EXPLORING "WHITE FLIGHT" VIA OPEN DATA AND BIG DATA

Jim Ridgway, James Nicholson and Sean McCusker
SMART Centre, Durham University, UK
jim.ridgway@durham.ac.uk

Open Data and Big Data have important implications for statistics education (and beyond). We sketch some properties of each, and provide examples of the use of social media (an aspect of Big Data) in understanding some of the uses of statistics in argumentation about social change, and the use of Open Data in exploring complex social phenomena. We explore some implications for statistics education and for statistics educators.

OPEN DATA AND ENLIGHTENMENT

Figure 1: Tweet with graphic: a rare example of someone sharing a data visualisation which illuminates their social experiences (from Smith, 2013).

OPEN DATA AND BIG DATA

The Open Data movement seeks to put large data sets gathered by national agencies into the public domain in a way that is accessible and useable by ordinary citizens. Many countries subscribe to this idea; a list of country and state open data sites can be found at http://www.data.gov/open-gov/. Open Data can be described as "curated" – collected in a systematic way for some purpose, in such a way that data collection, analysis and interpretation can be critiqued. In contrast, Big Data is often: a byproduct of our daily activities and interaction with digital services; automatically collected; geographically and temporally trackable; generated in real time, often responded to in real time; and unstructured (see Ridgway and Smith, 2013, for further discussion on Open and Big Data).

A huge array of data is being collected to illuminate almost all areas of human activity, and about our impact on the world – physical, environmental, social (and antisocial), and economic. Examples include the use of mobile phone locations in crime and terrorism cases, biometric data

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

from runners to improve fitness, and for airport security, sensors in domestic appliances to advise on future purchases, use of satnav locations to monitor traffic and identify jams, bus tracking information, mobile banking transactions to detect economic downturns, weather data, and satellite images for both prediction and to assess long-term changes, mining social media - food related conversations on Twitter are strongly correlated with food price inflation, analysis of online searches by users to identify sources of localized disease outbreaks before they become global epidemics (e.g. the Global Viral Forecasting Initiative (GVFI) claims they can successfully predict outbreaks up to a week ahead of global bodies such as the World Health Organization), crowdsourcing - NGOs such as Ushahidi use crowdsourcing (and provide freeware for others) to obtain, verify and disseminate real-time information about human induced and natural disasters and election monitoring.

There is increasing interest in recording the microstructure of human behaviour, especially for security, and for commercial purposes. The US National Security Agency has developed capabilities to take advantage of smartphone apps such *Angry Birds*, that transmit users' private information across the internet (see Guardian Newspaper 28 Jan 2014). The NSA are able to collect details of users' lives such as: home country, current location, age, gender, zip code, marital status, income, ethnicity, sexual orientation, education level, and number of children, as well as personal contact lists, using social media such as *Twitter*, *Facebook*, and *Google Maps* (see NSA, 2010). Google recently acquired *Nest*, a company developing smart home gadgets. Disney's *MagicBands* - rubber wrist bands, monogrammed and coded for each visitor - contain a RFID tag which allows the user's location, rides and food choices to be tracked. Foreman (2014) describes these developments by rival commercial giants as the "meat space data race" (adopting Gibson's (1995) use of "meat space" in *Neuromancer* to capture the idea of describing human activity from the perspective of an Artificial Intelligence).

Big Data is being used more and more in commerce, official statistics, and influences the daily lives of citizens; it merits an appropriate response from the statistics education community. A number of aspects need to be addressed. These include: understanding the very idea of Big Data, and understanding new forms data gathering, data representation, data analysis, and data interpretation and misinterpretation.

