SMARTCENSUS – MAKING SENSE OF CENSUS DATA

James Nicholson, Jim Ridgway and Sean McCusker
SMART Centre, Durham University, Durham, UK
j.r.nicholson@durham.ac.uk

Statistics Agencies (SAs) increasingly are putting data into the public domain. However, open access does not mean that data will be displayed and analysed in appropriate ways, or interpreted sensibly. SAs face a number of potential problems associated with getting data used widely and appropriately – there are issues of access to media, and of appropriate data interpretation – statistical literacy in the population is rather low. INCENSE is a collaboration between the SMART Centre at Durham University and the UK Office of National Statistics, designed to make census data more accessible through data visualization, and to understand how politicians and journalists used 2011 census data when it was released. This paper offers a critical account of some challenges of creating effective data visualisations. We will explore the benefits to SAs and to academics from collaborations of this nature.

INTRODUCTION

Pullinger (2013) in his presidential address to the Royal Statistical Society argues that statistics are essential to good decision making, providing a basis on which to make tough decisions. They cannot provide the moral compass on which the decision is taken but statistics can at least provide evidence which informs the decision making process. He argues that statistical evidence is not a substitute for judgment but those using evidence often seem to suspend common sense in their effort to use it.

In 2012 the UK Administrative Data Taskforce was formed with the aim of improving access to and linkage between government administrative data for research and policy purposes. Routinely collected administrative data are often high quality, nationally comprehensive resources which provide information about long periods of people's lives, and are relatively inexpensive to exploit, compared to the costs of establishing specially commissioned surveys. By unlocking the research potential of these data, the hope is that we can improve our knowledge and understanding of the action required to tackle a wide range of social, environmental, health and security issues.

The technical issues regarding safe, secure and accurate linkages between administrative data sets are substantial but the potential benefits in terms of making progress in tackling difficult issues in our society are considerable. Ridgway, McCusker & Nicholson (in press) report on a collaboration with a research team who are working with linked data sets, and provide some initial observations on the media coverage generated by the presentation to the Demos thinktank (Kaufmann & Harris, 2013) and the subsequent public discussion of those media reports.

AUTOMATION OF DATA VISUALISATIONS

Figure 1 shows a screenshot of the ONS visualisation of the census data on economic activity in 2001 and 2011. The user can select specific categories (e.g. part-time employee) or agggregated groups of categories (e.g. figure 3 shows all economically active people), and can use postcodes or the zoom facility to vary the geographical region on display.

Smith (2013) reports on the work of the ONS Data Visualisation Centre (DVC) in promoting user engagement with official statistics and in particular, the 2011 UK Census which has been the focus of much of their recent efforts. Many of the visualisations produced by the DVC have been picked up and used by media agencies specializing in 'data journalism' – the BBC, the Guardian and the Telegraph, and Smith identifies rewards including broader public outreach than the ONS website would ever generate itself, and increased understanding of official statistics.

There are tensions however in the constraints which require a single interface to be able to be pointed at a large of number of similar datasets, and to be able to react dynamically to requests for a different geographical area to be displayed, or a different response level from the same dataset. The attraction of being able to display different levels of geography from national down to quite local encourages user interest in these visualisations which is most encouraging, but there are some prices to be paid in terms of how accurately the public understand the displays. We identify three

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 [© 2014 ISI/IASE]

conceptual issues with the GIS based displays which allow comparison of key statistics in the 2001 and 2011 census releases.

