Data science: Innovative data, methods and models strand & Data quality session

Data science strand organiser: Jason Hilton (University of Southampton)

Data quality session convenor: Philip Humby (Office for National Statistics)

2.00pm Tuesday 14 September: Data science: New sources of data for demography

Digital traces of sexualities: Patterns of social media disclosure and the salience of sexual identity

Connor Gilroy[1], Ridhi Kashyap[2]; [1]University of Washington, [2]University of Oxford

Using data about the disclosure of sexuality on social media, we analyze the expression of sexualities in the contemporary United States. Through the Facebook advertising platform accessed through its marketing application programming interface (API), we collect aggregate counts encompassing 200 million Facebook users, 28% of whom disclose sexuality-related information. Stratifying by age, gender, and relationship status, we show how these attributes structure the propensity to disclose different sexual orientations. We find a large generational difference, where younger social media users share their sexualities at high rates regardless of sexual orientation. For older cohorts, marital status substitutes for sexual orientation; only unmarried individuals frequently disclose. Consistent with prior research on gendered expectations about sexuality, women more often express a bisexual interest in both men and women; men are more explicit about their heterosexuality. Overall, we consider how these data provide a unique -- albeit technically-constrained -- lens to analyse how sexuality is socially signalled at scale, with the potential to extend the sociology and demography of sexuality.

Email: ridhi.kashyap@nuffield.ox.ac.uk

From the stork to fertility apps

Francesco Rampazzo[1], Alyce Raybould[2], Pietro Rampazzo[3], and Ross Barker[4]; [1]University of Oxford, [2]London School of Hygiene and Tropical Medicine, [3]Technical University of Denmark (DTU), [4]Wittgenstein Centre for Demography and Global Human Capital (Austrian Academy of Sciences, International Institute for Applied Systems Analysis, University of Vienna)

The market for smartphone apps tracking fertility has grown in recent years. These apps brand themselves as empowering their users to reach their reproductive goals, claiming to help achieve a pregnancy more easily than through conventional medical channels. This paper offers the first comprehensive quantification of fertility tracking app users. Using data collected from Google Play Store and Apple App Store, we calculate the most downloaded apps, and the global distribution of use. We use a log-log model fitted on the Google Play Store data to predict the Apple App Store number of installations. Our findings show that 74% of downloads are for just three of 28 apps. The majority of the reviews are left by users in North America, Northern Europe, and Australia; but it is noteworthy that downloads are also widespread in the Global South. Ongoing work aims to investigate the most discussed topics in the reviews.

Email: ross.barker@oeaw.ac.at

Now-casting Romanian migration into the United Kingdom by using Google search engine data

Andreea Avramescu[1], Arkadiusz Wiśniowski[2]; [1]Alliance Manchester Business School, University of Manchester, [2]Social Statistics Department, University of Manchester

Traditional forecasts of international migration are often based on data that are incomplete, biased, and reported with delays. There is also a relative scarcity of migration forecasts based on combined traditional and new forms of data. This research proposes an inclusive and interdisciplinary approach of supplementing official statistics with the so-called Big Data from the Google search engine. These data are used to create composite variables depicting general interest of Romanians in migrating into the UK. These variables are further

assessed as predictors in time series models to forecast the immigration of Romanians to the UK in scenarios when official data are reported with a delay. We found that the proposed Google Trends Index related to employment and study, which exhausts all possible keywords and eliminates the language bias, matches the trend observed in the official statistics. However, the reduction in the prediction errors is only moderate. We further discuss the usability and limitations of the Google Trend data in now- and short-term forecasting and its potential to serve as an early-warning predictor of potential sudden changes in international migration.

Email: a.wisniowski@manchester.ac.uk

Creating a geodemographic classification for older people in England

Yuanxuan Yang, Frances Darlington-Pollock, Les Dolega; Department of Geography and Planning, University of Liverpool

he population are ageing: by 2041, it is estimated half of the UK's adult population will be aged 50 and over. Population ageing is often demonised as a looming crisis wherein older people are homogenised as dependent, frail and a burden. However, the characteristics, behaviours and needs of the older demographic are not uniform, tending to vary spatially. To better understand the social and spatial heterogeneity within the older population and thereby support effective policy development and targeted service provision, we developed an open access, multidimensional classification of the older population in England at small area (Lower Layer Super Output Areas) level. We combine novel data sources (e.g. NHS prescribing data) with more conventional sources (e.g. the Census) to capture characteristics of the environment pertinent to experiences of older people, and characteristics of older people across multiple domains. This includes demographic, socio-economic, health, digital, mobility, civic participation and environment. The classification was built using machine learning models (clustering). This is the first bespoke geodemographic classification of the older population (age over 50) in England. As a robust tool to inform evidence-based planning and policy interventions, it provides valuable insights into the nature and geography of need, vulnerability and opportunity in a population which will continue to age.

