Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem

Ridhi Kashyap

Published online: 13 Dec 2021.

Submit your article to this journal ⍐

View related articles ⍐

View Crossmark data ⍐

Citing articles: 1 View citing articles ⍐

Routledge
Taylor & Francis Group

Check for updates

Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem

Ridhi Kashyap 🄓
University of Oxford

Over the past 25 years, technological improvements that have made the collection, transmission, storage, and analysis of data significantly easier and more cost efficient have ushered in what has been described as the 'big data' era or the 'data revolution'. In the social sciences context, the data revolution has often been characterized in terms of increased volume and variety of data, and much excitement has focused on the growing opportunity to repurpose data that are the by-products of the digitalization of social life for research. However, many features of the data revolution are not new for demographers, who have long used large-scale population data and been accustomed to repurposing imperfect data not originally collected for research. Nevertheless, I argue that demography, too, has been affected by the data revolution, and the data ecosystem for demographic research has been significantly enriched. These developments have occurred across two dimensions. The first involves the augmented granularity, variety, and opportunities for linkage that have bolstered the capabilities of 'old' big population data sources, such as censuses, administrative data, and surveys. The second involves the growing interest in and use of 'new' big data sources, such as 'digital traces' generated through internet and mobile phone use, and related to this, the emergence of 'digital demography'. These developments have enabled new opportunities and offer much promise moving forward, but they also raise important ethical, technical, and conceptual challenges for the field.

Introduction

Writing on the 30th anniversary of the Population Association of America's flagship journal *Demography,* Eileen Crimmins remarked that 'the world of information processing in 1994 differs dramatically from that of 1964' (Crimmins 1993, p. 579). This largely exogenous factor, she claimed, had shaped 'changes in demographic analysis over the past 30 years' by affecting both the kinds of data that were available to demographers and the modes of analysis they used (Crimmins 1993, p. 579). At the time of Crimmins's piece, desktop computers—which had decentralized information storage, data processing, and analysis for demographers away from mainframe computing—had become widespread, but the (commercial) internet was still a nascent technology and about 3 per cent of the global capacity to store information worldwide was digital (Hilbert and López 2011). The 1990s, however, marked the onset of the digital revolution, which saw radical transformations in information storage, transmission, and computational power: by 2007, the world's information storage capacity was over 15 times greater than in the early 1990s, with 97 per cent of information storage in digital form. The most staggering increases were in computational power: compared with 1993, an average computer in 2007 was over 1,200 times faster (Hilbert and López 2011).

The digital revolution and its accompanying improvements in technological capacities ushered in what has been called the era of 'big data' or in other formulations, a 'data revolution'. Definitions of big data are often ambiguous, and the terms big data and data revolution had both been used prior to the late 1990s (e.g. looking at the relative

prevalence of specific words or combinations of words in English language books since 1800 with Google Ngrams, there are two peaks for the terms big data and data revolution, the first in the 1980s and the second in the 2000s (see Google Ngram viewer 2021)). However, numerous commentaries and papers in the new millennium heralded these developments as the beginning of a new era for the social sciences (e.g. Laney 2001; Lazer et al. 2009; OECD 2013; Sagiroglu and Sinanc 2013; Einav and Levin 2014; Kitchin 2014; Connelly et al. 2016; Billari and Zagheni 2017; Alburez-Gutierrez et al. 2019). While different commentaries emphasize different features of the data revolution, a useful formulation is provided in the United Nations (UN) report *A World that Counts*, written in the context of setting the post-2015 international development agenda; this defines the data revolution as 'an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the "internet of things", and from other sources, such as qualitative data, citizen-generated data and perceptions data' (IAEG 2014, p. 6).

This description highlights how data in this era are more widely available and timelier, with greater variety in terms of their producers but also the kinds of information they provide. The report envisions tremendous potential in the ability to advance our understanding of *all* populations ('no one should be invisible', p. 22) and to use this knowledge to inform policymaking in the service of global sustainable development.

