Government Statistical Service
Gwasanaeth Ystadegol y Llywodraeth

The 21st GSS Methodology Symposium

Celebrating 21 years of innovation

Westminster Conference Centre
London
6 July 2016

Welcome to the 21st GSS Methodology Symposium 2016

Methodology: 'the key to the door' of innovation.

In popular culture a 21st birthday is often celebrated as a pivotal point in personal development. Traditionally, at 21, the experience necessary for making the decisions needed to navigate an uncertain future is considered to be attained. Hence the saying: "21: The key to the door" of adult life.

In 2016 the GSSM Symposium celebrates its 21st birthday. However, it is fair to say that since William the Conqueror ordered the Domesday Book (c.1085), Florence Nightingale made clear the plight of injured soldiers during the Crimean war (c.1856), and Sir Ronald Fisher introduced analysis of variance to the world of research (1918), methodology has had a little longer to grow and mature.

And yet, with an ever increasing demand for faster and better statistics, coupled with an extraordinary data revolution, we stand at a pivotal point in GSS history; facing a call to draw upon all of our experience to help make the challenging decisions needed to navigate the risks and obstacles ahead.

"I fully expect that, in five years' time, what we will be doing will be radically different."
(John Pullinger, 2015; Strategy for UK statistics, 2015 to 2020).

The GSSM21 Team hope sincerely that the symposium will be an opportunity to celebrate and share research and ideas serving to keep the GSS at the heart of Government policy. Methodology: 'the key to the door' of innovation.

Table of Contents

Enquiries: The GSSM21 Team:
methodology@ons.gsi.gov.uk: Steve Rogers
Karen Wilson
Lisa Eyre

9.00	**Registration**			
10.00 – 11.00	**MORNING KEYNOTE SESSION, *Syndicate Room 1:*** *Chair: Tricia Dodd, Chief Methodology Officer, ONS* **John Pullinger, the National Statistician:** **"Methodology: Celebrating 21 years of innovation"** **Iain Bell, Director for Data and Education Standards Analysis, Department for Education:** **"The impact of technology on the GSS"**			
11.00 – 11.20	**Refreshments & Exhibits:** **Cutting the GSSM Symposium 21st birthday cake: John Pullinger** ***Room: Outside Syndicate Room 1***			
	Morning Syndicate Sessions:			
	Session 1, Syndicate Room 1 **Methodological reviews: In practice and in theory** *Chair: Pete Brodie, ONS*	***Session 2, Syndicate Room 2*** **Exploring the use of administrative data in statistics** *Chair: Simon Compton, CMA*	***Session 3, Syndicate Room 3*** **Maintaining quality and minimising respondent burden** *Chair: Catherine Davies, ONS*	***Session 4, Syndicate Room 4*** **Advances in analyses: Utilising open source software** *Chair: Arran Cleminson, BIS*
11.20 – 11.50	**Methodological review of the VOA's Private Rental Market Statistics** *Neville de Souza, Stephanie Astley; Information & Analysis, Valuation Office Agency*	**Producing household estimates from administrative data: methodology and analysis towards ONS Research Outputs 2016** *Claire Pereira, Paul Groom, Pete Jones, Office for National Statistics*	**Quality within ONS: Providing a framework for statistical producers and assurance for our users** *Jill Pobjoy; Office for National Statistics*	**Interactive mapping using R** *Liam Cavin; Department for Business, Innovation & Skills*
11.50 – 11.55	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
11.55 – 12.25	**Facilitating improved Judicial diversity** *William Spry, Aidan Mews; Ministry of Justice*	**Modelling enterprise level statistics using administrative data** *Megan Pope, Jonathan Digby-North, Gary Brown; Office for National Statistics*	**Quality changes when moving from a survey to an administrative source** *Daisy Hamer, Jill Pobjoy; Office for National Statistics*	**Mapping migration** *Bruce Mitchell; ONS Geography, Office for National Statistics*
12.25 – 12.30	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
12.30 – 13.00	**No such thing as a neutral model: Why statisticians shouldn't shy away from informative priors in Bayesian modelling** *Ewan Keith; Ministry of Defence*	**Towards an integrated census-administrative data approach to item-level imputation for the 2021 UK Census** *Fern Leather, Katie Sharp, Steven Rogers; Office for National Statistics*	**Developments in measuring the burden placed on businesses responding to statistical surveys** *Adam Tucker, Denise Williams, Megan Pope, Ria Sanderson; Office for National Statistics*	**Equality and diversity analysis of performance management outcomes: A natural way of presenting results from ordered logistic regression, and sharing methodology using GitHub and R** *Sumit Rahman; Department for Business, Innovation and Skills*
13.00 – 13.45	**Lunch & Exhibits** ***Room: Outside Syndicate Room 1***			