Email: yangyuanxuan@hotmail.com

9.00am Wednesday 15 September: New methods & data for mortality analysis

Covid Infection Survey: producing official estimates using multi-level regression modelling

Owen Gethings, Melissa Randall, Office for National Statistics

Following the onset of the COVID-19 outbreak in the UK it was crucial to understand how COVID-19 was spreading across the population in order to control the pandemic and its effects. Decisions regarding the continued need for control measures to contain the spread of SARS-CoV-2 rely on accurate and up-to-date information about the number of people and risk factors for testing positive. ONS set-up the Covid Infection Survey to do this, repeatedly swab participants and generate an official estimate of covid positivity for the UK. Analysis needed to deal with relatively small numbers of positive tests and ensure estimates were unbiased. Multilevel regression and post-stratification (MRP) is a statistical technique used for correcting model estimates for known differences between a sample population and a target population. It has been shown to be an effective method of adjusting the sample to be more representative of the population for a set of key variables; previously it has been used in polling to predict the US election results, but it isn't widely used in other settings. MRP consists of two steps. First, an MRP is used to generate the outcome of interest as a function of (socio)demographic and geographic variables. Next, the resulting outcome estimates for each demographic-geographic respondent type are post-stratified by the percentage of each type in the actual overall population. For the Covid Infection Survey, the result was to produce an estimate which takes account of differences by age, sex, time and region. This presentation will share further detail on the method and the resulting estimates.

Email: owen.gethings@ons.gov.uk

Forecasting intensive care unit demand during the COVID-19 pandemic: A spatial age-structured microsimulation model

Sebastian Kluesener[1,3,4], Ralf Schneider[2], Matthias Rosenbaum-Feldbruegge[1], Christian Dudel[3], Elke Loichinger[1], Nikola Sander[1], Andreas Backhaus[1], Emanuele Del Fava[3], Janina Esins[5], Martina Fischer[5], Linus Grabenhenrich[5,6], Pavel Grigoriev[1], André Grow[3], Jason Hilton[7], Bastian Koller[2], Mikko Myrskyla[3,8], Francesco Scalone[9], Martin Wolkewitz[10], Emilio Zagheni[3] , Michael M. Resch[2] ; [1]Federal Institute for Population Research, Wiesbaden, [2]High Performance Computing Center, Stuttgart, [3]Max Planck Institute for Demographic Research, Rostock, [4]Vytautas Magnus University, Kaunas, Lithuania, [5]Robert Koch-Institut, Berlin, [6]Charité - Universitaetsmedizin Berlin, [7]University of Southampton, [8]University of Helsinki, [9]University of Bologna, [10]University of Freiburg

The COVID-19 pandemic poses the risk of overburdening health care systems, and in particular intensive care units (ICUs). Next to vaccination campaigns, non-pharmaceutical interventions (NPIs), ranging from wearing masks to (partial) lockdowns are important mitigation measures. Political decision-making on NPIs is in need of reliable forecasts of COVID-19-related ICU demand under alternative scenarios of COVID-19 progression reflecting different levels of NPIs. Substantial sub-national variation in COVID-19-related ICU demand requires a spatially disaggregated approach. We implement a spatial age-structured microsimulation model of the COVID-19 pandemic by extending the Susceptible-Exposed-Infectious-Recovered (SEIR) framework. The model accounts for regional variation in disease dynamics including potential spatial diffusion pathways, and for regional variation in population risk structure (i.e. age). It is calibrated against ICU data to determine the current ICU-relevant disease dynamics for subnational regions, which serves as a base for scenario-based forecasts. We apply the model to Germany to provide forecasts over a 2-month period at the level of the 16 federal states. To illustrate the merits of our model, we present forecasts of ICU demand for three different stages of the pandemic during 2020. Our results provide evidence for substantial spatial variation in (1) the effect of the pandemic on ICU demand, and (2) the potential and need for NPI adjustments at different stages of the pandemic. The model is programmed in R and can be applied to other countries, provided that reliable data on the number of ICU patients infected with COVID-19 are available at sub-national level.

