OPEN DATA, OFFICIAL STATISTICS AND STATISTICS EDUCATION – THREATS, AND OPPORTUNITIES FOR COLLABORATION

RIDGWAY, Jim[1] & SMITH, Alan[2]
[1]SMART Centre, University of Durham, UK
[2]Data Visualisation Centre, Office for National Statistics, Titchfield, UK

Contact emails: jim. ridgway@durham.ac.uk, alan.smith@ons.gsi.gov.uk

ABSTRACT
Technology is shaping the ways that evidence is used to influence policy and public opinion. Developments include: the semantic web; the Big Data movement, new tools for data visualisation, and the rise of data-driven journalism. Such developments will have profound effects in terms of the nature of evidence that is gathered, the ways in which it is presented and used, and the skills that will be needed for its interpretation. As such they offer opportunities, but also pose threats to both National Statistics Offices (NSOs) and to statistics educators. A great deal is to be gained from collaborations between NSOs, statistics educators, and other groups. Here, we give examples of the ways that technology is influencing practice, and describe a UK collaboration between the Data Visualisation Centre within the Office for National Statistics and the SMART Centre at Durham University, which sets out to work with journalists and policy makers, and uses Big Data tools to explore success. The opportunities and threats presented by technological developments to NSOs and statistics educators are discussed. We discuss strategies for working effectively with journalists, and other data users.

KEY WORDS: Open Data, Big Data, Visualisation, Data Driven Journalism, Census, Statistical Literacy, Futures

INTRODUCTION

Developments enabled by technology are shaping the ways that evidence is used to influence public opinion, and policy. These include: the Open Data movement; increasing uses of Big Data; the emergence of new ways to visualise data, and the rise of data-driven journalism. Such developments are of particular interest to National Statistics Offices (NSOs) and to statistics educators. The Open Data movement aims to make high quality data collected by governments and non-governmental organisations accessible to citizens. Most NSOs share this vision. Ill-structured and opportunistic data derived from sources such as traffic on email, *facebook* and *twitter* and human activity patterns tracked via mobile phones or travel cards ('Big Data') – is being used more and more, usually for purposes unforseen by the creators of the data. Increasingly, journalists are making good use of rich data, and are creating excellent, data laden, websites. An exciting range of visualisations are being developed to allow users to explore large, rich, data sets. All these developments will be significant for NSOs and statistics educators.

OPEN DATA

As early as 1792, Condorcet asserted the importance of informing citizens about governance, and presenting evidence about the state of society, in order to increase awareness of injustices and structural social inequalities. He believed in *savoir liberateur* – knowledge that would enable people to free themselves from social oppression. More recent initiatives such as *data.gov* in the USA and *data.gov.uk* in the UK explicitly state political objectives, notably to promote the democratic process by giving citizens access to data which can stimulate debate and inform policy making. Economic advantage is a second driver. For example, in the UK, the Open Data Institute http://www.theodi.org/ claims that it "will catalyse the evolution open data culture to create economic, environmental, and social value. It will unlock supply, generate demand, create and disseminate knowledge to address local and global issues". The Open Data movement has had considerable success in recent years in persuading major data providers (such as NSOs, Eurostat, and governments) to make data available to anyone who wants it. The promises of open

data are obvious – but there are barriers to be overcome. There are considerable technical issues in synthesising data from different sources; interpreting multivariate data is non-trivial; there are powerful vested interests across the spectrum of 'advisers' (including civil servants) whose traditional role has been to both find and interpret evidence for policy makers, who might not welcome easier access to data. NSOs are challenged to present their data for new (often naïve) audiences; statistics educators are faced with the challenge of educating an entire population about seemingly difficult ideas, and in creating curricula which devote more attention to the interpretation of large scale data sets.

