INTEGRATING THE USE OF OFFICIAL STATISTICS INTO MAINSTREAM CURRICULA VIA DATA VISUALISATION

NICHOLSON James, RIDGWAY Jim and McCUSKER Sean

SMART Centre, Durham University,

United Kingdom

Contact email: j.r.nicholson@durham.ac.uk

ABSTRACT

There has been a great deal of concern in recent times about the capacity of social science students at all levels to cope with the demands of quantitative methods in the curriculum. The Nuffield Foundation funded a project Reasoning from Evidence to produce some data visualisations and associated curriculum materials to support the teaching of social science at Advanced-level (ages 16–19 in the UK), using data sets relevant to the Sociology curriculum but which have usefulness across other subject areas also. Social sciences deal routinely with contexts in which the population under consideration is not homogenous. The data used is often presented in aggregated form which disguises the characteristics of the subgroups – whether these are by ethnicity, age, socio-economic status, region or some other categorisation. This paper reports on the development of materials using data on health and on the UK public disorder of August 2011. We report on further development of data visualisations using the 2011 UK Census data.

INTRODUCTION

A logical approach to the role of data and any theory looking to explain it would be to examine the data, consider what the stories in the data are, and then look to provide theoretical perspectives which are consistent with the data. Similarly, if decision makers are proposing a new policy or business plan which is informed by evidence, allowing people access to the evidence, and providing the commentary on the rationale for the policy or business plan would be logical. The focus of this paper is on the use of these data visualisations in an educational context, but we believe that similar arguments pertain in both business and policy contexts.

Historically, it has not been possible to access data relevant to complex contexts, such as those routinely encountered in social sciences, and in business. Current sociology textbooks for UK pre-university courses ('A-level courses') aimed at 16-19 year old students report the associations between a variable of interest and individual factors one at a time - for example in discussions of educational attainment, performance of boys and girls is presented, then the performance of different ethnic groups, then the performance of pupils who are and are not eligible for free school meals (a proxy for socio-economic status). Several pages of text identify the trends over time, and discuss some theoretical explanations of each association. However mixed in with these two strands there is another in which the text identifies specific demographic groups whose behaviour is not consistent with the general patterns of associations between factors which have been described, and tries to provide theoretical explanations as to why these groups have their distinctive behaviour.

Many students in Social Sciences (and their teachers) do not feel confident in mathematics. Social data is more complex in its structure than the data that students work with in mathematics where they 'learn' statistics. If data is presented in tabular form, even one table has a lot of data and students often need to work with a number of related tables to get a good understanding of the evidence. Political and journalistic commentary often does not help by concentrating on single headline figures, even in situations where one would not accuse them of cynically focussing on an aspect of the data which supports the general argument they want to promote, while ignoring other aspects which do not support their stance.

The Nuffield Foundation funded a project *Reasoning from Evidence* to produce some data visualisations and associated curriculum materials to support the teaching of social science at Advanced-level, using data sets relevant to the Sociology curriculum. The Economic and Social Research Council (a UK funding body) is funding a current project to produce data visualisations of 2011 UK Census data, which offers further opportunities for course developers, and teachers, to integrate the use of real complex data into mainstream curricula and into college lecture theatres and school and college classrooms – where the emphasis is on the stories in the data as a precursor to trying to produce theoretical frameworks which are consistent with the observed patterns of behaviour.

Next, we discuss three examples of how particular multivariate data visualisations can be used in a variety of mainstream curriculum areas and discuss implications for curriculum development in mathematics and other curriculum areas.

CASE STUDY 1: USING HEALTH DATA TO PROVOKE DEBATE ON PUBLIC SPENDING

As part of the *Reasoning from Evidence* project the SMART Centre has produced visualisations of data on general health in the adult population and for young people, as well as more detailed data on alcohol consumption in both groups, and data on mental health issues in the population. These visualisations are immediately relevant to Sociology, where inequalities in health are explicitly considered, but a number of other subject areas can make use of them from different perspectives.

The UK health budget for 2012 – 13 is £108.9 billion with the NHS providing free (at the point of use) care for the 63.2 million people resident in the UK. That is an average expenditure of over £1,700 per person annually by the government. A person's health is one of the major factors in determining their quality of life, but health is a difficult concept to define or measure as it has many facets. The potential demand on the National Health Service (NHS) far outstrips the capacity of the NHS to deliver, so choices have to be made as to the allocation of scarce resources. Determining priorities for spending on health in a democracy where advances in technology and medicine are constantly creating possibilities of significant improvements in health is an extremely sensitive issue.