Web technologies facilitate the creation of documents that combine interactive content and video (as well as supporting text such as tables and metadata descriptions) in ways that can be accessed via mobile devices as well as by PCs. There is an increasing trend to create coherent stories rather than data visualisation tools that users explore for themselves. Examples include *100 years of census* (Office for National Statistics, 2014; BBC, 2014) and *100 years of commercial flight* (International Air Transport Association, 2014). Such displays challenge existing notions of "statistical literacy", because of the complex inter-relations between evidence and storytelling. An exciting (and excited) community is emerging comprising statisticians, computer scientists, mathematicians, psychologists, journalists etc. who define themselves in terms of their current intellectual interests, rather than their first university degree. There is an opportunity for the statistics education community to benefit from, and contribute to, these developments and this community. There have never before been better opportunities to develop the statistical literacy of citizens and politicians. Conversely, failure to engage with this brave new world is likely to leave "statistics" and "statistics education" as a sleepy backwater populated by "statisticians" devoted to the elaboration of models developed in the 1930s, and constrained to interpreting data sets that can be hosted on a single computer.

Big Data and Open Data mediated via the World Wide Web are providing rich content for a world that is increasingly data-hungry. The appetite for more and better data visualisation continues to grow, and offers both opportunities and challenges for statistics education. There are profound curriculum implications at all levels. These include a radical review of the approaches to modeling data that are taught, the constituency of "statistics teachers", and the emergence of a new and rapidly evolving body of "pedagogic content knowledge" that relates to the ways that students come to understand new models, new representations and new forms of data interpretation. Big Data (and Open Data) pose similar challenges to the development of the statistical literacy of citizens and politicians.

Here, we offer illustrations of the opportunities for statistics educators afforded by Big Data and Open data. First is a media account that contrasts two "folk wisdom" explanations of the migration of whites from London, along with comments from an open discussion. Second is a brief account of analyses of *Twitter* responses to some data visualisations. These are followed by an example of the use of Open Data by a leading researcher exploring migration in England and Wales by different ethnic groups.

WHITE FLIGHT OR ACHIEVING DREAMS?: CASE STUDY 1

Mark Easton wrote an article for the BBC website (2013) entitled *Why have the White British left London?* based on the 2011 census data. Here is the structure of his article.

- *Data:* there were 620,000 fewer whites in London than 10 years earlier; whites now comprise 45% of the population so they are in the minority for the first time
- *Argument:* conventional wisdom is "white flight" - BUT this is just part of the story – aspiration is another big part
- *Data:* non-London whites (i.e. whites in the rest of England and Wales) are up by 200,000 (and a statistical explanation – low birth rate and emigration explain the difference between the 620,000 fewer whites in London, and the 200,000 more whites in the rest of the country)
- *Use of Interactive Map:* to show areas with the biggest decreases and increases in the white population: big losses in London, and gains in a horseshoe of areas around London
- *Question:* is this attributable to the dream of living in the country – or is it white flight?
- *Social history:* Barking and Dagenham (BD) the area with the most dramatic demographic change (80% to 49% white in 10 years)
  - 1920-1930: migration from London slums to good quality social housing in BD
  - 1931: huge car factory created jobs (during a depression)
  - 1944-55: migration from bombed out London to BD
  - 1980-2000: social housing could now be bought by tenants at 30% of their commercial valuation – about 70% of houses were sold
  - 2000-2010: decline and closure of the car factory (total jobs in BD down by 25%)
- *Speculation about individual motivations:* no job, but valuable house (and perhaps redundancy money), so opportunity to move onwards and upwards
- *Revisit the Graphic:* expensive inner city London has more whites – but not British (e.g. Russians)
- *Interviews in BD:* whites planning to leave – 2 people each planning to move to a nearby seaside resort; and someone moving to a rural cottage with a large garden; family who swapped a 3 bed house for a 6 bed house further from the centre, but well connected by rail. There are towns with lots of ex BD residents.
- *Conclusion:* do people move because of changes in the culture of where they live? No, "It is a story of aspiration. It is a story of success".