BIG DATA

Big Data usually refers to data which is collected in real time, where the volume is huge, and the data are markedly varied (and often very noisy). Examples include: *google* searches for symptoms to predict flu epidemics; the use of data from satnavs to identify traffic jams or patterns of traffic flow; data from supermarket transactions to monitor changes in customer demand; analysis of social media to determine the 'mood of the nation'; and CCTV and mobile phone data for the prevention of terrorism. Big Data differs from Open Data in a number of respects. Open Data have been created for some purpose; measures are clearly defined; data are usually multivariate; the population being sampled is known, and the whole process of data generation and presentation has been subjected to extensive scrutiny. Big Data often have none of these characteristics. Further, it is often not 'open', but is owned by companies who seek economic advantage from its use; there is usually far too much data to store or to analyse via standard statistical techniques (so artificial intelligence (AI) tools are commonly deployed – such as neural nets to create models to determine credit risks); data cannot usually be grouped into social units of importance to public policy, such as families; new phenomena are measured; and measures are usually measures of behaviour, not internal states such as 'attitudes'.

As in the case of Open Data, there are challenges and opportunities for both NSOs and statistics educators. For NSOs, there are challenges that new methods of collecting data will be used to replace current methods, with associated dangers of poor quality control, and incompatibility between new and old data sources. There are also threats to the whole profession of 'statistician' – traditional methods of analysis cannot be applied, and new methods need to be developed collaboratively with computer scientists. For statistics educators, fundamental ideas such as the principles of measurement, data quality, and plausible inference in the face of uncertainly, need to be in ascendancy over the mastery of specific techniques. In addition, the skills base needs to be extended to include an understanding of some AI techniques.

DATA VISUALISATION

Major data providers are providing powerful visualisations in the hope of making their data more accessible (e. g. OECD's data is accessible via *Gapminder* and *eXplorer*). Key political targets such as the UN Millenium Development Goals are presented in the form of an interactive dashboard, to encourage public engagement. This work has a long history - Playfair (1786) can be credited with the invention of statistical graphics - and graphics have often been used to make social data accessible to a wide audience, exemplified by the Neuraths' work in the period 1930-45 designing graphical displays to demonstrate social inequalities (see Neurath, 2010).

The SMART Centre provides a number of interactive displays on topics such as educational attainment, health, and riots, along with a facility for users to embed their own data in interactive multidimensional displays (see http://www. dur.ac.uk/smart.centre/). Figure 1 provides an example. It shows data on educational attainment by students who are 16 years old and in receipt of free school meals in England and Wales in 2010, broken down by sex and ethnic group. Notice that girls in every group outperform boys; white children perform *worse* that black children, and Chinese students outperform others by a considerable margin. Different data sets are provided under different tabs ('White', 'Mixed' etc) and data can be explored by moving sliders and dragging variable names onto the axes.

