Data science

Strand organisers: Jason Hilton (University of Southampton); Arkadiusz Wiśniowski and Wendy Olsen (The University of Manchester)

15:15 - 16:45 Tuesday 10 September: Data science 1: Data, methods and models

An approach for creating age-sex population projections for multiple overlapping small area geographies via ultra-small area projections
Tom Wilson

The creation of population projections for multiple overlapping small area geographies presents forecasters with a challenge. If projections are prepared for different geographies separately, then inconsistencies are likely for those areas which happen to have the same boundaries in both geographies. Alternatively, projections can be prepared for one principal small area geography and converted to other target geographies using a population-weighted geographical concordance. But this is likely to introduce concordance inaccuracies if each area in the target geography is created from a small number of part-areas of the principal geography. A solution is to create projections for ultra-small areas which can be used as 'building bricks' that can be aggregated up to a wide variety of small area geographies. This paper presents an approach developed recently to produce internally consistent small area projections for multiple geographies by sex and age group. Projections for SA2 areas (generally with populations of 3,000 to 30,000) were created using a bi-regional cohort-component model. These were then broken down to ultra-small SA1 areas (which have populations mostly in the range 200-800) using a new hybrid projection approach combining a Hamilton-Perry cohort change model, a static age structure model, and a total population model. The resulting SA1 projections are shown to be sensible and plausible, and are fully consistent with the SA2 projections of population by sex and age group. Their use in creating population projections for local government areas and transport analysis zones for the Government of South Australia is illustrated.

Email: advanceddemographicmodelling@gmail.com

Beyond the Norm: Exploring Multimorbidity Risks in India's Aging Demographic
Ajay Kumar & Bharti Singh - International Institute for Population Sciences (IIPS), Mumbai

India is passing through the parallel phase of demographic and epidemiological transition with a double burden of NCDs (Non-Communicable Diseases) and CDs (Communicable Diseases). Unhealthy ageing has been identified as the primary driver of the multimorbidity and disease burden among older adults in India. The present study aims to assess the proportional distribution of morbidity, estimate multimorbidity risk for sociodemographic factors, and identify leading risk factors.

The study uses the data from the Longitudinal Ageing Study in India (LASI), Wave – 1, 2017–18, aged 45 years and above. First, we assess the morbidity burden and distribution of morbidity counts over age using a proportional distribution graph. Further, we applied the Random Survival Forest (RSF) model to estimate the risk of multimorbidity, associated with socioeconomic and demographic factors over age and finally conditional plots were utilized to assess the contribution of leading risk factors to morbidity counts.

The prevalence of multimorbidity is 42% among older adults in India. Eye disorders, followed by

CVDs, exhibited the highest proportional distribution over age. While Endocrine diseases, Gastrointestinal Conditions and Infectious diseases showed a concordantly decreasing proportional share in the aging population. The relative proportion of five or more morbidity counts increases significantly over age.

The median expected risk of multimorbidity in females (66 years) is significantly higher than in males (71 years). The study also provides empirical evidence that population with higher level of education, obesity, currently working, and who had poor childhood health were more prone to higher risk of multimorbidity at an early age. Further, obesity correlates with early multimorbidity onset and leads to a pronounced escalation in morbidity counts, particularly in female.

Email: rpvvajaykumar@gmail.com

A Bayesian Model to Estimate Male and Female Fertility Patterns at a Subnational Level

Riccardo Omenti - University of Bologna, Monica Alexander - University of Toronto, Nicola Barban - University of Bologna, - ,

Accurate subnational fertility estimates are crucial for shaping policy decisions across diverse sectors, including education, health care, and social welfare. One of the major challenges in producing these estimates is the presence of small populations, in which information about birth counts stratified by the age of the parent at the birth of the child may be lacking or inadequate. In this research paper, we describe a Bayesian model tailored to estimate the period Total Fertility Rates (TFR) for both men and women at a subnational level. Building on previous work by Schmertmann and Hauer (2019), the model utilizes population counts from age-sex pyramids and models age-specific mortality and fertility patterns allowing for uncertainty. We present a real data application focusing on fertility estimation in US counties for the historical period 1982-2019. Preliminary results reveal distinctive fertility trajectories for men and women across different US counties. Furthermore, the proposed model exhibits significant potential for the examination of male and female fertility behaviors across diverse regions and time frames in multiple countries.

