Data science: innovative data, methods and models

Strand organiser: Dr. Jason Hilton (University of Southampton)

Session organisers:

Spatial Modelling in International and Micro Context – I and II: Prof. Wendy Olsen and Dr. Arkadiusz Wisniowski (University of Manchester)

Digital footprint data for population science: Francisco Rowe (University of Liverpool)

1:30 - 3:00 Monday 11 September: Digital footprint data for population science 1

Becoming influencers: The second shift of full-time mothers in China
Mia Ruijie Zhong - UC Berkeley

China's full-time mother influencer community is rapidly expanding and offers attractive income prospects for top players. This phenomenon is paradoxical as it has emerged against the backdrop of China previously having the highest female labor participation rate worldwide and urban full-time mothers being marginalized. Moreover, members of this group present themselves as satisfied achievers of career and family balance, which raises questions about the interplay between work, family, and fertility. This study explores whether job flexibility and individual entrepreneurship provide a solution to the career and family dilemma, and how motherhood experiences and expertise are framed in knowledge, discourses, and practices. Using online ethnographic observations of over 200 influencer accounts and 45 in-depth interviews, this study investigates why and how young mothers choose to leave the labor force, devote themselves to childcare, and build personal brands as social media influencers. The findings reveal that job flexibility is a primary reason for well-educated women to opt out of the labor market and pursue an influencer career. However, flexibility and work-life balance may be an illusion created by successful influencers and massive invisible sharing labor. The rise of the full-time mother influencer group implies a transition from an adult-centered to a child-centered culture, where the value of high-quality childhood companionship and education dominates educated urban women's career and family choices. This study highlights the value of social media data in understanding decision-making processes, demographic behaviors, and changes in social norms and discourse.

Email: ruijie_zhong@berkeley.edu

Exploring the rise and nature of baby banks in the UK using news media coverage and natural language processing
Hannah Slocombe and Francisco Rowe - University of Liverpool

Rising levels of hardship since the introduction of austerity have rendered essential items unaffordable for many low-income families with young children. Baby Banks – organisations that freely provide essential items and equipment to those with, or expecting, babies and young infants – have grown in the last decade. To date, Baby Banks have received little academic attention and much of what is known about them comes from news coverage. News media plays a critical role in raising public awareness, shaping public opinion and attitudes, and influencing policy formation. Technological advances have allowed for news data to be stored digitally, giving researchers new opportunities to explore this influential data form. This paper draws on 384 news articles to explore the scale and nature of news article coverage of Baby Banks between 2009 and 2022 through sentiment analysis and topic modelling. Our results show that the number of articles written about Baby Banks has grown since 2009, with peaks in coverage during the COVID-19 pandemic in 2020 potentially reflecting their increasing number across the UK. Whilst sentiment towards Baby Banks within news articles has largely been positive, since 2019 there has been a rise in negative coverage due to an increase in articles critical of the growing number of people requiring Baby Banks. This is reflected in the underpinning narratives in articles, which has included a coverage of national political changes driving rising hardship.

Email: sghsloco@liverpool.ac.uk

Online social integration of migrants: Evidence from Twitter
Jisu Kim[1], Soazic Elise Wang Sonne[2], Kiran Garimella[3], Andre Grow[1], Ingmar Weber[4], Emilio Zagheni[1], [1]Max Planck Institute for Demographic Research, [2]World Bank, [3]Rutgers University, [4]Saarland University

As online social activities have become increasingly important for people's lives and well-being, understanding how migrants integrate into online spaces is crucial for providing a more complete picture of integration processes. We curate a high-quality data set to quantify patterns of new online social connections among immigrants in the United States. Specifically, we focus on Twitter, and leverage the unique features of these data, in combination with a propensity score matching technique, to isolate the effects of migration events on social network formation. The results indicate that migration events led to an expansion of migrants' networks of friends on Twitter in the destination country, relative to those of users who had similar characteristics, but who did not move. We found that male migrants between 19 and 29 years old who actively posted more tweets in English after migration also tended to have more local friends after migration, which indicates that migrants' demographic characteristics and language skills can affect their level of integration. We also observed that the percentage of migrants' friends who were from their country of origin decreased in the first few years after migration, and increased again in later years. Finally, unlike for migrants' friends networks, which were under their control, we did not find any evidence that migration events expanded migrants' networks of followers in the destination country. While following users on Twitter in theory is not a geographically constrained process, our work shows that offline (re)location plays a significant role in the formation of online networks.

