Government Statistical Service

The 23rd GSS Methodology Symposium

Better Statistics, Better Decisions through Innovation

Westminster Conference Centre
1, Victoria Street
London
Wednesday, 18 July 2018

Welcome to the 23rd GSS Methodology Symposium

Better Statistics, Better Decisions through Innovation

In its *Innovative* perspective, the Strategy for UK Statistics challenged us to deliver measurable benefits from continuous improvement and a spirit of curiosity. We should anticipate new opportunities and respond to unmet statistical needs using innovative, effective and efficient methods.

As the Strategy passes its half way mark, we will share work from across the Government Statistical Service at the GSS Methodology Symposium, reflecting on progress so far. This includes unlocking value through better use of data, improving the statistical methods we use, or rethinking our data collection methods. These are sometimes part of a larger programme of development, or an agile response to an urgent requirement.

Our keynote speakers will talk about the successes and key messages from the last year in methodology; Statistical innovation: the future of official statistics; and a policy department perspective - enabling innovation in statistics and data science.

The GSSM23 team hope you will enjoy the symposium, celebrating and sharing our achievements in methodology and looking forward to better statistics and better decisions through innovation.

GSS Methodology Symposium Programme

09:00 10:00 ***Registration –*** *Please arrive early if you can, to minimise queues*

10:00 10:50 **Morning Keynote Session**
Sarah Henry - *Welcome Address*
David Hand - *Statistical Innovation - the future of official statistics*

10:50 11:00 The GSS Methodology Advisory Committee

11:00 11:20 Exhibits, posters, refreshments

Morning parallel sessions

		AM1: Data Linkage	***AM2: Using VAT data***	***AM3: Managing innovation***	***AM4: Survey design***
		Pete Brodie	**Jennie Davies**	**Julie Brown**	**Charlie Wroth-Smith**
11:20	11:50	OSR's data linkage review findings: what would a safe and effective data linkage system look like?	Investigating methods of efficient detection of errors in VAT data	Necessity is the mother of invention: An innovative approach to survey design through collaboration	Moving from Face to Face Survey to Online Self-Completion Survey Methodology
11:50	11:55	Break			
11:55	12:25	Skills and resources for data linkage	Quarterly VAT turnover analysis: stagger comparison and calendarization methods	Statistics on Income and Living Conditions- Recent Survey Changes, and Methodological Considerations	Setting up an online panel to collect longitudinal data as part of the Taking Part Survey
12:25	12:30	Break			
12:30	13:00	Data linkage success stories	Smarter working – alternatives to using alternative sources of data	UK Trade: Delivering methodological change through an Agile approach	Using Machine Learning and Natural Language Processing to automate the Crime Survey for England & Wales (CSEW) offence coding process

13:00 14:00 Lunch

14:00 14:30 **Afternoon Keynote Session**
David Fry (BEIS) - A policy department perspective - enabling innovation in statistics and data science

Afternoon parallel sessions

		PM1: Geography & Coding	***PM2: Time series and indicators***	***PM3: Policy evaluation***	***PM4: Uncertainty & Data Quality***
		Nick O'Rourke	**Atanaska Nikolova**	**Simon Compton**	**Gary Brown**
14:30	15:00	Automated equal area cartograms - algorithm-led generation of equal area cartograms for anywhere, any period and any scale	Testing Alternative Distributions for Easter Effects in Seasonal Adjustment	The effect of early release from prison on reoffending using RDD	Using R for variance estimation in social surveys
15:00	15:05	Break			
15:05	15:35	Improving the Efficiency of Interviewer Workloads for ONS Surveys	Measuring discontinuities in the categorization of road accident data	Impact evaluation of the prison-based Core Sex Offender Treatment Programme	Use of Household Income Data in the Northern Ireland Deprivation Measures, 2017
15:35	15:40	Break			
15:40	16:10	Optimus – A NLP pipeline for turning free-text lists into hierarchical datasets	Developing new indicators of current and future business performance	A latent class analysis of loneliness	Identifying anomalies in categorical Census data

16:10 16:50 **Closing Panel Session:** Sarah Henry, Julie Stanborough and Rachel Skentelbery
Overcoming barriers to innovation

Keynote Speakers and Panel Discussants

Sarah Henry is the Executive Director of Methods, Data and Research in the Office for National Statistics. She recently completed her first year in post having previously been a GSS customer as Head of Intelligence and Performance at Manchester City Council and GM Connect.

