Small Big Data: Using multiple data-sets to explore unfolding social and economic change.

Emily Gray, Will Jennings, Stephen Farrall[1] and Colin Hay

Forthcoming in *Big Data & Society*.

We would like to acknowledge the support of the ESRC in the form of research grant ES/K006398/1.

[1] Corresponding Author; s.farrall@sheffield.ac.uk

Abstract

Bold approaches to data collection and large-scale quantitative advances have long been a preoccupation for social science researchers. In this commentary we further debate over the use of large-scale survey data and official statistics with 'big data' methodologists, and emphasise the ability of these resources to incorporate the essential social and cultural heredity that is intrinsic to the human sciences. In doing so, we introduce a series of new data-sets that integrate approximately thirty years of survey data on victimisation, fear of crime and disorder and social attitudes with indicators of socio-economic conditions and policy outcomes in Britain. The data-sets that we outline below do not conform to typical conceptions of 'big data'. But, we would contend, they are 'big' in terms of the volume, variety and complexity of data which has been collated (and to which additional data can be linked) and 'big' also in that they allow us to explore key questions pertaining to how social and economic policy change at the national level alters the attitudes and experiences of citizens. Importantly, they are also 'small' in the sense that the task of rendering the data usable, linking it and decoding it, required both manual processing and tacit knowledge of the context of the data and intentions of its creators.

Big questions

The shift towards the use of 'big data' made a seemingly 'explosive' entrance into the social sciences in 2011 (Burrows and Savage, 2014:1). While there is no definition of the term, it is typically used to denote data from online sources (i.e. web usage), public records (e.g. geocoded reports of crime incidents, ordnance survey data) or transactional data (i.e. phone calls to public services, financial expenditure or insurance claims) from commercial enterprises that is continuously updated in vast quantities (Savage and Burrows, 2007; Manovich, 2011). Beyond the epic contents of 'big data', it has brought with it fundamental questions about the nature of social science data, the quality of the knowledge generated from it and the epistemologies that underscore traditional scholarly enterprises. Such is the breadth and depth of 'big data' that Housley et al (2014) argued that it made for "uncomfortable" (2014:2) comparisons with the 'bread and butter' of more traditional data sources, such as episodically generated data-sets (c.f. Savage and Burrows, 2007; Mayer-Schönberger and Cukier, 2013).

While there is little doubt that the features of 'big data' compel social scientists to redefine the nature of social knowledge and the validity of our research methods (Savage and Burrows, 2007), national surveys and official statistics remain crucial to our enterprise. Conducting research on long-term attitudinal trends or patterns of crime for example, by definition, involves the close inspection of historical processes, of which the most reliable data is habitually derived from national surveys and official indicators – and for which 'big data' cannot be created, either due to the impossibility of retrospectively imputing measures of social attitudes or because the manual extraction of data from paper records is either too costly and time-consuming or where missing data may not be random. Furthermore, many large-scale national surveys, such as the Crime Survey for England and Wales (CSEW),[2] the British Social Attitudes Survey, the British Election Study and the Labour Force Survey continue to be updated on a regular basis. This means it is possible to use this data to understand *dynamic* interrelationships and to observe and model both rates of change and lagged processes over time (Pawson and Tilley, 1997). In recent years computational technology has broadened the scope of statistical techniques available to us (c.f. Mayer-Schönberger and Cukier, 2013). It is now possible to combine high volume[3] data-sets from a variety of sources, explore dynamic social processes through advanced quantitative methods and organise the data in such a way as to observe shifts at individual *and* aggregate levels. By collating data over large periods of time, it also allows for robust analyses of particular items where responses or subgroups may be rare, for example, male victims of domestic or sexual violence (Gadd et al, 2002) or to dissect three types of time-related effects such as age, period, and cohort analysis (Ryder, 1995)[4].

In sum, repeated cross-sectional surveys afford researchers distinctive opportunities to assess long-term temporal processes to address complex research questions. Attention to historical resources has been underlined by Rock (2005) who has stressed that criminological researchers – as well as other social scientists – should be aware of a manifest 'chronocentrism' that frequently "neglect[s]

[2] The CSEW was originally known as the British Crime Survey.

[3] The Crime Survey for England and Wales is updated on an annual basis and typically has in excess of 40,000 respondents per year, for example.