Debates about the personal effects of lifestyle choices and some of the reasons for making or avoiding difficult decisions belong (in the UK education system) within a personal, social and health education (pshe) framework, but broader issues of public policy, rights and responsibilities and the role of the NHS have more of Citizenship focus (Jerome, 20013). However, if one starts to consider some of the detailed issues it is clear that subjects like Politics and Government, Psychology, Economics, Critical Thinking and General Studies could all approach this debate from their own perspective, and the data visualisation provides an open entry into cross-curricular development as well as a valuable standalone resource within any of those subject areas.

Many important concepts are introduced within different curriculum subjects with a fairly theoretical and abstract explanation and definition. We contend that a more effective way to communicate some of these big ideas is through the use of data visualisations to provoke stories from the data about core subject content, and then retrospectively to examine the rationale for the choice of measurement, the reliability and validity of the data sources etc.

In the UK, data are collected about blood pressure, obesity, smoking, alcohol and participating in exercise, for males and females of different age groups, mostly for 2003 and 2010. Figure 1 shows a screenshot of the (self-reported) alcohol consumption levels of 16 – 24 males and females in 2003. Males drink more than females, even after allowing for different definitions of light, moderate and heavy drinking for males and females to reflect different physiologies. The panel beside the data explains the definitions used. Moving the slider pointers at the bottom of the display allows one to explore data for other age groups, and for 2010. In 2010, 16-24 year old men and women both reported drinking less alcohol than their counterparts in the same age groups in 2003. Moving the 'consumption', 'sex', 'age' and 'year' labels (via drag and drop) one you to display other comparisons directly.

Discussion about why *light*, *moderate* and *heavy* alcohol consumption correspond to different amounts of alcohol for males and females provides a much better introduction to issues of

measurement than an abstract, theoretical exploration of the challenges of measurement. Issues such as the difference between data based on self-report, and those based on clinical studies, are more meaningfully and memorably raised through the use of well chosen examples (as are issues relating to sample size). The richness of these data resources, and the accessibility of the stories the data has to tell within a short period of exploration, is the reason why they provide multiple curriculum opportunities. Here, each of the tabs contains a multivariate data set, which is itself richer than any data that students could routinely consider in print materials. Because health is multi-faceted the opportunity to look at a number of aspects (alcohol, smoking, blood pressure, obesity, activity participation) within a single resource offers insights into the complexity of the area, but without being intimidating.

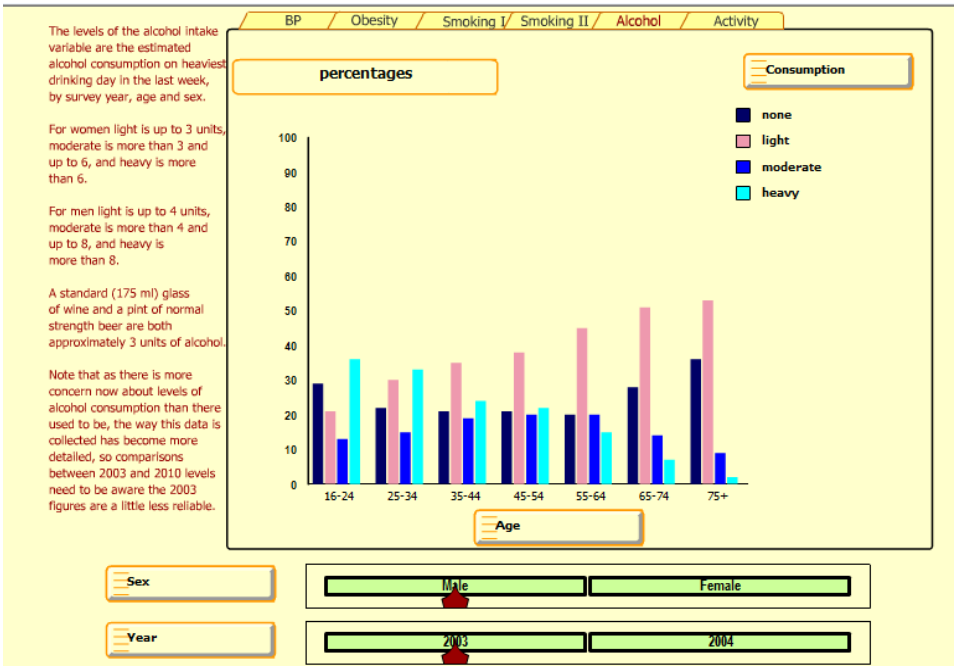

Figure 1: alcohol consumption levels for males in 2003 across different age groups.

