Big Data and the charitable sector: Research implications

Voluntary Sector and Volunteering Research Conference, New Researchers' sessions 2014

Diarmuid McDonnell, University of Stirling,
diarmuid.mcdonnell@stir.ac.uk

Date submitted: 04/08/2014

Abstract

This paper briefly considers the opportunities and limitations of Big Data approaches to the study of the charitable sector in the UK. First a consideration of the core features and concerns surrounding Big Data is provided. A number of research projects that are characterised as or analogous to Big Data techniques are then described. In particular, there will be a focus on the potential research use of administrative data held by the Scottish regulator of charities, OSCR, and national surveys such as the Scottish Household Survey. Finally, the paper reflects on further potential of Big Data approaches for research on the voluntary sector.

Introduction

Quantitative approaches to the study of the UK charitable sector have traditionally relied upon the contents of charities' annual returns and accounts, information often held by their respective regulators. However, there are significant limitations to the type of research and generalisation of findings that use these sources, including but not limited to: missing data, small sample sizes, inconsistency of information reported despite SORP, and differing reporting requirements for charities depending on their annual income (Morgan, 2011).

The 'explosion' of administrative and survey data held by public and private sector bodies has led to a sustained focus in academia on their potential use in research. Commonly known as 'Big Data', these datasets are becoming available to researchers as a result of open data. Concurrently, UK research councils are promoting the reuse of large secondary datasets in an effort to extend and expand the range of research in this area. As a result, Big Data and its associated methodologies, such as data linkage and causal inference, could enable social science researchers to develop more detailed and generalizable findings on the charitable sector (Savage & Burrows, 2007).

This paper briefly considers the opportunities and limitations of Big Data approaches to the study of the charitable sector in the UK. First a consideration of the core features and concerns surrounding Big Data is provided. A number of research projects that are characterised as or analogous to Big Data techniques are then described. In particular, there will be a focus on the potential research use of administrative data held by the Scottish regulator of charities, OSCR, and national surveys such as the Scottish Household Survey. Finally, the paper reflects on further potential of Big Data approaches for research on the voluntary sector.

Big Data: definition, features and issues

'The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of *understanding* the world.' [emphasis added] (Anderson, 2008)

Big Data approaches are sweeping through the public, private and third sectors. The National Science Foundation (2012) describes Big Data as a multi-dimensional concept, incorporating the 'scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous datasets.' Berman (2013) adopts a more parsimonious definition of Big Data, delineating it according to the three Vs:

- Volume – large amounts of data;
- Variety - the data comes in different forms;

- Velocity - the content of the data is constantly changing, through the absorption of complementary data collections, through the introduction of previously archived data or legacy collections, and from streamed data arriving from multiple sources.

Big Data is about more than large datasets however; it is also about the infrastructure that allows for complex data collection and analysis, quickly becoming known as computational social science (King, 2014).

These datasets are often one of the following types:

- Linked – two or more datasets are merged together to generate additional variables or cases;
- Administrative – datasets that capture the operational concerns of institutions (for example, local authority datasets on welfare programs);
- Transactional – datasets generated through consumer interactions with an institution;
- Open – datasets that are made available by their owners for third-party use; and
- Large, longitudinal – datasets, such as the British Household Panel Survey, that capture large amounts of complex data over a number of different time points.

From a sociological perspective, Big Data holds the promise of embracing descriptions and classifications of the social world that were previously unthinkable (Savage & Burrows, 2007).

Issues

It's important to be explicit about what Big Data cannot do. For instance, Big Data is very good at discovering correlations but not the meaning of these; data alone are no substitute for scientific understanding or other tools. 'The claim that causation has been "knocked off its pedestal" is fine if we are making predictions in a stable environment but not if the world is changing…or if we ourselves hope to change it.' (Harford, 2014) "Let the data speak for themselves" is often the maxim of Big Data proponents but this is wildly optimistic assessment of the lack of bias or subjectivity in datasets (Schutt & O'Neill, 2014). Analysts should bear in mind that a feature of bigger datasets is a related increase in the number of spurious patterns, often dwarfing genuine findings (Harford, 2014). Fung (2014) synthesises these core criticism of Big Data into his OCCAM framework, describing Big Data as:

- *Observational:* much of the new data come from sensors or tracking devices that monitor continuously and indiscriminately without design. 'The core challenge is that most big data

that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.' (Lazar et al, 2014: PAGENO)

- *Lacking Controls:* controls are typically unavailable, making valid comparisons and analysis more difficult;
- *Seemingly Complete:* the availability of data for most measurable units and the sheer volume of data generated is unprecedented, but more data creates more false leads and blind alleys, complicating the search for meaningful, predictable structure;
- *Adapted:* third parties collect the data, often for purposes unrelated to the scientific inquiries, presenting challenges of interpretation;
- *Merged:* different datasets are combined, exacerbating the problems relating to lack of definition and misaligned objectives.

