is likely to take would be to check variable distributions against an existing dataset and top of the list would be census data. The careful sampling procedures enabling derivation of an unbiased sample and statistical inference are crucial aspects that are not easily possible using Big Data. To check its utility for research, Thompson et al. (2011) compared opinion poll data with census microdata and large scale social survey sources finding the source not representative of persons outside of mid-life ages and of non-White ethnicity. In the quest for an appropriate denominator in crime rates, Malleson and Andresen (2016) compare Big Data sources with census workday populations concluding that the latter is the most appropriate population-at-risk measure.

The second reason is that, like administrative data, Big Data sources are influenced by inconsistencies that emerge from the process of collection / generation and the lack of researcher control / input to this process. For example, Google's analysis of search trends relating to flu revealed very strong correlations between the frequency of internet searches related to flu and actual diagnosed cases with the potential to predict rises of flu cases in particular areas enabling health practitioners and policymakers to respond quickly to a flu epidemic (Ginsberg et al., 2009). Subsequently, Lazer et al. (2014) demonstrated that in the US, the Google flu trend analysis was predicting more than double the proportion of doctor visits for influenza-like symptoms compared to the Centers for Disease Control and Prevention. This example provides a warning for the assumption that Big Data might substitute traditional data collection and analysis pointing out that quantity of data does not sidestep issues of measurement and construct validity. A key issue is that search engines are continually evolving to improve their performance, a process which is thought to influence what is actually searched for thus complicating the comparison of search trends over time.

A third reason is linked to issues of data ownership. Much Big Data is owned by commercial organisations (even social media data) rather than by the public (through the State) as is the case for the Census and Government-funded surveys. A further differentiation between conventional and Big Data sources is that the former are most likely to have had ethical consideration about the data collection, dissemination and usage but for the latter there have been fewer constraints. We would argue that the State has a role to play in monitoring and providing information on the circumstances of the general population and the inequalities in outcomes that are experienced across social, ethnic and demographic groups. Such issues should not be left entirely to the data provided by commercial organisations with a range of vested interests. Publically available data are needed by the general public to hold the government and institutions to account. While web-scraping techniques, data mining and the availability of APIs to tap in to huge quantities of data are well advanced, there is not, in all cases, an explicit licence to use data for research purposes. A comprehensive discussion of the access and use of Twitter data is offered by Puschmann and Burgess (2013: 7) where they assert that “rights to data cannot be inferred from technical availability alone”.

The fourth reason for the use of Big Data in combination with other sources looks to the future and the geographical frameworks used. Whilst the geographical projection system used for the geocoding of individual level point data is not problematical, it is the aggregation into zones which needs conventional sources. Any analysis will need to refer to existing geographies in some way to make sense of what is being analysed. Currently, census, electoral and administrative geographies are intertwined and delimited by headcounts (people and households) informed by formal data collection processes (predominately a census). In the UK, small area geographies of data dissemination defined for the census are then used for other datasets to be made available. The methods by which these census geographies are defined (Cockings et al., 2011) can be used with alternative data inputs but the census provides a recognised source.

There are a number of challenges facing traditional data sources such as the census and social surveys. These include low response rates (in surveys), a perception of data lagging events and increasing concerns around the costs of data collection. Undoubtedly, Big Data brings exciting new opportunities, but without careful comparison with more traditional data sources such as sample surveys and censuses it has the danger to confuse not clarify our understandings of the social world around us. Dorling and Simpson (1999) in ‘Statistics in Society: the arithmetic of politics’ note that ‘Statistics are pervasive and powerful, but often misleading or misunderstood, no more so than when they concern society’. We should take heed of this warning and the other issues we identify as Big Data sources become ever more available and influential. As Chris Yiu of the Policy Exchange points out (2012: 23), “the application of big data tools and techniques has the potential to improve public service delivery and efficiency. An abundance of data and computing power does not, however, automatically guarantee good decision making.”

Acknowledgements

The authors are grateful for feedback on a draft from the editors, especially Roy Carr-Hill, and from our colleague Andy Turner.

References

Ajebon, M. & Norman, P. (2016). Beyond the census: a spatial analysis of health and deprivation in England. *GeoJournal* 81(3): 395-410 DOI:10.1007/s10708-015-9624-8

Akinwale, B., Lynch, K., Wiggins, R., Harding, S., Bartley, M. & Blane, D. (2011). Work, permanent sickness and mortality risk: a prospective cohort study of England and Wales, 1971–2006. *Journal of Epidemiology and Community Health,* 65: 86–792 doi.org/10.1136/jech.2009.099325

Beatty, C. & Fothergill, S. (2005). The diversion from 'unemployment' to 'sickness' across British regions and districts. *Regional Studies.* 39(7): p837-854

Buckley, J. & King-Hele, S. (2014). What is Weighting? *UK Data Archive.* University of Essex

Carr-Hill, R. (2015). Non-Household Populations: Implications for Measurements of Poverty Globally and in the UK. *Journal of Social Policy*, 44(02), 255-275

Cockings, S., Harfoot, A., Martin, D. & Hornby, D. (2011). Maintaining existing zoning systems using automated zone design techniques: methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A* 43(10) 2399 – 2418. doi:10.1068/a43601

Crockett, A., Afkhami, R., Rafferty, A., Higgins, V. & Marshall, A. (2011). Weighting the Social Surveys. *ESDS Government.*

Dawes, P., Fortnum, H., Moore, DR, Emsley, R, Norman, P., Cruickshanks, K., Davis, A., Edmondson-Jones, M., McCormack, A., Lutman, M. & Munro, K. (2014). Hearing and vision in middle age: a population snapshot of 40-69 year olds in the UK. *Ear and Hearing* doi: 10.1097/AUD.0000000000000010

Dorling, D. & Simpson, S. (1999). *Statistics in Society: the arithmetic of politics*. Hodder. London.