[4] Specifically, age effects represent typical developmental changes in the life course; period effects arise via cultural and economic changes that are exclusive to the study-period, while cohort effects are the core of social change and represent the effects of formative experiences (Ryder 1965).

what is old" (2005:20), overlooks the accumulation of data and works against the collective structure of knowledge. Similarly, scholars in sociology and politics have argued that crucial social phenomena are best explained in terms of the temporal study of 'path dependence', that is to say how particular courses of action and development are alighted upon and become reinforced over time (Pierson 2000; David, 2011)[5].

The long view: capturing the legacy of Thatcherite social and economic policy on crime

As a research team, we were confronted with the methodological and theoretical considerations of 'big' data-sets after embarking on a project to understand the long-term impact of Thatcherite public policies from the 1980s to the present day. Our initial analysis had demonstrated, in line with a substantial field of research on the link between the economy and crime rates (e.g. Cantor and Land, 1985), that as levels of unemployment and economic inequality rose, so property crime rose (Jennings et al, 2012; c.f. Morgan, 2014). As property crime increased, so too did fear of crime and so too did government attention to the issue of crime (see Farrall and Jennings, 2012; Hay and Farrall 2011; Farrall and Hay, 2010). However, we wanted to further explore differences across different demographics, such as by gender, housing tenure and geography, and to model attitudinal shifts in relation to other types of crime (such as violence). Notably, scholars from related branches of social policy have also begun to conduct allied longitudinal investigations in housing policy (Dorling, 2014), opiate drug-use (Morgan, 2014), education policy (Berridge et al., 2001), and social attitudes (Duffy et al, 2013; Nacten, 2014), highlighting the need for us to build an integrated model of analysis.

Big small data: the construction of a multi-layered data-set

These ambitions necessitated the creation of a connected series of longitudinal survey data-sets (which incorporate demographic markers such as age, education, household income, region and gender) and aggregate statistics on policy, economy and society, such as indexes of unemployment and inflation rates or numbers of probation and police officers. The data are linked through common variables, most notably the observed time period (i.e. the year or month) but also by categories of respondent (e.g. age, ethnicity, income, region, employment status). The significance and challenge of building such a data-set became ever more apparent during their construction. Secondary data analysis may side-step the task of data collection, but, in our experience, such data was rarely ready to use 'off the shelf'. We therefore needed to manually check variable names and codings (against original documentation that did not render itself amenable to automated processing), the length and 'direction' of any scales employed and the consistency of survey question wordings and sampling techniques over time. This work, as laborious as it was, nevertheless made us intimately familiar with the data-sets at hand, and tacit features of the original data collection. As such, the longitudinal data-sets that we mined have required extensive 'cleaning' and adaption (such as the standardisation of variable names or the re-coding of variables to enable comparisons across cross-sections), which was far more intricate and resource-intensive than might typically be associated

[5] David (2011) defines path dependency as a "dynamic process whose evolution is governed by its own history…. The concept, thus, is very general in its scope, referring equally to developmental sequences (whether in evolutionary biology or physics) and social dynamics (involving social interactions among economic or political agents) that are characterized by positive feedbacks and self-reinforcing dynamics" (2011:88).

with a ‘big data’ approach to the processing of large datasets. It also involved the consultation of portable document format copies of original documentation, especially for older surveys, to identify variables and the response categories for survey questions where no readable electronic data existed. Such information cannot be automatically extracted from the digital record due to its format and variation in the way it was originally collected. What resulted from this manual process of integrating the data however, is a valuable and unique resource, which would otherwise not be readily accessible in an integrated and usable format. In the next section we describe the content and structure of our data-sets which will be deposited with the UK Data Service in late 2015. Full details will be provided in a technical manual, but some important components are identified and summarised below (see Table 1).[6]

Individual-level data

Victimisation: officially recorded crime statistics have long been held in suspicion by many criminologists (Maguire, 2007). Our data incorporates *self-reported* data on victimisation from the CSEW[7]. This records respondents’ experiences, within the preceding twelve months, of most forms of crime[8]. The CSEW also includes a series of questions on fear of crime, perceptions of anti-social behaviour in the local area, confidence in the police and attitudes towards punishment and the criminal justice system. The merged CSEW data-set that we have developed, combining 21 sweeps of the survey that ran between 1981 and 2013, consists of 599,517 respondents and over 150 survey items that have been asked in multiple surveys.

Social attitudes: our data on public attitudes towards crime and criminal justice, and many other domains of social and economic life, is taken from two main sources. First, we have drawn on the 28 waves of the British Social Attitudes Survey (BSA)[9], which provides measures of social attitudes towards sentencing, punitiveness and matters relating to welfare. Second, the British Election Study’s ‘Continuous Monitoring Survey’ (BES-CMS) that ran on a monthly basis between 2004 and 2013 includes a range of measures of socio-political attitudes, such as satisfaction with the criminal justice system, evaluations of government/party handling of crime, and emotions about crime.