13.45 – 14.45	**AFTERNOON KEYNOTE SESSIONS, *Syndicate Room 1:*** *Chair: Dr David Best, Director of Digital Services, Technology & Methodology, ONS* **Evelyn Ruppert, Professor of Sociology, Goldsmiths, University of London:** **"Trust in statistics"** **Ian Coady, Methodology & Statistical Infrastructure: ONS, Geography:** **"Connected and meaningful: how innovation can help us make more of statistics"**			
	Afternoon Syndicate Sessions:			
	Session 1, Syndicate Room 1 **Advances in linking and matching data** *Chair: Owen Abbott: ONS*	***Session 2, Syndicate Room 2*** **Improving the accuracy of statistical outputs** *Chair: Drew Hird, BIS*	***Session 3, Syndicate Room 3*** **Towards the management of multiple data sources** *Chair: Gareth James, ONS*	***Session 4, Syndicate Room 4*** **Improving quality and trust in statistical outputs** *Chair: Ruth Fulton, NISRA*
14.45 – 15.15	**Sampling procedures for assessing accuracy of record linkage** *Paul Smith; University of Southampton* *Shelley Gammon, Sarah Cummins' Christos Chatzoglou, Dick Heasman, Office for National Statistics*	**Modelling weather effects on road casualty statistics**** *David Mais, Daryl Lloyd; Department for Transport* *Jennifer Davies; Office for National Statistics*	**Developing an Integrated Business Survey System for Northern Ireland** *Dr. James Gillan; Northern Ireland Statistics and Research Agency*	**Reviewing aspects of quality reporting within ONS** *Sarah Tucker; Office for National Statistics*
15.15 – 15.20	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
15.20 – 15.50	**Use of Graph Databases to improve the management and quality of linked data** *Christos Chatzoglou, Theodore Manassis, Shelley Gammon, Nigel Swier; Office for National Statistics*	**Calibrating the 5-Quarterly longitudinal Labour Force Survey dataset** *Gareth Davies; Cardiff University*	**How Welsh Government worked with Schools, Local Authorities and MIS data suppliers to develop a system to collect and maximise the utility of their admin data** *Dani Evans; Welsh Government*	**A new process for assessing the quality of an output makes better quality statistics** *Catherine Davies, Catherine Bremner; Office for National Statistics*
15.50 – 15.55	*5 minute break*	*5 minute break*	*5 minute break*	*5 minute break*
15.55 – 16.25	**Improvements in methodology for matching the 2021 Census to the Census Coverage Survey** *Sarah Cummins, Peter Jones, Shelley Gammon, Office for National Statistics*	**Calculating weights for the BIS Self-employed survey, an LFS follow up** *Katie Connolly; Office for National Statistics*	**What role can harmonisation play in a changing statistical landscape?** *Suzanne Ellis, Steven Cooley; Office for National Statistics*	**Users' understanding and use of uncertainty measures to describe data quality** *Silvia Manclossi[1], Victoria Ayodele; Office for National Statistics* *[1] (on loan to the Welsh Government)*
16.25 – 16.40	**Refreshments & Exhibits**			***Room: Outside Syndicate Room 1***
16.40 – 17.00	**CLOSING SESSION, *Syndicate Room 1:*** *Chair: Tricia Dodd, Chief Methodology Officer, ONS* **Closing comments & announcements: John Pullinger**			

****Project supported financially by the GSS Quality Improvement Fund (QIF)**

21st GSS Methodology Symposium: Keynote Speakers

Morning Session, Syndicate Room 1

Chair: Tricia Dodd, Chief Methodology Officer, ONS

John Pullinger
UK National Statistician

John was appointed as the UK National Statistician, Head of the Government Statistical Service (GSS) and Chief Executive of the UK Statistics Authority in July 2014. The latter role includes executive responsibility for the Office for National Statistics (ONS).

In his early career John worked at the Department of Trade and Industry, the Department of the Environment and the Office of Manpower Economics. In 1992 he became Director of Policy and Planning at the then Central Statistical Office where he became project manager for the creation of the ONS. As Head of Profession John has chaired many GSS committees and led the neighbourhood statistics programme. John was involved in the development of statistical governance which led to the creation of National Statistics and the Statistics Commission in 2000.

In 2004, John became the 14th Librarian of the House of Commons. As a member of the Management Board, John has worked closely with the Speaker of the House and several committees in both Houses to develop the connection between Parliament and the public. John is a former President of the Royal Statistical Society and was the inaugural chair of the Royal Statistical Society's "getstats" campaign. John was appointed as a Companion of the Order of the Bath (CB) in the 2014 New Year's Honours for services to Parliament and to the community.

Methodology - celebrating 21 years of innovation

The GSS Methodology Symposium has come of age. John will review the innovations that have been discussed at the Symposium over the years. He will draw out those he feels have had the greatest impact on the GSS and draw some lessons to help guide the future development of methodology. Methodology has a particularly important role in the delivery of the Better Statistics, Better Decisions strategy as we seek to mobilise the power of the data revolution. John will conclude by considering the questions we will need methodologists to tackle if we are to maximise our potential to help Britain make better decisions.

Iain Bell
Director for Data and Education Standards Analysis, Department for Education

Iain is Director for Data and Education Standards Analysis in the Department for Education. He is responsible for data collection in the Department, school performance tables and policy-facing analytical support across schools, post-16 and teaching policy. He has been in the Government Statistical Service for 22.5 years (so slightly pre-dates the methodology symposium). His career in the GSS has spanned labour market, Government Finance, Transport, Justice, work, pensions and education statistics. Throughout his career he has specialised in exploiting the power of administrative data.

The impact of technology on the GSS

Iain reflects on 21 (and more) years of technological advancement and the impact this has had on the work of the Government Statistical Service; the things we do now that he would never have dreamed possible at the start of his career and his views on the lessons this holds for the GSS and the methodologies we use moving forward.