CASE STUDY 2: EXPLORING HEALTH PROBLEMS IN A LOCAL CONTEXT

Local communities often have issues which are of particular significance to them, but might not be considered an issue within the population at large. This is the case in Northern Ireland, where Protestant and Catholic religious affiliations are strongly aligned with political affiliations and aspirations - such as those for a United Ireland. Historically, the issue of funding levels across different communities has been a contentious matter and Northern Ireland continues to wrestle with appropriate ways to deal with the legacy of historical inequality.

Simpson's Paradox is often difficult to explain, but context can be a powerful ally in articulating difficult statistical concepts. It is obvious to almost everybody that younger age groups will have lower proportions of people suffering from long term health problems or disability than older age groups will have. It is widely known in Northern Ireland that there is a substantial difference in the age profiles of Catholic and Protestant communities, with larger numbers of Catholics in the younger age groups. This creates the sort of situation in which Simpson's Paradox can appear, and it does: there are more Catholics suffering long term health problems or disability than Protestants in every age group, yet there is a smaller of proportion when the population is considered as a whole.

Rich interactive displays can overcome some limitations inherent in static displays in helping citizens understand complex ideas. Figure 2 shows a graph relating age to health for different religious groups created by the Northern Ireland Statistics Research Agency (NISRA, 2013). The graph shows the proportion of people with *some* long term health problem or disability, for various

age groups and religious groupings. However, the published data actually shows these people in two groups – those whom this affects *a little* and those who it affects *a lot*, and it is also disaggregated by sex in the published table. The static display can not easily display the full scope of the available information. NISRA also made a choice that they would display a proportion with a certain characteristic (suffering some long term health problem or disability) for each demographic group – a combination of age range and religious affiliation. Figure 3 shows the interactive visualisation which allows all the information to be displayed and explored by the user. The flexibility of the interface also allows three versions to be available through the use of the tabs at the top of the display. In the first one, the actual numbers in each group are displayed; in the second (% - fixed) the comparator variable is fixed in the top right corner of the display so the user always sees the health distribution for each combination of sex, age and religious affiliation – which makes it easier for the user to keep clear what the basis of the percentage calculation is. In the third display the user is given the freedom to alter the position of the health status label, which allows the same comparison to be made as was made in the NISRA press release – with *none* being the complement of the *some* level used in it.

