Government Statistical Service
Gwasanaeth Ystadegol y Llywodraeth

The 22nd GSS Methodology Symposium

Methodology: Insight; Innovation; Implementation; Impact

Westminster Conference Centre
London
12 July 2017

#GSSM22

Welcome to the 22nd GSS Methodology Symposium 2017

Methodology: Insight; Innovation; Implementation; Impact

As researchers and methodologists we are driven quite naturally to look for the insight and innovation that leads to improved or better methods than those we already have. There is no doubt that this drive helps to ensure that the professional advice and statistical analyses provided by the GSS year on year remains at the heart of Government policy and decision making. And yet, we are often challenged about the impact of our work - 'it's all very interesting, but what difference does it actually make?' In some cases, the challenge may be valid, but in many cases, the significance of the research and methodology that underpins policy and decision making remains largely unrecognised simply because it does not make the press release or glossy public facing visualisation.

One of the aims of the 2017 GSS Methodology Symposium is to provide an opportunity for researchers and methodologists to share ongoing research, development, and innovative ideas. However, of equal importance, another aim is to provide just a moment of pause; to celebrate why the work that we do is important and how it has made, or has the potential to make, a significant difference in pursuit of providing society with better statistics for better decisions. Methodology: Insight; Innovation; Implementation; Impact.

We hope you enjoy the Symposium.

The GSSM22 Team

Table of Contents

Enquiries: The GSSM22 Team:
methodology@ons.gsi.gov.uk: Steve Rogers
Karen Wilson
Johanne Merriman
Lisa Eyre

	Registration			
10.00 – 11.00	**MORNING KEYNOTE SESSION, *Syndicate Room 1:*** *Chair: Pete Brodie, Head of Survey Methodology & Statistical computing, ONS* **Jane Naylor**, *Deputy Director for Methodology, Office for National Statistics:* **"What have the methodologists ever done for official statistics?"** **Professor Peter W. F. Smith,** *Director of ADRN and ADRC-E, University of Southampton:* **"ADRC-E: Making a methodological impact"**			
11.00 – 11.20	**Exhibits :** ***Syndicate Room 1:*** The Economic Statistics Centre of Excellence (ESCoE) ***Foyer*:** The Royal Statistical Society (RSS); Southampton University (MOffStat & MDataGov); The Learning Academy **Refreshments:** *Foyer*			
	Morning Syndicate Sessions:			
	Session 1, Syndicate Room 1 **Data science for public good** *Chair: Tom Smith, ONS Data Science Campus (DSC)*	***Session 2, Syndicate Room 2*** **Counting on developments in population estimation methods** *Chair: Gary Brown (ONS)*	***Session 3, Syndicate Room 3*** **Towards the future of time series and longitudinal data analyses** *Chair: Frances Pottier (BEIS)*	***Session 4, Syndicate Room 4*** **Sampling and sample design: Optimising the allocation** *Chair: Aidan Mews (MoJ)*
11.20 – 11.50	***11.20 – 11.25*** **Welcome, Data Science Campus overview** *Tom Smith; Managing Director, DSC* ***11.25 – 11.50*** **Economic statistics and data science** *Louisa Nolan; Head of economic statistics, DSC*	**What is the population of Afghanistan?** *Kim Bradford Smith, Nicola Wardrop, Francis Harper; UK Department for International Development*	**Anticipating Brexit effects in time series analysis** *Jennifer Davies, Duncan Elliott, Charlotte Gaughan, Atanaska Nikolova; ONS*	**The 2016 Annual Business Survey: A sample re-design** *Megan Pope, Jonathan Digby-North; ONS*
11.50 – 11.55	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
11.55 – 12.25	**Developing statistics from image data** *Tom Smith; Managing Director, DSC*	**Generalised Structure Preserving (GSPREE) models in Small Area Estimation: Estimating ethnic group size for local authorities** *Philip Clarke, Alison Whitworth; ONS*	**The impact of moving holidays in official statistics time series** *Bethan Russ, Tariq Aziz; ONS*	**Which business to price? Sampling Business Price Indices** *Matthew Mayhew; ONS*
12.25 – 12.30	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
12.30 – 13.00	**Building data science capability across government** *Alexis Fernquest, Gareth L. Jones; Data analytics apprentices, DSC*	**Measuring uncertainty in ONS population estimates: Tools, techniques and outcomes** *Louisa Blackwell, Katy Stokes, Paulina Galezewska; ONS*	**Increasing frequency and improving timeliness of unemployment estimates from the UK Labour Force Survey** *Duncan Elliott; ONS*	**Standard errors from audits** *Sumit Rahman; Department for Business, Energy and Industrial Strategy*
13.00 – 13.45	**Exhibits :** ***Syndicate Room 1:*** The Economic Statistics Centre of Excellence (ESCoE) ***Foyer*:** The Royal Statistical Society (RSS); Southampton University (MOffStat & MDataGov); The Learning Academy **Lunch:** *Foyer*			