Email: kim@demogr.mpg.de

The UK online debate on immigration: Polarization, key sources, and speed of content on Twitter
Andrea Nasuto and Francisco Rowe - University of Liverpool

Over the past two decades, immigration has been a prominent topic of discussion in the UK, and social media has increasingly played a crucial role in shaping public attitudes towards immigration. This study examined 220,870 immigration-related tweets to explore how Twitter is influencing the immigration debate in the UK. The study used Natural Language Processing (NLP) to classify tweets into different categories and Social Network Analysis (SNA) to explore the structure of users involved in the discussion. The study revealed that there is a high level of polarization in the online public debate surrounding immigration, with the anti-immigration community being 2.8 times denser than the pro-immigration community, indicating a higher level of engagement within the xenophobic network. Additionally, the study found that only 1% of both producers and spreaders of content generated a disproportionately high amount of content, particularly in the anti-immigration community. The study also found that anti-immigration content spread faster and reached a wider audience than pro-immigration content. The study suggests that identifying and monitoring highly prolific users could help mitigate the spread of anti-immigration sentiment. Moreover, the study's findings contribute to existing literature on measuring the speed of social media content and the role of bots in altering the speed which were found to play a minor role. Overall, this research provides valuable insights into the extent of polarization in the UK's immigration debate and highlights the need for systematic tools to stop online anti-immigration abuse promptly.

Email: sganasut@liverpool.ac.uk

4:45 - 6:15 Monday 11 September: Digital footprint data for population science 2

High-resolution forecasting of European population decline: A machine learning approach
Niall Newsham and Francisco Rowe - University of Liverpool

Population decline is becoming widespread across Europe, with the reversal of longstanding continental population growth imminent. Understanding where population declines will occur remains a considerable challenge, though is essential in preparing for unprecedented population change. Though population change is inherently a localised process, forecasts are often only produced at large spatial scales. Both traditional and more modern computational forecasting methods require highly detailed demographic data, which are typically lacking for small areas and making accurate forecasts difficult to produce. However, recent advancements in remote sensing have yielded spatially granular demographic estimates, enabling the production of high-

resolution population forecasts. Further developments in machine-learning based forecasting methodologies present an exciting opportunity to understand demographic futures. Making use of WorldPop gridded population count and age-structure data, we develop a Long Short-Term Memory recurrent neural network to forecast municipal level population change across the entire European continent to the year 2050. In constructing this predictive model of European population change, we produce a comprehensive overview of likely population decline futures. Specifically, we aim to analyse the geo-spatial distribution of future population decline, and discuss its likely consequences.

Email: n.newsham@liverpool.ac.uk

Internal and international migration of scholars worldwide

Aliakbar Akbaritabar, Maciej J. Dańko, Xinyi Zhao, and Emilio Zagheni - Max Planck Institute for Demographic Research

The migration of scholars has been often studied across countries, however, these studies have rarely focused on sub-national regions. We used data on 28+ million Scopus publications of 8+ million unique authors and geo-coded the affiliation addresses. Our results show that by focusing on the sub-national regions, the share of mobile scholars increases from 8% to 12.4%. We found that in all continents when a sub-national region is attractive for international migrants, it is also attractive for internal ones. The reverse is not true, though. For most continents, a depopulation is happening where scholars move abroad and their position is filled by scholars arriving from other sub-national regions inside the country. In the US, as an example, states in the mid-eastern area have the highest net rate of scholars leaving for other destinations inside the US, mostly on the west coast. In Europe, multiple countries show a similar trend that more developed provinces receive scholars from internal origins and send scholars to international destinations. Our results have implications for the global circulation of academic talent by adding more nuance to the generally accepted image of brain drain and brain gain. We highlight the interrelation between internal and international migration, specifically for regions constantly losing their academic workforce.

Email: akbaritabar@demogr.mpg.de

It's a match! A global view on the use of online dating applications

Francesco Rampazzo 1, Ross Barker 2, Allison Geerts 3, Doug Leasure 1, Pietro Rampazzo 4 1 University of Oxford 2 London School of Economics and Political Science 3 Stockholm University 4 Independent Researcher

The use and experience of dating apps are increasing globally, especially among younger generations, but a comprehensive view is lacking. This study proposes a method to estimate and distribute the number of installations of dating apps by country using market data from the Google Play Store and the Apple App Store. Contextualised topic models are used to identify reasons for using the apps from 1.9 million reviews in multiple languages. The market is primarily dominated by five apps (Tinder, Badoo, Bumble, Plenty of Fish, and Grindr), with high levels of installations observed in high-income countries, but indications of usage are also seen in low and middle-income countries. The most popular reason for use appears to be for casual relationships and friends, although a large proportion reported finding a serious partner through the apps. A separate analysis of LGBTQ+ apps (e.g., Grindr, Scruff, and Her) revealed complaints about non-LGBTQ+ members using these apps and reviews about distance to other users. Regression analysis with multiple predictors will be used to examine the association between national-level dating app usage and economic development, technological adoption, divorce legislation, and LGBTQ+ rights. These indicators may be associated with national-level dating app usage.