Sarah will deliver the welcome address, reflecting on the last year in methodology, the successes and key messages.

David Hand is Emeritus Professor of Mathematics at Imperial College, where he previously held the Chair in Statistics. He is a former President of the Royal Statistical Society, and has served on the Board of the UK Statistics Authority since 2013. He also serves on the European Statistics Advisory Board and the Advisory Board of the ONS's Data Science Campus.

Statistical innovation: the future of official statistics

We live in a time of unprecedented opportunity as regards data, data sources, and data analytic methodology. This yields huge scope for developments to improve official statistics, in terms of how well they describe conditions in society. But innovations can have unexpected side effects, and new tools do not always behave in the way expected. This talk examines some of the potentials and risks of the new opportunities.

David Fry is currently the head of Data Analytics and Business Statistics at the Department for Business, Energy and Industrial Strategy (BEIS) where he is responsible for a number of business surveys and leads the department's data science capacity. As Head of Profession for Statistics, David is responsible for leading and developing the analytical capacity of over 70 statisticians in BEIS. He is a Fellow of the Royal Statistical Society.

At the then Department for Communities and Local Government David led analytical support for troubled families, homelessness, local public services and local government finance. There he was responsible for the management and analysis of a major social survey - the English Housing Survey. His team also collected housing and planning data from local authorities, published the indices of multiple deprivation, analysed lettings of social housing and carried out sub-national analysis of well-being. Earlier in his career he led on the public-sector mapping agreement, briefed on the labour market, contributed to thinking on the national minimum wage, published a book of statistics on children and examined the pay of MPs.

A policy department perspective - enabling innovation in statistics and data science

In this session David will discuss statistics and data science at BEIS. He will look at the development of their data science strategy and how it has been implemented, illustrating this with examples of successful projects.

Rachel Skentelbery (left) has recently re-joined the Office for National Statistics as the Deputy Director for Methodology.

Julie Stanborough (right) is the Deputy Director responsible for the Best Practice & Impact Division.

Rachel and Julie will join a final panel discussion on how we can promote innovation across the GSS.

Abstracts for Parallel Sessions

AM1: Data Linkage

Unlocking value through greater use of data linkage

Catherine Bromley, Office for Statistics Regulation, Shelly Gammon, Office for National Statistics, Hayley Moore-Purvis, DWP Work and Health Unit, Claire Wainwright, Scottish Government

With robust safeguarding and governance in place, greater use of linked data will enable better statistics and better decisions. The Office for Statistics Regulation has been conducting a review of the UK statistical system's ability to maximise the value of linked data to provide insights to users. This panel will outline the actions we think might help remove some of the barriers we heard about and open new opportunities.

It will include three presentations:

1. OSR Data Linkage Review – how can the system make better use of linked data and how can the Office for Statistics Regulation and Code of Practice for Statistics support this aim? (Catherine Bromley)
2. Methodological skills and resources for data linkage, safeguarding and governance – what does the system need and what's currently available? (Shelley Gammon & Catherine Bromley)
3. Data linkage innovation success stories – what can we learn from successful examples of data linkage taking place in government?
    - Education and child and family justice - MoJ Justice Statistics Analytical Services
    - Health-led Trials (launch May 2018) – DWP Work and Health Unit
    - Dental health of looked after children OR out of hospital cardiac arrest – Scottish Government

AM2: Using VAT data

Investigating methods of efficient detection of errors in VAT data

Katie Davies, ONS

As part of Business Statistics Transformation, the Office for National Statistics (ONS) are investigating including administrative data sources such as VAT data to supplement or replace survey data to reduce burden. The Distributive Trade Transformation project is one such instance where this is being looked at. As the proposed plan is that data will be used to replace some survey data, investigation is required to remove errors from the data. However, due to the magnitude of the data, error detection methods need to be efficient minimising resources required.

Techniques investigated include selective editing and macro level outlier detection. Selective editing has been considered as it prioritises errors that have the greatest impact on the final estimates therefore reducing the number of errors that require treatment. The macro level outlier detection method uses time series methods to automatically detect and treat outliers at an aggregate level therefore also targeting those with the greatest impact on the final estimates. Research includes looking at the two methods individually to assess suitability, how they compare and if a combination of the two methods is required or viable.