13.45 – 14.45	**AFTERNOON KEYNOTE SESSIONS, *Syndicate Room 1:*** *Chair: Jason Bradbury, Programme Director for Data Collection, ONS* **John Pullinger,** *The National Statistician:* **"Methodology: Driving innovation in statistics"** **James Gillan,** *Economic and Labour Market Statistics, Northern Ireland Statistics and Research Agency:* **"Connect and Transform"**			
	Afternoon Syndicate Sessions:			
	Session 1, Syndicate Room 1 **Modernising methodology through transformation, collaboration, and harmonisation** *Chair: Gareth James (ONS)*	***Session 2, Syndicate Room 2*** **Linking into administrative data: Infrastructure, method, and application** *Chair: Sumit Rahman (BEIS)*	***Session 3, Syndicate Room 3*** **Statistical outputs: Enhancing the user interface and protecting confidentiality** *Chair: Nick O' Rourke (ONSG)*	***Session 4, Syndicate Room 4*** **Big Data analytics: Opening new avenues of innovative research and understanding** *Chair: Owen Abbott (ONS)*
14.45 – 15.15	**Transformation of methodology in ONS - the why and how** *Gary Brown; ONS*	**Addressing for Census and beyond** *Alistair Calder; ONS*	**The methodology of maps** *Bruce Mitchell, Jeremy Brocklehurst; Anna Harris; GIS Mapping and Spatial Analysis Unit, ONS Geography*	**Natural Language Processing (NLP) techniques for structuring large volumes of human text data** *Alessandra Sozzi, Kimberley Brett; ONS*
15.15 – 15.20	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
15.20 – 15.50	**GSS Harmonisation Programme: Past, Present and Future** *Ian Sidney, Becki Aquilina; ONS*	**Many-to-many linkage: Finding siblings in birth registration data** *Christos Chatzoglou, Charles Morris, Shelley Gammon, Julie Mills, Lynda Cooper, Theodore Manassis; ONS*	**Improving workforce analysis through geographical design** *Mark Baines; ONS Geography*	**Using machine learning techniques to analyse economic statistics: a case study with HMRC Trade in Goods statistics** *Andy Banks; ONS*
15.50 – 15.55	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
15.55 – 16.25	**Lessons from Anthropology** *Tacey Laurie; KAI Analytics HM Revenue and Customs*	**Household effects in Big Data research: Privacy-preserving methods in evaluating a Welsh Government fuel poverty scheme** *Karen Tingay; Sian Morrison-Rees; ADRC[1,4]; Sarah Lowe; KAS[2,5]; Charles Musselwhite; CIA[3,4]*	**Disclosure Control: Applying Cell-Key Perturbation to 2021 Census outputs** *Keith Spicer, Stephanie Blanchard, Iain Dove; ONS*	**Exploring mental well-being from prisoner case-notes using text mining** *Jo Lee; Advanced Analytics Unit, Analytical Services Directorate, Ministry of Justice*
16.25 – 16.40	**Exhibits :** ***Syndicate Room 1:*** The Economic Statistics Centre of Excellence (ESCoE) ***Foyer*****:** The Royal Statistical Society (RSS); Southampton University (MOffStat & MDataGov); The Learning Academy **Refreshments:** *Foyer*			
16.40 – 17.10	**CLOSING KEYNOTE SESSION, *Syndicate Room 1:*** *Chair: Simon Compton, Head of Statistics, Competition and Markets Authority* **Sarah Henry,** *Director of Methods, Data, and Research (MDR), Office for National Statistics:* **"Our Data Capability – United we stand!"**			
17.10	**CLOSE:** Please fill in your evaluation forms! Have a safe journey home			

[1]Admin Data Research Centre, Swansea, Wales; [2]Knowledge and Analytical Services; [3]Centre for Innovative Ageing; [4]Swansea University; [5]Welsh Government

22nd GSS Methodology Symposium: Keynote Speakers

Morning Session, Syndicate Room 1

Chair: Pete Brodie, Head of Survey Methodology & Statistical Computing, Office for National Statistics

Jane Naylor

Deputy Director for Methodology, Office for National Statistics

Jane has over 15 years' experience working as a methodologist and statistician at the Office for National Statistics (ONS). Her areas of expertise include small area estimation, statistical disclosure control and protecting confidentiality, population and migration statistics, and the use of administrative data for statistical purposes.

More recently, Jane has focused her expertise and research interests on the development of data science in the ONS. Jane established the ONS Big Data team in 2014, which aims to understand the impact of big data and data science on UK official statistics and position ONS to take full advantage of the opportunities these new avenues hold for the future. Jane now leads one of the divisions within Methodology at ONS.

What have the methodologists ever done for official statistics?

In keeping with the theme of the symposium Jane will celebrate the impact that methodologists have on official statistics. Often this impact is not recognised since methodologists are not always natural promoters and their work tends to be downstream of the high profile data collection exercise and upstream of the final outputs and publicity that goes with them. Jane will argue that we need to do more to ensure that methodological research is delivered and moved into implementation. Also, that we demonstrate and communicate the benefits and impact of the work and get the appropriate recognition. This is particularly important at a time when the increased use of alternative (administrative and big data) data sources require new methodological innovations to ensure appropriate use within official statistics.

Professor Peter W. F. Smith

Director of the ESRC Administrative Data Research Network and the Administrative Data Research Centre for England, University of Southampton

Peter is Director of the Economic and Social Research Council Administrative Data Research Network and the Administrative Data Research Centre for England. He is also Professor of Social Statistics at the University of Southampton, where he has worked for over 26 years.

Peter was awarded the Royal Statistical Society's Guy Medal in Bronze in 1999 and was Joint Editor of Series C of the Society's journal from 2013 to 2016. He was a member of the Government Statistical Service Methodology Advisory Committee from 2011 to 2016 and has advised the Office for National Statistics on a number of methodological projects.

Peter's research interests include the development of new statistical methodology and their application to problems in demography, medicine and health sciences. Current projects include: developing models for population forecasting, which include coherent probabilistic quantification of forecast uncertainty; modelling administrative data; and investigating the uses of paradata (field data) with the aim of improving data collection methods.

ADRC-E: Making a Methodological Impact

As part of its remit, the Administrative Data Research Centre for England is undertaking a programme of methodological research. Peter will describe a few examples from this programme including work on the representativeness of surveys using a unique data set linking call record paradata from three UK social surveys to

census auxiliary attribute information on sample households; guidance about the information that needs to be made available about the data linkage process by data providers, data linkers, analysts and researchers; and a new approach to linkage based upon weights derived using a scaling algorithm.

Afternoon Session, Syndicate Room 1

Chair: Jason Bradbury, Programme Director: Data Collection, Office for National Statistics

John Pullinger

National Statistician, UK Statistics Authority

John is the National Statistician, Head of the Government Statistical Service (GSS) and Chief Executive of the UK Statistics Authority since July 2014.

His role is to safeguard the production and publication of high quality official statistics by all departments, agencies and institutions in the UK. His priorities are to improve measurement of the economy, to bring better evidence to public policy debates and to enhance data capability.