Email: francesco.rampazzo@demography.ox.ac.uk

War and migration: Quantifying the Russian exodus with digital trace data

Athina Anastasiadou, Max Planck Institute for Demographic Research Artem Volgin, University of Manchester Douglas Leasure, University of Oxford

Following the Russian invasion of Ukraine, many Russian citizens have left their home country due to increasing repressions by the government, the fear of mobilization, or to escape the economic downturn. As of yet, reliable statistical data on those who left are not available. Hence, much remains unknown about the characteristics and scope of this population. Here we aim to fill this gap by identifying potential migrants using online search results. Our analysis combines search queries provided by Yandex.Wordstat with municipal-level data on socio-

demographic and geographic characteristics and examines the two exodus waves, the first one shortly after the war began and the second one shortly after mobilization efforts started. Extracting meaning from this data with a gravity approach we find that regional indicators for wage levels were less important for the second wave and the proximity of a country of interest increased fourfold in importance. This supports some of the dominant narratives in the media about the Russian brain drain and the scale of the Exodus. Beyond this specific study, we also argue that Yandex searches have the potential to inform migration research.

Email: anastasiadou@demogr.mpg.de

9:00 - 10:30 Tuesday 12 September: Data science: Innovations in demographic data

A machine learning approach to creating robust household-based synthetic populations
Daniel Kopasker[1], Vladimir Khodygo[1], Alisha Davies[2], Nik Lomax[3], Alison Heppenstall[1], S. Vittal Katikireddi[1], [1]University of Glasgow, [2]The Alan Turing Institute, [3]University of Leeds

Background. Synthetic populations are useful in health research as they can overcome issues with accessing observed data (e.g. privacy and timeliness). Methods to form synthetic populations aggregating individuals into households remain limited, despite important drivers of health operating at this level e.g. household income. Furthermore, existing methods often result in multiple clones from individuals underrepresented in observed data, potentially creating bias. Aims. This paper develops a novel and generalisable machine-learning approach to forming a fully open-source household-based synthetic population for the UK to support research on the economic determinants of health and health inequalities. Methods. The UK Household Longitudinal Study (Understanding Society) was used as the primary source of observed household and individual data from which synthetic data are generated using a conditional tabular generative adversarial network. Data are selectively extracted from the large pool of synthetic data in order to match the distribution of household types in census data, while ensuring no clones exist. The synthetic population is validated against external aggregate statistics, with selective replacement of households used for alignment, when necessary. Results. Over one billion households of synthetic data were produced, from which 22.3 million households were selectively extracted to exactly match the distribution of household types in UK census data. The synthetic population also provide a close approximation of the observed age-sex distribution within household types. Observed sex-specific median levels of psychological distress, the initial health focus, were also present in the synthetic population. Validation of economic data is ongoing.

Email: daniel.kopasker@glasgow.ac.uk

Estimating granular migration flows in the UK using consumer data
Olawale A. Ogundeji, Nik Lomax, Stephen Clark - University of Leeds

Migration is a complex phenomenon that significantly impacts a country's social, economic, and environmental landscapes. However, traditional methods of estimating migration flows, such as census data and administrative records, are often limited in their spatial and temporal resolution ability to capture the full extent of migration, particularly for short-term and intra-national moves. This can make it difficult to track the movement of people within countries and to understand the factors that drive migration. The use of consumer data to estimate migration flows has been gaining interest in recent times. Consumer data, such as property rental and transaction, mobile phone records and credit card transactions, can provide a more granular and timely view of migration patterns to evolve a good planning system, from health service provisions to housing and transport facilities. This research proposes to use property rental and transaction data to estimate granular migration flows in the UK. The study will use a spatial interaction model to estimate the probability of migration between two locations to account for the spatial and temporal relationships between migration flows. The model will be calibrated using property rental and transaction, and census data. The research will then be used to estimate migration flows for various periods and geographies in the UK. The results of this research will provide a more granular and timely view of migration patterns in the UK. This information can be used to improve our understanding of migration and to develop policies to address the challenges of migration. Keywords: migration, consumer data, spatial interaction model, UK

Email: o.a.ogundeji@leeds.ac.uk

Using machine learning to predict long-term international migration
Mingqing Wu, Michael Hawkes, Sam Butler, Ollie Pike and Issac Shipsey - Office for National Statistics

The ONS is transforming its official statistical production using innovative modelling methods and new data sources. This paper presents initial research into assessing whether supervised machine learning may be able to make more accurate record-level predictions of long-term international migration than the rules-based approach that is currently used in experimental admin-based migration estimates. Our results demonstrate the promising results of several machine learning algorithms on a subset of non-EU international immigrants recorded in Home Office Border Systems Data. We welcome feedback on this work in progress.