Afternoon Session, Syndicate Room 1

Chair: Dr David Best, Director of Digital Services, Technology & Methodology

Evelyn Ruppert

Professor of Sociology, Goldsmiths, University of London

Evelyn describes herself as Data Sociologist. She joined the Sociology Department at Goldsmiths, University of London in April 2013 from the Centre for Research on Socio-cultural Change (CRESC); a collaboration between the Open University and the University of Manchester. Evelyn is currently Principal Investigator of a European Research Council funded project called *Peopling Europe: How data make a people* (ARITHMUS, 2014-19). She is founding editor of a SAGE open access journal, Big Data & Society: Critical Interdisciplinary Inquiries, launched in June 2014. Prior to relocating to the UK, she worked for five years on the Canadian Century Research Infrastructure project, which involved building micro-databases from early 20th century Canadian census manuscript forms.

Trust in Statistics

Evelyn will reflect on insights from the ARITHMUS project, which involves the study of several National Statistical Institutes, UNECE and Eurostat as they remake methods of knowing populations. From questions of the politics of working with administrative and big data to technical and infrastructural barriers to innovating methods, she will discuss the role of trust in securing professional, political and public confidence in statistics. How is trust understood, demanded and performed? And what does it mean to trust statistics at a time when numerous new actors are generating competing data, methods and population knowledge?

Ian Coady

Methodology & Statistical Infrastructure: Office for National Statistics, Geography

Ian is Policy and Research Manager for the ONS Geography Team. Ian has been with ONS for 7 years having moved over from local government and an earlier career as a landscape archaeologist. Having originally been brought in to support the 2011 Census geographic maintenance and deliver the Office's commitments to the European Commission's INSPIRE Directive he has since covered a wide range of policy and research topics including the development of 2011 Census grids for Eurostat, the development of the Workplace Zone geography, the implementation of linked data and supporting the UN Global Geospatial Information Management Initiative. Ian is also a Director at the Association for Geographic Information and works to promote the use of geospatial data across government.

Connected and meaningful: how innovation can help us make more of statistics

The ways in which we analyse and interpret data are changing. Increasingly methodologies are built around big data, data linkage and data science but what lies beyond this? Ian will be looking at the concepts of 'linked data' and the 'semantic web' as methods of structuring and linking data so that it can be managed and analysed in more flexible ways. He will look at examples of the application of linked data in both a statistical and geographic context and consider what benefit this approach may have for methodologists in enabling them to extract greater value from data and deliver the modernisation of official statistics.

21st GSS Methodology Symposium: Parallel Sessions

1.1 Morning Session 1, Syndicate Room 1

Methodological reviews: In practice and in theory

Chair: Pete Brodie, ONS

1.1.1 Methodological review of the VOA's Private Rental Market Statistics

Neville de Souza, Stephanie Astley; Information & Analysis, Valuation Office Agency

The Valuation Office Agency (VOA) is an executive agency of HM Revenue & Customs. It carries out valuations of properties in England and Wales, to provide the Government with information required to determine taxation and benefits related to property. One strand of this is the collection of data by VOA rent officers on the private rental market in England.

The VOA uses the lettings data to produce Private Rental Market Statistics (PRMS) on a six monthly cycle. Following methodological work undertaken in the last five years by the VOA and Office for National Statistics, this rich source of information now underpins the Owner Occupied Housing element of the CPIH and Index of Private Housing Rental Prices. The data is also used by the Department for Work and Pensions to determine Local Housing Allowance rates for housing benefit claimants living in the private rented sector.

The VOA set out to improve the methodology for PRMS in 2015 and some of the work completed so far includes: an assessment of data sufficiency and implications for publication frequency and granularity; comparisons between the PRMS population and Census 2011 frequencies; investigation of key price drivers and regression modelling to estimate the marginal contribution of property attributes to the overall price; post-stratification. This methodological work is still in progress, but there are some interesting findings and lessons learnt in terms of carrying out a methodological project that we would like to share with the GSS community.

Key Words: Administrative data, Methodological change, Regression, Hypothesis test, Private rental market

1.1.2 Facilitating improved judicial diversity

William Spry, Aidan Mews; Ministry of Justice

Ministry of Justice statisticians have been investigating the impact of short listing tools (e.g. online tests) used in Judicial Appointment processes, with a particular focus on Black and Minority Ethnic (BAME) candidates. Key issues have included whether there are significant differences in the selection rates of BAME candidates, the contribution of other factors such as age and post qualification experience, and the impacts associated with the content and timing of the shortlisting tests. To address such issues a variety of analytical techniques were performed including hypothesis testing, calculation of effect size measures, regression, differential item functioning, Cronbach's alpha, corrected item total correlations and cut score modelling.

This session will explain the analytical work done with a focus on one of the Judicial appointment tests as an example and briefly outline the impact of the results in terms of policy follow-up.

Key Words: Judicial; Appointment; Test; Shortlisting; BAME

1.1.3 No such thing as a neutral model: Why statisticians shouldn't shy away from informative priors in Bayesian modelling

Ewan Keith; Ministry of Defence

Bayesian methods have achieved mainstream status over the past several decades. However, many analysts are still wary of using informative prior distributions in their modelling. This presentation makes the case for the use of such priors when they reflect analyst expectations. Two major developments in statistical modelling, hierarchical and

regularised modelling, are explored. These topics appear largely independent from a frequentist point of view. But from a Bayesian perspective they are both cases of using informative prior information in modelling. A simple non-Bayesian model will then be discussed briefly. Specifically, to show how even non-Bayesian methods rely on the analysts informal 'priors'. Informative priors are thus justified on pragmatic grounds, and on the grounds that they are no less (inherently) impartial than non-Bayesian analyses.