The first observation is that this is a good piece of "data driven journalism" (irrespective of whether one agrees with the conclusion). There is effective use of an interactive display, a clear structure to the argument, and a synthesis of multiple sources of evidence (one might question the representativeness of the interview data). There were 2065 comments on the BBC website in a 2 day period. Here, we explore the 20 highest rated comments. Of these, 3 were removed by the moderator. An example of a comment in response to the title of the article that was *not* removed is *Because it is an overpriced, immigrant driven, crime infested hell hole.*

There were challenges to the whole idea of trying to explain the data: *What's positive about this? You have said nothing more than white people have moved out of London. This is a sad story of mass displacement – there's no disguising it.* Most of the comments expressed opinions, with little justification. All the comments disagreed with Easton's conclusions, and a large proportion asserted that the article (and/or the BBC) was biased *A very sad article, not the subject but the writer and the organization that seeks to deny the obvious*. The dominant counter claim was that people were uncomfortable in a multiethnic environment; close behind was the assertion that

whites were being pushed out, and there were claims of aggression from incoming groups. Evidence cited included personal experiences, friends' experiences, and a radio recording of overt aggression. Only one respondent cited empirical evidence: …*Even the immigrants that are here dont* [sic] *want to compete for jobs against yet more immigrants. This is the reason why UKIP* [the UK Independence party] *have gone from 4% to 16% of the vote this year. This is likely to go higher next year when the Romanians and Bulgarians arrive. Yet more unskilled workers set against 17% youth unemployment.*

There was just one appeal for more data: *When's the BBC going to do some probing journalism to nail down the big demographic issues? What are the shifts in population ethnicity nationwide and regionally? What are the birth-rates for different cultural groups? What's a sustainable UK population size? The consistent failure to provide the public with answers to these questions appears almost conspiratorial.*

What are the implications for statistics education? The emergence of data driven journalism is, in itself, a welcome development. Perhaps the most obvious implication is that discussion forums provide an excellent place to promote statistical literacy. Forums offer a window into statistical conceptions and misconceptions that can feed into initiatives around statistical literacy. One contribution would be to emphasize the use of empirical evidence in informing debate, and to provide links to relevant data (for example, data visualisations on shifts in ethnicity and different birth rates demanded in *When's the BBC going to…* based on the 2011 census are available on line (http://www.smartcensus.org.uk/). The role and status of different sorts of evidence is interesting. It is unsurprising that personal experiences dominate, and these accounts are sources of information in themselves (albeit rather hard to evaluate in terms of authenticity or generalisability). The wealth of rival explanations poses its own challenges to an empirical approach to public policy. Accessing data in a form that can distinguish between rival accounts was almost impossible before the emergence of Open Data. For example, to explore the hypotheses proffered in comments about Easton's article would require data about migration that could be disaggregated by age, ethnicity, family composition, direction of migration, and knowledge about the demographics of different areas. We see how Open Data can help, later.

ANALYSING SOCIAL MEDIA: CASE STUDY 2

Alan Smith offers a selected sample of 24 Tweets relevant to the data visualisations created by the Data Visualisation Centre at the Office for National Statistics. The majority of Tweets offer a short description of the displays; most provide positive comments; *Congrats on the census stuff – my 11-year-old daughter started playing with the tool and immediately found stories. Impressed!* and encouragement to use the displays: *Have some fun with these Census visuals – compare boroughs versus regions – interesting data*. Very few actually offer conclusions: *Manchester's age distribution massively affected by large student pop'n* (and Georgina Orso's Tweet in Figure 1).

The absence of substantive comments in Tweets is unsurprising, given the limitations of the medium. However, this is consistent with an early analysis conducted to analyse the statistical insights and misconceptions associated with websites which presented a variety of data visualisations, where comments focussed on aesthetic aspects of displays, and there were almost no comments relevant to data interpretation (see Ridgway, Nicholson and McCusker, 2008).

ETHNIC DIVERSITY AND MIGRATION WITHIN THE UK

Kaufmann & Harris (e.g. 2013a) are researching the topic addressed by Easton – how the white population in Britain is responding to the changing ethnic profile in the country in terms of where they choose to live. National demographics are changing as a result of, *inter alia*, immigration, mixed race relationships, and differential fertility of different ethnic groups. Understanding the associated social changes does not lend itself to simple approaches or analyses.