John has represented the UK internationally in EU, UN, OECD and other forums. He was previously Chair (2015) and Vice-Chair (2016) of the United Nations Statistical Commission (UNSC).

In 2004, John became the 14th Librarian to the House of Commons. He continued to be involved in the statistical community and was President of the Royal Statistical Society from January 2013 to June 2014.

Methodology - Driving Innovation in Statistics

More than ever, there is a demand for timely statistics to help explain the world around us. Our users are changing too and we must be prepared to meet their needs. John will explore the link between how changes to methodologies can help bring better statistics to users more quickly, taking advantage of the data revolution. As technology grows, so too does the availability of more data and in richer forms. This helps drive new, innovative ways of using these data to keep up with the fast pace of a changing world. Sound, transparent methodologies underpin the quality of statistics and are the driving force behind improving the evidence used for decision making.

Dr. James Gillan

Economic & Labour Market Statistics, Northern Ireland Statistics and Research Agency (NISRA)

James is responsible for the production of NISRA's Economic & Labour Market Statistics and has led the development of NISRA's Integrated Business Survey System and its e-survey facility. He is currently preparing for the greater use of HMRC and other Big Data sources following the Digital Economy Bill and has been closely involved in producing NISRA's first set of Supply Use and Input-Output Tables. He has drawn heavily on the ONS's Methodological Advisory Service over the years and sits on the Methodology Advisory Committee. Current projects include developing a "Better Jobs Index" and a non survey based measure of vacancies. James previously worked as a NI-CO consultant in Russia and the Middle East and he was the Northern Ireland 2001 Census manager.

Connect and Transform

As a relatively small Statistics Agency NISRA has to be agile in trying to keep pace with the changes required by the Bean report in terms of transforming its data collection and economic statistics functions. This presentation focuses on NISRA's business survey data collection transformation and the development of its e-survey systems as an example of recent innovation. The talk will also ask how methodological developments can enable better decision making and cover NISRA's response to the increasing demand for more flexible, innovative and impactful economic statistics.

Closing Session, Syndicate Room 1

Chair: Simon Compton, Head of Statistics, Competition and Markets Authority

Sarah Henry

Director of Methods, Data, and Research (MDR), Office for National Statistics

Sarah is the newly appointed Director responsible for leading the Office for National's Statistics' use of Data and Methodology, supporting the creation of knowledge and research to inform better decisions. Her role is to enhance the statistical community's Data Capability by introducing new data sources and developing new methods to make good use of that data.

Sarah was previously Head of Intelligence and Performance at Manchester City Council where she worked closely with decision makers who were leading the way in local Devolution. Her role included developing the evidence base for flagship programmes that are changing the way local services are provided.

Sarah has worked as an analyst for more years that she cares to remember and has public, private and military experience.

Our Data Capability – United we stand!

Used well, data has the power to save lives, fuel the economy and improve society. Sarah will explore examples of imaginative use of data that turns conventional wisdom on its head, leading to new insight and better decisions in ways that were unimaginable before data became big and accessible. She will examine what the GSS can do collectively to inspire change, improve lives and shine the light on facts that are real and relevant for today's society.

22nd GSS Methodology Symposium: Parallel Sessions

1.1 Morning Session 1, Syndicate Room 1

Data science for public good

Chair: Tom Smith: Data Science Campus, Office for National Statistics

1.1.0 Welcome, Introductions, Data Science Campus Overview

Tom Smith; Managing Director; ONS Data Science Campus

1.1.1 Economic statistics and data science

Louisa Nolan; Head of Economic Statistics; ONS Data Science Campus

The first of three presentations from DSC will provide an overview of the economic statistics projects at the Campus, including work on: Leading indicators of GDP; Financial sector classifications; and Regional indicators of turnover by industry sector. The talk will outline data sources, methods and preliminary results.

Key Words: data science; economy; machine learning; classification

1.1.2 Developing statistics from image data

Tom Smith; Managing Director; ONS Data Science Campus

The second of three presentations from DSC will provide an overview of the rapid growth in developing indicators and statistics from image analysis. The presentation will showcase work underway at the ONS including: local street-level imagery to develop a 'natural capital urban tree' dataset at local level, based on neural network image segmentation & classification methods, and using satellite data to identify mobile homes for Census applications. The talk will also explore using remote earth observation satellite data for estimating population and economic data, including Sustainable Development Goal indicators.

Key Words: data science; image data; statistics; neural network; economic data

1.1.3 Building data science capability across government

Alexis Fernquest, Gareth L. Jones; Data analytics apprentices; ONS Data Science Campus

The third of three presentations from DSC: With the opening of the Data Science Campus ONS has established a major hub for data science capability with Government and this talk will overview the ways that the Campus and GSS Learning Academy are helping build capacity. The talk will also present projects carried out by the Campus apprentice group, working on the UK's first data analytics apprenticeship scheme.

Key Words: apprenticeship; cluster and factor analysis; Labour Force Survey; Shiny; business survey

1.2 Morning Session 2, Syndicate Room 2

Counting on developments in population estimation methodology

Chair: Gary Brown, Office for National Statistics

1.2.1 What is the population of Afghanistan?

Kim Bradford Smith; Nicola Wardrop; Frances Harper; UK Department for International Development

In fragile and conflict-affected states the availability and reliability of national data can be poor. In Afghanistan, the first and only census was held in 1979, although this covered only two thirds of districts due to insecurity. Further census enumeration was planned for 2008, then rescheduled to 2011, but was eventually cancelled due to ongoing insecurity. Current national and subnational population estimates are based largely on projected population counts from a 1979 baseline, updated with information from pre-census household listing activities, where available. There is an urgent requirement for up-to-date national population data in Afghanistan, but due to the infeasibility of conducting a national census, an alternative method is required.

