Innovative data, methods & models

Strand organizer: Jason Hilton (University of Southampton)

Innovation in data collection & processing – Tuesday 11 September 9.00am

Dispersal and data: methods for analysing asylum seekers and refugees in the UK

Sarah Nurse, Jakub Bijak; University of Southampton

The limited availability of data is a key challenge faced by those researching asylum seeking and refugee populations in the UK. In particular, analysing the dispersal policy, where asylum seekers requiring support are housed with no element of choice of location, is difficult with so few data sources. In this context, combining datasets that do not report dispersal status with those that do has the potential to open up a new range of variables and topics to be explored. The main research question addressed here is: how can combining datasets on asylum seeker and refugee populations help us further understand dispersal? Through an assessment of the additional information gains and the errors of estimation that are introduced, the feasibility of methods for combining data is explored. Three possible options for combining sources are considered: individual data linkage is presented with a discussion of the potential for analysis of a resulting dataset, an illustration of how information on individuals can be 'borrowed' shows how successfully dispersal status can be predicted and finally a method that utilises cell structures to combine information on aggregates and allows further analysis of social and economic outcomes is presented. Results show that augmenting the Annual Population Survey by predicting probability of dispersal, using model coefficients based on data from the Survey of New Refugees, can effectively add an indicator of dispersal to the existing source. Finally recommendations for how additional, targeted data collection could facilitate a more effective application of these methods are offered.

Email: s.l.nurse@soton.ac.uk

A longitudinal approach to investigate European migration to the UK using the Facebook Advertising Platform

Francesco Rampazzo[1], Jakub Bijak, Agnese Vitali[2], Ingmar Weber[3], Emilio Zagheni[4]; [1]University of Southampton and Max Planck Institute for Demographic Research, [2]University of Southampton, [3]Qatar Computing Research Institute, [4]University of Washington and Max Planck Institute for Demographic Research

International migration to the UK has become a hot topic both in research and in the media. Nevertheless, a huge limitation of this topic of research is the availability of data for measuring migration. Raymer and Willekens (2008), Poulain, Perrin and Singleton (2006), Bijak (2013) and Abel and Sander (2014) address the issue of improving existing statistics on international migration from a definition perspective, but also through statistical modelling. Detailed data on migrants' characteristics are much needed for producing accurate statistics and informing policy. We want to use online advertising data to provide a clearer picture of these migrants. This paper has two aims. First, it aims to create the first longitudinal geo-located dataset from Facebook's Advertising Platform through a weekly collection from mid-December 2017 onwards for two years. Secondly, it aims to give a weekly picture of immigration from the EU following the EU referendum. The dataset will be stratified by characteristics such as age, gender, country of origin, education level, and employment field. The analysis will complement traditional data sources provided by the ONS, with which we will compare with our estimates. Then, a time series analysis approach will be used to analyse the data. It will be important to observe the differences by country regarding the trends, cycles, and seasonal changes in the data. With this approach, it will be possible to make projections of future trends in migration using Facebook data. The preliminary results from Facebook show proportions close to the ONS estimates for 2016.

Email: f.rampazzo@soton.ac.uk

Creating a spatially-detailed synthetic population micro-dataset
Paul Williamson, University of Liverpool

An increasingly wide range of researchers find themselves requiring realistic population microdata as inputs to their various modelling approaches. These include those engaged in dynamic microsimulation, agent based modelling and transport modelling. A now standard way of attempting to address this problem is to calibrate a population survey to a known set of local area benchmark constraints. Two particular challenges stand out. First, constructing synthetic populations of individuals nested within households that optimally satisfy both local person-level and household-level estimation constraints. Second, the creation of synthetic populations that are integer- rather than real-weighted. This paper outlines a novel technique, 'Global Optimisation' (GO) that addresses these challenges. GO provides solutions that are 'globally optimal' in the sense that they maximise the possible fit to the local area benchmarks; and that are integer-weighted without introducing rounding bias. The results section of this paper presents the strengths and weaknesses of this approach, comparing it to the three main alternative approaches currently in use (Iterative Proportional Fitting (IPF); Generalised Regression (GREG) and Combinatorial Optimisation (CO). Each approach is evaluated in terms of its ability to fit local area benchmarks; and to estimate the known (but not benchmarked) distribution of age by sex by health. Finally the paper presents results demonstrating other aspects of GO, including its sensitivity to the size of the survey being calibrated and the spatial scale for which survey estimates are required. Lessons are drawn from this to the relevance of the approach for adjusting the Census for under- and over-count.