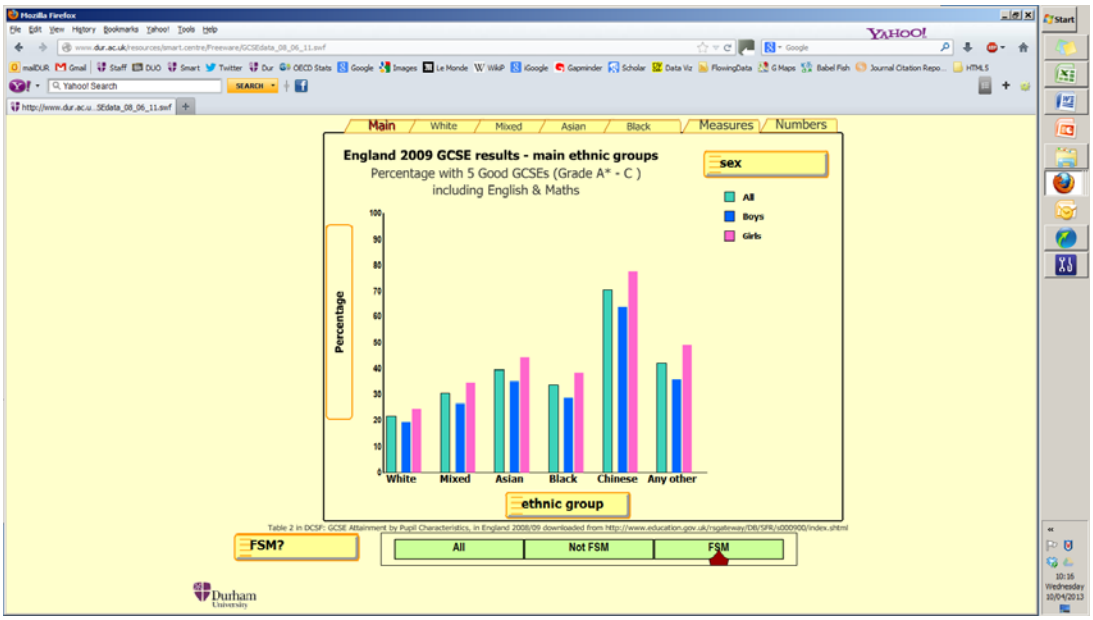

Figure 1: *Educational Attainment of Students in Receipt of Free School Meals*

The Data Visualisation Centre within the UK ONS has created a variety of exciting data displays such as a dynamic population pyramid, a dynamic map of commuter flow, and series of choropleth maps to display small area census data (see http://www.ons.gov.uk/ons/interactive/index.html). Figure 2 shows an interactive display of educational attainment of students who are 16 years old for the period 2008 to 1011, in different regions. Users can 'mouse over' the map to compare the attainment of students in any region with national student attainment, over time.

Figure 2: *Educational Attainment of Students in Different Regions*

Data visualisation presents opportunities and threats to both NSOs and statistics educators. The opportunities are clear – there are novel ways to present complex data in ways that facilitate exploration by users – both expert and novice. Development of visualisations has direct

implications for resources, and a great deal of institutional inertia has to be overcome in those NSOs who believe that their role is simply to gather data then make it available to people with sufficient technical skill to access and interpret it. User competence in working with data visualisations is largely undetermined (although there is some evidence that statistically naïve users can make sense of multivariate data presented in interactive displays – see Ridgway, Nicholson, and Mccusker, 2007). For statistics educators, there is the promise of documents that facilitate interactions which could be used to support statistics teaching, and direct access to rich, authentic, contemporary content to support (in particular) social sciences. There is also the challenge of teaching skills associated with the interpretation of rich data, and developing a curriculum that encourages and facilitates engagement with authentic data.

DATA DRIVEN JOURNALISM

An important cultural trend is the emergence of 'data driven journalism' (e. g. Bradshaw, 2010; Brooke, 2010; Rogers, 2011). Gray, Bounegru, and Chambers, (2012) discuss "*the new possibilities that open up when you combine the traditional 'nose for news' …with the sheer scale and range of digital information now available.*"

Early, influential examples of data driven journalism in the UK can be found in the coverage of the *Wikileaks* cables (see http://www.telegraph.co.uk/news/worldnews/wikileaks/), or the MPs' expenses scandal (see http://www.guardian.co.uk/politics/2009/dec/16/mps-expenses-what-we-learned). In both these instances, journalists had to work with large datasets (very often adopting 'crowd-sourcing' techniques – accessing distributed help across the Internet) to extract and publish high profile stories. The success and impact of these stories, both in the UK and the United States, encouraged a strong movement within mainstream media to build better data-driven reporting capabilities: The Daily Telegraph, FT. com/interactive, Guardian Datablog and BBC Visual Journalism team are all examples of this rising trend. In the United States, the New York Times interactive narratives are often cited as excellent practice in this area. In content terms, data journalism is typified by visually rich, often interactive, articles. Infographics feature heavily, while text is often abbreviated to support the visuals, in contrast to traditional forms of journalism, where often the reverse is true.