Supported by DFID and UNFPA, the Afghanistan Central Statistics Organisation has been conducting a “rolling census” since 2011. The Socio-Demographic and Economic Survey (SDES) is conducted in one province at a time, and includes a full household enumeration and detailed data collection for 50% of households. To date, two thirds of the Provinces have been enumerated by SDES. The resulting data have been used along with covariate datasets representing factors known to be correlated with population density (e.g. settlement areas), to predict population counts in areas which have not yet been enumerated using regression modelling. This novel approach has provided the current best estimate for the national and subnational population sizes for Afghanistan. There is scope to apply similar methods in other countries where comprehensive population enumeration activities are not possible, for example due to insecurity.

Key Words: census; population estimation; statistical regression; Afghanistan

1.2.2 Generalised Structure Preserving (GSPREE) models in Small Area Estimation: Application in estimation of ethnic group size for local authorities

Philip Clarke; Alison Whitworth; Office for National Statistics

The Office for National Statistics is looking at using more administrative and survey data to produce typical census outputs, in an Administrative Data Census. However, survey data often have very small or null sample sizes within areas, and administrative data may not cover the entire population. The census will continue to be a detailed source of

information for some time after it is carried out, but for intercensal estimates it would also become outdated. This leads to the use of data sources in combination to produce reliable estimates for small areas and sub-groups of the population for which there are small or no samples. Small Area Estimation (SAE) methods provide a framework for combining data sources in Official Statistics.

Many cases of SAE involve continuous variables such as income or simple binary categories such as poverty. For these multilevel regression models using correlating administrative and census variables are in common use. However, where the interest is in multiple categories and where particular correlating variables are harder to justify, then these methods are less flexible.

Generalised Structure Preserving Estimation (GSPREE) can combine low precision cross classified small area by variable-of-interest data (e.g. from sample survey) with higher precision data of the same cross classification (e.g. from administrative data) and with aggregate margin data in a model to produce good estimates for the categorical variable-of-interest. Such examples are population size by ethnic group or by household characteristics. The methods are flexible in that additional data sources can be introduced as available and existing sources excluded if they are no longer relevant.

GSPREE has been used to produce population estimates for ethnic groups by local authorities in England (Zhang and Chambers, 2004; Luna-Hernandez et al., 2015). The performance of the GSPREE estimator has now been assessed in a validation scenario where the true population distribution is known (i.e. March 2011 Census). A GSPREE model considered the contribution of different auxiliary information (2001 Census and 2011 English School Census) for various age groups, alongside survey data. Uncertainty in the GSPREE estimates was estimated using a bootstrap.

Key Words: small area estimation; GSPREE; administrative data; combining data; ethnicity

1.2.3 Measuring uncertainty in ONS population estimates: Tools, techniques and outcomes

Louisa Blackwell, Katy Stokes, Paulina Galezewska; Office for National Statistics

The Office for National Statistics (ONS) in collaboration with the Southampton Statistical Sciences Research Institute (S3RI) at Southampton University has developed methods for estimating the uncertainty associated with the annually-published local authority mid-year population estimates. By 'uncertainty' we mean the quantification of doubt about the estimates. This research and the uncertainty estimates that we produce are intended to support decision-making by users of these key National Statistics. This year we published uncertainty measures for the 2012-15 local authority mid-year population estimates for England and Wales. This included confidence intervals for each local authority, and an estimate of the contributions that the 2011 Census, internal migration and international migration made to the uncertainty for each local authority.

The methodology involved a simulation approach which replicates the cohort component methodology used to create the mid-year population estimates. This presentation describes the methodology in more detail and provides some illustrative results. We are now conducting research to assess how these methods and results can be used to quality assure the research outputs from the Statistical Population Dataset (SPD). The latter are new population estimates created from administrative data. We will describe how uncertainty in the mid-year population estimates is helping us to i) create target quality standards for the SPD population estimates, ii) assess the quality of the current SPD research statistics and iii) develop direct measures of uncertainty for the SPDs.

Key Words: uncertainty; population estimates; bootstrapping; quality assessment; administrative data

1.3 Morning Session 3, Syndicate Room 3

Towards the future of time series and longitudinal data analyses

Chair: Frances Pottier, Department for Business, Energy and Industrial Strategy

1.3.1 Anticipating Brexit effects in time series analysis

Jennifer Davies; Duncan Elliott; Charlotte Gaughan; Atanaska Nikolova; Office for National Statistics

The outcome of the European Union referendum and the exiting of the United Kingdom from the EU could have a widespread impact on economic time series produced by the Office for National Statistics. The impact could manifest itself in the form of outliers, level shifts and ramps at the end of a time series, which can be difficult to identify and categorise correctly. Failing to account for these correctly can lead to revisions to seasonally adjusted series and forecasts when they are later detected. We aim to minimise the impact of revisions to ensure an accurate and reliable estimate of the economy. This paper presents an empirical study of univariate options for modelling shocks, taking in to account previous research in the field. It then proceeds to examine a multivariate approach for forecasting, considering relationships between different time series, and utilising these as indicators. A final consideration is also given as to whether incorporating expert Economists views on Brexit effects in to models, is a worthwhile approach.

Key Words: time series, seasonal adjustment, forecasting, outliers.

1.3.2 The Impact of moving holidays on official statistics time series

Bethan Russ; Tariq Aziz; Office for National Statistics

A major challenge faced when seasonally adjusting time series is accounting for annual events that move dates from one calendar year to the next, for example, Easter. If these events are not accounted for appropriately it will impact on the estimation of seasonal factors, and leave systematic calendar related effects in the seasonally adjusted series. Currently the Time Series Analysis Branch (TSAB) tests for Easter effects and, if identified, estimates and removes them as part of seasonal adjustment. This method assumes that daily activity changes by a fixed amount or proportion for a given number of days before Easter Sunday and remains at this level until Easter Saturday.
There are other moving holidays celebrated in the UK, which may have an impact on time series despite not being public holidays. These are Chinese New Year, Ramadan, Eid al-Fitr and Eid al-Adha. Currently these holidays are not adjusted for in any seasonally adjusted time series published by the Office for National Statistics (ONS).
TSAB has undertaken research to test alternative windows for Easter effects and whether other moving holidays have identifiable effects on ONS time series.
This presentation will present the methodologies used in the research and the findings on a range of ONS time series.

