New Data Horizons

NCRM Autumn School 2017

22-24 November 2017

Highfield House Hotel, 119 Highfield Lane, Highfield,

Southampton, S017 1AQ

Programme

## Day 1 – Wednesday 22 November 2017

11.30 Registration opens

12.00-13.00 Lunch

13.00-13.15 Patrick Sturgis and Gabriele Durrant(NCRM, University of

Southampton) ***Introduction to the autumn school***

***Session 1: Changing nature of and innovations in surveys***

13.15-14.30 Patrick Sturgis (University of Southampton) *Changing patterns of Social Science data usage*

*Abstract:*

*There has been a great deal of excitement in recent years about the potential for new forms of data to transform empirical social science. Administrative, transactional, and other forms of ‘big data’ are, we are told, set to replace more conventional forms of ‘planned’ data collection such as censuses and surveys. To what extent are such narratives reflected in the actual practice of social science? In this talk, I shall draw on recent research which uses surveys of social scientists and content analysis of journal articles to assess the extent to which social scientists data usage has changed over the past fifty years, with a particular focus on the last decade.*

* + - 1. Joel Williams (Kantar Public) Address based online surveying

*Abstract:*

Throughout this decade the commissioners of social science research have been under pressure to reduce expenditure on surveys. This has led to many studies of the value-for-money of each component of expenditure. In particular, lots of work has been done to test whether lower cost methods can replicate the results obtained from higher cost methods. Kantar Public’s contribution has been to develop ‘address based online surveying’ (ABOS) – a flexible random sample method that may be used as an alternative to both higher-cost interview methods and lower-cost (non-random sample) online panels. This session introduces the method and the many studies that have been carried out to assess how well it works.

15.30-16.00 *Break*

16.00-17.00 Curtis Jessop (NatCen) *Developing the NatCen panel*

Abstract:

The NatCen Panel is the first open probability-based research panel in Great Britain. It is designed to produce robust estimates for the British population in a shorter time-frame and at a lower cost than the ‘traditional’ probability-based approaches currently available. This paper outlines the development of the NatCen Panel over the last two years, explaining the rationale for setting it up, and providing technical information on its methodology and key metrics on sample quality. It also summarises some initial findings from methodological experiments, and provides examples of how the Panel has been used over the last two years.

17.00-17.30 Interactive/discussion session

*18.30 Evening Dinner*

## Day 2 – Thursday 23 November 2017

### Session 2: Data linkage, administrative and census data

9.00-10.00 James Doidge (University College London) *Practical aspects and methodological challenges in research using linked data*

Abstract:

This session will start by covering some basic questions such as: What is linked data? Where does it come from? (where can you get it and how do you apply for it?) What is it good for? (examples of research questions that can be addressed using linked data). We will then consider some of the methodological challenges that are specific to research using administrative or linked data, focusing on the concepts of dynamic populations and linkage error.

10.00-11.00 Laura van der Erve (Institute for Fiscal Studies) *Using administrative data to investigate graduate earnings and beyond*

*Abstract:*

In this session a project with Jack Britton, Anna Vignoles, Lorraine Dearden and Neil Shephard is presented that links administrative tax records to Student Loan Company records. The challenges associated with the linkage and the pros and cons of the dataset are discussed. Results are shown focussing on earnings differences by subject, institution, socio-economic background and gender. Future administrative datasets are highlighted that will come available in the future and talk about some of the possibilities for future education research.

11.00-11.30 *Break*

11.30-12.30 Dave Martin (University of Southampton) *Transforming the Census*

*Abstract:*

The 2020-21 round of international censuses represent a major break with traditional practice, with a huge reorientation taking place towards internet data collection and the linkage of traditional census enumeration with administrative data sources. Census data provide key baseline sociodemographic and denominator data for a vast array of research and policy applications via a variety of different data products, and the new methods being adopted will have major implications for users. Taking the England and Wales Census Transformation programme as a model, this session will explore some of the major opportunities and risks for researchers.

12.30-13.00 Interactive/Discussion session

13.00-14.00 *Lunch*

### Session 3: Social media and big data

14.00-15.30 Leslie Carr (University of Southampton, Professor of Web Science in the Department of Electronics and Computer Science, Director of the Web Science Institute and Director of the Web Science Centre for Doctoral Training) *Title to be confirmed*

*(Matthew Williams who had been scheduled here had to pull out at very short notice following a bereavement.)*

15.30-16.00 *Break*

16.00-17.00 Susan Banducci and Iulia Cioroianu (EXPONet, University of Exeter) *Online Data Sources: Linking Old and New, Big and Small*

*Abstract:*

In this presentation we cover two areas: 1) tools to harvest online unstructured data (e.g. Twitter, comments, clickstream data) and 2) linking unstructured data to traditional data sources (e.g. surveys). The diversity of online media (mixed text/image/video content in heterogeneous formats) creates challenges in data collection and analysis. Using social media and online data also adds new challenges such as representativeness, measurement error and causal inference. We explore how to address some of these challenges by discussing tools to harvest data, data management and data linkage to traditional sources of social data. We illustrate some of these tools with examples from EXPONet -- a project designed to develop tools toward understanding information exposure in dynamic and dependent social networks.

17.00-17.30 Interactive/Discussion session

18.30 Evening Dinner

## Day 3 – Friday 24th November 2017

### Session 4: Biosocial data

09.00-10.30 Michaela Benzeval (University of Essex) *Integrating biosocial and social*

*science data*

*Abstract:*

*Understanding the interaction between people’s social and economic circumstances and their health across the life span is essential to develop policies not only to improve the nation’s health but also its social and economic capacities. To investigate these complex links, studies are required that bring together rich data on people’s social and economic lives with accurate measures of health and health functioning. Such data are available in Understanding Society: the UK Household Longitudinal Study, a major infrastructure investment for UK social science, which in recent years has been enhanced with the addition of biomarker, genetics and epigenetics information. This session will outline the biological information available in Understanding Society, some of the issues that need to be considered in analysing them and illustrate the ways in which they might add value to social science research.*

10.30-11.00 *Break*

* + - 1. Melinda Mills (University of Oxford) *Combining Social Science and Molecular*

*Genetic Research*

*Abstract:*

*Within the last decade there has been an explosion of data that includes both social science and molecular genetic data. The UK is a frontrunner in collecting this type of data (e.g. ALSPAC, 1958 Birth Cohort, ELSA, Understanding Society), including recent large samples such as the UK Biobank. This talk provides an overview of recent social science applications in this areas of fertility, education and wellbeing. Key techniques and applications will be covered including using polygenic scores from large-scale GWAS (genome wide association studies) and GREML (genomic-relationship-matrix based restricted maximum likelihood) techniques. The talk will also reflect on the relevance and possibilities of using molecular genetic data for social science research.*

12.30-13.00 Concluding discussion session

13.00  *Finish*