Data Science: Innovations in data collection & methodology

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

Synthetic data session convenor: Paul Norman (University of Leeds)

Systems Science in Public Health & Health Economics Research session convenor: Nik Lomax (University of Leeds)

1.30pm Monday 5 September: Systems Science in Public Health & Health Economics Research (SIPHER)

An overview of the Systems Science in Public Health and Health Economics Research (SIPHER) consortium: facilitating the move towards healthy public policy

Petra Meier[1], Corinna Elsenbroich[1], Julian Cox[2]; [1]University of Glasgow, [2]Greater Manchester Combined Authority

SIPHER is a large UK public health research initiative, established with funding from the UK Prevention Research Partnership (UKPRP) bringing together scientists from six universities, three government partners at local, regional and national level, and ten practice partner organisations. SIPHER is developing research to prevent physical and mental ill health, and tackle persistent health inequalities. SIPHER works across some of the most important social determinants of health: inclusive economic growth; decent and affordable housing and policies that promote mental wellbeing. Drawing on participatory systems mapping and evidence synthesis, SIPHER is developing complex system models and decision support tools for use in public policy settings. This talk provides context for the project and highlights how a multidisciplinary approach co-produced with policy partners can help us move towards healthy public policy.

Email: petra.meier@glasgow.ac.uk

An End to Policy Silos? Analysing policy efforts to use complex systems modelling to achieve more joined-up policymaking

Katherine E. Smith[1], Liz Such, University of Sheffield[2], Clementine Hill O'Connor[1]; [1]University of Strathclyde, [2]University of Sheffield

Breaking down departmental and organisational silos has long been a goal of government, whether through the doctrine of 'coordination' or 'joined up government' (Hood, 2005). In this paper, we will first review efforts to achieve this kind of policy co-ordination within public health, such as 'Health in All Policies' (HiAP) and 'intersectoral collaboration for health' (Such et al, 2022). We will explore this with reference to a complexity-informed systematic review of the HiAP literature. Next, we will draw on empirical analysis of how three policy organisations operating within the UK are trying to use complex systems modelling to achieve inclusive economy goals. Analysing multiple qualitative data sources (interviews with policy actors and researchers, policy documents and observational data), we will discuss three key tensions that emerge in all three contexts: (1) complexity versus usability; (2) rational, data-driven approaches to policy analysis versus strategic, politically driven decision-making; and (3) sophistication and rigour in analytic support for policy decisions versus a push for transparency and democratic engagement in policymaking processes. We will conclude by outlining some potential options for managing these tensions.

References:

Hood C. 2005, The Idea of Joined Up Government: A Historical Perspective. In Bognador, V (ed). Joined Up Government. Oxford University Press, Oxford.

Such L, Smith K, Woods H and Meier, P. 2022, Governance of Intersectoral Collaborations for Population Health and to Reduce Health Inequalities in High-income Countries: A Complexity Informed Systematic Review. International Journal of Health Policy and Management.

Email: katherine.smith.100@strath.ac.uk

Utilising Synthetic Microdata to assess the spatial distribution of the inclusive economy
Nik Lomax[1], Ceri Hughes[2], Ruth Lupton[2]; [1]University of Leeds,[2]University of Manchester

A barrier to undertaking detailed spatial analysis of policy is the lack of high-resolution data on the composition and distribution of the population across key variables of interest. In this paper we utilise a spatially representative individual level 'digital twin' dataset of the UK population that has been built from a range of datasets (census, survey and administrative) using spatial microsimulation methods. We are concerned with assessing the spatial distribution of variables which contribute to how inclusive the economy is at a fine spatial resolution. These indicators include access to services, jobs and affordable housing which all vary spatially and temporally. We create an area-based classification from these variables and compare area types to understand how (un)equal the distribution of inclusivity is within a number of UK cities. We briefly outline the data used, methods, strengths and limitations of both the digital twin population and the indicators set before presenting results from the classification exercise. SIPHER's inclusive economy indicators are designed for inclusion in models of relationships between policies for economic inclusion and health outcomes. The results will help policy makers by demonstrating where there is specific need across the inclusive economy domains

Email: n.m.lomax@leeds.ac.uk

A dynamic microsimulation of UK mental health outcomes
Rob Clay[1], Nik Lomax[1], Alison Heppenstall[2]; [1]University of Leeds, [2]University of Glasgow