Table 1. Summary of individual-level data

	BCS/CSEW	BSA	BES-CMS
Key/Sample Questions	Victimisation (multiple categories) Fear of crime Common problems Confidence in police/criminal justice system Attitudes on sentencing Burglar/car alarm	Role of government Unemployment vs. inflation Puntiveness and authoritarianism Likelihood of riots Attitudes on welfare state Trust in government	Crime situation Government/opposition handling of crime Emotions towards crime Sought crime assistance Satisfied with assistance Importance of crime as an issue

[6] Documentation for the data-sets is provided online. The survey data was collected via face-to-face and computer-assisted-personal interviews (CAPI).

[7] First conducted in 1982, the CSEW was commissioned by the UK government to measure the ‘dark figure’ of unreported crime incidents. The survey sampling is structured to be representative of two groups, namely residential households in England and Wales, and adults (aged 16 years and over) living in those households (Bolling et al, 2004).

[8] The survey focuses on types of property and acquisitive crime, and physical, sexual and domestic violence.

[9] The British Social Attitudes Survey series began in 1983. It is based on an annual random probability, face-to-face survey of approximately 3,000 Britons. The series is designed to act as a counterpart to other large-scale government surveys such as the Labour Force Survey or the General Lifestyle Survey, which provide data on behavioural actions and tangible ‘facts’. It has been conducted every year since 1983 (except 1988 and 1992).

			People trustworthy
N of variables (including demographics)	109	80	63
N of respondents	599,517	89,466	124,110
Period	1981-2013	1983-2012	2004-2013

Aggregate-level data

The longitudinal processes that we are interested in require us to examine trends over time. To do this, our individual-level data-sets are also recoded to aggregate level. Additionally the CSEW (between 2002 and 2013) and BES-CMS (between 2004 and 2013) data-sets can be aggregated to monthly intervals, to observe finer-grained trends. The key *contextual* variables considered in longitudinal studies of crime, such income inequality (measured using the Gini Coefficient)[10] or unemployment, are treated and measured as national-level constructs. We have collected over a hundred time series which are summarised in Table 2.

Criminal justice system: for comparison against the CSEW victimisation-data, and also for enabling a longer-term view of crime, our data includes official recorded statistics on crimes for England and Wales. Annual data on the size of the prison and probation population is taken from Home Office Probation and Prison Statistics England and Wales.

Socio-economic indicators: Data is also included on levels of inequality, poverty and incomes from the Institute for Fiscal Studies (www.ifs.org.uk). Standard measures of inflation and unemployment rates, the claimant count and rate, economic inactivity, average earnings, labour disputes, and GDP are drawn from official statistics of the Office for National Statistics (www.ons.gov.uk). Data on annual benefits expenditure (and specific categories of benefits) is taken from the Department of Work and Pensions (2014). We have also collated data on truancy from the Youth Cohort Study from 1985 and school expulsions from the late 1990s. To complete our measures of social conditions we have data on the number of children in care dating back to the 1960s.

Policy and politics: our data-set also includes measures of political attention to policy action on crime. We draw on data from the UK Policy Agendas Project (www.policyagendas.org.uk) to capture the amount of attention given to crime, and law and order, in the statement of policy intentions set out in the Queen's Speech and in Acts of Parliament (between 1945 and 2012).

Public opinion: finally, we have collated a number of aggregate-level measures of public opinion over an extended time period, enabling a long-term view of attitudinal shifts. This includes survey data on the "most important problem" facing the country, as collected by the Gallup Organization between 1944 and 2001 (see Jennings and Wlezien 2011). In addition, we include data on the public's preferences for left-wing or right-wing public policy ('public policy mood'), from Bartle et al (2011), and have constructed a measure of public punitiveness using survey items on capital punishment, sentencing and other aspects of criminal justice, using a method developed by Stimson (1991) and applied by Enns (2014) in the US.

Table 2. Summary of aggregate data

[10] The Gini Coefficient is a quantification of relative deprivation (Yitzhaki, 1979). It is a widely respected measure of inequality in the distribution of household income.