Key Words: Bayesian; Prior; Informative; Hierarchical; Regularisation

1.2 Morning Session 2, Syndicate Room 2

Exploring the use of administrative data in statistics

Chair: Simon Compton, CMA

1.2.1 Producing household estimates from administrative data: methodology and analysis towards ONS Research Outputs 2016

Claire Pereira, Paul Groom, Pete Jones, Office for National Statistics

The Office for National Statistics' Census Transformation Programme is responsible for taking forward three high-level deliverables:

- A predominantly online census of all 26 million households and communal establishments in England and Wales
- Development of alternative administrative data census estimates, compared to the 2021 Census
- Improved and expanded population statistics through increased use of administrative data and surveys

In October 2015, the Programme published its first set of Administrative Data Research Outputs, demonstrating progress on the second deliverable. Linked administrative data were used to produce population estimates at local authority by five-year age sex groups for 2011, 2013 and 2014.

The next release in autumn 2016 is expected to include estimates for the number of households at local authority level. This paper outlines the complexities of applying a census definition of households and their compositions when using administrative data. It also shows the progress made at ONS in the development of automated matching algorithms that can accurately assign Unique Property Reference Numbers (UPRNs) to addresses held on administrative sources.

We present an analysis of local authority household estimates derived from administrative data compared to official estimates. We also show how biases in the administrative data estimates can be reduced through combined use of a coverage survey using dual system estimation. The paper concludes by outlining our future aspirations to use administrative data to produce research outputs with distributions for household size and composition at local authority level.

Key Words: Linkage; Households; Administrative Data; Census; Dual System Estimation

1.2.2 Modelling enterprise level statistics using administrative data

Megan Pope, Jonathan Digby-North, Gary Brown; Office for National Statistics

The UK is required to provide estimates of Structural Business Statistics (SBS) to Eurostat on an enterprise reporting basis for the 2016 reference year. This represents a new output, as currently the Office for National Statistics (ONS) produces estimates on a Reporting Unit (RU) basis. For those enterprises with more than one RU, any intra-flows within the enterprise must be removed to produce consolidated accounts. A feasibility study was conducted to investigate whether it would be viable to estimate enterprise level statistics (e.g. turnover) using existing data.

Two external/administrative sources were identified as able to provide enterprise level data: FAME and VAT. FAME is produced by Bureau Van Dijk and VAT data is provided to ONS by HMRC. The data were matched to the Inter-Departmental Business Register (IDBR) and multiple regression modelling was used to derive relationships between RU level ABS/IDBR variables and the external enterprise level data.

In general, total RU turnover and employment were strong predictors of consolidated turnover. When applying the models to ABS data, consolidated turnover was found to be approximately 10-20% lower than the simple sum of RU turnover. Although the findings have been more or less consistent across data sources, there are data quality and implementation issues which require further work to resolve.

Keywords: Administrative data, Regression modelling, Enterprise level statistics

1.2.3 Towards an integrated census-administrative data approach to item-level imputation for the 2021 UK Census

Fern Leather, Katie Sharp, Steven Rogers; Office for National Statistics

In preparation for 2021 UK Census the ONS has committed to an extensive research programme exploring how linked administrative data can be used to support conventional statistical processes. Item-level edit and imputation (E&I) will play an important role in adjusting the 2021 Census database. However, uncertainty associated with the accuracy and quality of available administrative data renders the efficacy of an integrated census-administrative data approach to E&I unclear. Current constraints that dictate an anonymised 'hash-key' approach to record linkage to ensure confidentiality add to that uncertainty.

Here, we provide preliminary results from a simulation study comparing the predictive and distributional accuracy of the conventional E&I strategy implemented in CANCEIS for the 2011 UK Census to that of an integrated approach using synthetic administrative data with systematically increasing error as auxiliary information. In this initial phase of research we focus on imputing single year of age. The aim of the study is to gain insight into the performance of a conventional E&I strategy such as that used for 2011 UK Census compared to that using administrative data with increasing degree of error as auxiliary information.

Key Words: Census, Imputation, Administrative data, CANCEIS

1.3 Morning Session 3, Syndicate Room 3

Maintaining quality and minimising respondent burden

Chair: Catherine Davies, ONS

1.3.1 Quality within ONS: Providing a framework for statistical producers and assurance for our users

Jill Pobjoy; Office for National Statistics

A requirement of the European and UK Code of Practice is that National Statistics Institutes must define their quality policy and make this available to the public.

To ensure this requirement is met, ONS have launched an updated Quality Management Strategy which is publicly available on the ONS website to provide assurance to the users of our statistics and includes activities which monitor, improve and report on the quality of statistical products. It also serves as a useful framework for the producers of statistics within ONS.

The updated strategy reflects the activities that we have in place as an organisation to manage quality and sets out goals for improvement. It reflects the organisational approach and is relevant to all areas of the office. The strategy is supported by a statistical quality framework which sets out the quality initiatives in place for quality assurance, quality control, quality improvement and quality reporting.