Email: p.williamson@liverpool.ac.uk

Changing administrative sources in an Admin Data Census: measuring the risks, benefits and opportunities
Louise O'Leary, Bethany Fitzgibbon; Office for National Statistics

Reliance on any administrative data source to provide information about the size of population, characteristics of the population or households and families carries risk. A changing political climate, including Brexit, or a change in government and any resulting policy changes may impact the operational system used to collect the data, the stability of definitions (for example tax thresholds) or the frequency and timing of when the data is available, as well as the appetite to share these data. These all have an impact on the quality of the data available to researchers. Using case studies (including health, tax and benefits data), this paper seeks to explore how planned and unplanned changes to administrative systems impact the ability of the Integrated Data Division in ONS to produce reliable estimates of the population (both size and characteristics). It also assesses how we can begin to understand, quality assure and measure uncertainty around the impact of external changes to systems and what steps can be taken to minimise these. In conclusion, the Integrated Data Division at ONS is working towards a framework to understand and measure the uncertainty and challenges around using non-survey data (administrative, big and commercial) whilst also seeking to optimise the benefits and opportunities of new administrative systems to help provide new insights into society.

Email: louise.o’leary@ons.gsi.gov.uk

Innovation in modelling & forecasting – Tuesday 11 September 1.30pm

Projecting people and households at high spatial resolution
Andrew P Smith, Nik Lomax; University of Leeds

The Infrastructure Transitions Research Consortium, a collaboration of seven UK universities, is looking at future infrastructure demand across a range of sectors (transport, water, waste, energy and digital). The consortium requires high resolution (i.e. a very fine spatial scale) projections of population and housing in order to feed their demand models. Demographic projections for both people and households are produced by statistical agencies around the world. These projections are essential for planning the delivery of services and the allocation of resources to sub-national areas but with few exceptions, projections are limited to larger administrative areas (e.g. local authorities in the UK) because the geographical detail is not available, or is simply not required: for example national funding allocation is usually given to administrative areas, not small sub-administrative units. This paper outlines and compares two methodologies for producing consistent high resolution projections of people and households in Great Britain. Firstly a technique that uses a series of microsimulations constrained to official projections at wider geographies; secondly a dynamic microsimulation model which utilises survey and census data as well as supply data for housing stock. In both models, people are allocated to households, who are then distributed to physical housing units. In the second model, we discuss how the coupling between the time-evolution of populations and that of households can be captured.

Email: a.p.smith@leeds.ac.uk

The shelf life of sub-national projections, from Australia to England

Ludi Simpson[1], Tom Wilson[2] (Charles Darwin University), Fiona Shalley[2]; [1]University of Manchester, [2]Charles Darwin University

Wilson et al. (2017) measured the empirical distribution of the accuracy of projected population in sub-national areas of Australia, developing the concept of ‘shelf life’: the projection’s furthest horizon which remains within 10% inaccuracy for at least 80% of areas projected. The shelf life depended on size of area, being 9 Â½ years for areas of about 10,000 population and 13 Â½ years for areas of about 100,000 population. This paper extends the analysis to (a) report on official sub-national projections in England since 1974, and (b) take into account the user’s need for projections of specific horizons for different purposes. Since local government reorganisation in 1974, 19 official projections of the population of local government areas in England have been made. By comparing the published projection with the post-census population estimates, the empirical distribution of errors will be described, dependent on horizon and population size. Users of projections tend to have in mind a horizon and a required accuracy that is of relevance to each application. A shelf life of 10 years in the sense of within 10% inaccuracy for at least 80% of areas would not be sufficient if the user required that accuracy of a forecast 15 years ahead. The relevant shelf life must deduct the user’s horizon. One can expect shorter horizons to require greater accuracy. We explore the empirical performance of official English sub-national projections in this light. We propose questions about users’ perceived needs that will help focus the analysis of official projection accuracy.