The authors developed a measure of the ethnic diversity of a geographical area using one data set (the full census data from 2011), then used differet data sets to look at migration to and from high and low ethnically diverse locations, as a function of a variety of social and demographic features. Figure 2 shows data derived from the 1991, 2001 and 2011 waves of the ONS longitudinal survey. The sample is restricted to those who moved wards (small geographical areas) between the waves of the longitudinatal survey.

The data have been loaded into a SMART Centre interface (see http://www.smartcensus.org.uk/). Figure 2 shows the profiles of people who are moving from less diverse to more diverse areas. Here, the groups are people in their 20s, and people with dependent children (of course, these two groups overlap to some extent), for both white British and minority citizens.

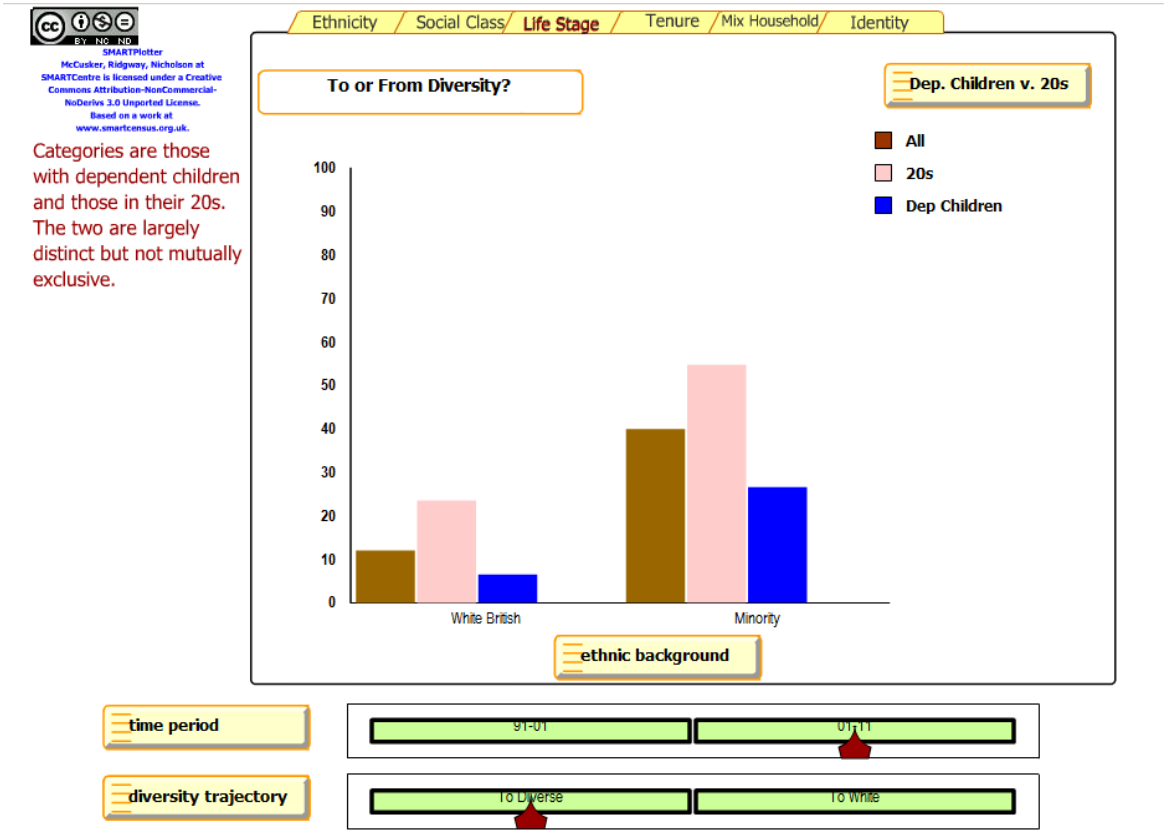

Figure 2: Movement (2001-2011) to more diverse areas by those in their 20s, and for those with dependent children, by ethnicity