Quarterly VAT turnover analysis: stagger comparison and calendarization methods

Jack Sim and Paul Labonne, ONS

We first present some work that has been done within the Retail Sales transformation program, where the VAT data is being examined as an administrative data source. VAT can be reported on a monthly, quarterly or annual basis, with three different staggers for the reporting of quarterly VAT. To assess any potential differences between the VAT staggers within the industries retail, wholesale and motor trade industries; the quarterly turnover was apportioned to calendar months by dividing by three. The median turnover values for each stagger in each month were plotted to explore any level differences between the quarterly staggers. Stagger comparisons were performed at a 2- and 4-digit standard industrial classification (SIC) level, with enterprises with employment less than 10, to examine the impact of using VAT data with the smallest businesses.

In the second part of the presentation we explore temporal disaggregation methods for the quarterly VAT data on a broader scale. The techniques we investigate consist of generating monthly estimates of turnover that are constrained to the quarterly totals provided by the VAT figures. Since the latter are noisy and subject to strong seasonal movements and potential stagger biases, we have developed state space methods that are well suited to account for these features of the data. We use nonlinear techniques to allow for multiplicative and logarithmic model specifications. The possibility to use the MBS for the largest businesses and the monthly VAT data as indicator series is also investigated.

Smarter working – alternatives to using alternative sources of data

Gary Brown & Megan Pope ONS

For business statistics the potential benefits of the data revolution are a reduction in survey costs and in respondent burden – both through complementing survey data with administrative data, and replacing them where appropriate. However, these benefits can also be realised through alternative survey strategies – specifically using cut-off sampling to replace survey data, rather than direct replacement using administrative data. This presentation explores this issue from a theoretical standpoint, then gives a real-world example where the two approaches are compared directly. The costs and benefits of the two approaches will be discussed – as these are the key decision-making criteria – as will the outcome: are the statistics better?

AM3: Managing innovation

Necessity is the mother of invention: An innovative approach to survey design through collaboration

Madeleine May; Sport England, Olivia Christopherson & Genevieve Mitchell; Department for Digital, Culture, Media and Sport, Graeme Sinnott; County Sports Partnership Network, Margaret Blake; Ipsos MORI

Sport England’s vision is that everyone in England, regardless of their age, background and level of ability, feels able to take part in physical activity. The Active Lives Children and Young People survey was launched in September 2017 to provide an understanding of children's attitudes and behaviours around physical activity.

Key to monitoring the Government’s strategy for sport (Sporting Future) and informing policy decisions, the project is ambitious with an annual target sample size of 100,000 children aged 5 to 16, local authority level coverage and multiple stakeholders. The project has taken a collaborative approach from the outset bringing together multiple Government Department, academics in specialist areas such as physical literacy and other stakeholders such as the Swim Group during the development.

The survey takes place in schools but rather than taking a traditional approach to recruiting schools, Sport England has partnered with a network of County Sports Partnerships who already play an important role in joining up local authorities, schools and other organisations to maximise the impact of physical activity on wider values. Two terms into the first year of data collection, the survey has gathered responses from over 85,000 children with a response rate of 32%. First results will be published towards the end of 2018.

We will explore the benefits of taking a collaborative and innovative approach to survey design and recruitment such as reduced costs and lower participant burden but also some challenges, including how to avoid selection bias and ensuring consistency at a national level.

Statistics on Income and Living Conditions- Recent Survey Changes, and Methodological Considerations

Alexandra Pop, ONS

The European Union Statistics on Income and Living Conditions (EU-SILC) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions.

In 2017, EU-SILC in the UK underwent major methodological changes. To better meet these requirements the survey would:

- Increase to 5 waves (ultimately 6 waves) and will require five-year longitudinal weights.
- Replace FRS (Family Resource Survey) with the use of SLC (Statistics on Living Conditions) as EU-SILC Wave 1.
- Introduce a child cross-sectional weight for childcare to comply to policy requirements.

The presentation will consider the methodological strategy to account for these changes:

- The change from 4 waves to 5 waves and how we changed the longitudinal weighting strategy accordingly.
- And finally, an overview of using SLC compared to FRS as the first Wave for EU-SILC.

The aim of the presentation is to give an overview of the EU-SILC structure, how this structure will change with the additional requirements, and the methodological considerations taken to account for these changes.