At the outset it would seem that demographers—who have always been quite data savvy and whose professional outputs have long been shaped by available information processing and analysis technology—would have much to benefit from the data revolution. But how has the data revolution been a *new* revolution for demographers, in the way that has been heralded for other social sciences? This is a question worth asking with at least some degree of critical reflection, because many of the features used to define the data revolution were arguably not novel for demographers, even in the 1990s. On one hand, if we think of the data revolution in terms of the proliferation of big data—as defined in terms of their volume or scale—demographers, in their quest for population measurement, have been using big data based on full enumeration (e.g. censuses or

complete-coverage vital registers) for a long time, even for centuries in some European countries. Even demographic sample surveys have tended to have large samples, given the need for population generalizability. On the other hand, we may think of the data revolution in terms of the increase in volume and variety of 'found' (Connelly et al. 2016) or 'ready-made' data (Salganik 2019), such as those emerging as by-products of government, administrative, or digital transactions, which increasingly social scientists seeking a unifying definition of big data have tended towards. From this perspective the idea of repurposing data originally meant for something other than research is not new to demographers, who have long been avid consumers of administrative data, such as population registers or, before the state became the monopolist of record-keeping, parish registers. Through the use of these data and other types of often-imperfect data, demographers had already by the middle of the twentieth century developed skilful techniques for repurposing data, whether in the tradition of indirect estimation following on from the work of William Brass (Brass 1996) or in the meticulous linking of parish registers by Henry in France (Henry 1967) or Wrigley, Schofield, and the Cambridge Group for the History of Population Structure in England (Wrigley and Schofield 1983). These examples may lead us to think that demographers were using big data before it became fashionable, and the data revolution in demography perhaps happened well before it did in the other social sciences.

So—*did* the data revolution happen in demography before it happened elsewhere? Or looking at the pages of *Population Studies* and other demographic journals, do we see signs of a 'new' data revolution in the field in the past 25 years? This paper attempts to address this question by reflecting on how the data ecosystem for demographic research has developed and how these developments have affected what demographers are able to measure, describe, and explain. My discussion is anchored around two dimensions of these developments. First, I outline changes in the kinds of data collected in and available from long-established ('old') sources that have conventionally informed demographic research: censuses, population registers, surveys, and civil registration systems. Next, I describe how demographers are beginning to use 'new' big data sources—enabled through the use of the internet and mobile phones and the growing digitalization of social life—and the emergence of the area of 'digital demography'. I conclude by discussing the

Data paradigms in demography

Before reflecting on how the landscape of demographic data has changed since the 1990s, it is helpful to understand where the field was in the mid-1990s, on the occasion of the 50th anniversary edition of *Population Studies*, and the changes that had occurred in the preceding decades. Demography has experienced key shifts in data paradigms over time (Billari and Zagheni 2017). These shifts are visible on the pages of *Population Studies* through the years. Prior to the mid-1970s, most papers in the journal relied on aggregate-level data, generally drawing on data sources such as censuses or other administrative records, and were focused on measuring and describing population-level processes and parameters. Often data (e.g. from censuses) were only available to demographers in tabulated form, and even if individual-level data were available, the data were deployed to characterize population-level patterns and, in many cases, examined with careful scrutiny to check their plausibility and quality. While description is often derided in other social sciences, a lot of the work published in the journal was precisely that—good description of population-level indicators and processes—although, as Samuel Preston noted in his contribution reviewing changes in mortality research in the journal's 50th anniversary edition, the distinction 'between description and analysis was in any event imprecise' (Preston 1996, p. 529). Where data permitted, differentials or heterogeneity were studied, but opportunities for multivariate analyses were limited because of the aggregated nature of available data.

By the last quarter of the twentieth century, however, a second type of data paradigm—one that relied on individual-level data from sample surveys—had become well established, having benefited from the development in statistical sampling techniques in the post-war period. The World Fertility Survey (WFS), the predecessor to the ongoing and widely used Demographic and Health Survey (DHS), was conducted across 62 countries between 1974 and 1986 and was a first-of-its-kind attempt at fielding a comparable, cross-national survey to study fertility and its determinants (Cleland 1996). With the availability of surveys such as the WFS in the 1970s and 1980s, individual-level (statistical) analyses of survey data, including event history analyses of retrospective birth or life histories within them, became increasingly common. This shift in the demographic data paradigm was duly noted in a number of contributions in the 50th anniversary edition (e.g. Cleland 1996; Coale and Trussell 1996; Preston 1996).