Email: sebastian.kluesener@bib.bund.de

Bayesian reconstruction of multi-state mortality

Andrea Tamburini[1], Dilek Yildiz[2]; [a]OeAW, Wittgenstein Centre (IIASA, OeAW, Univ. Vienna) [b]IIASA, OeAW, Wittgenstein Centre (IIASA, OeAW, Univ. Vienna)

The connection between education and mortality has already been widely analysed. And the idea of including the level of education among the main demographic dimensions is not new. At the same time, the availability of data concerning the stratification of mortality according to level of education is rare and often limited to developed countries and recent years. Considering the interplay between mortality and education, the focus up to this point has mainly been on how education, often seen as a proxy for socio-economic status, affects life choices and habits and thus, indirectly, mortality. When mortality has been considered directly, the focus has been more on life expectancy in specific countries and for specific ages, without engaging in a systematic study with a broad geographical and temporal spectrum. Leaving aside the education dimension and focusing on the use of Bayesian models for estimating mortality rates, the focal point has been primarily on predicting future mortality rates and not on reconstructing past ones. In this paper we propose a mortality-specific model that is able to combine patchy data in order to produce estimates of past mortality curves according to the level of education. The proposed model is a hierarchical Bayesian model that uses data on known mortality age schedules by education attainment and combines them with data from DHS to produce age, sex and education-specific mortality curves. The model is applied in a case study for the female population of Turkey.

Email: andrea.tamburini@oeaw.ac.at

Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth

Andrea Nigri[1], Susanna Levantesi[2], and José Manuel Aburto[3,4]; [1]Department of Agricultural Sciences, Food, Natural Resources and Engineering, University of Foggia, [2]Department of Statistics, Sapienza University of Rome, [3]Leverhulme Centre for Demographic Science, Department of Sociology and Nuffield College at University of Oxford, [4]Interdisciplinary Centre on Population Dynamics, University of Southern Denmark

Life expectancy is one of the most informative indicators of population health and development. Its stability observed over time has made life expectancy appealing to predict or forecast. However, predicted or estimated values of life expectancy do not tell us about age-specific mortality. Reliable estimates of age-specific mortality are essential in the study of health inequalities, well-being and to calculate other demographic indicators. However, this task comes with several difficulties including lack of reliable data in many populations. Models that relate levels of life expectancy to a full age-specific mortality profile are therefore important but scarce. We propose a model (DNN) to derive age-specific mortality from observed or predicted life expectancy leveraging deep learning algorithms akin to demography's indirect estimation techniques. Out-of-sample validation was used to validate the model, and the predictive performance of the DNN models was compared with two state-of-the-art models designed to do the same thing. Out-of-sample validation indicate that the DNN model provides reliable estimates of age-specific mortality for the USA, Italy, Japan and Russia using data from the Human Mortality Database. We further show how the DNN model could be used to estimate age-specific mortality for countries without age-specific data using neighbouring information or populations with similar mortality dynamics.

Email: andrea.nigri88@gmail.com

Midday Wednesday 15 September: Data science: Bayesian methods in demography

Extending the Integrated Model of European Migration

Georgios Aristotelous, Peter W. F. Smith, and Jakub Bijak; University of Southampton

In many countries, migration patterns are the key determinant of population change. Accurate estimates of place-to-place population migration flows are essential for making population policy estimates or projections. However, there are many difficulties inherent to estimating migration flows: for example, countries may underreport migration, use different migration definitions, or have different data-collection systems. We report on work undertaken as part of the Quantifying Migration Scenarios for Better Policy (QuantMig) project, funded by the European Union's Horizon 2020 programme. This work is extending the methodology developed in the Integrated Modelling of European Migration (IMEM) project. It will provide harmonised migration estimates for the flows within the 32-country EU+ system, and flows into and out of Europe, by origin, destination, age and sex, from 2009 to 2019 with a statistical assessment of their uncertainty. Furthermore, the flows will be disaggregated into EU nationals or third-country nationals, to enhance their utility to policymakers, given the different legal status of these groups within the EU and the differences in the associated channels of migration. The estimation utilises a hierarchical Bayesian approach based on the IMEM model. We use migration flow data collated by Eurostat, and incorporate covariate information and information provided by experts on the effects of undercount, measurement and accuracy of data collection systems. We also specify a migration model to relate the true unknown flows to the covariates and a measurement model to relate the observed flows to the true unknown flows, correcting for the inconsistencies and inaccuracies in the observed migration.

