British Academy & British Society for Population Studies

Policy Forum: 'New' Data for Policy
8th July 2015

This is a summary of a discussion on 'New' Data for Policy *held at the British Academy under the Chatham House Rule. Presentations were given by David J Hand, Mark Birkin, and Jane Naylor on the opportunities and challenges of big data and data mining, the opportunities and challenges of linked data and the Consumer Data Research Centre, and the ONS Big Data Project*

Big data has been identified as a significant strategic priority for the Government, and it is estimated that it will create 58,000 jobs by 2017. The Government has made a £73m investment through the ESRC, recognising the potential of big data to transform public services through greater understanding of people and their choices, behaviour and movement. However, big data is not yet mainstream, and the challenges that surround its validity, quality and ownership, need better understanding in both the research and policymaking communities in order to fully realise its potential.

The opportunities for new data

New forms of primarily administrative data (also including 'big data') provide significant opportunities for policymakers and demographers. Whereas survey data tell researchers what people *say* they do, or think, administrative data, not collected initially for the purposes of demographic research, shows us what people *actually* do: how they spend their money, travel across a city, use a piece of technology, show preferences on social media, or communicate via e-mail. It is collected automatically by companies, institutions both public and private, and in huge volume.

Researchers and policymakers are rightly enthusiastic about the sheer volume of information generated by administrative and transactional processes, the rate at which it is acquired, its scope, and the potential to learn from linked and merged data sets. It is hoped that we will be able to develop a much richer understanding of the ways in which people live; modern data are much closer to the social reality.

This 'new' data is already feeding into policymaking, often providing a supplementary picture to more traditional kinds of analysis. Prices are being 'scraped' from the web to produce research outputs, with the potential to produce economic outputs in future. The rate of acquisition is crucial here: the usual timeframe for the measurement of GDP in the UK is around 85 days, with the first estimate made after 25 days with only 44% of the data available. Big data, which can be scraped from the web in real time, can be used to supplement early estimates and boost their reliability.

In another example, the Department for Business, Innovation and Skills has used the careers website LinkedIn to understand better the career trajectories of IT graduates. Similarly, the ONS has used Twitter to better track student movements into and out of university towns.

There has been notable growth in attempts to use big data – collected often from social media – to understand populations in relation to place, particularly around urban mobility and patterns of environmental criminality. For example, geo-located tweets can be used to track population flow, and also combined with the content of tweets to gain an understanding of people's preferences for particular buildings or places.

Challenges

However, this automatic collection method, whilst a significant benefit of new forms of data, also provides a challenge – the use of administrative data for research purposes is necessarily *secondary* analysis, exposing it to a greater number of potential flaws, which ordinarily would be ironed out through survey design. The notion that we escape errors by using "all" the data is misleading.

Suitability
Data may not be ideally suited to the research itself. There would have been little opportunity, in most cases, for researchers to help shape the data collection or its design.

Consent
There may be issues of consent if the data is being used for something other than its original purpose. The size of the data set, and the numbers of individuals' choices or movement patterns, for example, make retrospective consent near impossible[1].

Selection distortion
Selection distortion is a big problem – the website LinkedIn, for example, contains a significant male bias, and similarly, Twitter presents an age bias.

Size
The size of the data set is most problematic when researchers may wish to 'eyeball' the data in order to better understand any anomalies. The data is also often not static – growing in real time.

Changes to the collection procedures
Not only does the continuous nature of the collection of administrative data produce challenges in size, but also in changes to the data collection, which may make two portions of the data set incompatible.

Changes in definitions
The definitions of the collected data may change over time, so that discontinuities may appear in time series.

Top tips when using administrative data

It is essential that those involved in the secondary analysis of administrative data have a strong understanding of these quality issues. The following five tips are taken from the UKSA Administrative Data Quality Assurance Toolkit (2015):

1. Don't trust the safeguards
When coming into contact with administrative data, get to know the safeguards that were put in place when it was collected, and also check if safeguards are functioning effectively.

2. Get involved

[1] Responsible Use of Data, House of Commons Science and Technology Committee, 19/11/2014. http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/245.pdf.

It is crucial that researchers get close to the organisations carrying out the data collection, to understand their methods, and develop a common understanding. There are opportunities for secondments, and many webinars.

3. Raise a red flag
If you identify potential data quality issues, investigate anomalies and keep the community informed.

4. See the big picture:
Identify what investigations and audits have been conducted and what they found.

5. Corroborate the evidence:
In some ways the most important element of using administrative data – corroborate using other surveys to confirm or challenge what is derived from the administrative data[2].

The Consumer Data Research Centre

Many of the challenges above can be overcome or accounted for through greater engagement with the collectors of big data, and more consistent awareness of quality issues. The Consumer Data Research Centre (www.cdrc.ac.uk), funded from the ESRC investment, is taking a major role in both elements of this – in acquiring and analysing big data, undertaking and stimulating research, and building capacity in the research base.

A backlash from a data-savvy public?

There remains a significant piece of public awareness raising activity to be undertaken. In a vast majority of cases, the individual's record within a vast data set is not what is of use or interest in secondary analysis. However, the focus in public discourse has been on the ethical issues such as the lack of control of big data, informed and implicit consent, and privacy. The need for public trust to facilitate big data research is noted elsewhere[3].

A risk remains, however, of a public backlash against this kind of automatic data collection. The ONS detected a large unexpected drop in volume of geo-located tweets in September 2014. This coincided with a new Apple Operating System, which made an 'opt-out' option from geo-location much clearer for users. There is also concern that the public will begin to manipulate the data once they become more aware of its secondary usage. For example, data on local crime hotspots are used on property websites, which may affect local residents' propensity to report criminality. It is dangerous, therefore, to approach administrative data as if it were without flaws and provided a clear reflection of the real world.

[2] Lovelace, R., Birkin, M., Cross, P., Clarke, M. (2015) From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows, Geographical Analysis, published online 19/08/2015. doi: 10.1111/gean.12081.

[3] The UK Administrative Data Research Network: Improving Access for Research and Policy, Report from the Administrative Data Taskforce, Economic and Social Research Council, December 2012.