Comparing the brown bars (All) it can be seen that white British are far less likely to move to more divese areas (about 10% compared with about 40%), and that this is more pronounced for those with dependent children (the blue bars – about 5% compared with about 25%). Each tab at the top of the display allows the user to explore migration with the population disaggregated by other factors. Kaufmann's view of his current research is that the stories are more nuanced than in many of the other areas in which he has worked (personal communication). SMART are collaborating with his group because he felt that the dynamic data visualisation would allow the stories in the data to be conveyed to a policy and journalism audience more effectively than the alternative methods (raw numbers or regression outputs) he has previously used, and he felt this was reflected in a more nuanced reporting of his findings, though this feeling is not based on any formal analysis or comparison. Unfortunately, our impressions of the media coverage generated by the presentation to the Demos thinktank (Kaufmann & Harris, 2013b) is that they are very heavily based on the text-based press release, and have not engaged with the data visualisation tool (or the complex interactions). More work is needed in this area to understand how to communicate effectively to policy makers and journalists where the stories in the data are nuanced.

This brief case study illustrates a number of ideas. Open Data facilitates the use multiple data sources, and more detailed exploration of patterns in data. Here, one data set was used to develop a measure (ethnic diversity of an area) that was then used to understand patterns in other data sets (longitudinal data on migration). The data visualisations facilitate data explorations and

conjecture testing that are difficult to do in other ways. They offer the prospect (but not yet the reality!) of better understanding of complex social phenomena by policy makers, and in the media.

REFERENCES

BBC Census animation: *100 years of population growth* http://www.bbc.co.uk/news/uk-18854073 downloaded 21st January 2014.

Easton, M. (2013). *Why have the white British left London?* http://www.bbc.co.uk/news/uk-21511904

Foreman, J. (2014). *You don't want your privacy: Disney and the meat space data race.* http://gigaom.com/2014/01/18/you-dont-want-your-privacy-disney-and-the-meat-space-data-race/

Gibson, W. (1995). *Neuromancer*. London: HarperCollins.

Global Viral Forecasting Initiative http://globalviral.org/

Guardian Newspaper (2014). http://www.theguardian.com/world/2014/jan/27/nsa-gchq-smartphone-app-angry-birds-personal-data

International Air Transport Association (2014) http://www.flying100years.com/

Kaufmann, E., & Harris, G. (2013a). *Exit, Voice or Accommodation? White working-class responses to ethnic change in Britain.* http://www.sneps.net/research-interests/whiteworkingclass/project-description

Kaufmann, E., & Harris, G. (2013b) *"White Flight" in London and the UK?* Presentation to Demos. Slides available to download from http://www.sneps.net/research-interests/whiteworkingclass

National Security Agency (2013). *Converged Analysis of Smartphone Devices.* Slides leaked by Edward Snowden. http://www.nytimes.com/interactive/2014/01/28/world/28mobile-annotateA.html

Office for National Statistics. *100 Years of census: England and Wales 1911–2011* http://www.ons.gov.uk/ons/interactive/ vp1-story-of-the-census/index.html Retrieved 21st January 2014.

Ridgway, J., Nicholson, J., & McCusker, S. (2008). Reconceptualising "Statistics" and "Education". In C. Batanero (Ed.*). Statistics Education in School Mathematics: Challenges for Teaching and Teacher Education*. Springer.

Ridgway, J., & Smith, A. (2013). *Open Data, Official Statistics and Statistics Education – Threats, and Opportunities for Collaboration.* Keynote Talk: Proceedings of the first Joint International Association for Statistics Education and the International Association for Official Statistics, Statistics Education for Progress, Macau.

Smith, A. (2013). Data visualisation and beyond: A multi-disciplinary approach to promote user engagement with official statistics. *Statistical Journal of the IAOS, 29*, 173-185.

Ushahidi http://www.ushahidi.com/