Email: G.Aristotelous@soton.ac.uk

Modeling international migration flows by integrating administrative and household survey data

Emanuele Del Fava[1], Arkadiusz Wiśniowski[2], Emilio Zagheni[1]; [1]Max Planck Institute for Demographic Research, [2]Social Statistics Department, University of Manchester

Migration has become a significant source of population change at the global level, with broad societal implications. Although understanding the drivers of migration is critical to enacting effective policies,

theoretical advances in the study of migration processes have been limited by the lack of data on flows of migrants, or by the fragmented nature of these flows. In this paper, we build on existing Bayesian modelling strategies to develop a statistical framework for integrating different types of data on migration flows. We offer estimates, as well as associated measures of uncertainty, for immigration, emigration, and net migration flows among 31 European countries, by combining administrative and household survey data from 2002 to 2015. Substantively, we document the historical impact of the EU enlargement and the free movement of workers in Europe on migration flows.

Email: a.wisniowski@manchester.ac.uk

Bayesian Poisson regression for reconstructing fertility rates
Afua Durowaa-Boateng Afua, Yildiz Dilek; IIASA, OEAW, Wittgenstein Centre (IIASA, OeAW, Univ. Vienna)

Many researchers have investigated the relationship between education and fertility decline. It has been showed that the stall in fertility decline in Sub-Saharan Africa was due to a stall in education, thus stressing the need to include education in fertility studies. The main source of detailed fertility data in many developing countries is sample surveys. However, due to sample size, incompleteness, and irregular survey times, it is not always adequate to estimate (age specific) fertility rates by education. Fertility data by level of education of mother is quite rare to obtain especially for developing countries. Data on education specific fertility rates available are either incomplete or not frequently estimated. Due to the unavailability of complete education specific fertility rates, we propose a method to estimate fertility rates by education. We estimate fertility rates by level of educational attainment using Bayesian modelling techniques on a Poisson regression model on DHS data.

Email: afua.durowaa-boateng@oeaw.ac.at

Combining data sources to develop a Bayesian fertility projection model for England and Wales
Joanne Ellison[1], Ann Berrington[1], Erengul Dodd[1], Jonathan J. Forster[1 2];[1]ESRC Centre for Population Change, University of Southampton, [2]Department of Statistics, University of Warwick

Fertility projections are a key determinant of population forecasts; they are also vital to anticipate demand for maternity and childcare services, for example. It is standard practice for fertility projection models to use aggregate population-level data alone, e.g. from vital registration - this ignores the rich inferences that individual-level data can provide. To this end, we develop a Bayesian parity-specific fertility projection model for England and Wales that combines individual- and population-level data sources. Individual-level data informs the base of our model, namely fertility histories and additional information collected from women in Wave 1 of Understanding Society. Through fitting logistic Generalised Additive Models (GAMs), we learn about the smooth dependence of fertility on age, cohort and time since last birth, as well as the effect of qualification. Embedding our chosen GAMs into a Bayesian framework, it then becomes possible to incorporate population-level parity-specific fertility rates from vital registration. We achieve this via a marginalisation process in which we weight the contributions of the two data sources according to our prior beliefs about their relative importance to overall inference. The ability to retain individual-level covariates such as qualification allows us to generate plausible forecasts for population subgroups determined by these variables, despite the population-level data not informing about them. This work demonstrates how inferences from detailed individual-level data can be combined with coarser population-level data for the purposes of demographic forecasting.

Email: J.V.Ellison@soton.ac.uk

4.30pm Wednesday 15 September: Data quality

Working with local authority data on looked after children - learning about data quality through data linkage
Cecilia Macintyre; Scottish Government

ADR Scotland is a partnership between Scottish Government and Scottish Centre for Administrative Research (SCADR). The data acquisition team have produced the longitudinal Looked After Children dataset, which can track an individual's sequence of placements through the care system. This innovative work links eleven years of data, giving an understanding of the possible pathways enabling linkage to other outcomes to better understand this population.   'Looked after children' are children in the care of their local authority. Some children are cared for at home, with regular contact with social services. In other cases, the child or young person is cared for in placements away from their normal place of residence by foster or kinship carers, prospective adopters, or in residential care homes, schools or secure units.   The data on looked after children is based on annual returns which are made to Scottish Government by local authorities. Individual level data has been collected since 2008 with details of each placement. Each return provides a full history of any episode (a series of consecutive placements) of care which has occurred within the reporting period. Data returns were combined, resulting in almost 60,000 children and around 140,000 placements.   The presentation will outline the approach to producing a linkage ready dataset, and the accompanying documentation produced to help users and ensure transparency of the process. The presentation will discuss how the Scottish Government has worked with partners – local authority, academic and other public bodies - to improve the quality of the data used in this project.