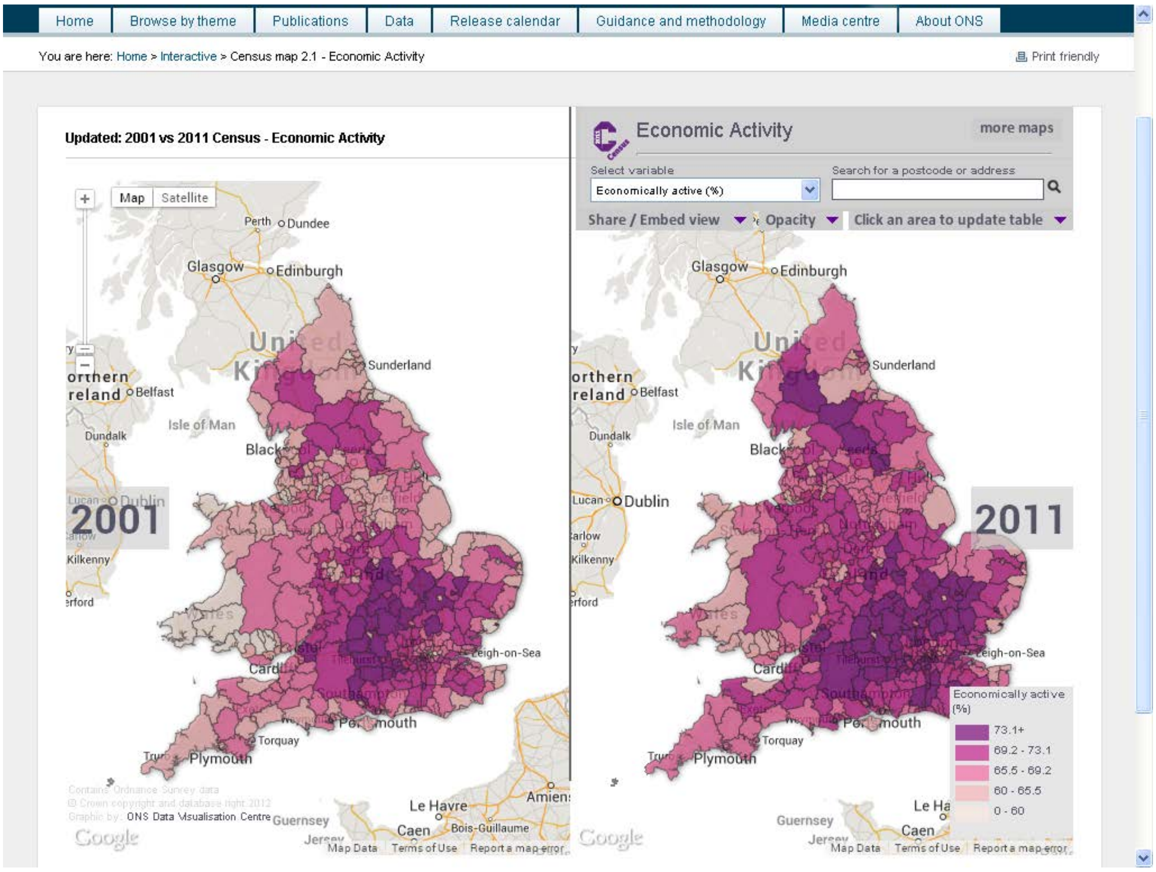


Figure 1: ONS visualisation of economic activity in England and Wales 2001 and 2011

Conceptual Display Issues (1)

Figure 1 shows GIS chloropleth maps where the colour representations on both maps are generated by quintiles taken from the 2001 census, although this is not made explicit in the key. The same boundaries are used for the 2001 and 2011 maps. The 2001 map shows five colour levels which show the *relative* level of economic activity – so it is only the geographic relative distribution that is presented in the visualisation – one has to go to the key to get some of the information about actual values (and you only get the full measurement limits of the 3 groups which are not at either extreme).