The quality management strategy also strengthens the governance for quality management within ONS. A network of quality champions provide regular reports to the Quality Centre and any issues relating to quality management are fed up to a senior committee within the organisation twice each year

This paper/presentation will provide more detail on the quality strategy and provide additional information around the quality framework. We will report on the how things have progressed since the launch of the strategy and how the quality management culture is being embraced in the ONS.

Key Words: Quality; Strategy; Framework

1.3.2 Quality changes when moving from a survey to an administrative source

Daisy Hamer, Jill Pobjoy; Office for National Statistics

National Statistical Institutes are increasingly looking to use administrative data for statistical purposes, often replacing survey questions or whole surveys with such data. The purpose of this research is to look at possible changes in quality that result from moving from a survey to an administrative source. The first step to carrying out this piece of research was to conduct a literature review looking at what previous investigations have found and see whether any practical advice could be drawn from this. The second step was to find case studies, working with areas at the Office for National Statistics (ONS) which are, have, or are considering replacing survey data with administrative data and investigating quality changes that result. Research is still ongoing but current findings point to some benefits and some issues that arise as a result of replacing survey with administrative data. These are areas which need to be considered in the development of guidance for statistical producers. This paper will summarise findings from the literature review and the research conducted so far.

Key Words: Administrative Data; Quality; Research

1.3.3 Developments in measuring the burden placed on businesses responding to statistical surveys

Adam Tucker, Denise Williams, Megan Pope, Ria Sanderson; Office for National Statistics

The Code of Practice for Official Statistics specifies the need to report annually on the burden placed on respondents to surveys of businesses and households. Whereas information on the time taken for a household to respond to a questionnaire can be measured at the point of collection, it is more challenging to measure the time and cost to businesses of responding to surveys used to compile official statistics. A traditional approach to measurement of surveys conducted using paper questionnaires, is to send a short review questionnaire to a sub-sample of businesses. This gathers information both on the time taken to respond to the main survey but also who in the business provides this information; this can then be used to estimate the financial costs to the business.

Such reviews ceased at ONS in 2012 and information on respondent burden was collected through a self-assessment tool used by survey managers to assess the quality of statistical outputs. However, it proved difficult to collect high quality information on respondent burden without the data from these review surveys. Therefore, a shortened review process has been piloted, to balance the burden placed on respondents by this process and to make the process more efficient, we tested the use of statistical modelling to estimate respondent burden for surveys with similar characteristics.

In this paper, we report on the pilot exercise carried out, including the methodology, results and conclusions of this work. Also considering the implications for the future measurement of respondent burden placed on businesses.

Key Words: Respondent Burden; Code of Practice; Future Measurement

1.4 Morning Session 4, Syndicate Room 4

Advances in analyses: Utilising open source software

Chair: Arran Cleminson, BIS

1.4.1 Interactive mapping using R

Liam Cavin; Department for Business, Innovation & Skills

Everything happens somewhere.

Maps are an engaging tool that can help us to look up information for locations and investigate spatial patterns. This presentation is a case study of taking a traditional map produced as a poster, and turning it into an interactive web-based tool that allows users to interrogate the map.

A common method for producing interactive maps is to use the open-source JavaScript library Leaflet. This is a powerful and flexible tool, which allows you to create 'slippy' maps like those popularised by Google maps. However, Leaflet requires a considerable input of time to master. I didn't use Leaflet - I cheated. Having stumbled upon the R package leafletR, I was producing simple interactive maps within a few hours.

Key Words: Mapping; Data Visualisation; R; Leaflet

1.4.2 Mapping migration

Bruce Mitchell; ONS Geography, Office for National Statistics

ONS Geography have been investigating spatial trends in migration to and within England and Wales. We have used published 2014 data at district (LAD) level. The population studied were aged between 25 and 64, an age band considered to best represent the mobility of the labour force.

- Low levels of both international and internal migration correlate with the ONS Area Classifications 'Mining Heritage', and 'Manufacturing Traits'.
- Districts attracting low levels of **international** migration are most likely to be in Wales, the East and West Midlands, the North East and the North West.
- Districts generating low levels of **internal** migration tend to be in Wales, the North East and the North West and Yorkshire and the Humber.
- Internal migrations are most common occurred in London, the South East, South West, or East of England.

We used the statistical programming language R to produce flow maps, chord, Sankey and small multiple diagrams. We demonstrate that the LAD geography is too coarse and too variable to support the identification of internal migration catchment areas. Results of our analysis of average migration distances were often artefacts of the district boundary set rather than informative on labour mobility.

Projecting international data down below LAD level (e.g. MSOA) would not be possible, but it might prove possible for some internal migration data. This would be very challenging and resource-intensive, and potential benefits could be undermined by uncertainties in the data that underlie the distribution of migration down to lower levels. But if these were overcome, a more granular spatial analysis might solve many of the problems encountered in calculating catchment areas.

Key Words: Migration; International; Internal; Mapping; GIS; R

1.4.3 Equality and diversity analysis of performance management outcomes: A natural way of presenting results from ordered logistic regression, and sharing methodology using GitHub and R

Sumit Rahman; Department for Business, Innovation and Skills

We use ordered logistic regression to analyse performance management outcomes in BIS, looking to present the model's estimated effects to senior non-analyst colleagues in a way they understand. The key purpose of the analysis is to establish if a person's age, ethnicity, gender etc affect their probabilities of receiving the various scores in the performance management system. In this paper I show a way of deriving these probabilities objectively that focuses on the 'pure' impact of each individual demographic variable after correcting for the other confounding effects. Thanks to this method, everyone in BIS gets to see the different probabilities for women and men (say), which are consistent with the overall observed distribution of box marks (traditional 'marginal probabilities' do not achieve this) but which are also consistent with the effects found in the regression. People can easily see which demographic variables have relatively large impacts.