UK Trade: Delivering methodological change through an Agile approach

Katie O’Farrell, Office for National Statistics

As the UK leaves the EU, the importance of high quality trade statistics for users has significantly increased. The work of the Methods team within the UK Trade project is meeting this challenge head on and looks to provide measurable benefits to publicly available trade statistics. The Methods team conducted a comprehensive review of sources and methods underlying statistical and conceptual adjustments applied to Trade in Goods data, identified methodological improvements in line with best practice guidelines, prioritised the work against internal and external requirements and deadlines, and worked alongside other workstreams in the project to develop a programme of delivery to fulfil both immediate and long term aims. Delivering Agile change within a more traditional waterfall environment has not been without its challenges, but the results have demonstrated the team’s ability to deliver efficient, effective change with a flexible, positive mindset using project management techniques that have been shared within ONS as examples of best practice.

AM4: Survey design

Moving from Face to Face Survey to Online Self-Completion Survey Methodology

Rosanna White, Statistician; DCMS

The Community Life Survey (CLS) is a key evidence source for government, covering community cohesion and engagement in England.

It is the first large government household survey to move from face-to-face data collection to an online/paper self-completion methodology.

Thorough testing to investigate the feasibility of moving the survey online began in 2013, when the survey was managed at the Cabinet Office. The new methodology was tested alongside the existing methodology for three years before moving fully to self-completion methods for the 2016/17 collection. The first Official Statistics based on the new method were published by DCMS in July 2017.

Considerations during the testing phase included determining the best way to conduct within household sampling; online response rates; the importance of offering an alternative paper questionnaire; which incentives were most cost effective; the impact of mode vs sample effects and implications for the time series.

The change in methodology has delivered substantial cost savings to the department alongside larger sample sizes which are not restricted to geographical clustering.

There are many lessons learned of wider relevance, given the interest in moving surveys online. As well as the methodological considerations during testing, the results show some interesting patterns, such as a systematic impact on responses to some personal questions, including the four ONS well-being questions.

Setting up an online panel to collect longitudinal data as part of the Taking Part Survey

Author Alex Björkegren; Olivia Christophersen; DCMS

Taking Part is DCMS's flagship survey, collecting evidence on participation in key DCMS sectors.

As part of the five year Taking Part strategy (published March 2016) the survey underwent a major redesign which separated the longitudinal element into a new web panel running alongside the main, cross-sectional, face-to-face household survey. This methodological change was designed to improve headline estimates whilst providing cost effective longitudinal data via short, quarterly online surveys for adults, youths and children.

Recruitment to the web panel started in April 2016. For the adult web panel, numbers are close to target, with lower than expected recruitment balanced out by lower attrition. The system for contacting panel members is effective and, while the first full dataset is not due until October 2018, response rates so far suggest it will provide a wealth of data to help understand changes in behaviour, which is a key policy interest at DCMS.

There have also been many lessons learned in the first two years of the panel. These include:

- the complex issues associated with processing and analysing a dataset based on quarterly modules with a flow sample
- agreeing the most appropriate data structure to meet different user priorities

- resource implications of designing/refining adult, youth and child surveys for each cohort every quarter
- engaging a sufficient number of young people to sustain a viable youth panel.

This work has also raised interesting questions such as the potential use of differential incentives to improve the demographic profile of the panel.

Using Machine Learning and Natural Language Processing to automate the Crime Survey for England & Wales (CSEW) offence coding process

Alessandra Sozzi; Shannan Greaney; Office for National Statistics

The Office for National Statistics (ONS) Big Data team and Crime Statistics team have been collaborating to explore the benefits of using Machine Learning and Natural Language Processing (NLP) techniques to automate the Crime Survey for England & Wales (CSEW) offence coding process.

Currently, all offences recorded as part of the CSEW are coded manually by an external contractor with a random 10% being 'dual coded' by researchers in the Crime stats team. This quality assurance task absorbs significant time and resource, especially in instances where ambiguous cases might require the agreement of multiple persons in the team. The current process is subject to frequent quality reviews, which assure consistency and high accuracy of both external and internal coders.

Using the closed responses along with respondents' textual descriptions about the crime from the survey, we produced an offence classifier which automates the categorization of different crimes.

Model confidence on predictions is used to keep only those cases for which the model feels highly confident, allowing us to assign the correct offence code to at least a consistent 30% of cases with nearly perfect accuracy.