Key Words: time series; seasonal adjustment; calendar effect; moving holiday; regARIMA

1.3.3 Increasing frequency and improving timeliness of unemployment estimates from the UK Labour Force Survey

Duncan Elliott; Office for National Statistics

Unemployment is estimated from data collected in the Labour Force Survey (LFS). While data is collected continuously the survey design is structured in such a way as to provide quarterly estimates. These quarterly estimates are published each month as *rolling quarterly* estimates and have been assessed to be of sufficient quality to be designated as *National Statistics*. ONS also publish monthly estimates of unemployment, but these are designated *Experimental Statistics* due to concerns over the quality of these data. Due to the sample design of the LFS, monthly estimates of change are volatile as there is no sample overlap. A state space model can be used to develop improved estimates of monthly change, accounting for aspects of the survey design, and so providing increased frequency and a slight improvement in timeliness. Additional sources of information related to unemployment could be used within such a framework to help improve timeliness further. In this paper an overview of the modelling approach is provided and issues associated with regular publication of results are discussed.

Key Words: time series; state space models; unemployment; repeated surveys

1.4 Morning Session 4, Syndicate Room 4

Sampling and sample design: Optimising the allocation

Chair: Aidan Mews, Ministry of Justice

1.4.1 The 2016 Annual Business Survey: A sample re-design

Megan Pope; Jonathan Digby-North; Office for National Statistics

In terms of the number of variables collected, the Annual Business Survey (ABS) is the largest business survey run by the Office for National Statistics (ONS), sampling approximately 62,000 businesses in Great Britain. It covers the production, construction, services and distribution industries and provides important indicators of economic activity, including estimates of total turnover, the total value of purchases of goods, materials and services and approximate Gross Value Added, much of the information it provides feeds directly into the UK National Accounts.

The ABS sample was last optimised by Methodology in 2010; over this time the target population had increased in number by around 25%, both due to natural evolution and the (first-time) introduction of certain single source (PAYE only) businesses into the target population. Other requirements included the selection of 'Sharing Economy' businesses and ensuring that all reporting units (RUs) within a multi-RU enterprise were selected, to enable the delivery of robust enterprise level statistics to Eurostat in 2017. A survey re-design was therefore considered essential.
The method used to allocate the sample to strata was a 'power' allocation, which essentially seeks to simultaneously optimise the precision of estimates at both the overall and lower (e.g. strata) levels. The performance of the new allocation was assessed on multiple years of past data, enabling us to quantify the expected impact on the quality of the estimates at various levels of aggregation.

This presentation will discuss the survey, details of the redesign and provide an update on the enterprise level work.

Key Words: sample design; Annual Business Survey; sample allocation

1.4.2 Which business to price? Sampling Business Price Indices

Matthew Mayhew; Office for National Statistics

While price data collection for the Consumer Prices Index, is conducted by the price collector walking into a shop and recording the prices of items, price quotes for the Business Price Indices cannot be collected as easily. Therefore a different collection method is required. Surveys are sent to businesses asking them to provide price quotes for the products or services they provide. These surveys need to be allocated efficiently to allow for coverage of the appropriate sector of the economy, to maximise accuracy for the all items index, and to minimise the cost of the surveys. This allocation is done to two stages, first the number of businesses to sample in each area of the relevant economic sector, and second the number of quotes each business is asked for, with a burden of constraint. Recently all four Business Price Indices have undergone a sample reallocation, with an increase in sample size for Services, Exports and Imports to improve the quality of these indices, and allowing for harmonisation of the methods. This paper describes the method and challenges overcome in the reallocation of the samples.

Key Words: price indices; sampling; production; services; trade;

1.4.3 Standard errors from audits

Sumit Rahman; Department for Business, Energy and Industrial Strategy

In this paper we discuss two examples of where the author was asked to derive standard errors following audits (of funding and loans for further and higher education) based on random sampling. In both cases the sampling methods were not explicitly related to the estimators being used in the audit and some investigation was required to produce standard errors for the estimates.

In the first example, the standard error was straightforward to produce, as it was simple to show the main estimate from the audit was a form of ratio estimator. We were able to make recommendations to improve the sample in future years to make estimation more efficient (i.e. reduce the standard error without increasing the size of the sample or the cost of the audit). The impact of this work was to improve the quality of the department's accounts and the quality of the audit.

In the second example, it was much harder to produce the standard errors at first as there was a variety of sampling methods used in different strata and in one stratum no relationship between the sampling method and the estimator. We show how a transformation of the sampled values allowed us to arrive at an answer, and then discuss how the

sampling method could be improved and (briefly) whether model-based estimation would be worth pursuing in this case.

Key Words: standard errors; ratio estimators; Neyman allocation; design-based estimation

2.1 Afternoon Session 1, Syndicate Room 1

Modernising methodology through transformation, collaboration, and harmonisation

Chair: Gareth James, Office for National Statistics

2.1.1 Transformation of methodology in ONS - the why and how

Gary Brown; Office for National Statistics

The Office for National Statistics was historically the leading 'survey house' in the UK – we are now transforming to be the leading data house, regardless of the source: surveys; Big data; administrative data. To unlock the potential of new data sources requires an upgraded technology base, an unshakeable foundation of data architecture, and new innovative methods for utilising the new sources in production: alongside and in place of existing sources. In order to deliver the new methods required, the whole methodological function in ONS was reviewed by an external expert and a wide range of recommendations were made. These recommendations have been turned into a transformation plan which will make methodology in ONS more efficient, more effective, more timely, more innovative, and more responsive to the emerging needs of users and the UK statistical system. This talk focuses on the principles driving the transformation, and explains what benefits the changes will bring to the GSS.

Key Words: innovation; data revolution; methods as a service

2.1.2 GSS Harmonisation Programme: Past, Present and Future

Ian Sidney, Becki Aquilina; Office for National Statistics

The Harmonisation Team works across the Government Statistical Services (GSS) to produce harmonised definitions, questions and outputs. These harmonised principles are reused across national statistics allowing users to more easily compare data from different sources and makes our statistics easier to understand.