Email: riccardo.omenti2@unibo.it

Developing Joint Bayesian Projections of Male and Female Fertility

Joanne Ellison - University of Southampton, Jason Hilton - University of Southampton, Jakub Bijak - University of Southampton, Erengul Dodd - University of Southampton, Jonathan J. Forster, University of Warwick, Peter W. F. Smith - University of Southampton

The fertility projections literature focuses almost exclusively on births by mother's age. This is partly because male fertility tends to be less well documented, with data that is collected generally having lower quality resulting from a comparatively high proportion of missing paternal ages. Male fertility estimation, however, is a rapidly growing research area, with methods being developed to enable and enhance the comparison with female fertility. These developments make it possible to jointly model and forecast fertility by age of mother and father, which is the aim of this paper. Such an approach accounts for the mechanism by which families are formed: typically a child has a mother and father, and the majority are in a relationship which has particular age characteristics. Incorporating this information where it is available has the potential to achieve more reliable fertility projections.

Preliminary work has focused on modelling male and female fertility rates within the same model for England & Wales and the USA, using vital registration data. We apply Bayesian Generalized Additive Models to estimate a smooth age-period rate surface for each sex. We will investigate methods to borrow strength across sex, such as by sharing parameters or modelling the smooth difference between the surfaces. We will also compare our projections with existing methods as well as independent male and female fertility projections, to determine the impact of the joint modelling approach. The outputs are of substantive interest in and of themselves, however a key secondary use is to inform kinship projection models.

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

09:00 - 10:30 Wednesday 11 September: Data Science 2: Uncertainty in demographic forecasting and now-casting

Spatial Elements in Poisson Regression
Diego Andres Perez Ruiz - University of Manchester

Title: Spatial Elements in Poisson

Regression Using Bayesian

Methods: Applying the Besag-York-

Mollié Model

This presentation introduces the Besag-York-Mollié (BYM) model in STAN, a probabilistic programming platform that does full Bayesian inference using Hamiltonian Monte Carlo (HMC). We start by reviewing the definitions and the calculation of the intraclass correlation coefficient (ICC) for the Poisson estimation of a BYM model using labour force data from India. We studied what routes gender affects the risk of youth in India as active/inactive in the labour market. We observed that STAN efficiently fit our multivariable BYM model using a different set of variables and taking into account special variations in norms.

Email: Diego.perezruiz@manchester.ac.uk

Assessing the quality of data on international migration flows in Europe: the case of undercounting
Maciej Jan Dańko - Max Planck Institute for Demographic Research; Arkadiusz Wiśniowski - University of Manchester; Domantas Jasilionis - Max Planck Institute for Demographic Research; Dmitri A. Jdanov - Max Planck Institute for Demographic Research; Emilio Zagheni - Max Planck Institute for Demographic Research

Undercounting presents a crucial challenge in migration statistics, introducing bias due to insufficient reporting requirements and enforcement issues. Typically, information sources on undercounting include accompanying metadata in official statistics and expert opinions. However, these sources may overlook critical details in migration data shared by various countries, such as changes in methodologies or retrospective updates post-census. To tackle this, our work introduces a methodological solution with three primary objectives regarding undercounting in international migration data. First, we offer an overview of available metadata and expert opinions on undercounting in European migration flows. Second, we propose a novel data-driven approach that incorporates year-specific and duration-of-stay-adjusted classifications. The proposed methodological solution relies on comparisons of flows in the same direction reported by a given country with high-quality data reported by another set of countries. We use bilateral migration data provided by Eurostat, UN and selected national statistical institutes. Duration-of-stay correction coefficients are derived through an optimization model or borrowed from the literature. Metadata and expert opinion scores can also be integrated to classify undercounting. Finally, we provide a dynamic undercounting classification for 32 European countries spanning 2002 to 2021, accessible via an online Shiny application. This platform offers adaptability and flexibility. Our findings reveal significant undercounting in new EU member states like Bulgaria, Latvia, and Romania. Interestingly, even countries presumed to maintain reliable population statistics display notable periods of undercounting.