As with other developments, there are opportunities and threats for NSOs and statistics educators. Media have their own agendas that are not necessarily aligned with those of NSOs or educators. There are opportunities for a much wider dissemination of data, and discussions around their interpretation and use. This optimism needs to be tempered by moves by newspapers (Times, New York Times and others) to create 'paywalls' wherein users pay for full access to materials. Data driven journalism has the potential to promote statistical literacy, because journalists want to communicate ideas that are grounded in data. Gal (2002) states that statistical literacy refers to " the need for people … to develop the ability to comprehend, interpret, and critically evaluate messages with statistical elements or arguments conveyed by the media and other sources". It follows that SE needs to equip learners with the ability to understand and interpret new data visualisations. There are associated pressures for the curriculum to place more emphasis on existing core themes in statistics, such as the reliability of the source, the authenticity of the data, and plausible data interpretations.

INCENSE – AN ON-GOING COLLABORATION

Open Data, Big Data, data visualisation, and data driven journalism all converge to offer NSOs and statistics educators new opportunities to further their ambitions. A collaboration between the DVC within the ONS and the SMART Centre at Durham University sets out to promote the use of data from the 2011 census for public debate and for policy making. The work is funded by the UK Economic and Social Research Council. The collaboration is grounded in common interests. Both groups have developed innovative visualisation tools; both groups are keen to have more and better use made of census data to inform policy and debate; both groups want to make good use of current and emerging tools to understand and use emerging patterns of influence in modern societies. There are good reasons for collaboration; NSOs have resources and expertise of considerable value to statistics educators; academics can do things that ONSs cannot. For example, we have designed the INCENSE website (http://www.smartcensus.org.uk/) with

naïve users in mind, untrammelled by the constraints of either the university website or the ONS website. Users are offered easy ways to navigate a variety of visualisations of census data. We can also present content and opinion for discussion that would not be appropriate for ONS to display (such as comments that reflect political views, or are critical of government policy), and we can support blogs and forums targeted on specific aspects of the census. Statistics educators also have time to explore patterns of data use, and the reasoning associated with particular data sets. We are exploring the use of some of the tools being developed as part of the Big Data movement (such as sentiment analysis) to look at patterns of influence in social media.

Census data are collected to inform policy –notably in health, education, and social care – and data are also collected that are relevant to understanding important changes in demography and culture, such as country of birth, ethnicity and religious affiliation. The key idea in the INCENSE project is to work with people who shape opinion and policy, in order to get census data discussed and used wherever it is appropriate. We are in conversation with a number of distinct groups that include politicians, people working in health, and in local authorities. We are working with people who set out to inform public and political opinion. These people include researchers devoted to the analysis of census data for a variety of purposes, such as studies on the problems of an aging population, poverty, ethnicity, and the like, and in particular people who create briefing documents and policy documents targeted on decision makers. It is important that results from work of this sort should be of some long term benefit, and so we are collaborating with the Royal Statistical Society's (RSS) *Getstats* campaign, amongst others. Here, we describe just the programme of work with journalists.

WORKING WITH JOURNALISTS

Collaborations with journalists have the potential for public engagement which goes way beyond anything that can be achieved by either NSOs or statistics educators working alone. An example of this is the ONS' engagement with journalists that is intended to increase outreach for the content of the 2011 census. Interactive graphics developed by the DVC have been syndicated to a range of media outlets, which supplemented them with other journalistic content, and presented them on their websites. There were no special arrangements made for individual organisations – all visualisations were available to everyone who wanted to use them. Notable users have been the BBC and the Guardian and Telegraph newspapers, both on TV and in print, and on their respective websites. An example is shown in Figure 3.