The GSS Harmonisation Programme has developed a range of harmonised definitions, questions and outputs for a range of key social topics over the past ten years or so. However, it is now looking at developing business harmonised principles as well.

The Bean Review, EUROSTATs Framework Regulation Integrating Business Statistics (FRIBS) and the requirements of the UK Government's Digital by Default initiative are providing an ideal opportunity to harmonise business survey questions and definitions. The Office for National Statistics (ONS) is aiming to move all business surveys from paper to an electronic data collection mode (EDC). The EDC programme gives rise to a unique opportunity to review all ONS business survey questions and to develop harmonised business survey definitions, questions and outputs and through a process of rationalisation reduce the number of surveys wherever possible. This has the potential to reduce cost of collection and respondent burden. It will potentially improve the quality, timeliness and accuracy of data and could increase response rates from certain groups.

This presentation will focus on what has been achieved to date with social and business harmonisation, how we work across the GSS to develop and encourage the use of the harmonised principles, how key programmes such as the 2021 Census are instrumental to the harmonised programme and how you can help to develop and promote these harmonised principles.

Key Words: harmonisation; electronic data collection; census; business variables; social principles,

2.1.3 Lessons from Anthropology

Tacey Laurie; KAI Analytics HM Revenue and Customs

Working with the data generated by HMRC's tax platform presents unique challenges, many of which are methodological. What do you do when you are unable to access the data statistics come from? When the compilation process is not available? When there are no data sources for benchmarking?

This presentation presents insights gained from my secondment to the HMRC Digital Data Centre Operations Team. Using an approach borrowed from social anthropology, it examines differences between statisticians and the people who build our online services.

It offers practical advice on how to create better working relationships, communicate more effectively, set boundaries to allow work to move forward and avoid getting an ulcer while at the same time demystifying the working practices of engineers, technicians and developers.

Key Words: digital; online; data; analysis; communication

2.2 Afternoon Session 2, Syndicate Room 2

Linking into administrative data: Infrastructure, method, and application

Chair: Sumit Rahman, Department for Business, Energy and Industrial Strategy

2.2.1 Addressing for Census and beyond

Alistair Calder, Office for National Statistics

As in 2011 the address register is right at the centre of the 2021 Census design and impacts on virtually every census process. A high quality address list is used to deliver internet access codes, to allow targeting of follow-up for non-responding households and to underpin estimation and the production of outputs.

The over-all approach will build on the 2011 design and experience but will be extended to include the use of more address-level intelligence from administrative, commercial and open-data sources. By knowing more about addresses it will be possible to better target resources by, for example, identifying those areas where we are likely to find vacant properties, second residences or access problems.

A register of communal establishments (university accommodation, prisons, caravan parks, army camps, etc) will form an integral part of the address register.

As well as being central to the field operation, the address register forms the spine for the reuse of administrative data for 2021 outputs and beyond. High quality matching is essential throughout. ONS have already developed expertise in matching between addresses and, following discussions with Government Digital Services (GDS), have been leading on developing an address look-up and matching service for use across ONS – and potentially across government.
This session will describe current thinking for the census and beyond and describe the innovative work currently being carried out applying data science thinking to address matching.

Key Words: addresses; census; matching

2.2.2 Many-to-many linkage: Finding siblings in birth registration data

Christos Chatzoglou; Charles Morris; Shelley Gammon; Julie Mills; Lynda Cooper; Theodore Manassis; Office for National Statistics

This presentation provides an overview of the exploratory work on the production of a 'pregnancy spine' from birth registrations data, and presents the new challenges experienced when performing many-to-many linkage of multiple siblings to their mother. The resultant sibling dataset will be used for ONS and ADRC-E projects relating to health outcomes, child and maternal health.

Data relating to birth registrations are processed by, and held in ONS. This dataset contains 3.7 million unencrypted live and still birth registrations for the years 2000-2005. Probabilistic matching techniques were applied to link siblings to their mother based on the mother's (maiden) name, date of birth, country of birth and address, as mothers' NHS numbers were not recorded pre-2005).

The birth registrations dataset was linked to itself using probabilistic techniques (Fellegi-Sunter). Traditional methods only allow one-to-one matching and thus focus on finding the best match for each record. Here a new method was needed to identify many-to-many links, corresponding to multiple siblings, and to create familial clusters.
The effectiveness of graph structures and graph-theory methods to extract and analyse linked siblings' data is explored. Graph clustering metrics are investigated for their potential to optimise linkage quality.

Further work will be done to assess linkage quality. It is hoped, this project may be expanded to create more complete groups of maternal siblings through inclusion of data from earlier years, and to investigate the possibility of linking in other sources of maternity data.

Key Words: data linkage; probabilistic matching; graph clustering

2.2.3 Household effects in Big Data research: Privacy-preserving methods for grouping individuals into household groups, and its use in evaluating a Welsh Government fuel poverty scheme

Karen Tingay; Sian Morrison-Rees; Administrative Data Research Centre, Wales, Swansea University
Sarah Lowe; Knowledge and Analytical Services, Welsh Government
Charles Musselwhite; Centre for Innovative Ageing, Swansea University

The effect of the social environment on physical health and wellbeing has long been an area of study. Extrapolating this would potentially reduce confounding effects in health-related research. While household-level data can be obtained through surveys, doing so using routine administrative linked datasets, such as those held in the Secure Anonymised Information Linkage (SAIL) databank, held at Swansea University, allows for a broader population to be measured using a wider variety of metrics. However, routine datasets often have issues with data quality. Additionally, grouping individuals into households, especially when linking to other data sources, increases the risk of re-identification. A method for securely and accurately modelling household composition on the population as a whole, using routinely collected administrative data, are of benefit to public health research.