Email: cecilia.macintyre@gov.scot

Data collection changes due to the pandemic and their impact on estimating personal well-being
Sarah Coates and Hannah Aston; Office for National Statistics

Throughout the Coronavirus (COVID-19) pandemic, the Office for National Statistics (ONS) has published estimates of personal well-being using both the Annual Population Survey (APS) and the Coronavirus module of the Opinions and Lifestyle Survey (OPN). In this presentation, we will show how the ONS considered the impact the pandemic has had on data collection, the extent to which it has influenced estimates compared with pre-pandemic and reviews the comparability of estimates between the APS and the OPN.   Both surveys showed substantial worsening of personal well-being at the start of the pandemic. Both surveys have since shown significant improvements in average scores of happiness and anxiety. The same significant improvements were not seen in life satisfaction and feeling that things done in life are worthwhile, and data from the OPN suggest that both life satisfaction and feeling that things done in life are worthwhile have worsened throughout the pandemic.   Personal well-being is reported slightly more favourably ion the APS than the OPN. Where possible, statistical analyses were used to establish the effect of data collection differences on the estimates. Whilst it was not possible to quantify the influence of the differing modes of data collection, it is believed that this is the main driver behind the differences between the estimates.   The ONS will continue to publish estimates on a quarterly and annual basis, using the APS to allow for the longer time series. Personal well-being will also continue to be monitored throughout the pandemic, on a weekly basis, using the OPN.

Email: sarah.coates@ons.gov.uk

Assessing mortality registration in Kerala: The MARANAM Study
Aashish Gupta and Sneha Sarah Mani; Univeristy of Pennsylvania

Complete or improving civil registration systems (CRS) in sub-national areas in low- and middle-income countries (LMICs) provide several opportunities to better understand population health and its determinants. In this paper, we provide an assessment of vital statistics in Kerala, India. Kerala is home to more than 33 million people and a comparatively low-mortality context. We use individual-level vital-registration data on more than 2.8 million deaths between 2006 and 2017. We compute age-specific death rates from the CRS mortality records and population exposures estimates and compare the rates to those observed by the Sample Registration System (SRS). In robustness tests, we calculate mortality estimates using alternative population exposures; compare infant mortality estimates from the SRS,NFHS, and the CRS; compare un-adjusted and Gompertz mortality rates in single-year ages 40-90; and compare the relationship between child mortality and old-age mortality for Kerala and the HMD countries. We do not find evidence that the CRS under-estimates mortality. Instead, CRS rates are smoother across ages and less variable across periods. In particular, the CRS records higher death rates than the SRS for ages where mortality is low, and for women. Using these data we provide the first set of annual sex-specific life-tables for any state in India. We find that life expectancy at birth was 77.9 years for women in 2017, and 71.4 years for men. Although Kerala is unique in many ways, our findings strengthen the case for more careful attention to mortality records within LMICs, and for their better dissemination by government agencies.

Email: sneham@sas.upenn.edu

Where do couple's report of the year of first marriage differ? An international evaluation of couple's reported first year of first marriage
Emmanuel Olamijuwon; University of St Andrews

Accurate reporting of the timing of marriage is crucial for interventions that address the consequences of marital status on the health outcomes of women and families. However, the accurate reporting of the onset of marriage may be significantly influenced by a myriad of structural factors that stigmatizes unmarried women or single mothers. This study investigates the magnitude of inconsistent reporting of the first year of marriage by couples in their first union by comparing the reported year of first marriage by women and their husband. I also evaluate how inconsistent reporting of the first year of first marriage intersects with socio-demographically relevant characteristics. The analysis is based on a pooled sample of 220,505 couples drawn from the demographic and health surveys of 47 low-and-middle-income countries in seven regions/sub-regions. The results suggest that only about 45% of the couples had a matching first year of first marriage. The result also highlights vast regional heterogeneity in the magnitude of inconsistent reporting of marriage - very high in the Asia region, moderate in sub-Sahara Africa and low in Latin America and the Caribbean. More than 80% of couples in the Kyrgyz Republic, Rwanda, Cambodia, and Afghanistan had a matching reported first year of first marriage compared to less than 35% in Zimbabwe, Angola, Benin, and Indonesia. Finally, I observed that inconsistencies in the reporting of the first marriage are correlated with socio-demographic characteristics. These findings altogether highlight variations in the magnitude of inconsistent reporting of the first marriage, where these variations might be significant and highlight important considerations for dealing with data from couples with an inconsistent report of marriage onset

Email: eoo1@st-andrews.ac.uk