Data, big data & statistical uncertainty strand

Strand organisers: Jason Hilton (University of Southampton), Phil Humby (Office for National Statistics), Louisa Blackwell (Office for National Statistics)

Statistical uncertainty. Tuesday 15 September 11.00am

Longitudinal linkage of administrative data; design principles and the total error framework - ***Sarah Cummins, Louisa Blackwell, Nicky Rogers, Eleanor Fordham, Office for National Statistics***

The Office for National Statistics (ONS) is committed to increasing its use of administrative data in the production of population statistics. This paper seeks to offer a template for the statistical design of datasets that are derived through the linkage of administrative data, to produce longitudinal datasets. While the focus is on administrative data linkage, it could also be applied to other data blends, for example microdata linkage of survey and administrative data. We emphasise the importance of careful statistical design, prior to implementation. The idea is that the design should represent the optimal balance of user requirements, design features and statistical error management. To support this, we propose an error framework. This is intended as a helpful taxonomy of the potential errors that are integral to administrative data sources, to be aware of in either evaluating admin-based sources or in designing new, integrated, longitudinal datasets. Some of the error that we draw attention to are conceptual, resistant to quantification, in any precise way, at least. Some error can be quantified, and we highlight the use of scaling factor analysis, linkage error investigations, edge effects and the analysis of residuals. We are still developing our thinking regarding error management and the quantification of statistical error and uncertainty, so this paper reports on work that is still in progress. We provide an example from our application of this framework to help us understand administrative data, for estimating international migration.

Email: Louisa.Blackwell@ons.gov.uk

Measuring statistical uncertainty in population estimates - ***Adriana Castaldo, Steve Martin-Drury, Paulina Galezewska, Louisa Blackwell, Gemma Hanson, Office for National Statistics***

The Uncertainty Project was established as part of the Migration Statistics Improvement Programme (MSIP 2008 to 2012). Working in collaboration with Southampton Statistical Sciences Research Institute (S3RI), this project aimed to provide users of Office for National Statistics (ONS) local authority mid-year population estimates with more information regarding their quality. This report summarises the methodology for deriving the statistical measure of uncertainty associated with the local authority mid-year population estimates. We use the cohort component approach to create the local authority mid-year population estimates. The cohort component method uses the 2011 Census for the population base and then incorporates natural change (births less deaths), net international migration and net internal migration, and other adjustments (for example, asylum seekers). Initial work identified the census base, international migration and internal migration as having the greatest impact on uncertainty, and our measure of uncertainty is a composite of uncertainty associated with these three components only. The methodology for deriving the statistical measure of uncertainty applies the same processes to derive simulated distributions for each of these components, which are then combined using the cohort component formula. We derive measures of statistical uncertainty empirically from the ranked simulated composite values.

Email: Gemma.Hanson@ons.gov.uk

Understanding the statistical properties of integrated data for estimating the usual population of England and Wales - ***Louisa Blackwell, Adriana Castaldo, Amy Large, Office for National Statistics***

The Office for National Statistics (ONS) is committed to maximising its use of administrative data and reducing reliance on the decennial census. In 2019 a third version of Admin-Based Population Estimates (ABPE) for England & Wales was published as research statistics. The ABPEs are created through the linkage, both deterministic and probabilistic, of administrative data sources including the GP Patient Register, DWP

Customer Information System, Higher Education Statistics Agency (HESA) data, and School Census data for England and Wales. The ABPEs then use activity indicators to determine record inclusion in the population estimate. ‘Activity’ refers to an individual interacting with an administrative system, for example by paying tax, collecting benefit or changing address. In this presentation we provide insights on ABPEs quality through the application of the error framework for longitudinal administrative data sources. We also describe how we derive measures of statistical uncertainty for the ABPEs. Here we define statistical uncertainty as the quantification of doubt about an estimate. By bringing these two measures together we are enhancing our knowledge of statistical uncertainty that may appear in administrative data sources, and on the datasets we produce when they are combined through record linkage. We consider how we might optimise the design of integrated datasets going forward using these new insights and quality indicators.

Email: Amy.Large@ons.gov.uk

Big data & machine learning. Wednesday 16 September 9.30am

Identifying the under-five pneumonia, diarrhoea, and malaria prediction characteristics’ differences between urban and rural residents: a machine learning predictive modelling approach - ***Muhumuza Rornald Kananura, London School of Economics and Political Science***