Validation of government policy is increasingly assisted by 'What-If?' population projection scenarios. To build these scenarios, we use dynamic microsimulation models which operate at the level of the individual. The impact of changing a variable in the system can be estimated from longitudinal datasets, such as panel surveys, and these are used to create probability distributions from which a random draw can be applied to each individual in the model. These random draws are used to transition the state of each individual at discrete time steps, for example an individual might transition from good to poor health. In this paper we make use of Understanding Society, a longitudinal survey dataset that is representative of the UK population over 11 waves. The population is evolved forwards into the future under hypothetical policy strategies. We look specifically at the example of what impact changes in household disposable income might have on mental health outcomes via a set of pathways, including housing quality, employment and environmental quality. We assess the spatial and between group impacts of these hypothetical policy experiments by applying the distributions to a spatially reweighted "digital twin" population of the UK. These projections provide useful insight to policy makers who need to make decisions about how to allocate constrained budgets and better target specific policies to improve health outcomes

Email: gyrc@leeds.ac.uk

4.45pm Monday 5 September: Synthetic data

Linking individuals to areas: protecting confidentiality while preserving research utility
Paul Norman[1], Jessie Colbert[2] and Dan Exeter[2]; [1]School of Geography, University of Leeds, [2]School of Population Health, University of Auckland, New Zealand

A tension exists between making data available which protects the confidentiality of individuals while containing sufficiently detailed geographic information to underpin the utility of research. We test this by investigating the interplay between two geographical entities (points for the observations and polygons for area attributes) at a variety of scales, using a synthetic population of 22,000 people. We do this for individuals located by postcodes and by postal sector and postal district centroids and link these to a variety of census geographies. We also 'jitter' postcode coordinates to test the effect of moving people away from their original location.   We find a smoothing of relationships up the geographical hierarchy. However, if postal sector

centroids are used to locate individuals, linkages to Lower / Medium Super Output Area scales and subsequent results are very similar to the more detailed unit postcodes. Postcode locations jittered by up to 500 metres in any direction are likely to allow the same conclusions to be drawn as for the original locations.

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

A simulation setup allowing the comparison and development of small-area mortality estimation models
Benjamin-Samuel Schlüter; Catholic University of Louvain (DEMO)

Within-country variations in mortality are an important dimension of inequality. Studying mortality at a subnational level comes with an additional difficulty: the presence of small populations in which the stochastic variation in death counts is relatively high, making the underlying mortality unclear. Several small-area mortality estimation models have been developed addressing that statistical challenge. They are commonly estimated in a Bayesian framework allowing to easily smooth data over space (and time), incorporate demographic knowledge and handle uncertainty. Here we propose a simulation setup allowing to compare models' performances while knowing the true age-specific mortality rates present in small-areas. Simulating further allows assessing models' performances according to different scenario such as population sizes and spatial structure of mortality within a fictious country.   We created the simulation by randomly drawing deviations from standard mortality age schedules- representing different European countries- using the Brass relational model. Having different mortality standards allowed us to tune the spatial structure of mortality in the simulation. We further applied time series model on the Brass parameters, calibrated on real data, to simulate changes in small-area mortality levels over time. Simulation outputs suggest that we obtained a diversity in mortality levels while having a realistic range of life expectancy at birth. Furthermore, it enforces an overall mortality decrease while imposing a certain stability in best and worst performing small-areas over time.   Using the simulation, three models (Schmertmann & Gonzaga (2010), Congdon (2014) and Alexander et al. (2017)) are compared according to their root mean squared error and coverage.

Email: benjamin-samuel.schluter@uclouvain.be

Employing population synthesis for updating neighbourhood geodemographic in the context of the 2021/2 Output Area Classification
Jakub Wyszomierski[1], Paul A. Longley[1], Christopher G. Gale[2]; [1]University College London, [2]Office for National Statistics

This paper reports ongoing work to develop a framework based upon individual names and addresses to enable annually updateable geodemographic classifications with UK-wide coverage. We propose a method for improving spatio-temporal granularity of geodemographic segmentations cognisant of the possible demise of UK censuses and wider use of administrative data (Coleman, 2013). Our innovation is to use observed individual level name and address data from linked consumer registers (Lansley et al 2019) to residualise reliance upon synthetic data derived using conventional Iterative Proportional Fitting techniques to aggregate census data. Individual level name-based models of ethnicity, age, sex and household structure are used to further limit reliance upon microsimulation and the vagaries of inferring micro scale demographics from macro-scale surveys. Our fundamental premise is that the best models of social dynamics are grounded at the level of the human individual and that this approach is more conducive to effective representation of local context and place-specific effects. We also enhance the quality of microsimulation at highly granular spatial scales by georeferencing all known individuals and using observed and modelled individual characteristics to infer the characteristics of missing individuals in interstitial locations.   Our methods use and model individual level Personal and Sensitive Personal Data. As such they are developed and applied in trusted research environments. This also enables understanding of local area residential dynamics and transitions independent of administrative boundaries of aggregated open data.   We present the results of preliminary analysis that demonstrate the value of the approach in developing and updating geodemographic segmentations