	Crime and criminal justice	Employment	Macroeconomics	Welfare/Other	Politics/Policy
Selected data series	Official recorded statistics (total/violent/property) Convictions (total/as % of recorded crimes) Prison population Police force strength	Unemployment rate (national/by region/males 16-17; 18-24) Economic activity rate Claimant count (national/by region) Average weekly earnings Labour disputes (days lost)	Interest rates Public spending GDP Inflation Inequality Poverty Child Poverty	Total benefit expenditure (real/nominal terms/% of GDP) Unemployment/incapacity/ housing benefit (real/nominal terms/caseload) Suicide rates Children in care Council house sales Truancy and school expulsions Drug addicts	Queen's Speech Acts of Parliament Parliamentary questions (e.g. referring to "crime rate", "burglary", "anti-social behaviour")

Our enterprise raises questions about the degree to which longitudinal shifts in social behaviours and public attitudes are accurately captured by newer forms of Big Data. Our view is that Big Data (that is, transactional data, administrative records or web data) cannot effectively capture behavioural or attitudinal patterns that occurred before the move of much social, economic and political economic activity online (post 20th century), potentially limiting us to 'chronocentric' data. Moreover, it is likely that what data is available in this format will - at this point in time – often be disparate and unprocessed, and require considerable effort to peg new automatically-collected measures against traditional survey-based instruments. Measures of criminal activity or public fear of crime, for example, would have to link and calibrate existing survey data to untried indicators and assess their face validity. Retrospective construction of measures over time is a substantially more complex task than the compiling of repeated cross-sectional survey data over an extended period of time. As such, we see the current research agenda promoting the usefulness of Big Data as welcome when it is used alongside (rather than as an alternative to) rigorously designed, sampled and collected survey data. In this way we do not, at least for the foreseeable future, imagine that Big Data will replace social survey data (which has the added advantage of extending back in time to the 1970s and beyond, enabling long-term trends to be observed in a consistent way). The next steps for those interested in advancing the cause of Big Data may include, therefore, figuring out how Big Data and existing social survey data may be integrated in order to combine the advantages of both.

An adaptable resource

It is important to acknowledge the limitations to what we are able to do. 'Big data', no matter how sizeable or how well-sharpened, is no magic bullet, even if it were integrated with social survey data. There are issues which we are interested in (such as the experiences of homeless people in the 1980s) and for which no data set exists. In sum, the sorts of experiences and attitudes which we are able to analyse with historic data reflect the sorts of preoccupations of an earlier generation of researchers. This is a perennial problem for those conducting secondary data analyses (Dale, 2004). Nevertheless, we have employed traditional "small" data and amalgamated them into what we believe is now a vast, broad and dynamic group of data-sets, with the potential to answer significant 'big questions' about the effects of specific social and political policies on behaviour and public sentiments over time. It is significant that the processes involved in rendering the data usable were "small", in terms of the manual extraction of data and the specific knowledge required for handling survey data where there exists no *clean* digital footprint of variable names or contents (i.e. electronic versions of data might be unlabelled or coded in different ways across time that would lead to errors in automatic processing, without closer inspection of the original documentation). Despite such datasets being "big" in the sheer scale of data points (with close to three quarters of a million respondents to surveys included in our data-sets), their merging and standardisation relied upon traditional methods of manual processing to create a resource for large scale data analysis.

These data-sets have been constructed to be used by other researchers. Our project is funded by the UK's Economic and Social Research Council (award number ES/K006398/1, for more information on the project see http://www.sheffield.ac.uk/law/research/projects/crimetrajectories), meaning all of the data which we have collated will be deposited at the UK Data Archive at the end of the project (Autumn 2015). New users may utilise or adapt the data as they see fit. For example, others can update the data-set as new sweeps of surveys are released to the public, as well as customising it to answer questions substantially different to our own. In this sense we hope our data could become a 'platform' for others to build upon, using for their own research projects, PhD studentships and teaching purposes.

References

Bartle, J., Dellepiane-Avellaneda, S., and Stimson, J.A. (2011). 'The Moving Centre: Preferences for Government Activity in Britain, 1950-2005.' *British Journal of Political Science* 41(2): 259-285.

Berridge, D., Brodie, I., Pitts, J., Porteous, D. and Tarling, R. (2001) *The Independent Effects Of Permanent Exclusion from School on the Offending Careers of Young People RDS Occasional Paper No. 71*. Home Office: London.

Bolling, K., Clemens, S., Grant, C., Smith, P. and Brown, M. (2004) *2003-04 British Crime Survey (England and Wales): Technical report.* London: BMRB.

Burrows, R. and Savage, M. (2014) 'After the crisis? Big Data and the methodological challenges of empirical sociology', *Big Data and Society*, June 1:1 .