Dugmore, K., Furness, P., Leventhal, B. & Moy, C. (2011). Beyond the 2011 Census in the United Kingdom. *International Journal of Market Research*, 53(5), 619-650

Elias, P., Martin, D., Norman, P., Rees, P. & Simpson, S. (2013). Save Our Statistics: Beyond 2011 Independent Working Group. [Accessed December 2016]. Available from http://popgeog.org/beyond-2011-independent-working-group/

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature.* 457. doi:10.1038/nature07634

Harper, G. & Mayhew, L. (2012). Applications of population counts based on administrative data at local level. *Applied Spatial Analysis and Policy*, 5(3), 183-209

Holdsworth, C., Finney, N., Marshall, A. & Norman, P. (2013). *Population and Society*. London: Sage Publications Ltd.

Ipsos MORI (2014). Perceptions are not reality: Things the world gets wrong. [Accessed December 2016] available from https://www.ipsos-mori.com/researchpublications/researcharchive/3466/Perceptions-are-not-reality-Things-the-world-gets-wrong.aspx

Kitchin, R. (2015). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*. 3(3) 262–267.

Lazer, G., Kennedy, R., King, G. & Vespignani, A (2014). The parable of global flu: Traps in Big Data analysis. *Science*. 343: 1203-1205

Lomax, N. & Norman, P. (2016). Estimating population attribute values in a table: 'get me started in' Iterative Proportional Fitting (IPF) *Professional Geographer* 68(3): 451-461 DOI:10.1080/00330124.2015.1099449

Malleson, N**.** & Andresen, M.A. (2016). Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice* 46: 52-63 DOI: 10.1016/j.jcrimjus.2016.03.002

Marshall, A., Plewis, I. & Norman, P. (2013). Development of a relational model of disability. *European Journal of Population.* DOI 10.1007/s10680-013-9300-y

Norman, P. & Bambra, C. (2007). Unemployment or incapacity? The utility of medically certified sickness absence data as an updatable indicator of population health. *Population, Space & Place* 13(5): 333-352

Norman, P. (2013). Whither / wither the census? *Radical Statistics* 106: 13-17

Norman, P., Simpson, L. & Sabater, A. (2008). Estimating with Confidence and hindsight: new UK small area population estimates for 1991. *Population, Space and Place* 14(5): 449-472

ONS (2014). The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority. [Accessed December, 2016]. Available from http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes---projects/beyond-2011/index.html.

Pearson H (2016). *The Life Project.* Allen Lane / Penguin Random House

Puschmann, C. and Burgess, J. (2013). The politics of Twitter data. HIIG Discussion Paper Series No. 2013-01. Available from https://ssrn.com/abstract=2206225

Smith T. W., (2013) Survey-Research Paradigms Old and New," *International Journal of Public Opinion Research* 25(2): 218–29, doi:10.1093/ijpor/eds040.

Stillwell, J. and Thomas, M. (2016) How far do internal migrants really move? Demonstrating a new method for the estimation of intra-zonal distance, *Regional Studies, Regional Science*, 3:1, 28-47, DOI: 10.1080/21681376.2015.1109473

Thompson, C., Stillwell, JCHS, Norman, P. & Clarke, MC (2016). Exploring the utility of Acxiom's Research Opinion Poll data for use in social science research. DOI: 10.13140/RG.2.1.3723.9922

U.N. (2016). Population and housing censuses: Alternative approaches to census design [Accessed January 2017]. Available from: http://unstats.un.org/UNSD/demographic/sources/census/alternativeCensusDesigns.htm

Yacyshyn, A.M. & Swanson, D.A. (2011). The costs of conducting a national census: rationale for re-designing current census methodology in Canada and the United States. [Accessed December 2016]. Available from http://cssd.ucr.edu/Papers/PDFs/Yacyshyn_Swanson_JOS_Aug26_2011.pdf

Ye, X., & He, C. (2016). The new data landscape for regional and urban analysis. *GeoJournal*, 81(6), 811-815.

Yiu, C. (2012). The Big Data Opportunity: Making government faster, smarter and more personal. *Policy Exchange*. [Accessed December 2016]. Available from https://policyexchange.org.uk/publication/the-big-data-opportunity-making-government-faster-smarter-and-more-personal/

Paul Norman, School of Geography, University of Leeds, p.d.norman@leeds.ac.uk

Alan Marshall, School of Geography & Geosciences, University of St Andrews, alan.marshall@st-andrews.ac.uk

Nik Lomax, School of Geography, University of Leeds, n.m.lomax@leeds.ac.uk