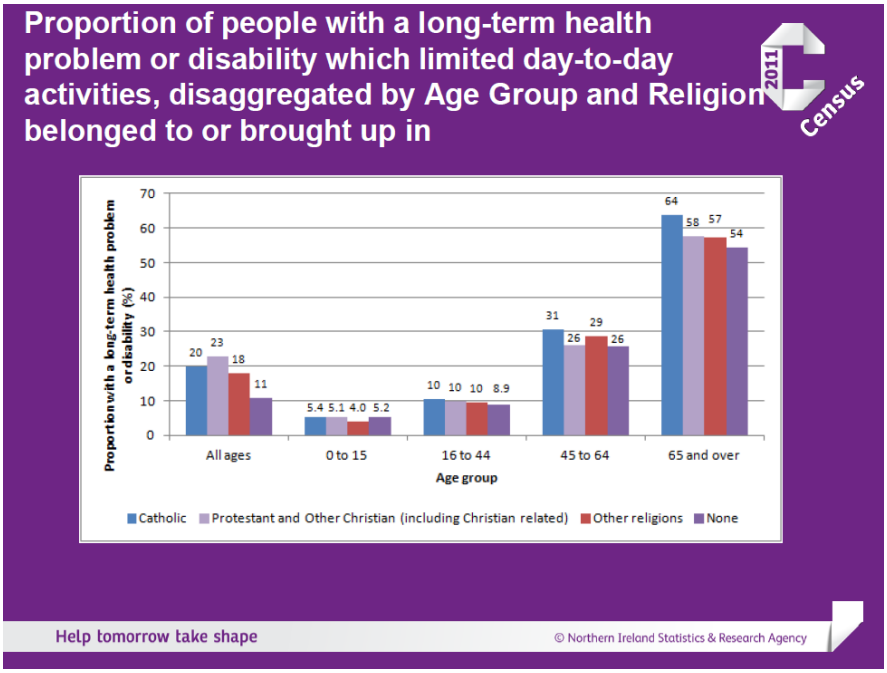

Figure 2: NISRA press release graph showing health data in Northern Ireland

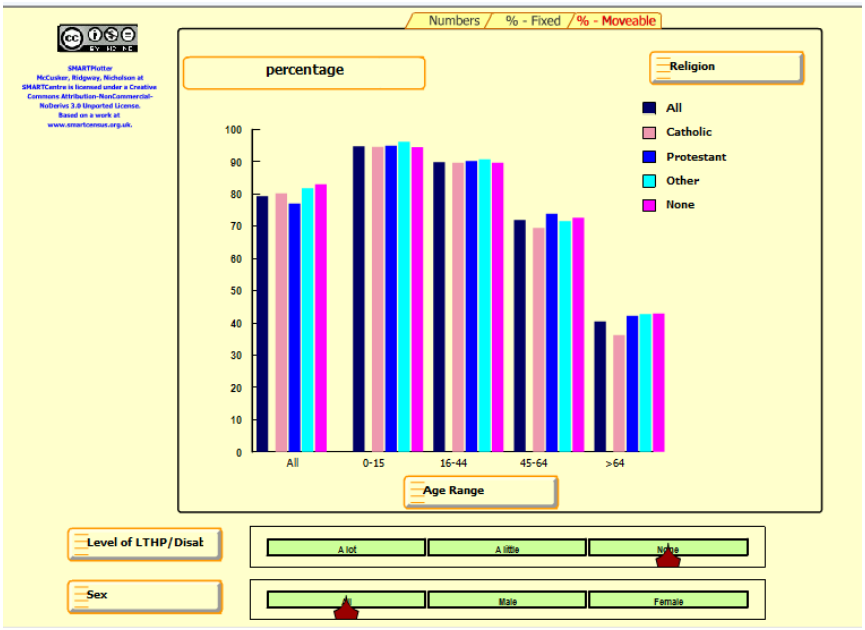

Figure 3: screenshot of the interactive visualisation of the same data set as in Figure 2.

There are some non-trivial issues here. Essentially when the data is displayed as a percentage, it is akin to a conditional probability – here it is the proportion with a particular level of health given a particular combination of age, sex and religion. However, one could also look at the distribution of religion affiliations at each combination of age, sex and level of health, or the gender balance within each combination of age, religion and level of health, or indeed the age distribution for each combination of sex, religion and level of health.

These different conditional constructions offer possibilities of quite different insights from the data, so which should be displayed? Currently, we do a modest amount of intermediate processing in order to create a particular display - but it does mean that the choices have to be made in advance, and often not by the person who will use the visualisation to interpret the data for a particular purpose. Ideally, the user should be able to choose the basis for construction of percentages, aggregation of groups etc., and be able to explore the relations between associations in an interactive display.

CASE STUDY 3: UK PUBLIC DISORDER IN 2011

In early August 2011 there was a 4 day period characterised by serious public disorder in a number of English cities. There was considerable public debate about the reasons for the disorder – an unlawful killing by the police? Outrage at political actions targeted on poor people? - and also about the profile of those involved in the disorder – hardened criminals simply being opportunistic? Law abiding citizens driven to extreme political actions? The Ministry of Justice published a special Statistical Bulletin a month later, which generated public debate, and much political commentary from all sides. They published four further bulletins over the following twelve months, providing an opportunity to consider the development over time of the pattern of criminal prosecutions.

The implicit assumption made by political and journalistic commentators when the first bulletin appeared was that the patterns of criminal behaviour observed in that data (the cases which were brought to court in less than a month) were representative of the totality of the criminal behaviour which had taken place. Figure 4 shows that this is not the case – only about half of the cases of robbery had appeared in the first bulletin, but a much higher proportion of the violent disorder cases were dealt with very quickly. It is not difficult to postulate reasons why more of the violent disorder cases were brought before the courts very quickly, but those reasons mean that commentators might be expected to appreciate that the data in such a statistical bulletin is likely to have some systematic bias.