Figure 3: *A Dynamic Display Created by ONS Embedded in a Media Website*

The underlying content was also available on the ONS website; however, syndication helped boost page views by around 1,000% for the first release of census data. The first release of data visualisations (*100 years of census* and the *twin population pyramids*) has 12,000 views in the first week on the ONS website, then 120,000 views when syndicated to the BBC, Guardian and Telegraph. The process of syndication is not without its challenges. Materials have to be of high quality, and robust. There can be issues around the format of the content (some media sites will not accept graphics created in *Adobe Flash*, because they cannot be read by many mobile devices – users with such devices represent an ever-increasing component of their readership). Consequently, there are increasing pressures to create websites and graphics that are 'responsive'; that is, the content changes its form appropriately according to the nature of the device it is loaded on - desktop computers, tablets, or smartphones. Such challenges in authorship represent a significant issue for NSOs. There can also be issues of attribution and branding – readers often believe that the visualisations have been created by the owners of the host website.

In our experience, journalists are keen to engage with NSOs – but not by recycling simple statements of findings ('the area with the highest proportion of people born outside the UK is…') – but rather by presenting rich data visually, coupled with insightful narrative. They are also keen to provoke and support debate, and can have considerable success. For example, a BBC article by Mark Easton based on 2011 census data, entitled *Why have the white British left London?* was commented on by more than 2000 people in a single day.

Journalists are also often willing to share data on access to their websites, such as that generated by *google analytics.* This can provide information on page hits and the time users spend exploring data.

IMPLICATIONS FOR NSOs

For those NSOs bold enough to engage with a brave new world and to adapt their organisations to a new scheme of things, there are opportunities for a huge boost to their influence and range of activities. The alternative is stagnation (or even atrophy), and increasing irrelevance in shaping future societies. New developments offer opportunities, but also pose threats to NSOs. The opportunities are clear – conventional ways of data gathering can be supplemented with

innovative methods; there are new ways to present high quality data gathered to inform policy, in a form which makes them accessible to a wide audience, and new ways to reach these wider audiences. However, as more and more people, agencies and interest groups offer commentaries on data, data is likely to be misrepresented (for a variety of reasons). Careful analyses prepared by NSOs can be overlooked. NSOs should engage actively in public debates by offering professional commentaries on the quality of data analyses. Collaborations should be developed with a wide variety of partners. Informal agreements (for example, with a range of media, via an open invitation to the journalistic community to engage) are more appropriate (in the UK at least) than formal partnerships. Such agreements must be based on mutual respect and a genuine understanding of shared interests and points of departure. Syndication can help improve outreach massively – but content has to be of high quality, and must be adapted to the technical demands of potential host sites.

Data-driven storytelling requires extra skills in addition to the traditional ones associated with journalism. NSOs should support initiatives designed to develop journalists' skills in dealing with evidence. For example, in the UK, the RSS has been working to ensure that journalists are offered career development opportunities. The RSS holds regular workshops for journalists to help them with statistical issues. Since 2007, there has been an annual award for Excellence in Statistical Journalism.

There is a pressing need to improve public statistical literacy, and the statistics education community can help. One direct approach is to include interpretative commentaries alongside data; another is to provide links to tutorials on interpretation. Both of these approaches could be problematic for different reasons: offering interpretations of data runs the risk (for NSOs, but not for journalists or academics) of encroachment into the political domain; linking to tutorials involves judgements about the quality of provision, and implicit endorsement (and may prove to be effective for just a handful of users). ONSs could offer pointers to websites that offer tutorial guidance on a range of topics, such as the Open Data Handbook http://opendatahandbook.org/en/, or resources from the UK Open University (e. g. http://www.open.edu/openlearn/science-maths-technology/mathematics-and-statistics).

IMPLICATIONS FOR STATISTICS EDUCATION

There is a broad consensus that the numeracy skills of the adult population are very poor, confirmed by evidence from international surveys such as the international Adult Literacy and Lifeskills survey (see http://nces.ed.gov/surveys/all/). The statistics education literature provides extensive evidence of the failures that students experience when tackling problems involving (amongst other things) tables (e. g. Watson and Nathan, 2010), graphs (e. g. Swan and Philips, 1998), and box plots (e. g delMas, Garfield, Ooms and Chance, 2007). The plethora of visualisations may well present further challenges for citizens. Open access to data does not mean that data will be displayed and analysed in appropriate ways, or interpreted sensibly. The provision of powerful tools does not necessarily lead to empowered citizens, and so a good deal of work needs to be done to develop representations that are intelligible to statistically naïve users, to demonstrate their effectiveness, and to develop curricula that enable students to interpret and critique novel visualisations.