The classifier can successfully save researchers' time and save the Office resource in the production of crime statistics.

PM1: Geography & Coding

Improving the Efficiency of Interviewer Workloads for ONS Surveys

Andy Clarke; Donna Viney; ONS Geography

The Office for National Statistics runs the Labour Force Survey and Annual Population Survey. This involves interviewers visiting addresses all over Great Britain (excluding areas north of Caledonian canal) to conduct the survey with respondents. Each interviewer needs to know which area they are responsible for.

The original workloads were created in 2003/4 and so have been many changes to the address distribution across Great Britain since then, resulting in workloads becoming uneven. This led to inefficiencies and difficulties managing work.

ONS identified an opportunity to improve the design of the interviewer areas to even out workloads across the country and reduce travel costs. The geography team implemented an innovative method for creating workloads using Model Builder, network analyst, and Python scripts along with University of Southampton's Automated Zoning tool. The new method calculates the "effort" involved in visiting the addresses based on the country they are in, whether they are in an urban/rural/or mixed urban rural area, and the sampling fraction applicable to that area. We then equalised this "effort" across workloads whilst also using the road network to ensure it was practical to travel around each workload and minimise travel costs.

The spread of sampled addresses from past quarters of the survey show the work is now much more evenly spread using these new workloads. This new method will result in the field interviewers being allocated workloads which optimise the number of interviews they can complete and reduce their travel time, which in turn will make operational savings.

Optimus – A NLP pipeline for turning free-text lists into hierarchical datasets

Steven Hopkins, Gareth Clews, Cerys Hopkins, Lucy Gwilliam, Data Science Campus, Office for National Statistics

Many data sets contain variables that consist of short free-text descriptions of items or products. The Data Science Campus has been working with DEFRA to understand shipping manifests of ferry journeys that record short descriptions of cargo on boarding lorries. The huge variation in detail, scale of description and how items are recorded (such as incorrect spellings or syntactic differences for identical products) make it difficult to automatically clean the data to a structured state that is ready for aggregation and analysis.

The Data Science Campus has implemented an NLP pipeline that utilises a Subword-Information Skipgram (SIS) model to retrieve vector representations of item descriptions and allow tiered-grouping of syntactically and semantically similar descriptions. The individual relationships between words within each group are assessed and labels are generated automatically in an appropriate way. The final output is a structured data set where each item is classified across multiple hierarchical tiers; data can therefore be aggregated at different levels or linked to existing classification taxonomies.

The pipeline is being generalised into a tool that can be applied to other text classification applications.

Automated equal area cartograms - algorithm-led generation of equal area cartograms for anywhere, any period and any scale

Mitchell, Bruce JW, Tzelepis, George, ONS, Geography Branch

Conventional geographical maps are subject to ‘the tyranny of large areas’: they may be dominated by zones that are large in areal extent but are often poor in data interest.

Cartograms can get around this by resizing areas on the basis of the variable of interest, and there are a number of online tools for this purpose. But these have to be redrawn afresh for every variable, and the distortion of original shapes renders both navigation and comparison of areas of similar statistical value difficult.

Equal area cartograms (EACs) render each area as an identical geometric symbol, which allow all areas to be evaluated on an equal footing. The most appropriate symbols are hexagons and squares (which additionally permit the display of percentage data).

ONS Geography is developing a tool to automate the generation of EACs. It is entirely open source (Python) and is built upon a robust and dependable methodology. It is intended to publish the tool via GitHub.

The ONS tool may be the first to provide UI-fronted and parameter-led method for users to create EACs for any 2-D space, for any period and at any scale.

PM2: Time series and indicators

Testing Alternative Distributions for Easter Effects in Seasonal Adjustment

Sam Jukes; Time Series Analysis Branch, Office for National Statistics

Moving holidays in time series can cause problems when a seasonal adjustment procedure is performed, as the holiday in question can change between calendar months and quarters and thus not be picked up by standard seasonal adjustment algorithms. This can lead to problems of residual seasonality. In the UK, Easter is the most prominent moving holiday occurring in either March or April. The Office of National Statistics currently uses three different default Easter regressors which assume the influence of Easter is the same for the entire window covered, and this window always covers a period before Easter Sunday. Three new distributions were recently created by the Time Series Analysis Branch, attempting to consider the UK school holidays. This presentation explores new distributions with flexible shapes. The tested distributions consider different attributes suspected to be influential of Easter. These include whether there is a 0, 1 or 2-week build-up, whether the Easter weekend itself has full constant influence, if the regressor is complex in nature (i.e. distribution based on real data) and more. The regressors are tested (including the six older regressors) on many series known to have Easter effects. The aim of this project is to test if there is an alternative Easter distribution that performs better than the current options.