Having described the method, I then explain how I've implemented it in the R programming language by writing an R package and making it available on GitHub (a popular open repository for code). The method is available to anyone in the GSS who wants to use it, examine it – or improve it. Once it is seen how simple it is to (a) use someone's R package, (b) build on work that has been shared in GitHub and (c) make your own package in R, GSS colleagues will see that this is an ideal way of sharing methodologies and building on them.

Key Words: Ordered logistic regression; R packages; GitHub

2.1 Afternoon Session 1, Syndicate Room 1

Advances in linking and matching data

Chair: Owen Abbott, ONS

2.1.1 Sampling procedures for assessing accuracy of record linkage

Paul Smith; University of Southampton
Shelley Gammon, Sarah Cummins' Christos Chatzoglou, Dick Heasman, Office for National Statistics

The use of administrative datasets as a data source in official statistics has become much more common as there is a drive for more outputs to be produced more efficiently. Many of these rely on linkage between two or more datasets, and this is often undertaken in a number of phases with different methods and rules. In these situations we would like to be able to assess the quality of the linkage, and this involves some re-assessment of both links and non-links. In this paper we discuss sampling approaches to obtain estimates of false negatives and false positives with reasonable control of both accuracy of estimates and cost. Approaches to stratification of links (non-links) to sample are evaluated using information from the 2011 England and Wales population census.

Key Words: Record Linkage; Accuracy; Sampling

2.1.2 Use of Graph Databases to improve the management and quality of linked data

Christos Chatzoglou, Theodore Manassis, Shelley Gammon, Nigel Swier; Office for National Statistics

Some major challenges experienced with linked data are:

- Requirements for different data linkage quality from different users (even when linking the same datasets)
- Need of a targeted clerical resolution for the difficult cases
- Need to update linked data over time considering additional datasets, time-points and data amendments
- Violations of transitive closure when linking more than two datasets
- Long data load, querying time and storage memory

The effectiveness of Graph Databases as alternative tools (from a traditional Relational Databases approach) to address the aforesaid challenges is explored in our pilot project. Several synthetic linked datasets with known their truth match status, are stored and queried in a Neo4j Graph Database while their importing and querying times are recorded. Graph clustering metrics are investigated for their ability to optimise the linkage quality.

A Graph Database stores records in the form of "nodes" in a graph and any relationships (links) between the records as "edges". Relationships may be expressed using record pairs' similarity scores and they can be re-used and analysed at anytime.

Graph Databases proved to be very efficient in loading, storing and querying the given synthetic linked data while improving the linkage quality. Further work needs to be done to test the robustness of Graph Databases using real and big volumes of data.

Key Words: Data Linkage; Linkage Error; Graph Databases; Neo4j

2.1.3 Improvements in methodology for matching the 2021 Census to the Census Coverage Survey

Sarah Cummins, Peter Jones, Shelley Gammon, Office for National Statistics

Since 2001, the Census has been matched to a Census Coverage Survey (CCS) to facilitate the estimation of non-response. The quality requirements for this matching exercise are very high, since errors in matching will impact the estimates of the population. For the 2011 Census to CCS matching exercise, 70% of person matching was done via automated methods and the rest were matched clerically; the resulting false positive rate was <0.01% and false negative rate was <0.25%.

Since the 2011 Census to CCS matching exercise, progress has been made in the field of automated matching methods due to the requirement to match large admin datasets in a secure environment. This research investigates whether the cost and processing time of the 2021 Census to CCS matching exercise can be reduced by finding more matches using automated methods, without incurring unacceptable levels of error. Matching methods were tested using matched 2011 Census and CCS data as a 'gold standard', taking the links made in 2011 as the 'true' match status.

A hierarchical matching strategy was adopted using deterministic 'match keys', probabilistic matching and associative matching. Overall the matching left just under 4% of matches yet to be found at a 0.25% false positive error rate. This would substantially reduce the clerical resource needed for the 2021 Census to CCS matching exercise, but further work needs to be done to investigate the impact of a higher level of error and potential biases in that error rate.

Key Words: Matching; Census; Error; Estimation

2.2 Afternoon Session 2, Syndicate Room 2

Improving the accuracy of statistical outputs

Chair: Drew Hird, BIS

2.2.1 Modelling weather effects on road casualty statistics

David Mais, Daryl Lloyd; Department for Transport
Jennifer Davies; Office for National Statistics

The Department for Transport publish statistics on the number of people killed and injured in road accidents in Great Britain that are reported to the police. It has long been known that the weather impacts on year-on-year changes in these figures. It is important for road safety policy making to be able to assess how much the weather has contributed to these changes.

A cross government group looking at how the weather impacts on different types of statistical series can be assessed was established. This led to the development of a statistical model to produce a weather-adjusted road casualty series. The presentation would summarise the regARIMA modelling approach that has been used, as well as discussing the adjustments made to the figures in recent years. There was an increase in road deaths between 2010 and 2011, but when the weather is accounted for this increase no longer exists.