There is an urgent need to rethink the statistics curriculum, and the development of statistical literacy. Statistical literacy involves a wide variety of skills and dispositions. In the context of Open Data and Big Data these include a sophisticated approach to data provenance (e. g. awareness of potential problems with metadata; plausibility of data), and to measurement (including the politics of measurement). The technical content in curricula also needs to be reviewed. For example, the logic of the analysis of large scale multivariate data sets is rather different from the logic of drawing inferences from small samples that are then applied to populations. Key activities for analysis are: assessing effect sizes; looking for (non-linear) functional relationships; and mapping interactions. Statistical ideas that require more curriculum emphasis include: modelling functional relationships; confidence intervals; effect size; and Simpson's paradox.

The internet offers opportunities to study statistical conceptions and misconceptions on a grand scale. Blogs, media websites, and social media such as *twitter* and *youtube* are all potential

sources of data (and tools are available to analyse traffic in a variety of ways). Analysis and direct engagement with these sources should become an important part of statistics education.

CONCLUSIONS

ONSs and statistics educators need to respond positively to the opportunities provided by Open Data, Big Data, and computer-based visualisations. The alternative is to see NSOs marginalised, and the increasing irrelevance of a static statistics curriculum that offers little help in interpreting the data that affects all our lives. Changes to methods of presenting data need to quite radical; new ways of approaching evidence need to be adopted. Strengthening the links between ONSs and statistics educators is one way in which we might make good use of current opportunities (and respond to threats).

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council [grant number ES/K004328/1]

REFERENCES

Bradshaw, P. (2010). http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide

Brooke, H. (2010). *The Revolution will be Digitised: Dispatches from the Information War*. London: Heinemann.

Condorcet, J. (1994). *Foundations of social choice and political theory*. Aldershot and Brookfield, VT: Elgar (original work published in 1792).

data.gov/ http://www.data.gov/

data.gov.uk http://data.gov.uk/

delMas, R. , Garfield, J. , Ooms, A. and Chance, B. (2007). Assessing Students' Conceptual Understanding After a First Course in Statistics. Statistics Education Research Journal, 6(2), 28-58. http://www.stat.auckland.ac.nz/serj

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

eXplorer http://ncva.itn.liu.se/explorer/openexp?l=en

Gal, I. (2002). Adult statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Gapminder http://www.gapminder.org/downloads/

Getstats http://www.getstats.org.uk/

Gray, J. , Chambers, L. , and Bounegru, L. (2012). The Data Journalism Handbook. O'Reilly Media. http://datajournalismhandbook.org/

HM Government (2012). Open Data White Paper: Unleashing the Potential http://www.official-documents.gov.uk/. Downloaded 18 March 2013.

Neurath, O. (2010) *From Hieroglyphics to Isotype: A Visual Autobiography*. London: Hyphen Press.

Playfair, W. (1786, 2005) *The Commercial and Political Atlas and Statistical Breviary*. Cambridge University Press

Ridgway, J. , Nicholson, J. and McCusker, S. (2007). Reasoning with Multivariate Evidence. *International Electronic Journal of Mathematics Education 2*(3), 245-269.

Rogers, S. (2011) How to get to grips with data journalism. http://www.journalism.co.uk/skills/how-to-get-to-grips-with-data-journalism/s7/a542402/

Swan, M. and Phillips, R. (1998). Graph interpretation skills among lower-achieving school leavers. Research in Education, (60), 10-20

Watson, J. and Nathan, E. (2010). Assessing the Interpretation of two-way tables as part of statistical literacy. Proceedings of ICOTS8