Email: jakub.wyszomierski.16@ucl.ac.uk

Understanding the demographic determinants of mortality crisis-related kin loss: An application to COVID-19 in developing countries
Mallika Snyder; University of California, Berkeley

Mortality crises leave bereavement crises in their wake. The steadily growing toll of COVID-19-related kin loss provides a trenchant example: each United States COVID-19 death is mourned by nine close kin (Verdery et al., 2020), an estimate that may vary across countries and contexts, but which indicates the large number of individuals at risk of the adverse health and socioeconomic impacts known to be associated with kin loss (Wang et al., 2021; Fletcher et al., 2013). As efforts to estimate numbers of affected individuals continue (Hillis et al., 2021, Snyder et al., 2022), little is still known about the pandemic's impact on bereavement experienced in developing countries, in which younger and larger kin networks, combined with differences in the age distribution of excess mortality from that observed in higher-income countries (Demombynes et al., 2021), may have led to experiences of kin loss substantially different from those in high-income countries for which most estimates have been developed. In this paper, I will draw on demographic microsimulation with the SOCSIM platform and formal demographic approaches to better understand the sensitivity of crisis-mortality-related kin loss to excess mortality and population age and kinship structures. Not only will this allow for predicting the impact of COVID-19 on kin networks in developing countries, but also provide a flexible framework for better understanding the demographic determinants of a country's experience of kin loss after a mortality crisis in a wide variety of contexts around the world.

Email: mallikasnyder@berkeley.edu

2.45pm Tuesday 6 September: Data science: Innovative sources of demographic data

The value of culture for predicting migration: Evidence from Facebook data
Carolina Coimbra Vieira, Sophie Lohmann, Emilio Zagheni; Max Planck Institute for Demographic Research

One of the strongest empirical regularities in spatial demography is that flows of migrants are positively associated with population stocks at origin and destination, and inversely related to distance. This pattern, observed in the 19th century, was formalized into what are known as gravity models of migration. Traditionally, distance is measured geographically, however, other measurements including economic and cultural factors have also been found to be relevant to explaining migration flows. In particular, we believe that cultural distance may be one important form of distance to explain migration flows because it is dynamic. In this paper, we propose a scalable approach to obtain proxies for cultural similarity across countries by using data from the Facebook Advertising Platform. Our results show that our new measure of cultural similarity adds over and above standard predictors in predicting migration, opening new opportunities to understand determinants of migration.

Email: coimbravieira@demogr.mpg.de

Stop, in the name of COVID! Estimating the effects of COVID-19-related travel restrictions on migration using social media data
Jordan D. Klein[1], Ingmar Weber[2], Emilio Zagheni[3]; [1]Princeton University, [2]Qatar Computing Research Institute, [3]Max Planck Institute for Demographic Research.

In the wake of the COVID-19 pandemic, the International Organization for Migration has postulated that international migrant stocks have fallen short of their pre-pandemic projections by nearly 2 million due to travel restrictions. However, this is not testable with migration data from traditional sources due to the negative impacts the pandemic has had on census data collection operations, and key migration stakeholders, including the IOM, have called for the use of data from alternative sources, including social media, to fill in these gaps. Building on previous work using social media data to estimate trends in migrant stock following external shocks, we test the hypothesis that migrant stock was reduced by COVID-related travel restrictions using Facebook data, adjusting them for bias using an established set of methods. Our preliminary results

support this hypothesis, suggesting social media data may help fill critical gaps in traditional migration data sources during the COVID-19 pandemic.