The Charitable Sector: Big Data examples

Exploring rural/urban differences in volunteering

Using a Big Data approach allowed academics from the University of Stirling to produce a more nuanced account for urban/rural differences in volunteering (Rutherford et al, 2014). The project combined data from six years of the Scottish Household Survey (2006 to 2011) over which the volunteering questions were relatively stable. Individual responses to these questions were then linked to the Scottish Charity Register in an effort to explore the effect of the number of charities in an individual's area has on the likelihood of volunteering. The model including charity data explained more of the variation on urban/rural differences in volunteering participation than simpler models that just used SHS data.

Risk and resilience in Scottish charities

Risk is an everyday part of charitable activity. Charity trustees are responsible for managing risk to ensure that their charities achieve their objectives and protect the organisation's funds and assets. In Scotland, the Office of the Scottish Charity Regulator (OSCR) has responsibility for implementing The Charities and Trustee Investment (Scotland) Act 2005, and ensuring that charities and their trustees comply with the law. One of the challenges for the regulator is ensuring that their action is appropriate, and that they balance enforcement of The Act against placing an undue burden on charitable organisations. Administrative data from OSCR is currently being used to examine the nature of risk in the charitable sector. OSCR holds historical data for all charities that have been registered in Scotland, as well as detailed 'metadata' in the form of case files and internal notes. This research will explore the extent to which this data could be used to measure risk. Quantitative

techniques will be used to develop risk indicators which could be used as warning signs that OSCR can adopt.

Future research

Many of the Big Data approaches applied to social research so far have focused on the study of networks and flows (Giles, 2012). These approaches often necessitate the use of innovative or novel methodologies, such as in the use of Twitter data for the study of political protests to tuition fees in the UK (Tinati et al, 2014).Similar approaches could be used to study the social networks of voluntary sector organisations, perhaps by analysing the connections between these organisations on Twitter (for example, analysing followers). Advocacy or fundraising initiatives could also be researched through aggregating social media data for the voluntary sector.

Another potentially rich area of research relates to the annual reports of voluntary organisations. There are accountability and public benefit concerns in the UK charitable sector at the moment (Thompson & Williams, 2014; Cordery and Morgan, 2013; Philips, 2013). Regulators do not possess the capacity to conduct a thorough assessment of the ability of charities to achieve public benefit, despite the availability of the data necessary to do so (which is contained in the Trustees' Annual Report). Academic studies such as Morgan & Fletcher's (2013) have utilised these data sources to good effect, producing detailed findings on the use of narrative reporting as a basis for demonstrating accountability. However, they are inevitably hampered by the sheer volume of data available (more than 150,000 registered charities submitting a TAR each year to the Charity Commission for England & Wales) and could only focus on a sample representing one percent of registered charities; they study also employed numerous research assistants in order to analyse the 1500 or so reports. Big Data approaches such as data mining or machine learning could be of use in this scenario, as they have been used before in the analysis of corporate reports.

Bibliography

Berman, J. J. (2013) *Principles of Big Data: Preparing, Sharing and Analyzing Complex Information*. Waltham, MA: Elsevier.

Cordery, C. and Morgan, G. G. (2013) 'Special Issue on Charity Accounting, Reporting and Regulation' *Voluntas*, volume 24, issue 3: pp. 757–59.

Fung, K. (2014) 'Google Flu Trends' Failure Shows Good Data > Big Data' *Harvard Business Review.*

Giles, J. (2012) 'Computational social science: Making the links' *Nature*.

Harford, T. (2014) 'Big data: are we making a big mistake?' *Financial Times.*

King, G. (2014) ‘Restructuring the Social Sciences: Reflections from Harvard’s Institute for Quantitative Social Science’ *The Profession*, volume 47, issue 1: pp. 1–8.

Morgan, G. G. (2011) ‘The use of UK charity accounts data for researching the performance of voluntary organisations’ *Voluntary Sector Review*, volume 2, issue 2: pp. 213–30.

Morgan, G. G. and Fletcher, N. J. (2013) ‘Mandatory Public Benefit Reporting as a Basis for Charity Accountability: Findings from England and Wales’ *Voluntas*, volume 24, issue 3: pp. 805–30.

Phillips, S. D. (2013) ‘Shining Light on Charities or Looking in the Wrong Place? Regulation-by-Transparency in Canada’ *Voluntas*, volume 24, issue 3: pp. 881–905.

Savage, M. and Burrows, R. (2007) ‘The Coming Crisis of Empirical Sociology’ *Sociology*, volume 41, issue 5: pp. 885–99.

Schutt, R. and O’Neil, C. (2014) *Doing Data Science*. Sebastopol, CA: O’Reilly Media.

Thompson, P. and Williams, R. (2014) ‘Taking Your Eyes Off the Objective: The Relationship Between Income Sources and Satisfaction with Achieving Objectives in the UK Third Sector’ *Voluntas*, volume 25, issue 1: pp. 109–37.

Tinati, R., Halford, S., Carr, L. and Pope, C. (2014) ‘Big Data: Methodological Challenges and Approaches for Sociological Analysis’ *Sociology* [online].