The presentation will describe the methods used in creating the households, issues with using routine administrative data to link anonymised individuals in this way, and how the methods were used the evaluate a Welsh Government intervention to improve the health of those living in fuel poverty: the Warm Homes Nest Scheme. Research findings demonstrated a positive effect on respiratory health, and suggested an improvement on cardiovascular conditions, although the latter figures were too small to be significant and will be revisited at a later stage. These findings were used to inform the Warm Home Nest successor scheme. More broadly, these methods have the potential to inform more effectively focused government schemes across a wide spectrum of health and social areas.

Key Words: linked population data; anonymisation; evaluation; households; fuel poverty

2.3 Afternoon Session 3, Syndicate Room 3

Statistical outputs: Enhancing the user interface and protecting confidentiality

Chair: Nick O'Rourke, Office for National Statistics Geography

2.3.1 The methodology of maps

Bruce Mitchell; Jeremy Brocklehurst; Anna Harris; GIS Mapping and Spatial Analysis Unit, ONS Geography

Perhaps it is easy to forget that any map is actually *designed* and issues from a 'decision tree' which accommodates and balances the demands of many competing and often incompatible objectives. The designer has to pick a way through the tree to attain the best result. At each decision point on the tree, the draft map changes and a new array of choices fans out – and sometimes none of these is appropriate, forcing you to step backwards and seek an alternative route.

The decision tree starts with the question of what precisely the map is required to display – what is the core message that it is supposed to convey? Other basic decision points relate to the framing of the geography, the geographical projection used and the scale – all of which are intimately linked with the size of the map – and this has to align with the requirements of the publication medium.

Only then should the designer address the visual aspects of the map, such as figure-ground, balance, emphasis and colour choice and (for statistical maps) the representation of the data, or theme, which may be counts, or rates, or classes.

For this paper, we will be looking at the design decisions involved in the creation of a single map of Europe. Because our objective was to create a template for long-term use and not merely the satisfaction of a single commission, we created and traversed a large decision tree before getting to our publication version.

Key Words: mapping; design; cartography; art

2.3.2 Improving workforce analysis through geographic design

Mark Baines; ONS Geography

For the 2001 Census there were only 3 tables published for workplace statistics at Output Area (OAs) level. As OAs are based on where people live and not where people work they are not suitable for publishing detailed workplace statistics as they would be disclosive (enable users to identify workers or workplaces from the statistics).

The presentation will explain how a workplace geography (based on the working population) was created for the 2011 Census for England and Wales to solve this problem. It will cover the methodology of how OAs were used as the building blocks to create Workplace Zones (WZs), a small area geography that is not disclosive with a consistent number of workers. This approach increased the number of tables that could be published to 30 for the 2011 Census.

Finally, the presentation will explain how the WZ methodology was implemented to address the challenges posed by the Scottish and Northern Irish workplace data. It will concentrate on explaining how this enabled the Classification of Workplace Zones (COWZ - a UK area classification) to be developed and how it helps businesses and employers to conduct data analysis at a more granular level. The impact of this has been to help local government make more informed decisions around transport planning and enable businesses to make decisions on where to open a new store.

Key Words: Geography; Workplace; Spatial Analysis;

2.3.3 Disclosure Control: Applying Cell-Key Perturbation to 2021 Census outputs

Keith Spicer; Stephanie Blanchard; Iain Dove; ONS Disclosure Control

In 2011, ‘pre-tabular’ protection in the form of record swapping was applied directly to the census microdata. After the standard releases, outputs could be requested by users, which then had to be individually created and assessed for disclosure risk by the SDC team. This process took considerable time, and after the redesign of tables for protection, some users were faced with outputs that did not meet their needs.

For Census 2021, ONS would like to allow much faster access to data and allow users to define/ create their own outputs. To remove the need for individual checking of tables, a post-tabular method of protection, ‘cell-key perturbation’, will be needed alongside record swapping. This perturbation will be required to protect against ‘differencing’ of tables but will consequently have an impact on outputs that will need to be explained to users. This talk will describe and demonstrate the proposed ‘cell-key perturbation’ protection method, the possible outputs system for 2021, how this differs from previous census methods and the resulting benefits and trade-offs for users.

Key Words: Disclosure Control; Census; Perturbation; Cell-key method

2.4 Afternoon Session 4, Syndicate Room 4

Big Data analytics: Opening new avenues of innovative research and understanding

Chair: Owen Abbot, Office for National Statistics

2.4.1 Natural Language Processing (NLP) techniques for structuring large volumes human text data

Alessandra Sozzi; Kimberley Brett; Office for National Statistics

The Big Data team in the Office for National Statistics (ONS) have been exploring the benefits of using Natural Language Processing (NLP) techniques. NLP is concerned with using computer algorithms to understand, and sometimes classify, large volumes of unstructured human text. The speakers will start by providing a short overview of some NLP techniques before covering how these have been used in two projects:

- Using descriptions of properties in housing website data to identify caravan properties and whether they are more likely to be residential or holiday homes. It is vital to understand where people may be living for the Census in 2021 to ensure that everyone in the country is counted. Caravan properties, and whether people are living permanently in them, are not recorded well in many data sources so the team used descriptions from the Zoopla website to add insight about these properties. This work will improve the efficiency of the Census, the largest peacetime operation carried out in the country.
- Applying sentiment analysis techniques to analyse the content of messages for the presence of defined expressions describing whether they were positive/negative, or displayed different emotional states e.g. joy, sadness, fear, anger. Two examples of applications include:
    - Automating feedback from internal staff seminars to understand sentiment.
    - Monitoring the level of daily social sentiment from Facebook comments towards events/topics. As part of the Eurostat task force in Big Data, the project aimed to identify new sources and techniques that can help towards understanding the level of daily satisfaction of citizens.

Key Words: Big Data, natural language processing, free text analysis, housing data, survey data

2.4.2 Using machine learning techniques to analyse economic statistics: a case study with HMRC Trade in Goods statistics

Andy Banks; Office for National Statistics

Recent developments in machine learning and data visualisation techniques have led to growing interest in their application to statistical and econometric analysis.