Email: jdklein@princeton.edu

Urban exodus? Understanding human mobility in Britain during the COVID-19 pandemic Using Facebook data
Francisco Rowe, Alessia Calafiore, Daniel Arribas-Bel, Krasen Samardzhiev, Martin Fleischmann

Existing empirical work has focused on assessing the effectiveness of nonpharmaceutical interventions on human mobility to contain the spread of COVID-19. Less is known about the ways in which the COVID-19 pandemic has reshaped the spatial patterns of population movement within countries. Anecdotal evidence of an urban exodus from large cities to rural areas emerged during early phases of the pandemic across western societies. Yet, these claims have not been empirically assessed. Traditional data sources, such as censuses offer coarse temporal frequency to analyse population movement over short-time intervals. Drawing on a data set of 21 million observations from Facebook users, we aim to analyse the extent and evolution of changes in the spatial patterns of population movement across the rural-urban continuum in Britain over an 18-month period from March 2020 to August 2021. Our findings show an overall and sustained decline in population movement during periods of high stringency measures, with the most densely populated areas reporting the largest reductions. During these periods, we also find evidence of higher than average mobility from highly dense population areas to low densely populated areas, lending some support to claims of large-scale population movements from large cities. Yet, we show that these trends were temporary. Overall mobility levels trended back to pre-coronavirus levels after the easing of non-pharmaceutical interventions. Following these interventions, we also found a reduction in movement to low density areas and a rise in mobility to high density agglomerations. Overall, these findings reveal that while COVID-19 generated shock waves leading to temporary changes in the patterns of population movement in Britain, the resulting vibrations have not significantly reshaped the prevalent structures in the national pattern of population movement.

Email: fcorowe@liverpool.ac.uk

"Tell me what you buy, and I will tell you who you are": Geo-demographic characteristics of migrants through online retail data
Francesco Rampazzo[1,2], Jason Bell[1], Ridhi Kashyap[2,3], Micol Morellini[3,], Melinda Mills[2,3], Andrew Stephen[1];
[1]Saïd Business School, University of Oxford, [2]Leverhulme Centre for Demographic Science, University of Oxford, [3]Department of Sociology, University of Oxford,

Marketing data is rich with demographic information at finer granularity in time and space. In this research, we aim to create a geo-demographic profile of Italian migrants in Europe using MammaPack consumer data in combination with other data sources. MammaPack is an online retail shop that targets Italian migrants in 21 European countries. MammaPack consumer data are analysed in combination with data from Facebook Advertising Platforms, Google Analytics, as well as data from the Register of Italians living abroad (AIRE). The period of analysis is January 2019 to December 2021. Our aim is to combine such traditional and non-traditional data sources in the development of migrants' geo-demographic profile. Preliminary results identify consumers that have recently become parents (i.e., they buy baby food). We are planning to complement this information by surveying MammaPack customers with an opt-in questionnaire. The survey will provide additional demographic information on the customers (e.g., origin and destination of migration trajectories). By training a model on those respondents for which we have both survey and customer data, we will be able to infer and predict the characteristics of the rest of the MammaPack customer base. This research will contribute to the use of consumer data in demography in combination with digital trace data and register data.

Email: micol.morellini@nuffield.ox.ac.uk

5.30pm Tuesday 6 September: Innovations in data collection and methodology

A workflow for cross-cultural research design across multiple study locations

Anushé Hassan[1], Laure Spake[2], Susan B. Schaffnit[3], Jainaba Badjie[4], Emma Hawkins[5,6], Lusako Mwalwanda[5], Shekinah Munthali[5], Rebecca Sear[1], Mary K. Shenk[3], Richard Sosis[7], John H. Shaver[2]; [1]London School of Hygiene and Tropical Medicine (LSHTM), [2]University of Otago, [3]Pennsylvania State University (PSU), [4]MRC Unit The Gambia at LSHTM, [5]Malawi Epidemiology and Intervention Research Unit (MEIRU, [6]University of Edinburgh, [7]University of Connecticut

Preparing for data collection in a new location is no easy feat. Preparing to conduct simultaneous cross-cultural data collection is even harder. We present a coordinated research design process employed by the Evolutionary Demography of Religion project, investigating the links between religion, co-operation, and child outcomes across five settings. Our team of 10 researchers from 4 institutions is running a mixed-methods study in five countries, working with research partners at: International Centre for Diarrhoeal Disease Research (Bangladesh), Medical Research Council (The Gambia), the Birbhum Health and Demographic Surveillance System (India), Malawi Epidemiology and Intervention Research Unit (Malawi), and Pennsylvania State University's Survey Research Center (USA). Our primary challenge in study design was striking a balance between generalisability and specificity to allow comparability across sites, while still capturing variation within each setting. After some trial and error, we settled on a workflow that made use of a range of online tools (Microsoft Teams, OneDrive, Dropbox, ODK, Github, Zoom) to collaborate on study design, developing data collection tools, and data analysis. Our process typically consisted of creating a generalised "master" copy of each instrument needed and adapting this to each study location post-piloting. We used standard operating protocols for sampling, piloting, conducting surveys and focus groups, and using instruments (e.g., tablets, ODK suite, anthropometry kits) to reduce response variation due to differences in survey administration across countries. We reflect on the advantages and disadvantages of our approach and emphasise the key role of our research partners in the research design process.