Environmental, socio-economic, and demographic characteristics are known as major determinants of morbidities’ occurrence. However, their effect is a complex process of interrelated mechanisms as these variables are usually many and correlated. Because of the classical statistical approaches’ limitation in handling such highly dimensional and correlated variables (collinearity assumption) some important variables are usually dropped and thus limiting the realistic design of interventions. In this study, I applied Machine Learning (ML) modelling approach to identify the predictors of suspected malaria, pneumonia, and diarrhoea among children with a focus on urban-rural differences. Using the 1988-2016 Uganda Demographic Health Survey data, I applied 4 ML techniques (logistic regression, random forest, gradient-boosted decision tree, deep neural network). I split data into 70% and 30% of training and testing datasets, respectively. Using the testing dataset, I selected the model with the highest average area under the receiver operating characteristic (ROC) curve. For all morbidity outcomes, gradient-boosted decision tree was the best model with the average ROC of at least 70%. For both outcomes, child age, maternal age, sharing of toilet, age of the household head, number of kids in the household were the most important predictors for both rural and urban dwellers. Ownership of income generating assets (animals), the incomplete roof and floor house structure were additional important predictors for rural dweller. Although both rural and urban have common enormous number of predictors of suspected pneumonia, diarrhoea, and malaria, effect elements of socio-economic position that predict these morbidities in rural and urban dwellers were different.

Email: r.m.kananura@lse.ac.uk

Measuring the impacts of health inequalities on the health economy of a deprived inner urban area of London using personal health records and other large administrative datasets - ***Les Mayhew[1], Gillian Harper, [1]City University***

Health professionals have long been aware of the potential influence of factors such as low income, poor education, housing and a range of other factors collectively termed ‘deprivation’ on people’s health. Until now it has not been possible to quantify the interplay between specific determinants such as low income, a person’s health and the subsequent use made of health services, including the financial burden on the health and social care economy of health inequalities. This paper reports the results of a major study in east London, completed in the last six months, involving the linkage of hundreds of thousands of personal medical records including three years’ worth of hospital admissions and outpatient data, attendances at accident and emergency centres, immunisation records and data on treatment costs. These data in turn are linked to other information at both person and address level such as household type, eligibility for welfare benefits, educational achievement and others. The legal background to the use of the data in this way is covered by the Data Protection Act 2018, which is the UK's implementation of the General Data Protection Regulation (GDPR).

The data set itself is fully pseudonymised with strictly applied time-limited access to a very small number of authorised users working in a secure environment. The project, approved by the Department of Health, involves various partners and stakeholders in the local health economy including the local council, hospital trusts and commissioning groups. This paper will provide an overview of the broad data sources, the algorithms used to link together multiple data sources at a person and household level. So far, the data set has been used in around 30 different case studies using a range of statistical techniques some results of which will be provided. Otherwise the results are being used by health providers the public health department and the local council to inform both health commissioning and strategy

Email: lesmayhew@googlemail.com

Using machine learning and twitter data to profile attitudes towards immigration - ***Francisco Rowe[1], Yerka Freire-Vidal[2], Eduardo Graells-Garrido[3], [1]University of Liverpool, [2]Universidad del Desarrollo, Chile, [3]Barcelona Supercomputing Center***

Immigration is a key ingredient for social cohesion and economic development. Yet, it is often portrayed as a major threat to national identity, values, economic stability and security, resulting in acts of intolerance, discrimination, racism, xenophobia and violent extremism. Understanding how misperceptions towards immigration are formed and shaped is key to address combat mis-representations of immigrants. Typically attitudes towards immigration are studied based on qualitative and nationally representative surveys but they offer low population coverage, coarse geographical resolution and slow data collection. Social media offers dynamic and open space to better understand experiences and public opinion about immigration. While some bias exists, social media data are produced at unprecedented temporal frequency, geographical granularity and is accessible in real time. This paper aims to measure and better understand attitudes towards immigration in Chile using Twitter data, topic modelling and sentiment analysis. Key findings indicate that negative attitudes emerge from a reduced number of users, and are more commonly manifested and intensify during negative immigrant news reflecting arguments of job competition and stricter immigration regulation. Positive attitudes are expressed by a more diffused number of users and are predominantly express to manifest support during specific events reflecting supportive arguments for immigrants' human and civil rights.

Email: F.Rowe-Gonzalez@liverpool.ac.uk

Data quality. Wednesday 16 September 1.00pm

Do we know what proportion of non-British citizens resident in the UK are applying to the EU Settlement Scheme? - ***Jo Zumpe, Becca Briggs and Jay Lindop, Office for National Statistics***

Following the European Union (EU) referendum in June 2016, the UK officially left the EU on the 31st January 2020 and is in a transition period until the end of the year. Citizens from the EU, other EEA countries and Switzerland, who are already living in the UK will have the right to remain, but they will need to apply for a new legal status in order to prove this. Consequently, in March 2019, the Home Office opened the EU Settlement Status (EUSS) Scheme, where eligible individuals can apply online for pre-settlement or settlement status within the UK. The proportion of the eligible population who have already applied or still have to apply, has been of much interest within and outside of the UK government. What do we know, what don't we know and why is this straightforward question so hard to answer? Estimates of the usually resident population of the UK by nationality are available from ONS and reports on the number of applications to the EUSS are updated monthly by the Home Office. But methodological differences in the survey and administrative data make it very difficult to make any true like-for-like comparisons. This paper builds on the note of caution in the Home Office published reports on EU application numbers, providing insights into the statistics and how they differ. It identifies gaps in the information currently available and discusses ways National Statistics may evolve.