Email: danko@demogr.mpg.de

Advancing Forecasting Methods for International Student Applications: A Comparative Analysis and Scenario Exploration
Ruth Neville - University of Liverpool, Francisco Rowe - University of Liverpool, Emilio Zagheni - Max Planck Institute for Demographic Research,

The UK is the second largest destination for international students in the world. However, recent declines in

applications from EU member states and other countries of origin pose challenges to its standing in the international education market. Reliable forecasts of future international student applications are crucial for understanding the trajectory of the UK higher education sector. In this study, we aim to enhance current forecasting methods by comparing the effectiveness of ARIMA time-series models, vector auto-regressive (VAR) time-series models, and machine-learning approaches. Secondly, we compute several 'what-if' scenarios – seeking to understand how applications of students in the UK will be affected by demographic, economic, and social changes. Through this work, we seek to not only improve the accuracy of forecasting models but also gain insights into the dynamic nature of international student mobility. Our findings have implications for researchers, institutional stakeholders, and policymakers, providing valuable guidance for strategic planning and decision-making in the higher education sector.

Email: ruth.neville@liverpool.ac.uk

Data science

Measuring the Information Value of Education and Likert Scales with Realist Ontology and Entropy
Wendy Olsen & Ziyang Zhou - University of Manchester

We can explicitly arrange and use ordinal variables better by using mathematical and logical methods. The Shannon entropy of groups of variables can be obtained with a mix of continuous, binary, and ordinal variables. Taking a realist approach, ordinal education as 'highest level of education' involves a cumulation of educational experiences. Its measurement scheme is ordinal-cumulative. A Likert scale is, by contrast, ordinal-distinct. For regression models, we can then consider ordinal-cumulative education as an independent variable. In other cases, a Likert scale can be a dependent or an independent variable. Random sample data from India is used (2022/23 Periodic Labour Force Survey and 2019 Asian-Barometers-India) to illustrate choices along the way. Retroduction is explained and demonstrated. Applying a depth ontology for an ordinal variable involves a real-world assessment, not just a data assessment (see also Borsboom, et al., 2003). There are four types of ordinal variable: entity-distinct; entity-cumulative; process-distinct; or process-cumulative. We have created entropy sums and done regressions for multiple sets of 5 variables, eg effects of education on using internet services in India. To validate our approach, we detected errors of interpretation by comparing alternatives in detail. In summary, data science has tended to ignore ordinality, but we have offered explicit ways forward.

[1] Borsboom, Mellenbergh, and van Heerden (2003) The Theoretical Status of Latent Variables, Psychological Review, DOI 10.1037/0033-295X.

[2] Watts and Crow (2022), The Shannon Entropy of a Histogram, Preprint, arXiv:2210.02848.

Email: wendy.olsen@manchester.ac.uk

Bayesian nowcasting subnational populations in Ukraine to integrate multiple data sources
Andrea Aparicio Castro, Douglas Leasure, Edith Darin - Demographic Science Unit (DSU), Nuffield Department of Population Health (NDPH), University of Oxford

The ongoing Russian military intervention in Ukraine has led to massive displacement and vulnerability among the remaining population due to loss of infrastructure and essential services. To address the need for accurate, up-to-date population data in this context, we have developed a Bayesian hierarchical model that integrates multiple data sources for real-time population nowcasting. This model combines an N-mixture model for imperfect observations with a population time-series model in a Bayesian state-space modelling framework. This model leverages real-time counts of active users on Facebook and Instagram with more traditional data from humanitarian surveys to estimate daily subnational population sizes across Ukraine with measures of uncertainty. Our method is particularly valuable in scenarios where conventional data gathering is not possible due to ongoing crises in the field. Our model was developed using simulated populations to assess the quality of estimates, compensating for the lack of validation data in Ukraine, before being applied to real data for population estimation. This methodology enhances our understanding of population dynamics in conflict zones and demonstrates a scalable technique for subnational population assessments in similar crises worldwide.

Email: andrea.apariciocastro@demography.ox.ac.uk