Email: ludi.simpson@manchester.ac.u8k

Forecasting UK fertility using Bayesian Parametric Mixtures

Jason Hilton, Erengul Dodd, Jon Forster, Peter W.F. Smith; University of Southampton

Fertility is a dynamic social process that is influenced by a wide range economic and cultural factors. This complicates the process of forecasting future numbers of births, as the direction and magnitude of changes in fertility rates are consequently much less predictable than they are for mortality. The current paper presents a Bayesian approach to fertility forecasting that adapts and develops existing approaches to modelling fertility age structures with parametric mixtures. Taking cohort as the primary forecasting axis, completed family size is forecast using time series methods, and for each cohort, age specific rates are obtained by decomposition into two mixture components. Both the mixture parameter and the location and scale of the components themselves are further modelled and projected using time series methods. This approach is parsimonious and has several other advantages. Firstly, the Bayesian framework allows all sources of uncertainty to be included in predictions of future fertility. Secondly, the parameters to be forecast have potentially meaningful demographic interpretations. Model comparison methods are used to compare the efficacy of the various functional forms of the mixture components. Additionally, change-point models and approaches incorporating stochastic volatility are investigated. The final efficacy of the approach is determined through its ability to predict 'held-back' observations not used during the fitting process.

Email: j.d.hilton@soton.ac.uk

Event history analysis of births to women in the UK using Generalised Additive Mixed Models

Joanne Ellison, Jon Forster, Erengul Dodd; Centre for Population Change, University of Southampton

Aggregate UK fertility data available at the population level can only give limited information about the patterns of variability of age-specific fertility rates. Survey data provides a rich source of information through fertility histories of individuals and their corresponding characteristics, which can help to gain a better understanding of the underlying variability of fertility rates. By modelling the fertility histories of 5,789 women surveyed in the 2006 General Household Survey, we investigate the dependence of birth events on selected covariates as well as information derived from the fertility histories themselves. Generalised Additive Mixed Models (GAMMs) allow the incorporation of covariates as smooth terms and provide a straightforward way to account for the clustering of observations within each woman through random intercepts. Fitting parity-specific GAMMs to the survey data, we learn about the variability of fertility as a function of age, cohort, education, ethnicity and birth interval. There is the potential to combine inferences from this detailed individual-level data with the coarser population-level data for the purposes of forecasting.

Email: je4g13@soton.ac.uk

Innovation in methodology – Tuesday 11 September 4.45pm

Probabilistic methods for combining internal migration data

Guy Abel[1], Guillermo Vinue Visus[2], Dilek Yildiz[3], Arkadiusz Wisniowski[4]; [1]Asian Demographic Research Institute, Shanghai University, [2]Wittgenstein Centre (IIASA, VID/OEAW, WU), [3]Vienna Institute of Demography, University of Manchester

In order to fully understand the causes and consequences of population movements, researchers and policy makers require timely and consistent data. Migration data are commonly obtained from censuses, registers or surveys. Each of these data sources can vary in their measurement of accuracy, coverage of population, undercount and definitions of a migration event. This paper proposes a Bayesian probabilistic methodology to harmonize migration data from different sources. In particular, we build a hierarchical model for combining migration data sources in the USA between 1980 and 2016. The model allows for estimates of true migration flows that explicitly compensates for the inadequacies in each data source and provides one-step ahead forecasts of bilateral migration patterns.

Email: guy.abel@oeaw.ac.at

Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman

Amira Al Aamri[1], Graeme Hornby[2], Abdullah AL Maniri[3], Li-Chun Zhang[1], Sabu S. Padmadas[1]; [1]Department of Social Statistics and Demography & Southampton Statistical Sciences Research Institute, University of Southampton, [2]GeoData, University of Southampton, [3]Oman Medical Specialty Board, Road Safety Research Program, The Research Council, Sultan Qaboos University