Email: Jo.Zumpe@ons.gov.uk

The Up-Series Generation in the Office for National Statistics Longitudinal Study - ***Alison Sizer[1], Oliver Duke-Williams[2], [1]CeLSIUS, Department of Information Studies, University College London, [2]Department of Information Studies, University College London***

Aim: This paper describes the ONS Longitudinal Study (LS) and its use to examine whether a convenience selected sample (participants of the Up-Series of documentaries) was representative of seven-year-olds living in England and Wales in 1964 in terms of their socio-demographic life courses. Methods: We used descriptive analysis of longitudinal socio-demographic data (gender, social class, education, National Statistics Socio-Economic Classification (NSSEC), tenure) on 14,900 LS members and compared this with the fourteen participants of the Up-Series. Data: Two data sources were used, the ONS LS and the Up-Series of documentaries, The ONS LS, covers England and Wales and is built around samples drawn from the decennial census on the basis of four birthdays. It now includes up to 40 years of data on over 1 million individuals. The Up-Series of documentary films, documented the lives of 14 participants at seven-year intervals, starting at age 7 in 1964. Results: On all but gender, the Up-Series participants were representative of the same cohort in the LS. In terms of their socio-demographic outcomes, LS members who grew up in high social class households had more advantageous life courses than those who grew up in working class households. In comparison, the Up-Series participants displayed more extreme socio-demographic life courses. Conclusions: The Up-Series gives the impression that social mobility was rare; the childhood circumstances of its participants largely determined the "men" they became. Analysis of the LS suggested social mobility was more common. The expansion of women's employment and a wider restructuring of the labour force were some of the reasons for this.

Email: a.sizer.11@ucl.ac.uk

Geography of UK Biobank: considerations for avoiding bias in measures of association - ***Harry Taylor[1], Paul Norman[2], [1]University of Manchester, [2]University of Leeds***

UK Biobank represents an opportunity to study a very large sample of the population yet there may be data limitations. Differing from the general population in sex, ethnicity, affluence and overall level of health, Biobank also displays urban bias due to its data collection strategy. Despite its non-representative sample, Biobank is regarded as suitable for studying disease-exposure relationships. The present research explores how the effect of an influencing factor (area-level deprivation) on an outcome (likelihood of hearing aid use) differs for participants living in urban and rural areas. The relationship between Townsend deprivation and hearing aid use was modelled correcting for age, sex, ethnicity, education and employment status, with predicted probabilities generated for hearing aid use across the range of observed Townsend scores. Deprivation was associated with an increase in hearing aid use from least to most deprived areas in urban (>10k population) areas. However, in rural/town areas, the association between deprivation and hearing aid use was stronger. To check for potential bias, equivalent models were run for general health outcomes in both Biobank and the Health Survey for England (HSE); a representative sample. Although significant health differences were observed in Biobank, the effect of urbanity was less pronounced than in HSE. Relationships between outcomes and predictors in the UK Biobank risk being inaccurate, unless geography is taken into account. The implications of the sampling and data collection methods should be understood and explained as part any research using this dataset.

Email: p.d.norman@leeds.ac.uk

Own-household fathers and part-time residence - a challenge to language, categories and survey questions - ***Rebecca Goldman[1], Paul Bradshaw[2], Adrienne Burgess[1], and Konstantina Vosnaki[2], [1]the Fatherhood Institute, [2]Scotcen***

Birth fathers who do not co-reside full-time with their child/ren ('Own Household Fathers'/ OHFs) are often termed 'non-resident'. Yet between a third and a half of these fathers regularly have their child to stay overnight; and others provide substantial daytime care. Challenging the usual binary classification of fathers as ‘resident’ or ‘non-resident’, we argue that ‘overnight care’ fathers are part-time resident with their child/ren. Our presentation will report findings from a study funded by the ESRC in which we developed and cognitively tested survey questions to identify distinct sub-categories of OHFs based on overnight stays and types of interaction with their children. Our question development drew on our scope of relevant questions asked on UK and international studies. Our cognitive test involved eleven mothers participating in the Growing Up in Scotland cohort study. We found that routinely used ‘frequency of OHF-child contact’ questions are difficult to answer accurately, consistent with published critiques by family demographers. Capturing details of the frequency, quantity (how much time) and pattern of interaction is complex. Using a calendar as a visual aid may reduce measurement error. All respondents considered it appropriate to be asked for contact details of the OHF for researcher follow-up. Our questions on overnight stays, episodes of time together, and virtual communications (including social media) could be adapted for use on quantitative studies investigating separated families. Part-time residence questions may be relevant to statistical surveys, which rarely identify children living part-time in two parental households, nor men living part-time with their children.

Email: r.goldman@fatherhoodinstitute.org