Email: mingqing.wu@ons.gov.uk

1:00 - 2:30 Tuesday 12 September: Session A "Spatial Modelling in International and Micro Context -- I"

Population projections for Germany: subregional development according to varied internal and external migration scenarios
Laura Cilek[1], Elke Loichinger[1], Frank Swiaczny[1], Claus Schlömer[2], Jana Hoymann[2], Steffen Maretzke[2], [1]Bundesinstitut für Bevölkerungsforschung (BiB), [2]Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR)

Due to factors such as population aging and dynamic migration patterns, the projected development of Germany's population at the local level is of particular interest. However, current official forecasts at the district level only include one scenario, failing to capture the entirety of potential population development. As the birth rate in Germany is persistently low and only modest changes in mortality are expected in the coming years, internal and international migration patterns are of great and increasing importance as a parameter determining population trends, especially at sub-regional levels. Using official data from the German statistical office, we create internal migration scenarios on past trends (e.g. suburbanization, urbanization, movements from structurally weak to strong subregions) and combine these with additional high, medium, and low variants of net immigration. We then use a cohort component method with 401x401 internal migration matrices to project the population yearly from 2021 until 2070 by single years of age and sex at the county level. Of particular note, this means that we must also pay special attention to the distribution, both spatially and by sex and age, to the Ukrainian war refugees. While many counties are expected to lose population in the future, our results show large regional differences in population development across all scenarios. The current age structure and number of international migrants play a large role in net growth or decline at the county level, while internal patterns lead to more delicate differences in population across the country.

Email: LauraAnn.Cilek@bib.bund.de

Spatial elements in poisson regression using Bayesian methods: Applying the Besag-York-Mollié model
Diego Perez Ruiz, Wendy Olsen, Arkadiusz Wiśniowski - University of Manchester

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. 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

Fertility and teenage pregnancy in Ghana using time-to-event models and cultural explanations over regions
Wendy Olsen, Jihye Kim - University of Manchester

Giving birth whilst a teenager occurred among 12% of Ghanaian women aged 15-19 years (Ghana Demographic and Health Survey, 2019 – data for 2017). Early pregnancy as a policy issue raises questions over whether national universal policy initiatives might fail due to micro-regional variation in cultural norms underpinning

some of the early pregnancies. ● For Ghana, we explore reasons for the regional differences in teenage pregnancy. ● Social and cultural determinants are involved in the risks of teenage pregnancy. ● Gender norms and the distance to the hospital nearest also have a role in a multilevel model of teenage pregnancy. The dataset for this research is the Ghana Socioeconomic Panel Survey Wave 3, 2018 – 2019 (GSPS). Younger women under age 36 have much more education, hence a higher chance of controlling their fertility, than their parents' generation. Religious and cultural groups were associated with large effects. We test the hypothesis among younger women under age 36 that they will give birth for the first time at a later age if they are in a region where gender norms are strongly pro-women. We assess strengths and weaknesses of a time-to-event model. Possible weaknesses arise from skewness of the dependent variable, and from endogeneity. Past survival models showed localisation of the effects of Ghana's religions.

Email: wendy.olsen@manchester.ac.uk

Toward a geospatial model of urban deprivation using remote sensing, street imagery and survey data
Atsumi Hirose[1,2], Santosh Bhattarai[2] - [1]Imperial College London, [2]UCL

With global urbanisation, many of the global health and development goals will not be achieved without considering the needs of growing urban slum populations. However, to adequately address urban slum populations in global development efforts, global development and health goal metrics stratified by urban sub-populations are necessary. Demographers have used the "slum households aggregation method" to stratify by urban sub-populations, as suggested by Fink et al in 2014. This method operationalises the slum definition by UN Habitat and identifies survey clusters or enumeration areas as 'slum clusters' based on the proportion of 'slum households' lacking access to sanitation, water, durable house and sufficient sleeping space in the particular clusters. While this approach can be easily implemented with survey data, it has limitations including scalability. In the remote sensing community, the availability of geospatial big data with data driven machine learning methods has contributed to development of an alternative approach to identifying slums. While the innovation can be incorporated into population health research to understand demographic processes and health profiles of urban sub-populations, it has not been applied widely. This cross-disciplinary study has three aims. Firstly it will use remote sensing imagery and Google street level imagery with machine learning methods to identify slums. Secondly, a layer of household data will be added into the model to incorporate additional slum features. Finally, model results will be fed into population-based Demographic and Health Survey data analysis to understand demographic processes. Neonatal mortality in Kampala Uganda will be used as the case study.