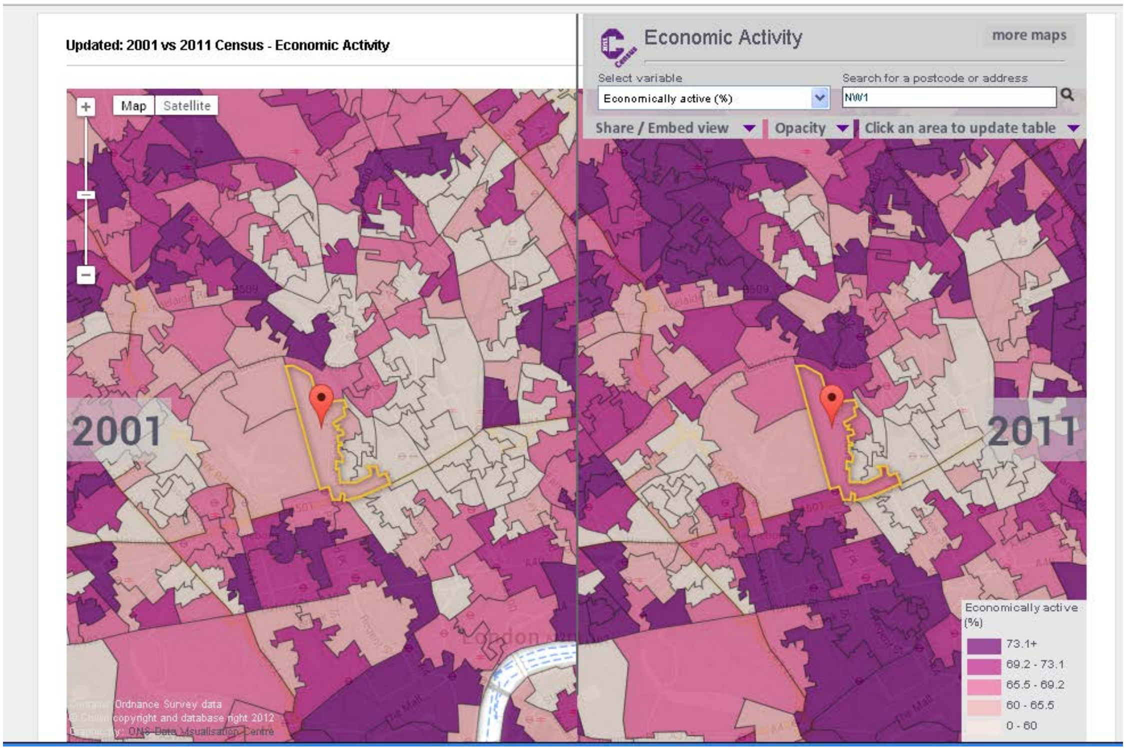
The data being displayed are themselves percentages (of the population in an area who are economically active), so for example the third quintile represents the 40th – 60th percentiles of the distribution, and the key tells you that the areas shown range from having 65.5% to 69.2% of the population who are economically active. This creates perceptual problems analogous (in some ways) to auto-scaling in Excel. Here it is compounded further by the mapping of unequal measure intervals onto equal entities (the quintiles). This is even more difficult because the 'measures' are themselves percentages of the population who are in a particular classification.

Conceptual Display Issues (2)

The second map uses the the quintile boundaries taken from the 2001 distribution of economically active people. The proportions represented by each colour do not each represent 20% of the 2011 distribution, and there is no way to tell what proportion of the distribution falls into each interval. The areas on which the data is based are chosen to represent roughly equal numbers of people, but are shown by their geographical size, which vary considerably. Large sparsely populated areas therefore occupy more space than the densely populated areas in cities which is why even the map for the 2001 data does not look as if there are 5 groups of regions, each with the same number of members, on the map. Comparing the colours on the maps for 2001 and 2011 therefore requires a deep understanding of the display – any change represents a shift in the *absolute* level of economic activity in that area between the two census counts. This mix of relative and absolute measurements is unlikely to be grasped by many viewing this visualisation.

Conceptual Display Issues (3)

There is an interesting comparison between the visual impressions of the two maps in figure 1 those in figure 2 where the 'data area' no longer has the surrounding 'not relevant' area which enables the reader to see two distinct entities for 2001 and 2011. The visualisation shown in figure 2 has no physical separation between the two maps: the reader has to work really hard to see two distinct entities, which makes making comparisons between them somewhat problematic. This is perhaps the one of the 3 display issues which is relatively easy to fix because the design could build in a gap between a pair of maps that would allow one to see two different maps for comparison in cases where there is not a surrounding 'not relevant' area.