Road traffic crashes (RTCs) are a major global public health problem and cause substantial burden on national economy and healthcare. There is little systematic understanding of the geography and spatial correlations of RTCs in the Middle-East region, particularly in Oman where RTCs are the leading cause of disability-adjusted life years lost. The foci of this study is to: (1) identify high density crash zones in the Muscat Governorate (2) explore the characteristics of crash hot-zone, and (3) examine the spatio-temporal patterns of RTCs in the study area. We applied an adjacency network analysis integrating GIS and data of five years (2010-2014) of RTCs in Muscat Governorate using robust estimation techniques including: Kernel Density Estimation (KDE) of both 1-D and 2-D space dimensions, Network-based Nearest Neighbour Distance (Net-NND), Network-based K-Function, Random Forest Algorithm (RF) and spatiotemporal Hot-zone analysis. The findings demonstrate evidence of spatial clustering of RTC hot-zones on long roads demarcated by intersections and roundabouts. Findings from RF algorithm and Wilcoxon tests show that hot-zones are associated with higher level of road traffic and with higher numbers of exits and entrances and shorter distance between junctions, while posted speed limits has no significant effect in determining the crash risk on road zones. The spatio-temporal analysis provides evidence of the consistency in the positions of crash hot-zones in the study area. The results from GIS application of NRTC data are validated using the sample data generated by iMAAP database.

Email: aka1e14@soton.ac.uk

Self-discovery for supervised measurement: An application to the concept of 'productive ageing'
Ginevra Floridi, London School of Economics

Quantitative social science studies often measure key concepts by combining a set of indicators into a scale. Unsupervised measurement methods use observed correlation structures to identify scales that best explain variation in the indicators, but may fail to measure the desired concept. Supervised measurement methods use expert judgements to aggregate indicators, but require explicit decisions about aggregating activities that are difficult to make and to assess. We propose a 'self-discovery' method for measurement supervision that uses the form of a conjoint experiment, with reference to the demographic concept of 'productive ageing' in Italy and South Korea. We consider older adults' participation in paid work, volunteering, grandchild care and informal care as indicators of productive ageing. We take these indicators as measured in major ageing surveys, and ask Italian and South Korean academics with a research interest in productive ageing to complete a series of pairwise comparisons on hypothetical profiles of older adults participating in different combinations of activities, and to different extents. By ranking profiles based on their level of productivity, the experts implicitly indicate the relative weights to place on each activity. We model responses on the full set of activities, revealing their relative weights. Results indicate a high level of inter-coder reliability in the definition and measurement of productive ageing. When comparing Italian and South Korean academics, we find some evidence of contextual differences in the definition of the concept, with Italian experts putting more weight on grandparental care and less on volunteering relative to their Korean counterparts.

Email: g.floridi@lse.acf.uk

Statistical archaeology to retrieve Small Area Statistics from the UK 1961 Census
Justin Hayes, University of Salford

Censuses provide the statistical bedrock for current and historical demographic analysis. The 1961 Census of Great Britain was the first UK census to use electronic computing for the processing of inputs and production of outputs. This enabled the production of detailed Small Area Statistics (SAS) for areas with populations of only a few hundred households for the first time, in addition to traditional published outputs for much larger areas. Small area outputs from UK censuses from 1971 onwards are available in digital form from various online sources without restriction. Unfortunately, no digital versions of the 1961 SAS survive, but the information content of the 1961 SAS has been preserved in a set of over 140,000 digital images held by the Office for National Statistics (ONS), taken from microfilm images of the paper prints on which the SAS were originally supplied. The Pattern Recognition and Image Analysis Research Lab (PRImA - http://primaresearch.org/) at The University of Salford has collaborated with ONS on a 'statistical archaeology' project that has developed methods to retrieve and integrate digital values from the 1961 census images in order to make the 1961 SAS available in digital form for analysis using modern tools and techniques. The work has been automated as far as possible so that most values have been digitised using optical character recognition, integrated, and then validated through a series of quality assurance processes. A small proportion of values have identified as containing errors, due mainly to quality issues with some of the images, and are being corrected using an innovative crowd-sourced approach. This paper will describe the methods and processes developed to retrieve the 1961 SAS, present results and outputs to date, and discuss the potential for similar retrieval of other historical statistical data.

Email: j.hayes1@salford.ac.uk