Although the greater availability of sample surveys, especially for low-income countries where data were otherwise lacking or defective, was seen as a favourable development for demographic research, I get a sense reading the 1996 issue that the authors also viewed this development of the micro-level paradigm based on surveys within demography with some degree of caution. Cleland, in his paper on demographic data collection, claimed that while the WFS/DHS model had made 'immense contributions to our understanding of the processes of family formation, proximate determinants of fertility and socio-economic correlates of child mortality and childbearing', he wondered if something less expensive—'a simple household survey with larger samples'—might have better served the provision of demographic measures (Cleland 1996, p. 445). While this was a question about data, ultimately it was also a question about the aims of demographic research or what they ought to be—measurement and description at the macro level, the pursuit of explanation at the individual or micro level, or something else?—a debate that has continued within the field (Billari 2015). Preston (1996, p. 535) envisioned an imminent decline in 'the type of individual-level analyses of childhood mortality, so prominent in the 1980s', as most of the relationships had been uncovered and found to be 'remarkably consistent across time and space'. He saw diminishing returns to the standard regression-type statistical approaches unless some other research designs were applied. Coale and Trussell (1996), in their contribution on demographic models, wondered if the shift to micro-level analyses with survey data meant that issues of data quality and lessons learned from the use of macro-models for checking the validity and consistency of data were being neglected.

Fast forward 25 years and the reliance on individual-level survey data analysis within the demographic data ecosystem has definitely persisted but, as I outline in what follows, surveys have also changed in significant ways. The incorporation of different types of data (e.g. biosocial data) within surveys, along with the ability to link surveys with contextual data sources (e.g. geospatial and

administrative data), have generated opportunities to address research questions with novel research designs. The increasing complexity of surveys and growing interest in new types of digital data sources have meant that issues such as selection, representation, and bias have yet again come to the forefront, and linked to this, questions of data quality, validity, and consistency are of renewed significance.

'Old' population data, new features

Censuses and large-scale administrative microdata

Starting in 2000, an exponential increase occurred in the public availability of individual-level records of 'old' big data sources, such as censuses, administrative registers, and large-scale surveys. Ruggles has described this in terms of an explosion in the availability of 'big microdata for population research' (Ruggles 2014). While about 100 million individual-level records were publicly accessible to the research community in 2000, by 2018 this number was estimated to exceed 2 billion, covering over 100 countries (Ruggles 2014). The availability of large-scale, even complete-enumeration, census samples has clearly benefited from the improvements in information storage, data processing, and extraction provided by the digital revolution. These technological improvements facilitated the development of several key data infrastructure projects that emerged in the late-1990s, including the Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota, the North Atlantic Population Project (NAPP) for historical eighteenth- to early-twentieth-century European and North American census samples (Ruggles et al. 2011), and the Integrated Census Microdata (I-CeM) project for historical Britain (Schürer and Higgs 2014).

The idea of census records being released at the individual level, rather than in aggregated or tabulated form, was itself not new in the 1990s. From 1962 the United States (US) Census Bureau provided a 1-in-1,000 sample of long-form records from the 1960 Census to researchers, on computer tapes, and as costs of information storage and processing fell in the 1960s and 1970s, the scale of census microdata from the US Census Bureau continued to increase. In 1974 Statistics Canada made public use microdata files available for the Canadian 1971 Census, and a sample of anonymized

records from the UK's 1991 Census was made publicly available in 1993. Other statistical agencies in a handful of countries also made internal microdata available to academic researchers by special arrangement (Ruggles 2014). Before 2000, however, statistical agencies in most countries 'had no systematic program for preservation or reuse of census microdata', and as a result, 'most machine-readable census data from the 1960s and 1970s had already disappeared by the 1990s' (Ruggles 2014, p. 289).