Updated: 2001 vs 2011 Census - Economic Activity

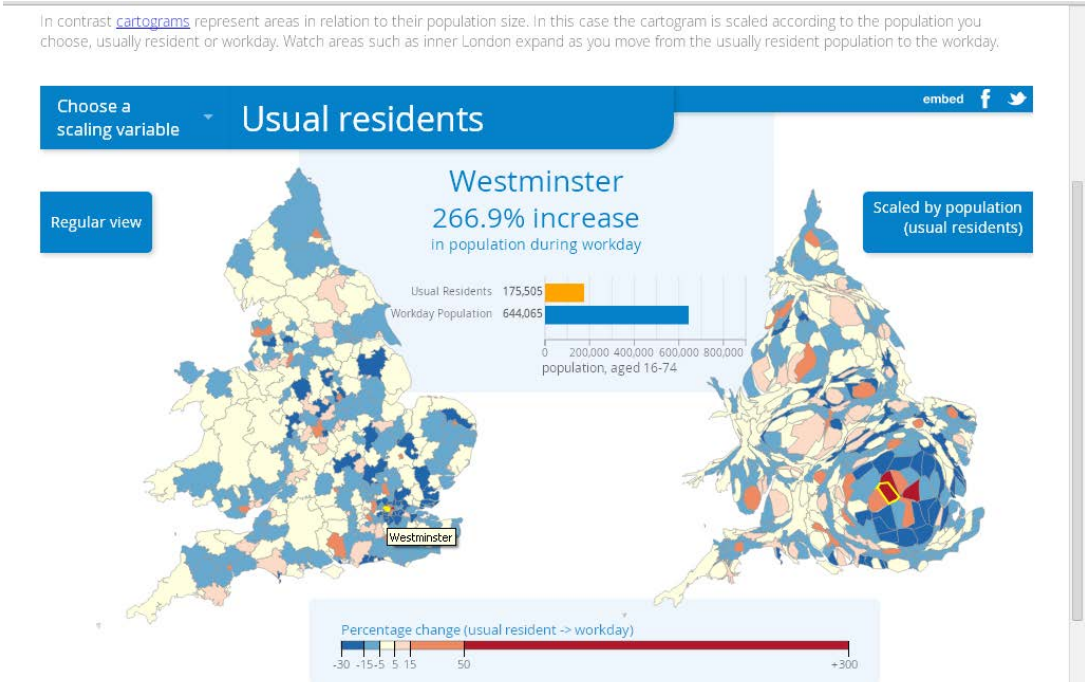
Economic Activity

Select variable: Economically active (%)

Search for a postcode or address: NW1

Figure 2: ONS visualisation of economic activity around London NW1 2001 and 2011

Recent innovations in visualisations have developed cartograms which seek to address some of the underlying difficulties discussed in the second issue above – namely that the geographical area is not proportional to the part of the population it is representing. Figure 3 shows the ONS visualisation of the percentage change in the population during the workday, using the usually resident population as the base for each area. The representation on the left is the typical chloropleth GIS map, with a key presented rather differently from that used in figures 1 and 2. The cartogram on the right scales each area in proportion to the size of the population in that area. Of necessity this means that the 'map' is distorted – Westminster has been highlighted on both maps to illustrate the point – in the map it is hardly visible while in the cartogram it is substantial, but the price to be paid for this improvement is the loss of the familiar orientation of the geography of England and Wales. However, shown together as they are in the ONS visualisation, one might hope that the user can synthesise a coherent understanding of the data story from the two images.



In contrast cartograms represent areas in relation to their population size. In this case the cartogram is scaled according to the population you choose, usually resident or workday. Watch areas such as inner London expand as you move from the usually resident population to the workday.

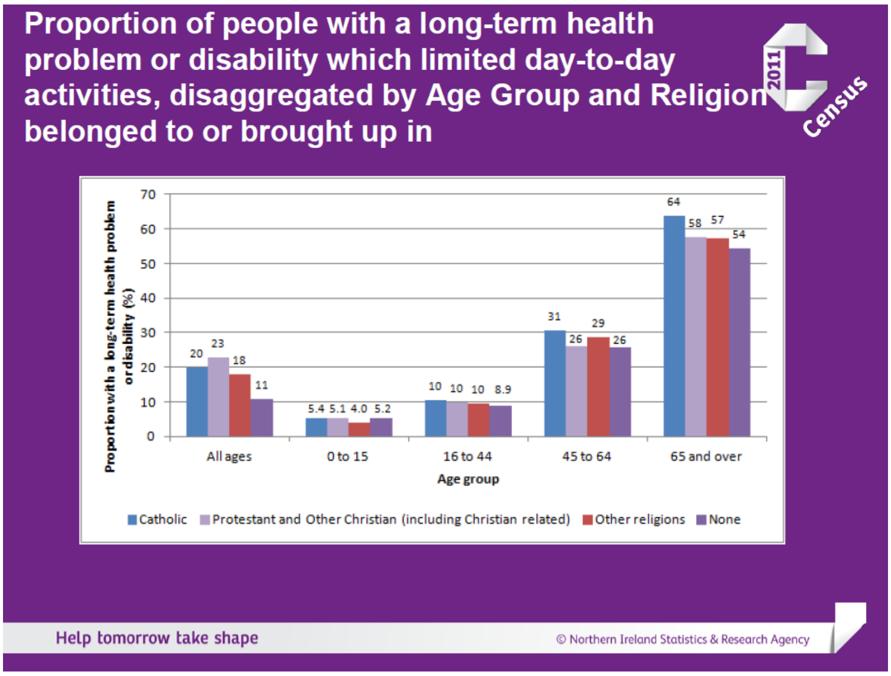
Figure 3: ONS visualisation showing the percentage change in population during the workday