This paper examines two potential applications of machine learning in analysing economic statistics: its use in regression analyses and outlier detection. The paper applies both to a HMRC administrative dataset that measures UK exports of products to the rest of the world.

First, we consider the use of wide and deep neural networks to produce linear regression models. These techniques are tested on the administrative data to produce predictive models of UK exports to other countries, and are tested against more 'general' regression techniques.

Secondly, we test unsupervised machine learning algorithms to show their effectiveness in measuring outliers in trade administrative data. This can be used to communicate the impact that a small number of outliers may have on headline aggregate statistics.

Finally, the advantages and limitations of a range data mining techniques are discussed.

Key Words: trade statistics, linear regression, outlier detection, unsupervised machine learning, neural networks

2.4.3 Exploring mental well-being from prisoner case-notes using text mining

Jo Lee; Advanced Analytics Unit, Analytical Services Directorate, Ministry of Justice

Evidence on people with mental health and substance misuse is costly to commission, yet these individuals are over-represented in the prison population. The most recent Adult Psychiatric Morbidity Survey of prisoners, published almost 20 years ago (Singleton et al., 1998), found that over 90% had one or more of five studied psychiatric disorders (psychosis, neurosis, personality disorder, hazardous drinking, and drug dependence).

To address this gap in knowledge, text mining techniques were explored on prisoner case-notes, where prison officers enter free text to describe prisoners' progress. The case notes are written in an ad hoc fashion, recording interactions that range from formal interviews/meetings to chance encounters. The case-notes contain information about prisoner well-being which can be explored to determine mental health despite gaps in timelines for prisoners, a large variation in length or detail of case notes, and no pre-defined coding or structure of information.

Here, I will explore the challenges faced by handling such large datasets (5 million case-notes are recorded each year), and how to explore the relatively vague topic of mental health in text, as recorded by non-specialists. Mental well-being is not easily categorised, but if done properly can provide rich context for prisoners throughout their time in custody. Having encoded mental health issues from the case notes, we are using them to improve predictive analytics (e.g., the risk of committing violent acts while in custody) and implementing an R Shiny application allowing operational staff to view a summary of issues logged across a prisoner's case note history.

Key Words: text mining; NLP; R Shiny; predictive analytics

GSSM22 Exhibitors and announcements

Please make time to visit the exhibitor's who will be only too happy to answer questions or have an informal chat. There are exhibitor's stands in Syndicate Room 1 and the Foyer.

In the Foyer: The Royal Statistical Society (RSS)

The RSS is a world-leading organisation promoting the importance of statistics and data - and a professional body for all statisticians and data analysts. Membership is for anyone interested in data and will give you a voice to shape decisions and promote the role statistics play in society. Learn more about the RSS at www.rss.org.uk

The RSS Excellence Awards

On the evening of the 12th of July, following GSSM22, the RSS will be hosting their annual Excellence Awards and summer reception at 12 Errol Street, London EC1Y 8LX (register here: https://www.statslife.org.uk/stats-excellence-awards). This ceremony, sponsored by the UK Statistics Authority and the Economic and Social Research Council, will be commending the good work of those in the official statistics community and media in their use of data and statistics, with the award for official statistics being presented by National Statistician, John Pullinger. **All are welcome**

In the Foyer: University of Southampton: MOffStat and MDataGov

The MSc in Official Statistics (MOffStat) programme is a collaboration between the University of Southampton and the Office for National Statistics (ONS) which is designed to provide you with the specialist skills and knowledge which are central to the conduct of professional statistical work in government. The new MSc in Data Analytics for Government (MDataGov) shares some of the statistical materials, but adds data science related topics, and will be on offer from 2017/18.

Many of the skills taught on these two programmes, such as survey methods and data analysis, are also in great demand by employers outside government and the programme provides relevant training for professional positions in a wide range of organisations conducting large-scale statistical work.

To find out more about the MSc in Official Statistics and MSc Data Analytics for Government Programmes, come along to our exhibit to collect a copy of the latest programme information and to speak to the Programme Director, Paul Smith, who will be on hand to answer any of your questions.

In the Foyer: The Learning Academy (Analytical/Data Science Branch)

The Learning Academy, based in the Office for National Statistics, aims to provide a flexible learning programme to support you in strengthening and updating your professional skills and knowledge. The courses and offerings available provide a useful overview of the most commonly used analytical tools and methodologies in official statistics. New offerings focus on new programming languages (R and Python) and their application within the analytical field.

Come along to our exhibit on the 6th July to meet the team and collect the latest copy of the Learning Directory!

In Syndicate Room 1: The Economic Statistics Centre of Excellence (ESCoE)

ESCoE is an investment by the Office for National Statistics (ONS) in response to the findings of the Bean Review. It is a consortium of leading institutions led by the National Institute of Economic and Social Research (NIESR) with King's College London, innovation foundation Nesta, University of Cambridge, Warwick Business School (University of Warwick) and Strathclyde Business School.

ESCoE's ambitious core research programme consists of thirteen projects in three broad areas: National Accounts and Beyond GDP, Productivity and the Modern Economy, Regional and Labour Market Statistics. To find out more about our work please visit the stand and pick up a copy of our research programme.

For more information about upcoming events and to take a look at our recent output, including blogs and podcasts please visit our website: http://www.escoe.ac.uk/

The GSS Methodology Advisory Committee (GSSMAC)

Would you benefit from support and advice on a methodological problem or issue?

The Government Statistical Service methodology advisory committee (GSS MAC) has two main aims. To provide:

- a forum to allow government statisticians to obtain advice on methodological issues from a group of interested and experienced professional statisticians from outside government, and
- an opportunity to build and strengthen links between the Government Statistical Service (GSS) and the rest of the statistical profession.

The committee meets twice yearly (May and November) to discuss statistical methodological issues relevant to the production and presentation of Official and National Statistics.

Information about the Committee, including full documentation from past meetings, can be found on the Committee's internet page:

http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/index.html

For more information please contact: methodology@ons.gsi.gov.uk