Email: anushe.hassan@lshtm.ac.uk

Improving the methodology behind homeless death statistics
Daniel Burns, National Records of Scotland

National Records of Scotland released experimental statistics on deaths of people experiencing homelessness for the first time in 2020 and are working towards having the methodology for these statistics accredited by the Office for Statistics Regulation as Official Statistics. To do this, improvements to the methodology need to be made and new ways of collecting and quality assuring the data need to be investigated. Currently, there are five different sources we collate to try and identify known homeless deaths in Scotland, these sources are compared using a capture-recapture process, which models the likely number of homeless deaths not identified by the sources. We will present the latest work to improve this methodology, as well as going into future plans for improvement.

Email: Daniel.Burns@nrscotland.gov.uk

Gendered patterns of global scholarly migration: Evidence on mobile researchers from bibliometric data
Xinyi Zhao[1,2], Ridhi Kashyap[2], Aliakbar Akbaritabar[1], Emilio Zagheni[1]; [1]Max Planck for Demographic Research, [2]University of Oxford

The increasing feminisation of migration and the exceptional rise in the share of high-skilled migrants are two key features of contemporary international migration. The joint analysis of the two issues lacks sufficient attention, however. A better understanding of the gender dimension in academic mobility is important for pursuing a more balanced gender representation in academia and ensuring the sustainability of academic careers for researchers. Using an exhaustive set of over 33 million Scopus publications, we offer a nuanced view of gender inequalities and dynamics in global scholarly mobility, by mapping the disparities in the volume, diversity, and tendency of transnational migration across the globe. We found that globally, females' representation in the group of international mobile researchers grew faster than that of the researcher

population. However, the gender balance is still a long way off for a majority of countries. The migration diversity patterns in academia indicated female researchers were more likely to depart from a narrower array of origin countries towards a broader range of destinations. Over time, the level of origin and destination country spreads have converged. It leads to a more symmetric migration trend among researchers of both genders. Moreover, despite the US remaining an important academic destination that received a considerable share of talents of both genders, female and male researchers respectively became more evenly dispersed across destination countries. It also suggested that the worldwide academic migration was becoming less skewed.

Email: zhao@demogr.mpg.de

11.30pm Wednesday 7 September: Data Science: Innovations in modelling and forecasting

Assessing future recuperation paths in Finland using a novel non-parametric cohort fertility forecasting approach

Julia Hellstrand[1,2], Jessica Nisén[2,3], Mikko Myrskylä[1,2]; [1]Centre for Social Data Science and Population Research Unit, University of Helsinki, [2]Max Planck Institute for Demographic Research, [3]INVEST Research Flagship, University of Turku

Most cohort fertility forecasting approaches systematically underestimate uncertainty. This challenge of too narrow confidence intervals is problematic particularly in times of abrupt trend changes in fertility, such as in Finland in the 2010s. To assess possible recuperation paths for cohorts still at childbearing age without strict modelling assumptions – e.g. restrictions on the smoothness or shape of the fertility schedules – a new non-parametric cohort fertility forecasting approach was developed. The approach was based on existing techniques to forecast period fertility and mortality levels, but modified to fit our purpose of estimating completed cohort fertility. Calculation of this approach relied on fertility histories from the Human Fertility Database (HFD), the largest high-quality fertility dataset to date. For a cohort with observed age-specific fertility rates up to age x, the universe of fertility changes for ages above x was calculated from past data and added to the most recent year's cumulative age-specific rate. We use fertility rates from 1975 onwards to derive a universe of possible recuperation paths for cohorts born between 1975 and 1987, resulting in a total of between 910 and 1,342 possible fertility recuperation schedules for these cohorts. To derive non-parametrically a probabilistic distribution of potential future fertility trajectories, we resample from the universe of fertility recuperation schedules with replacement (10,000 samples) for each cohort. The results showed that in order for future cohort fertility to remain stable in Finland, women at childbearing age would require stronger recuperation than ever seen in the HFD countries.