Dale, A. (2004) 'Secondary Analysis of Quantitative Data' in M.S. Lewis-Beck., A, Bryman. and T. F. Lioa (eds) *The Sage Encyclopaedia of Social Science Research Methods,* Sage: Thousand Oaks, CA.

David, P.A. (2011) 'Path dependence: a foundational concept for historical social science' in (eds) Zumbansen, P. and Calliess, G.P. (eds) *Law, Economics and Evolutionary Theory*. Edward Elgar: Cheltenham.

Department for Work and Pensions. (2014). *Benefit expenditure and caseload tables*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/310483/outturn-and-forecast-budget-2014.xls

Dorling, D. (2014) *All That Is Solid: How the Great Housing Disaster Defines Our Times, and What We Can Do About It.* Allen Lane: London.

Duffy, B., Hall, S., O'Leary, D. and Pope, S. (2013) *Generation strains*. Demos: London.

Enns, P.K. (2014). 'The Public's Increasing Punitiveness and Its Influence on Mass Incarceration in the United States.' *American Journal of Political Science* 58(4): 857–872.

Farrall, S. and Jennings, W. (2012) Policy Feedback and the Criminal Justice Agenda: an analysis of the economy, crime rates, politics and public opinion in post-war Britain, *Contemporary British History*, 26(4):467-488.

Farrall, S. and Hay, C. (2010) 'Not So Tough on Crime? Why Weren't the Thatcher Governments More Radical in Reforming the Criminal Justice System?'*British Journal of Criminology* 50, no. 3 (2010): 550-69.

Gadd, D. Farrall, S., Dallimore, D. and Lombard, N. (2002) 'Domestic Abuse Against Men in Scotland', *Scottish Executive Central Research Unit Report*, Scottish Executive: Edinburgh.

Hay, C. and Farrall, S. (2011) Establishing the ontological status of Thatcherism by gauging its 'periodisability': towards a 'cascade theory' of public policy radicalism*, British Journal of Politics and International Relations,* 13(4): 439-58.

Housley, W., Procter, R., Edwards, E., Burnap, P., Williams, M., Sloan, L., Rana, O., Morgan, J., Voss, A. and Greenhill, A. (2014) 'Big and broad social data and the sociological imagination: A collaborative response', *Big Data and Society*, December: 1–15.

Jennings, W., Farrall, S. and Bevan, S. (2012) The Economy, Crime and Time: an analysis of recorded property crime in England and Wales 1961-2006, *International Journal of Law, Crime and Justice*, 40(3):192-210**.**

Jennings, W. and Wlezien, C. (2011). 'Distinguishing between Most Important Problems and Issues?' *Public Opinion Quarterly* 75(3): 545–555.

Cantor, David, and Land, Kenneth C. (1985). 'Unemployment and crime rates in the post-World War II United States: A theoretical and empirical analysis.' *American Sociological Review* 50: 317-332.

Maguire, M. (2007) 'Crime data and statistics' in Maguire, M., Morgan, R. and Reiner, R. (eds) *Oxford Handbook of Criminology, 4th Edition*, Oxford University Press: Oxford.

Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data', in M.K. Gold (ed) *Debates in the Digital Humanities*, The University of Minnesota Press: Minneapolis.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data – A Revolution That Will Transform How We Live, Think and Work*, John Murray: London.

Morgan, N. (2014) *The heroin epidemic of the 1980s and 1990s and its effect on crime trends – then and now. Research Report 79*. Home Office: London.

Nacten (2014) *British Social Attitudes 31,* Nacten: London. Accessed online October 2014: http://www.bsa-31.natcen.ac.uk/media/38202/bsa31_full_report.pdf

Pawson, R. and Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Pierson, P. (2000) 'Not just what, but when: timing and sequence in political process', *Studies in American Political Development*, 14: 72-92.

Rock, P. (2005) Chronocentrism and British Criminology, *British Journal of Sociology*, 56(3): 473-91.

Ryder, N. B. (1965) 'The Cohort as a Concept in the Study of Social Change', *American Sociological Review*, 30:843-61.

Savage, M. and Burrows, R. (2007) ‘The Coming Crisis of Empirical Sociology’, *Sociology*, 41.5: 885-899.

Stimson, J.A. (1991). *Public opinion in America: Moods, cycles, and swings*. Boulder, Co.: Westview Press.

Yitzhaki, S. (1979) ‘Relative deprivation and the gini coefficient’, *Quarterly Journal of Economics*, 93, 231-324.