The SMART Centre have produced visualisations of each of the five statistical bulletins on the public disorder which were released between September 2011 and September 2012 as well as the one shown in Figure 4 which summarises the data in all five bulletins and shows the development over time. Individually and collectively these provide resources which we believe could make an important contribution to a number of subject areas.

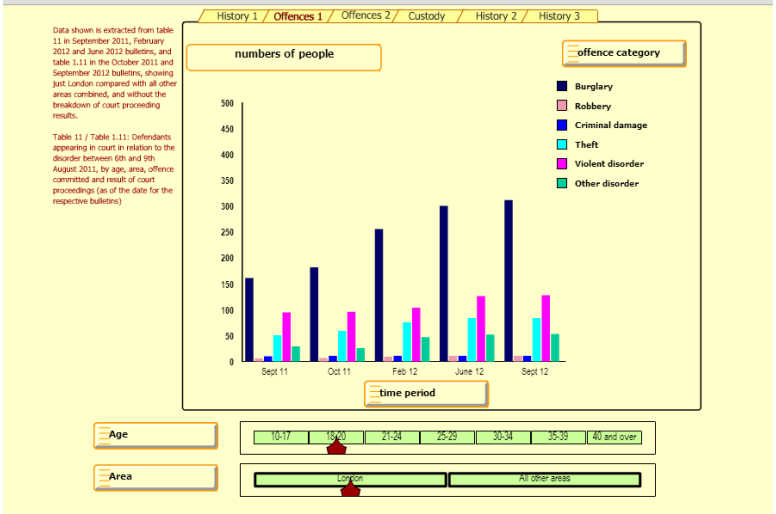

Figure 4: the numbers of different offences by 18 – 20 year old males in London in the public disorder

DISCUSSION

The internet provides access to a wide variety of resources to support debate across a range of issues and from a number of different perspectives. There is also unprecedented access to real data pertinent to these important debates, but often that data is not in a form that the stories in the data are accessible. We believe appropriate data visualisations offer opportunities for students to use real data to support plausibly valid arguments and to dismantle demonstrably false arguments, and consequently offer a new dimension to education in many social sciences. More detailed discussion of the health and public disorder data visualisations can be found in Nicholson, Ridgway & McCusker (2012, 20013), and the visualisations themselves are available to download at www.dur.ac.uk/smart.centre/nuffield. Visualisations of detailed characteristic census tables are available at www.smartcensus.org.uk.

We believe that the use of these visualisations offers opportunities to develop key statistical literacy skills and to develop some capacity to evaluate statistical evidence. Direct access to metadata, and the ability to reorder data in order to ask new questions leads directly to a questioning approach to data – how was it collected? and by whom? is there enough data to provide robust evidence? what is being measured (e.g. what is *poverty*)? does the data support the argument being made? However, there are major barriers to curriculum reform in the formal statistics curriculum in the UK. A disproportionate amount of time is devoted to the calculation of summary statistics and drawing by hand of graphs, both of which are always automated in the real world. There is not even an attempt to expose students to multivariate data (and so no attention to key associated ideas). The innovations implemented in the New Zealand and South African curriula need to be reflected in changes to the UK curriculum in the near future. In the near future, it may well be the case that students receive their grounding in the statistical ideas that are most important, and most relevant and directly useful to daily life, in subjects other than statistics and mathematics.

ACKNOWLEDGEMENT

This research was funded by the Nuffield Foundation Grant EDU/33713, *Reasoning from Evidence*, and ESRC Grant ES/K004328/1 *INCENSE*. The views expressed here are entirely those of the authors. Free teaching materials can be viewed at http://www.dur.ac.uk/smart.centre/ using the links on the left to freeware and to Nuffield.

REFERENCES

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

Nicholson, J., Ridgway, J & McCusker, S. (2012*)* The Summer Riots 2011: Lessons to be learnt. *Teaching Citizenship, 34, pp 24 – 27*

Nicholson, J., Ridgway, J & McCusker, S. (2013*)* Health, Wealth and Lifestyle choices: Provoking discussion on public spending. *Teaching Citizenship, 36, pp 23 – 27*

NISRA (2013) Detailed Characteristics for Northern Ireland on Identity, Religion and Health: press release published 16/05/2013