POTENTIAL BENEFITS FROM AUTOMATING THE PROCESS

Nicholson, Ridgway & McCusker (2013a) reported on a collaboration between the SMART Centre and the Northern Ireland Stratistics Research Agency (NISRA) to produce a visualisation of the Northern Ireland 2011 Census table DC2309NI, (long term health problems by religion by age by sex). The data in this table is an example of Simpson's Paradox – in all age bands, and for both males and females, there is a higher proportion of Catholics suffering long term health problems or disability than there is of Protestants, yet in the population as a whole the proportion of Protestants suffering these problems is higher than the proportion of Catholics. The reason being the demographic shift where the proportion of Catholics is larger in younger age groups – where the rate of incidence of long term health problems is lower. NISRA had produced a static graph in the press briefing (NISRA, 2013) which accompanied the data relaease which showed this as effectively as is possible with a single static, but wanted to see how much more was possible with a dynamic, interactive display.

Many 'do it yourself data visualtion' sites offer little in the way of safeguards to ensure that the user will display data in a sensible way. When the SMART Centre have been constructing bespoke visualisations we have been greatly exercised at times at how to make the stories in the data accessible while mimimising the potential for misinterpretation, while still providing as much flexibility as possible. One of the mechanisms by which we did this was where the data showed the proportions in each of several categories of a characteristic, we fixed that characteristic as the comparator variable in the display, so the user always saw the 'distribution' across the categories of that variable. The rationale was that it was much easier for the user to keep a clear understanding of what the percentages were of – because each group of bars in a block always added to 100% - the display always showed how the group identified by specific levels of the other demographic characteristics was distributed across the different levels of the comparator variable. In table DC2309NI, the variable of interest to NISRA was the proportions who suffered 'a little', 'a lot' or 'not at all' from long term health problems. However, because they wanted to display different age groups and different religions, NISRA chose to display only a single value combining 'suffer a little' and 'suffer a lot' categories together (figure 4). The display presents the most relevant information possible from that table but the reader does need to understand that the bars each represent a proportion of a different group (a specific combination of age and religion).

The SMARTplotter is able to display more information, so we were able to keep the breakdown by sex that the NISRA graphic omitted, and we could keep the 3 levels of the response to the long term health enquiry (i.e. ‘a little’, ‘a lot’ and ‘not at all’) separate by the use of two sliders (see Nicholson, Ridgway & McCusker, 2013a for fuller details).



Proportion of people with a long-term health problem or disability which limited day-to-day activities, disaggregated by Age Group and Religion belonged to or brought up in

Help tomorrow take shape

© Northern Ireland Statistics & Research Agency

Figure 4: NISRA press release graph showing health data in Northern Ireland

The difficulties with respect to automating the visualisation of tables such as these lie in the range of options available to the user of the data – data is available at a number of geographic levels, and it may be of particular interest to compare a number of regions of similar size, or it might be of particular interest to make comparisons of nested areas. For some data sets there may be an obvious characteristic to use as the base for the percentage calculation, but for others there may be more than one of the characteristics where useful insights could be gleaned from seeing the distribution across the levels of response – for example, sex, the age distribution, or ethnic distribution could all be of interest in considering health data or employment data. The interface to the census data currently offers a choice of geographic level, and the option to take all or only some levels of other characteristics. For example one can choose to have all age levels or only some, and for ethnicity there are some broadly aggregated groups which are also available at a disggregated levels. Once these choices are made, the data is returned as counts, and any calculation of proportions or rates has to be carried out by the user.