Email: a.hirose@imperial.ac.uk

Missing females and the discourse of violence – evidence from Indian subcontinent
Purbash Nayak and Suddhasil Siddhanta - Gokhale Institute of Politics and Economics

This paper argues that the latent phenomenon connecting violence and missing females in the marriageable age cohort resulting from previous generation skewed child sex ratios is the society's lust for status. Further, empirical analysis using event bigdata from news articles, GDELT 2.0, illustrates three routes through which aggressive social behaviour is engulfing the physical health of Indian societies – Missing Females, Lack of Cooperation and Lack of Effective Negotiations. Investigation using Multi Scale Geographically Weighted Regression (MGWR) technique further supports the path dependency argument which relates martial pride to sticky aggressive behaviour as well as to their historical association with female infanticide. However, social change can act as corrective by constructing a space for negotiations but the process is slow and requires cultural forces.

Email: purbash.nayak@gipe.ac.in

2:45 - 4:15 Tuesday 12 September: "Spatial Modelling in International and Micro Context -- II"

Small area estimation of multidimensional poverty: The case of Cambodia
Karina Acosta, Central Bank of Colombia

Many public policies in underdeveloped and developing countries are spatially blind. Among the drivers of this blindness is the lack of up-to-date geographical breakdown of data which limits the ability of policymakers to develop and assess impact of policies to tackle poverty at a granulated subnational scale. In this context, this study aims to contribute to the estimation and mapping at a subnational scale of multidimensional poverty in Cambodia for the years 2000, 2005, 2010 and 2014. Moreover, it aims to identify the dimensions that contribute to the largest disparities of multidimensional poverty. The methodological framework to overcome data limitations encompasses techniques from Small Area Estimation (SAE). Specifically, Bayesian spatial hierarchical models which allow us to combine information provided by surveys and additional sources of data to estimate poverty indicators that are not formerly targeted by national surveys. Additionally, this study compares geostatistical models with the most commonly used method for poverty mapping proposed by Elbers, Lanjouw, and Lanjouw (2003).

Email: kacostor@banrep.gov.co

The balance between population growth and land consumption in a dynamic space: An assessment of SDG-11.3.1 in peri-urban villages of Gujarat, India
Ankit Sikarwar - INED and Aparajita Chattopadhyay - IIPS, India

Population dynamics and land use changes are among the pivotal factors that control sustainable development patterns, specifically in developing setups. These factors are extremely dynamic and complex in the peri-urban space, where rural and urban characteristics overlap. On the one hand, such parameters' estimation, monitoring, and maintenance are important. On the other, evaluating global standards (such as the SDG targets related to these parameters) is instrumental for feasible sustainability. In this study, we compare population growth with land consumption, also manifested as SDG target 11.3.1 (Land Use Efficiency), for 615 peri-urban villages surrounding Ahmedabad city of India. The changes are assessed over two decades (i.e., 1991-2011) by integrating remotely sensed data with population census data. The methodology also demonstrates the importance of spatially derived proximity factors (villages' distance from the main city, major towns, and highways) in analyzing the balance between population growth and land consumption. The methods involve spatial analyses and correlation metrics. The results indicated an unusual balance pattern between population growth and land consumption, which also contradicts the recommended levels of SDG target 11.3.1. These trends were explainable in the context of the studied proximity factors. We also discuss the need to evaluate the validity of such global targets in a dynamic space. In addition, related socio-economic and environmental parameters illustrated the challenging situation of villages. Therefore, the study recommends inclusion of peri-urban villages in the urban planning framework.

Email: sikarwar.ankit-kumar@ined.fr

Geographical variation in females' first and second birth in China
Kuoshi Hu, Hill Kulu and Julia Mikolai - University of St Andrews

Geographical variation in fertility levels can be observed at different macro-levels in China. Previous studies have used socio-economic development and the implementation of family planning policies to explain this fertility variation. However, most of the previous research analysed how these macro-level factors explain TFR. Due to ecological fallacy, whether there is geographical variation in parity and how these factors influence females' births cannot be fully explained. We fill this gap by investigating the geographical variation in females' first and second birth among 25 provinces in 7 geographical regions in China. We also investigate how do socioeconomic factors and family planning policy influence females' first and second births in different regions. To do so, we use event history models and individual level data from the China Family Panel Studies. The preliminary result shows that there is geographical variation in the hazard of females' birth, especially in their second birth. Females in Northeast and East China are less likely to have a second child while those in Northwest and South China are more likely to do so. There is a variation in births by cohort too; the hazard of having a second birth for females born 1990-1999 is higher than in other cohorts in East, Central, and Southwest China, whereas females born 1960-1969 are more likely to have a second child in North and Northeast China.