It is hoped that the presentation would generate some ideas on how the methodology could be improved in the future as well as encouraging the audience to consider potential weather impacts on the statistics they use.

Key Words: Weather; Road; Safety

2.2.2 Calibrating the 5-Quarterly longitudinal Labour Force Survey dataset

Gareth Davies; Cardiff University

The 5-Quarterly Labour Force Survey (LFS) dataset is used to produce estimates of change in employment status. The dataset is formed using people who respond to the Labour Force Survey for all five quarters in question. Because of this, the 5-quarterly datasets are relatively small in comparison to quarterly cross-sectional LFS datasets.

Each respondent is assigned an initial design weight that can, informally, be thought of as how many members of the population that respondent represents. Whilst the design weights may be sufficient in certain estimation problems, a

procedure called calibration is used to adjust the design weights. The design weights are amended to give estimates that are consistent with the known totals for employment status in each of the five quarters being considered.

The calibration procedure seeks to minimize the deviation between the new, calibrated weights and the initial, design weights. The calibrated weights are then used to form calibration estimators of the change in employment status.

There are many 'distance functions' that can be used to assess the deviation between the design weights and the calibrated weights. We shall discuss the properties of several of these functions, and illustrate that the choice of distance function affects the calibrated weights, the corresponding estimates, and the variance of these estimates.

Key Words: Calibration; Labour Force Survey; Weighting; Longitudinal Data; Optimization

2.2.3 Calculating weights for the BIS Self-employed survey, an LFS follow up

Katie Connolly; Office for National Statistics

ONS conducted and delivered results for a telephone survey of 1500 self-employed individuals in a short space of time for the Department for Business, Innovation and Skills (BIS). The results of the survey would then be used to contribute to the evidence base of an independent review of self-employment being conducted by government.

Participants of the Labour Force Survey (LFS) are interviewed for five quarters about their employment status. This allows us to identify self-employed individuals. There is a list of LFS participants who have consented to re-contact. This was used as a sampling frame to allow for efficient sampling. Households with one or more individuals identifying themselves as being self-employed were sampled and from these a single self-employed individual was interviewed. This sample design has been reflected in the weighting strategy.

This presentation will highlight how the innovative sampling methodology employed helped enable us to have "impact at pace" and will provide the methodology used to calculate the weights.

Key Words: Sampling; Weighting; Innovation

2.3 Afternoon Session 3, Syndicate Room 3

Towards the management of multiple data sources

Chair: Gareth James, ONS

2.3.1 Developing an Integrated Business Survey System for Northern Ireland

Dr. James Gillan; Northern Ireland Statistics and Research Agency

The recently developed Northern Ireland Integrated Business Survey System (IBSS) has transformed the business statistics environment for a relatively small statistical office. The achieved solution has taken business surveys from a disparate array of databases onto a single shared IT platform, reducing survey silos, improving coherence between estimates, introducing more automated processing and leading to considerable efficiency gains. A combination of "off the shelf" software has been used to support a multi mode data collection strategy including OCR scanning of paper forms, Telephone Data Entry and Electronic Data Collection.

Novel features include the integration of a SAS layer alongside commercial case management software to help standardise the use of statistical methods at different stages of the Generic Statistics Business Process Model (GSBPM). Solving the problems associated with survey integration and automation has led inevitably to a Statistical Data Warehouse (S-DWH) solution. The IBSS can also handle a range of reporting unit structures, whether these are local sites or individual employees within a business. The IBSS was designed to integrate Administrative databases, and in principle could also be used for social survey purposes. The presentation includes discussion of the efficiencies gained, and the cultural changes needed to support the implementation.

Key Words: Integrated Business Survey System, Electronic Data Collection; Statistical Data Warehouse.

2.3.2 How Welsh Government worked with Schools, Local Authorities and MIS data suppliers to develop a system to collect and maximise the utility of their admin data

Dani Evans; Welsh Government

Abstract:
The use of admin data to produce official statistics is becoming more and more common as it already exists and so there's no extra cost for collecting the information, collections are regularly updated, it is not intrusive and it does not place a significant burden on the provider. However, as the data is not specifically being collected for statistics there are often limitations to the data for statistical purposes. Welsh Government worked with Schools, Local Authorities and MIS data suppliers to develop and build an information management system which meets the needs of all involved parties maximising the utility of the data collected. This presentation will give a background on the development process, how long it took to build and who was involved, risks and issues that occurred, and how they were resolved, the current structure of the system and how it works on a day to day basis.

Key Words: Administrative data**;** Information Management Systems**;** School Statistics

2.3.3 What role can harmonisation play in a changing statistical landscape?

Suzanne Ellis, Steven Cooley; Office for National Statistics

The Harmonisation Team works across the GSS to produce harmonised principles for questions, concepts and definitions. Greater harmonisation allows users to more easily compare data from different sources and makes our statistics easier to understand. It can assist with the UKSA's plan to make the space to innovate by being efficient in our use of time, money and other resources through taking a more coordinated approach across the GSS on key topics.

The UK Statistics Authority business plan outlines the need to work more innovatively with a move to increase the use of administrative data in line with the "collect once, use many times" data management principle. The harmonisation of this administrative data is a new and challenging area of work.

EUROSTATs Framework Regulation Integrating Business Statistics (FRIBS) along with the move towards survey data collection to be digital by default via the Electronic Data Collection programme are providing an ideal opportunity to focus harmonising business survey questions and definitions over the coming years.