Email: julia.hellstrand@helsinki.fi

Improving methods for fertility forecasting through the incorporation of parity information
Joanne Ellison, Ann Berrington, Jakub Bijak, Erengul Dodd; Centre for Population Change, University of Southampton

Fertility projections are a key determinant of population projections; they are also vital to anticipate demand for maternity and childcare services. While there is a diverse literature concerning fertility forecasting models, these tend to use aggregate population-level data indexed by age and time (period/cohort) alone. Therefore they neglect to include parity information, despite parity-specific data being collected by many countries and the evidence supporting greatly differing determinants of childbearing by parity. This omission risks ignoring a crucial mechanism of fertility dynamics and consequently producing biased predictions. To this end, in this paper we develop a Bayesian fertility projection model that incorporates parity information and combines multiple data sources, applying our methods to the constituent countries of the UK.   Preliminary work has focused on the use of Bayesian Generalized Additive Models to model parity-specific rate estimates for England and Wales from the Office for National Statistics (ONS). To reflect our prior knowledge that summary measures such as completed family size and parity progression ratios change slowly over time, we use time

series priors to provide suitable constraints, finding that they can increase forecast precision and accuracy. We develop this model by integrating rate estimates from the ONS Longitudinal Study, which links census and vital events data. Comparing our forecasts with the best-performing models in the literature, we quantify the impact of including parity information on predictive accuracy. If the impact is positive, the resulting projections could aid government policymakers and planners in their decision-making and enable more tailored, parity-specific, policy solutions.

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

Quantifying spatial inequalities in cause-specific mortality: A Bayesian hierarchical spatial model estimating cause-specific SMRs for Belgian municipalities in 2009-2013

Benjamin-Samuel Schlüter[1,] Martina Otavova[1], Bruno Masquelier[1], Brecht Devleesschauwer[2]; [1]Catholic University of Louvain (DEMO), [2]Sciensano

Spatial disparities are an important dimension of inequality in mortality. Within-country variations in mortality are indicators of national development, societal fairness, social justice and represent a major challenge for public health sectors worldwide. Previous studies showed that there are substantial differences between European regions in their all-cause and cause-specific mortality rates. Studying cause-specific mortality at a subnational level comes with an additional difficulty: the presence of small populations in which the stochastic variation in death counts is relatively high, making the underlying mortality levels unclear. Here we propose to model cause-specific standardized mortality ratios (SMRs) at the Belgian municipality level with a Bayesian hierarchical spatial model. The SMR is useful to study spatial disparities in cause-specific mortality as it is a relative measure and controls for the population age structure. The model smooths noisy SMRs estimates by pooling information from neighbouring municipalities. It further allows to evaluate the role of socioeconomic factors. In this study, we look at the association between the Belgian Index of Multiple Deprivation 2011 and broad causes of death considered at the municipality level in the 2009-2013 period.   The magnitude of spatial disparities greatly differs according to the cause of death considered. In addition, municipalities experiencing the highest mortality level differ across causes calling for cause-specific analyses.   Bayesian hierarchical spatial models open new perspectives in the study of spatial inequality of cause-specific mortality by allowing spatial smoothing in addition to the inclusion of socioeconomic variables.

Email: benjamin-samuel.schluter@uclouvain.be

An Age-Period-Cohort model in a Dirichlet framework: A coherent causes of death estimation.

Andrea Nigri[1] and Marco Bonetti[1,2]; [1]Bocconi University, [2]Carlo F. Dondena Research Center

Demographic studies often require the modelling of multiple outcomes. Although many tasks can be accomplished by modelling the various outcomes separately, some others need the different outcomes to be treated jointly in a unified framework. As a result, the use of models with multivariate and correlated responses has increased.   Here we shall focus on the case of compositional data (non-negative proportions with unit-sum), which are relevant in particular in longevity analysis, where the overall mortality can be described as the composition of several causes of death (CoDs). The compositional framework occurs whenever one needs to partition a measurement into percentage contributions from its components, hence the unit-sum constraint. In order to exploit traditional multivariate statistical models, Aitchison introduced the logratio analysis to model compositional data, by inserting the unit-sum constraint. Aitchison's approach has been criticized since it seems to give excessive importance to components that have little overall relevance for a meaningful understanding of the composition. For compositional data, Dirichlet regression models represent a solid alternative. By using appropriate link functions, Dirichlet regression provides a GLM-like framework that relates compositional data to other relevant variables of interest. Note that the Beta regression model is a special, and effectively univariate case of the former, with only two categories. Along these lines, we propose an Age-Period-Cohort (A-P-C) model within the Dirichlet Generalized Additive Model framework, with specific interest in its use for modelling longevity with multiple causes of death.

Email: andrea.nigri@unibocconi.it