Email: kh256@st-andrews.ac.uk

Geospatial inequalities in caesarean deliveries in Ghana
Fiifi Amoako Johnson - University of Cape Coast, Ghana

Background Caesarean section is aimed to save the lives of mothers and their newborns, but can results in adverse outcomes when not medically indicated. In Ghana, studies have reported inequalities in caesarean section use, however, geographical differentials at the district level where health policy interventions are implemented and monitored have not been systematically studied. This study examined geographical inequalities in caesarean deliveries at the district level in Ghana, investigating geospatial associations of pregnancy complications and birth risks, access to health services, and affluence with inequalities in caesarean section uptake. Data and Methods The data for the analysis was derived from the 2017 Ghana Maternal Health Survey. A log-binomial Bayesian Geoadditive Semiparametric regression technique was used to examine the extent of geographical clustering in caesarean deliveries and their correlates at the district level. Results The results show that 16.0% (95% CI = 15.3, 16.8) of deliveries in Ghana were via caesarean section. The analysis revealed a strong spatial dependence in caesarean deliveries, with a clear north-south divide. Low caesarean deliveries were observed among districts in the northern part of the country, while those in the south had high caesarean deliveries. The predominant factor associated with the observed spatial differentials was affluence (wealth status), rather than pregnancy complications and birth risk factors and access to services. Conclusions Strong geographical inequalities in caesarean deliveries exist in Ghana. Targeted and locally relevant interventions including health education and policy support are required at the district level to address the overuse and underuse of caesarean sections.

Spatial variation in gender inequality and occupational segregation among younger women in India
Amaresh Dubey - Jawaharlal Nehru University

In contemporary world, there is an evolving comprehension that development objectives cannot be accomplished until gender inequalities are obliterated. In South Asia in general and India in particular the causes of the persistence of gender inequalities that has noticeable variation across regions in include social stratification that define the roles and duties apportioned to men and women with distinctive valuation of the tasks executed by them. This has resulted in women being perceived as housewives and mothers rather than the active participants in the labour market. Not only such apprehensions have implications for labour supply by the women, it has a significant impact on the gender based occupational segregation whose attributes represent gender stereotypes in the society resulting in occupational segregation. In this paper we delineate the impact of regional and social diversity in gender norms on the labour supply of the Indian women for the age group 18-29 years. We deploy household level employment and unemployment data collected by the Indian National Sample Survey Organisation for the years 1983, 1993-94, 2004-05, 2011-12 and 2020-21. We find women's labour force participation has changed direction from secular decline until 2004-05 to a rise since 2011-12. We also find that spatial and social fixed effects along with education play a significant role in explaining the spatial and social variations in labour supply. These findings have important implication not only on perceptions about patriarchy and gender inequality but also on intrahousehold bargaining on time allocation and fertility in the Indian context.

Email: dubey.amaresh@yahoo.com

5:30 - 7:15 Tuesday 12 September: Data science: Estimation and forecasting

Balancing theory-, data-, and model-driven results in the Bayesian estimation of demographic quantities
Marija Pejcinovska and Monica Alexander - University of Toronto

The past decade has seen a rapid increase in the use of Bayesian models to estimate demographic quantities of interest. These methods are particularly useful in contexts where data availability is sparse or come from one or many imperfect sources. In these models, structural patterns and trends in the outcome of interest are often data driven, with appropriate adjustments made based on the quality or representitaveness of the data available. Bayesian approaches to demographic estimation differ from classical demographic models, which often rely on mathematical models or strong empirical regularities to make deterministic inferences about patterns in populations where good-quality data are not available. In this paper, we discuss the two approaches to demographic estimation and argue for the increased consideration and integration of classical demographic

knowledge in Bayesian models. In particular, we focus on the case study of estimating neonatal mortality rates (NMR) in all countries worldwide. UNICEF currently uses a Bayesian model to estimate NMR, which centers on modeling the ratio of NMR to other child mortality as a function of the under-five mortality rate (U5MR). The specific functional form of U5MR is derived from patterns observed in the available data. In this paper, we explore other functional forms of the model, which are informed by empirical regularities observed in child mortality patterns from high-quality data sources. We discuss how resulting estimates of NMR are impacted, how interpretations differ, and conclude by discussing the broader implications for demographic estimation using statistical models.

Email: monica.alexander@utoronto.ca

Developing Bayesian projections of subnational fertility for the UK
Joanne Ellison[1], Jason Hilton[1], Jakub Bijak[1], Erengul Dodd[1], Jonathan J Forster[2], Peter W F Smith[1] - [1]University of Southampton, [2]University of Warwick

Subnational fertility projections (SNFPs) are an important driver of subnational population projections (SNPPs), which are vital for national and local governments and businesses to distribute funding and anticipate future demand for resources, products or services. In the UK, SNPPs are published separately by the constituent countries, each with differing assumptions and variants. In this paper we develop a Bayesian SNFP model for the UK that borrows strength across the local authorities (LAs) within the four countries, appropriately quantifies uncertainty, and incorporates expert opinion regarding future national age-specific fertility rates (ASFRs). At the UK level, preliminary work has focused on clustering local schedules of ASFRs to identify groups of LAs with similar fertility patterns. It appears that around four clusters are required, corresponding to early and later ages of childbearing at varying intensities. The subsequent work on projections has concentrated on Scotland due to data availability. We apply Generalized Additive Models to estimate smooth cluster-specific age and year effects, and also investigate the addition of LA-specific random intercepts and age slopes. We find that in terms of predictive accuracy, our proposed model outperforms both naïve freezing of the local ASFRs and a simple extrapolation method, which are important baselines. We will compare our Bayesian SNFPs with existing methods from the literature and official projections to further assess predictive performance. By unifying the projections for the four UK countries within a probabilistic framework, our proposed SNFP methodology has the potential to improve projection reliability and therefore aid government planners in their decision-making.

