SP4b1 (linked with SP4b): New Developments in "Data" – Humphrey Southall

Twenty years ago, "social statistics" meant overwhelmingly the analysis of data from social surveys: small surveys carried out by individual researchers or larger surveys mostly obtained through the UK Data Archive but actually carried out, or at least funded, by the government – the largest and most widely used being the Census of Population. So what has changed?

- **Government surveys** still exist, but they are threatened by government cost cutting and popular resistance. In Radstats we tend to discuss the former much more than the latter, but I think we need to recognize that there are real issues with statistical surveys being seen and resisted as part of state surveillance. One consequence is declining response rates, but another is ONS developing more and more restrictive rules on statistical disclosure. There are issues here both of government policies and popular attitudes to statistical surveys. The future of the census is a major issue in itself.

- **Administrative data:** Some of the earliest applications of automated data processing were to statistical surveys, Hollerith counter-sorters being used to analyse the 1890 US census. As more and more aspects of the economy and public administration have been computerized, more and more data capable of statistical analysis have been created through the routine operations of companies and government, although additional technical, organizational, legal and ethical barriers have limited and slowed analytic use. Two examples: the vast bodies of data generated in retail, aided by loyalty cards which associate purchases with customers – but who can access the results; and the use of Police Recorded Crime data, and their pros and cons relative to the National Crime Survey.

- **Freedom of Information, open data (and privatized data)**: Who owns data, especially data produced with public funds? Anyone who remembers the attempts to maximize data licensing income from the 1991 census knows that some things have got much better: vast amounts of data is genuinely freely downloadable under the Open Government License which allows you to do just about anything with it. However, privatization means that some important data sets (especially in the health sector?) are now private property. There are large issues with access to trials data from drug companies (and car manufacturers?). There are also some continuing issues with access to data from academic research: in the US, if it is publicly funded it is in the public domain, but in the UK universities are supposed to make a profit from research outputs if they can – which rarely happens, but often leads to access being blocked.

- **"Big Data"** is an inherently vague term, and most clearly refers to techniques for managing very large bodies of data, using software such as Hadoop, rather than to

any particular set of analytic techniques. “Big data” mostly covers relatively unstructured non-quantitative data: text, images, even sound; but of course statistics can be derived from any large collection of digital stuff. The great superficial attraction of “big data” is that, as the world generated more and more digital stuff through its normal operation, we can find out what is going on in society not by expensively doing surveys but just by dipping our fingers in some vaguely relevant part of this digital flow: these days, even very powerful computers are far cheaper than survey teams – and less “intrusive”. “Administrative data” were noted above, often involve very large data sets but analyzing large quantities of welfare benefit payments, say, has little to do with the “big data” hype. More relevant is, say, Google Flu Trends: analyzing vast amounts of very current textual information to provide early warnings of epidemics. There is a level at which this clearly works, but it does not really allow us to scrap formal systems of disease notification.

- **Volunteered data:** The internet has clearly created new ways of gathering in data from many different people in many different places, and Wikipedia and the Citizen Science Alliance are very different examples of this being used in arguably democratic ways. Statisticians, on the other hand, do not seem to have done much to exploit this potential – and I think worries about self-selecting samples are only a partial excuse: more could be done. Maybe the clearest counter-example is YouGov, but I don't know as much about this as I should.

- **Crap data:** I don't know what else to call this, but think we should cover it: the very dodgy area of virtually fraudulent surveys carried out for marketing and PR purposes, often online. Petra Boynton could write something great about this, and there was a wonderful example given at the recent RSS statistics and the media meeting of a company specializing in doing this kind of “survey”. Important to write about partly because for many people these are what “surveys” are, and leads to all the terrible “surveys I did on Facebook” I find myself marking.