For users there is currently a modest investment of time required to process the data into the rates which the user wishes to consider, but much more time is required to make sense of the stories in the data – the data we used for the visualisation of DC2309NI was 9 tables of 5 rows by 5 columns. The SMARTplotter allows the exploration of the multiple comparisons across different characteristics to be carried out much more quickly than can be done in table format. We believe that if the process can be automated so the output of the analysis is displayed in the SMARTplotter as an alternative to tabular form, it will allow researchers to be much more effective in exploring the sort of rich data sets which linked administrative data promises to deliver.

DISCUSSION AND IMPLICATIONS FOR CURRICULUM DEVELOPMENT IN THE UK

The production of official statistics requires a huge resource investment in any country, with most of the resources utilised in the collection and validation of the statistics. The wealth of information contained in these statistics is underused because there is not sufficient capacity available to analyse and interpret them. The automation of data collection is accelerating the rate at which data is accumulating, mostly large observational data sets with multiple factors.

There is an urgent need for improving the ways in which we can make sense of data with these properties, but they currently do not form part of the UK curriculum which may mean that the difficulty in making use of these resources will continue for many years. New curricula introduced in New Zealand and South Africa include some multivariate reasoning embedded from an early age, yet in the UK it does not appear anywhere. There is an opportunity currently as the government seek to develop a new qualification for 16 to 18 year olds (Department for Education, 2013), with one of its aims being to support quantitative reasoning for students in the humanities and social sciences who are not taking a specialist mathematics course.

The British Academy (2012) produced a report articulating the rationale for improving the quantitative reasoning of these students. However, the usefulness and relevance of traditional inferential statistics for this group is dubious - hypothesis testing is often abused within social sciences where the data collected rarely comes close to meeting the structural requirements for conducting such tests, and then researchers compound the problem by often stating 'conclusions' in inappropriate language - so they make claims that the logic of hypothesis testing is incapable of supporting, no matter what the data.

Traditional inferential statistical methods are both inappropriate and not helpful when dealing with very large observational data sets - not helpful because, with the very large data sets available now, almost all differences will be significant irrespective of whether they are important differences. Concepts such as effect size are much more relevant than significance in this context. Experience of looking at disaggregated groups can illuminate concepts such as effect size and interactions – for example using data visualisation tools - where one can see if the 'data story' for an aggregated population stays broadly the same when it is disaggregated by different characteristics (age, ethnicity, sex, social or economic status etc.).

Nicholson, Ridgway & McCusker (2013b) and Jerome (2013) offer some examples of materials which could substantially improve quantitative reasoning skills in these groups while being accessible to the teachers who are expected to deliver such a course in school, many of whom are likely to be outside the mathematics department. Ridgway & Smith (2013) offers a more detailed analysis of the benefits of making the most of the opportunities offered by open data initiatives to improve statistical literacy in the general population and at school level. There is a need for the broad community of creators and users of statistics to shape the curriculum so that it reflects contempoary good practices, and enables people to better understand their worlds.

REFERENCES

British Academy (2012) *Society counts*. British Academy, London. (Available from http://www.britac.ac.uk/policy/Society Counts.cfm)

Department for Education (2013). *Introduction of 16 to 18 core maths qualifications*. Policy statement available for download at www.gov.uk/government/publications

Jerome, L. (2013). Teaching about health and the NHS. *Teaching Citizenship, 36*, 22.

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

Nicholson, J., Ridgway J., & McCusker, S. (2013a). Integrating the use of Official Statistics into mainstream curricula via data visualisation. *Proceedings of the first Joint International Association for Statistics Education and the International Association for Official Statistics, Statistics Education for Progress, Macau.*

Nicholson, J., Ridgway J., & McCusker, S. (2013b). Health, wealth and lifestyle choices - Provoking discussion on public spending. *Teaching Citizenship, 36,* 23 – 27.

Pullinger, J. (2013). Statistics making an impact. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(4), 819-840.

Ridgway J., McCusker, S., & Nicholson, J. (in press). *Research on big data to enhance statistics education.*

Ridgway, J., & Smith, A. (2013). *Open data, official statistics and statistics education – Threats, and opportunities for collaboration.* Keynote Talk: First Joint International Association for Statistics Education and the International Association for Official Statistics, Macau.

Smith, A. (2013) Data visualisation and beyond: A multidisciplinary approach to promote user engagement with official statistics. *Statistical Journal of the IAOS, 29*(3), 173–185.