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

Estimating population in data-scarce contexts: To which extent can we leverage satellite imagery?
Edith Darin and Douglas Leasure - University of Oxford

Granular data on population counts are crucial to support the planning of public infrastructures, ensure political representativity and inform about a population wellbeing. Traditionally it is collected through census exercises mobilising massive logistical and financial resources. Logistical challenges can therefore become insurmountable in countries where political instability, conflicts and natural disasters impede the work of the enumerators. As a consequence, the most vulnerable regions often have either outdated or partial data on their population sizes. Outdated and partial census data have been recently complemented by leveraging new data sources – satellite imagery and geospatial covariates – and innovative modelling approaches – Bayesian hierarchical models to account for the uncertainty in extrapolating population count from a sample of locations to the entire country. The approach coined as bottom-up population models (Wardrop et al. 2018) has been first developed for Nigeria (Leasure et al. 2020) and then expanded to produce population estimates for five provinces of the Democratic Republic of Congo (Boo et al. 2022) and to complement the Burkina Faso census that was impeded by a surge of violence in 20% of the country (Darin et al. 2022). The validity of the bottom-up population modelling approach and more broadly of leveraging satellite imagery to predict population sizes has however to this date never been assessed in a context of complete and reliable population data. We accessed the detailed 2018 Colombian census to study the impact of different settlement maps, sample sizes and sampling designs on population prediction accuracy and precision at different spatial scales.

Email: edith.darin@demography.ox.ac.uk

Migration flow proportions by age and educational attainment
Dilek Yildiz - International Institute for Applied Systems Analysis and Vienna Institute of Demography and Guy

Abel - Shanghai University

In Global North, post demographic transition countries, where fertility and mortality rates have been declining and the population growth is stabilised, migration plays a key role in the change of population size, structure, and characteristics. Especially, due to the refugee crises in 2015 and escalating climate migration, the past decade has seen a growing interest in quantitative research estimating and predicting migration flows. Considering the significant differentials in fertility and mortality, when projecting future population size and structure, it is important to consider the age and education breakdown of migration flows. However, compared to fertility and mortality, quantifying migration is more complicated and there is a lack of global, comparative, and good quality migration flow data stratified by age and education. The majority of research that include the characteristics of migrants have been restricted to migrant stocks which are easier to collect than migration flows. Availability of migration flows broken down by age are limited to high income and only available for the recent years (1990 onward). Therefore, a formal modelling approach is required to produce global estimates. In this paper we propose a methodology to estimate the proportions of age and education specific migration flows. We employ random forest models and Rogers Castro migration age schedules to predict the proportions and we validate our estimates with available data.

Email: dilek.yildiz@oeaw.ac.at

Modelling the age-sex profiles of net international migration
James Raymer, Qing Guan and Tianyu Shen, Australian National University; Sara Hertog and Patrick Gerland, Population Division, Department of Economic and Social Affairs, United Nations

In our paper, we present a methodology to infer the age and sex profiles of net migration. This research supports the United Nations Population Division's estimation and population projection procedures for producing the World Population Prospects (WPP). Age and sex profiles of net migration are required as inputs into demographic accounting models for population estimation and projection. However, most countries in the world do not directly measure migration and residual methods for inferring the patterns have proven inadequate, due to errors in the measures of populations, births and deaths. As net migration does not exhibit regularities across age and sex, we developed a strategy to first estimate flows of immigration and emigration by age and sex, which do exhibit regularities. Differences from these estimates are then calculated to obtain net international migration by age and sex. Based on empirical tests, using data from Sweden, South Korea, Australia, Canada and New Zealand, the methodology shows great promise for overcoming a major data limitation in countries around the world. Further, we apply the model to countries where the age and sex patterns of net migration are unknown and show the results. The paper ends with a discussion of next steps and further extensions.