This work will focus on what has been achieved to date with harmonisation, including the development of a harmonised question library with the UK Data Service, harmonised principles for surveys and also what remains to be done. It also outlines the benefits of harmonising and details the issues and challenges faced when attempting to harmonise.

Key Words: Harmonisation; Administrative data; Electronic Data Collection; UK Data Service

2.4 Afternoon Session 4, Syndicate Room 4

Improving quality and trust in statistical outputs

Chair: Ruth Fulton, NISRA

2.4.1 Reviewing aspects of quality reporting within ONS

Sarah Tucker; Office for National Statistics

ONS publish Quality and Methodology Information (QMI) reports for every Statistical Bulletin. QMIs contain information on how the output is created and report against the European Statistical System (ESS) quality dimensions. They also contain information on the strengths and limitations of data which help users decide upon suitable uses.

We recently reviewed some aspects of how quality information is communicated to users. Were there gaps in what is provided compared to current user needs? What could be done to extend the use of this information?

The primary purpose of quality reporting has been to help users decide upon suitable uses of the data. Through our review, we realised that we should first help users reduce the risk of misusing data. Research was undertaken to look at how ONS can first help users understand how to not misuse the data, and then understand suitable uses for the data.

This paper will discuss work that has been carried out to reduce the risk of users misusing data. The findings from internal focus groups and meetings will be presented, including further work to create a new quality reporting product for user testing. The paper will explore the results of user testing and discuss how we're implementing this product. This paper will also briefly mention how the research for this product is informing changes to other quality reporting products and how these can work together to provide a layered approach to quality reporting for our users.

Key Words: Quality reporting; Users; Communication; Strengths; Limitations

2.4.2 A new process for assessing the quality of an output makes better quality statistics

Catherine Bremner; Office for National Statistics

There is a requirement under the UK Code of Practice for Official Statistics to ensure official statistics are produced to a level of quality that meets users' needs and to seek to achieve continuous improvement in statistical processes by undertaking regular reviews.

Quality Centre assess the quality of an output within the Office for National Statistics (ONS) through the Regular Quality reviews (RQR) process where the data producer can discuss their output in the context of the five quality dimensions and Generic Statistical Business Process Model with a methodologist. The process also results in recommendations tailored to improving the quality of the statistical output.

To date 50 RQRs have been successfully completed. Customers feel the new process is proportionate and the recommendations useful. Quality Centre monitors recommendations and ensure they are implemented.

In this paper, we describe the new process, present feedback from business areas, discuss the common themes around recommendations and show how outputs have improved as a result of the RQRs.

Key Words: Quality; Improvements; Reviews; Assessments

2.4.3 Users' understanding and use of uncertainty measures to describe data quality

Silvia Manclossi[1], Victoria Ayodele; Office for National Statistics
[1] (currently on loan to the Welsh Government)

Abstract:
The UK Code of Practice for Official Statistics (UK Statistics Authority; 2009) indicates that users must be informed about the quality of statistical outputs against the European Statistical System dimensions of quality (relevance, accuracy and reliability, timeliness and punctuality, accessibility and clarity, coherence and comparability). For sample surveys, which form the basis of many ONS outputs, the typical measures of uncertainty that are recorded are Standard Errors (SEs), Confidence Intervals (CIs), Coefficients of Variation (CVs) and statistical significance.

Quality Centre at the Office for National Statistics (ONS) has carried out work to establish a better understanding of how users interpret information on quality, specifically measures of uncertainty, when using official statistics. We were motivated by a review of current practices for ONS statistical outputs and through exploring the approach used by other National Statistical Institutes.

Our work has mainly focussed on how data are used based on the quality information that is provided and whether presenting information in a different way, or using some standard definitions, would improve users' understanding. We have identified the CV in particular as a concept to explore with users, as it might be more difficult to understand compared to other measures, but we have also considered understanding and interpretation of SEs, CIs and statistical significance. Once work has been completed, its findings will help to inform how we report on quality in the future. This paper will set out the main stages of this project and present the main findings and recommendations that have been identified to date.

Key Words: Uncertainty measures; Quality; Users

GSSM21 Sponsors and announcements

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. The RSS website can be found at the following link:

http://www.rss.org.uk/

The Royal Statistical Society Excellence Awards

On the evening of the 6th of July, following GSSM21, the RSS will be holding their annual Excellence Awards and summer reception. This ceremony, sponsored by UKSA, 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. For more information and to register for the event please see the RSS website:

https://www.statslife.org.uk/events/events-calendar/eventdetail/665/-/2016-statistical-excellence-awards-ceremony

Southampton University: MOFFSTAT

The MSc in Official Statistics programme is a joint collaboration between Southampton University 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.

Many of the skills taught on the programme, 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 Programme, come along to our exhibit to collect a copy of the latest 2016-17 prospectus and speak to the Programme Director, Paul Smith, who will be on hand to answer any of your questions.

The GSS Capability Team

The GSS Capability Team, 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 statistical techniques used by statisticians and provide information on how and why to choose the most appropriate methodology.

Come along to our exhibit on the 6th July to meet the GSS Capability Team and collect the latest copy of the GSS Learning Curriculum 2016-17!

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 the secretary (isabella.wheeler@ons.gsi.gov.uk).