Measuring discontinuities in the categorization of road accident data

David Braunholtz, Duncan Elliott & Bethan Russ, ONS

Road accident data published by the Department for Transport is collected from police forces across the UK. Different police forces have been introducing a new method of categorising accidents as either slight or severe. The new method of categorising accidents has the potential to cause a discontinuity in the published time series recording the number of accidents by type of accident for different areas of the UK, different types of road users, different types of roads etc. Time series methods have been used to estimate the impact of the change in categorisation method in order to estimate a time series without a discontinuity which is important for policy analysts and users of the data. Methods for estimating the discontinuity using X-13ARIMA-SEATS, the GSS recommended software for seasonal adjustment, are discussed and the impact of changes in categorisation method presented.

Developing new indicators of current and future business performance

Philip Wales, Gaganan Awano, Ted Dolby, Jenny Vyas & Russel Black, ONS; Nicholas Bloom, Stanford University; Paul Mizen, Rebecca Riley & Tatsuro Senga; ESCoE,

The current socio-political landscape of the UK in terms of its subdued pace of productivity growth, as well as the level of uncertainty among businesses has been unprecedented. Developing an understanding of the undercurrents of the business environment is essential for effective policy making.

The ONS has taken an innovative approach to deliver new insights on firm performance, (a) using old data in new ways, and (b) collecting new data and using efficient methods to deliver effective outcomes.

Creative data linking techniques has extended the scope and range of benefits the ONS derives from existing survey and administrative data on Trade, by unlocking the potential for firm-level research in uniquely new dimensions. This linked HMRC-IDBR dataset supports detailed firm-level trade analysis by product, volume, origin and destination dimensions.

Also, collaborative work with the Economic Statistics Centre of Excellence (ESCoE) in developing the Management and Expectations Survey has delivered new insight on the relationship between management structure and firm performance across firms in the production and services industries. The survey also provides new insight on business managers expectations of UK's future economic growth, as well as expectations of their own business performance for 2017 and 2018.

These new datasets are poised to provide vital information to policy makers on the scope and scale of UK business' trade interdependencies within and outside the EU, including flows of intermediate inputs used in production, as well as the impact of current uncertainty on business performance and decisions on future output, employment and investment growth.

PM3: Policy evaluation

The effect of early release from prison on reoffending using RDD

William Spry; Aidan Mews; Ministry of Justice

Ministry of Justice statisticians have investigated the impact that releasing offenders early from prison on home detention curfew (HDC) has on reoffending. The driver for the analysis was to support a government policy to release more prisoners on HDC, and thus take pressure off the prison system. The reoffending impact of HDC is difficult to measure as early release tends to only be given to those considered to have a low risk of reoffending. The analysis used a Regression Discontinuity Design (RDD) approach. RDD can be used where there is a natural cut-off on a continuous variable which determines at least to some extent whether individuals are placed in the treatment or control group. In this study, the RDD attempted to exploit the fact that only prisoners with a sentence length of between 3 months and 4 years are eligible for HDC, assessing whether there were discontinuities in the reoffending rates at these thresholds. The presentation will cover the many issues encountered and include a discussion of the strengths and weaknesses of RDD.

Impact evaluation of the prison-based Core Sex Offender Treatment Programme

Aidan Mews; Laura Di Bella; Ministry of Justice

The presentation describes how the high-profile impact evaluation of the prison-based Core Sex Offender Treatment Programme (England and Wales) was tackled. A cross-disciplinary team at the Ministry of Justice overcame many methodological challenges commonly found in the field, ensuring a high-quality quasi experimental product. Such challenges included identifying a counter-factual which was done using propensity score matching, and catering for individuals switching from the comparison to the treatment group. A variety of sensitivity analyses were performed including an assessment of the magnitude of bias from unmeasured factors that, if present, would change the results of the study, and to detect whether the main sexual reoffending effect of treatment was robust across multiple scenarios. Working closely with experts and practitioners, the team produced a narrative that was mindful of strengths and limitations when communicating the controversial findings.