Email: james.raymer@anu.edu.au

9:00 - 11:00 Wednesday 13 September: Data science: Modelling kinship

Analysing biases in genealogies using demographic microsimulation
Liliana P. Calderón-Bernal - Max Planck Institute for Demographic Research & Stockholm University; Diego Alburez-Gutierrez - Max Planck Institute for Demographic Research; Emilio Zagheni - Max Planck Institute for Demographic Research

The long-term analysis of demographic dynamics is often challenging, especially for questions concerning generational and kinship relations, which depend on vital events and kinship networks spanning decades or centuries. Genealogical data, both offline and online, show great promise for this kind of analysis as they could allow us to link human populations over time, space, and multiple generations. So far, an incomplete understanding of the biases that affect their representativeness has slowed their full exploitation. Here, we conduct a series of experiments on synthetic populations aiming to assess the effect on demographic estimates of some of the most typical biases in ascendant genealogies. We simulated populations using the SOCSIM microsimulation program and data for Sweden (1751-2021) retrieved from the Human Fertility Database (HFD) and the Human Mortality Database (HMD). We analyse three types of biases: the survival and extinction of certain lineages, the inclusion and exclusion of given kin ties, and the under-representation of sub-population groups. We evaluate the size and effect of these biases by comparing usual demographic measures estimated

from ‘perfectly recorded’ and ‘bias-infused’ synthetic populations. Some of the experiments show that the accuracy of demographic estimates based on ascendant genealogies of survivors is affected by the exclusion of extinct lineages, the number of generations and the extent of the kinship network. Yet, such effects are not necessarily linear or unidirectional as their importance likely varies over time. This work seeks to contribute to unlocking the power of genealogies to conduct research in historical and kinship demography.

Email: calderonbernal@demogr.mpg.de

Estimating death rates in complex humanitarian emergencies using the network survival method
Casey Breen[1], Saeed Rahman[2], Christina Kay[2], Joeri Smits[2], Steve Ahuka[3] and Dennis Feehan[1], [1]UC Berkeley, [2]Impact Initiatives, [3]University of Kinshasa

Reliable estimates of death rates in complex humanitarian emergencies are critical for assessing the severity of the crisis and effectively allocating resources. However, in many humanitarian settings, logistical and security concerns make conventional methods for estimating death rates infeasible. In this study, we develop and test a new strategy for estimating death rates in humanitarian emergencies. Our method is based on the network survival method, a promising new approach that uses reports about deaths in survey respondents’ social networks. To test our method, we collected original data in a setting where reliable estimates of death rates are in high demand: Kalemie Province of the Democratic Republic of the Congo. In Kalemie, qualitative work suggested basing death rate estimates on respondents’ reports about deaths among immediate neighbors and kin. We will evaluate our method by benchmarking against a contemporaneous retrospective household mortality survey (N = 2,970). The results will allow us to decompose different sources of error in our network survival estimates. Through this systematic validation, we aim to demonstrate the promise of applying the network survival method in humanitarian disasters.

Email: caseybreen@berkeley.edu

Modelling bias in crowd-sourced genealogies
Nathaniel Darling - LSE

Crowd-sourced genealogies such as Familinx are a novel source for historical demographers with significant potential to overcome problems of other demographic sources. They cover long time spans, have large sample sizes, and are a transnational data source which enable individuals to be linked in spite of migration. However, it is well recognised that these genealogies are biased. They overrepresent past groups with higher net fertility, as these groups are more likely to leave descendants, and are therefore more likely to appear in genealogies. The magnitude of this bias is not well understood, so studies using crowd-sourced genealogies to date have not been able to adjust for it. In this paper I propose to share the results of a simulation using SOCSIM to model this bias. The simulation will use demographic parameters based on those of England and Wales across the demographic transition in order to compare the fertility and mortality characteristics of individuals who do and do not leave descendants. This will enable the bias in crowd-sourced genealogies to be modelled for a range of assumptions about the distribution of fertility and mortality over time.

Email: n.t.darling@lse.ac.uk

Projections of kinship for all countries
Diego Alburez-Gutierrez - Max Planck Institute for Demographic Research, Ivan Williams - Universidad de Buenos Aires and Hal Caswell - University of Amsterdam

Demographers have long attempted to project the future size and age-sex composition of populations but have largely ignored what these changes mean for the size, structure and age composition of family networks. Kinship structures matter because family solidarity, a crucial source of informal care for millions around the world, is conditional on kin availability. Here, we present probabilistic projections of biological kin (nuclear and extended family) for all countries in the world over the 2023-2100 period using novel methods from mathematical demography. We outline potential future kinship structures and discuss what they imply for the availability of informal care. Kinship networks will shrink, with important differences over age: individuals will have fewer relatives when they are young and more kin when they are old. For population in median age, the largest losses will be from horizontal kin (siblings, cousins, nieces) and the highest gains from vertical older kin (grandparents, parents, uncles). The speed of demographic change matters. Populations with a history of high fertility will enjoy

larger and younger kinship networks. Changes in kin supply matter to individuals, who are providers and consumers of informal care. They also matter for policy makers, as ever-slimmer kinship structures put increased pressure on systems of social support in the context of rapidly ageing populations.

Email: alburezgutierrez@demogr.mpg.de