A latent class analysis of loneliness

Ed Pyle; Dani Evans; ONS

Loneliness can seriously damage health and well-being (Holt-Lunstad, 2015). Though it can affect those of all ages, from different backgrounds, we know that individuals with particular characteristics and circumstances are more vulnerable. In January, the Prime Minster announced the development of a strategy to alleviate loneliness and requested that ONS develop national measures of loneliness. As part of this, we published *Loneliness – What* *characteristics and circumstances are associated with feeling lonely?* using measures and data already available.

We used Latent Class Analysis (LCA) to classify people experiencing loneliness most often into groups of people with similar characteristics and life-circumstances. This produced different loneliness ‘profiles’ – combinations of characteristics that group together in ways associated with loneliness – and allowed us to identify particularly lonely people and find out more about them. Three profiles were identified: (i) Widowed older homeowners living alone with long-term health conditions; (ii) Unmarried, middle-agers with long-term health conditions; and (iii) Younger renters with little trust and sense of belonging to their area. This information helps explain why people feel lonely and could be useful for informing policies and interventions aimed at ameliorating loneliness.

Analysis was conducted using R software (R Core Team, 2017) and the package poLCA (Linzer and Lewis, 2011), we applied LCA to respondents of the Community Life Survey 2016-17 (DCMS, 2017). This presentation will focus on describing our methodology and steps taken. It may be of interest to anyone interested in using R software for data analysis, LCA, or investigating loneliness.

PM4: Uncertainty & Data Quality

Using R for variance estimation in social surveys

Eleanor Law; Vahé Nafilyan; Ria Sanderson; ONS

Variance estimation is an important part of the production of official statistics and their accuracy, accessibility and comparability. Currently the primary tool used in sample design and estimation for social surveys at ONS is a SAS implementation of the generalised linearised jackknife method of variance estimation.

A greater focus on innovation and cost effectiveness across the Government Statistical Service is supported by increased use of open source software. To aid re-use and transparency of methods, we have implemented this method of variance estimation for complex survey designs in R.

We have used the Wealth and Assets survey as a case study to test this R implementation. This survey features cluster sampling, stratification, and calibration so offers a useful comparison between the R method and the existing approach. In this presentation, we will give an overview of the method, how it is implemented in R and the tests that we have carried out. We will discuss next steps, where we are looking to implement more flexibility, and the ability to estimate variance of change between different time points.

Use of Household Income Data in the Northern Ireland Deprivation Measures, 2017

Dr Tracy Power, Northern Ireland Statistics & Research Agency; Michelle Furphy, Department for Communities, Northern Ireland

Relative spatial deprivation measures have been developed in Northern Ireland since the 1970s. The most recent release by the Northern Ireland Statistics and Research Agency (NISRA) was in November 2017. Whilst this was an update of the 2010 Measures rather than a review, several new and innovative indicators were incorporated for the first time due to the availability of new data. This paper covers the use of income data in deprivation measures for the first time in the whole of the UK, derived from data held by the Department for Communities through data linkage at the household level. It also details the new dissemination tools developed designed to increase public engagement.

Identifying anomalies in categorical Census data

James Dawber; Paul Smith; Zoheir Sabeur; Gianluca Correndo; Galina Veres; University of Southampton

The population census will be online first in 2021, and this will make substantially more data available digitally quickly after the census. This means that data can be explored to help to identify recurring anomalies. In this paper we report some preliminary exploration for this problem from two standpoints.

The first is classical statistics, using multiple correspondence analysis (MCA), a dimension reduction method similar to Principal Component Analysis (PCA), but for categorical data. MCA can identify key patterns in a data set with a large number of categorical variables, providing insight into the relationships between categories, variables and individual records based on underlying dimensions. We apply MCA to categorised 2011 Census data, and demonstrate how the identified dimensions can be useful tools to generate propensity scores for use in anomaly detection.

The second approach is through data science, where a variety of unsupervised methods have been used, including KAMILA (k-means for mixed large data), and methods based on support vector machines (SVM). These methods are applied to the same categorised 2011 Census data, and they demonstrate how unusual records can be identified.

We discuss potential further uses and research in this area to support Census processing and analysis.