In light of this, it is a remarkable achievement that over the past two decades IPUMS has built a public web-accessible repository that now provides the research community access to anonymized census and/or large-scale survey samples for nearly 100 countries, many of which are low- and middle-income countries (LMICs) where data availability was previously much more limited. Through the IPUMS data storage, processing, and extraction systems, data are interoperable and harmonized across time and space, and data sets can be downloaded swiftly from the internet. The availability of census samples has fuelled an increase in publications for parts of the world (e.g. Latin America, Africa) where previously the DHS was the only available demographic data source. Another category of microdata that has experienced tremendous growth is historical census microdata, which have become available through collaborations between researchers and internet-based genealogical organizations such as ancestry.com or findmypast.com. As historical samples face few confidentiality restrictions—often including information on names and dates of birth—attempts at linking individuals across census samples have also been made. In both historical and contemporary samples, individuals can be placed and related to others within their co-residential household context.

The proliferation of individual-level, high-density census samples has facilitated at least three types of opportunities for demographic research, examples of each of which are now visible in journals, including *Population Studies*.

First, the large, often complete, coverage of census samples has been helpful in analysing subpopulations with greater precision and has provided researchers with the flexibility to define and explore heterogeneity on their own terms for specific subgroups, thereby capturing previously understudied differentials that could not be analysed with aggregated group-level data or using survey-based samples. For example, studies using census data have allowed for better statistical visibility and

analyses of non-traditional family forms, such as same-sex couples (e.g. Festy 2007; Rosenfeld 2010) and non-cohabiting marriages (e.g. Ferrari and Macmillan 2019). The wider availability of individual-level samples for non-Western countries and historical populations has enabled multivariate analyses for populations where such analyses were much less widespread. However, the greater availability of census samples has also led to the return and more detailed applications of indirect estimation methods originally developed for macro-level measurement, such as the own-children method (OCM) for estimating fertility. For example, Reid et al. used historical census data and OCM to estimate age-specific fertility and analyse differentials in age-specific fertility by place of residence and social class in the 1911 Census in England and Wales (Reid et al. 2020). Guilmoto used microdata from census samples to apply OCM to analyse gender bias in reproductive behaviour by reconstructing family relationships (e.g. sibling order) and sex from household information provided in census records (Guilmoto 2017). Confidence intervals around indicators commonly used to study son preference in reproductive behaviours (e.g. conditional sex ratios) tend to be large due to random variation arising from small sample sizes in surveys; this complicates the meaningful detection and interpretation of gender bias. Larger census samples offer ways to overcome these limitations.

Second, census data have enabled the incorporation of contextual, or more macro-level, characteristics—either at the household level, such as through information on co-residential members, or at different geographical levels—to understand and explain variability in demographic outcomes. They have also provided ample opportunities for the study of living arrangements and household contexts, thus becoming a primary resource for family demographers. They have helped to explore interactions between individual and geographical characteristics for understanding variations, for example in fertility over the course of historical fertility transitions through the use of multilevel models (e.g. Dribe et al. 2014; Klüsener et al. 2019). In addition, they have facilitated more geographically precise estimates of indicators linked to fertility (e.g. Schmertmann et al. 2013), internal migration (e.g. Rodríguez-Vignoli and Rowe 2018), and human development (e.g. Permanyer 2013) at subnational levels, thus helping to shift the conventionally national focus of demographic analysis. In general, the use of data with lower levels of aggregation that can be aggregated to higher levels allows for

the adoption of a multilevel paradigm for demographic analysis and provides an opportunity to integrate micro- and macro-level traditions (Courgeau et al. 2007). Inferences at one level of aggregation may not generalize to other levels—but to be able to understand and explain these differences, we first need to be able to detect them.

Third, and perhaps the most striking feature of the big population microdata made available through data infrastructures such as IPUMS is that they are harmonized and comparable across time and space, and significantly easier to link with other spatio-temporally referenced data. In the case of some historical census samples, linking individuals and their children across censuses has also been carried out (largely for fathers, however, as linkage typically involves names, and women tend to change name on marriage); this has led to the emergence of multigenerational research (Ruggles et al. 2018). The growing accumulation of cross-national, harmonized, and standardized individual-level data is a development that is not restricted to census microdata, but also applies to widely used survey data sets, such as the DHS and several health and ageing studies. Comparative cross-national research and studies covering changes over decades of time have flourished across demographic journals as a result of these developments. The accumulation of multiple samples across time and space has led researchers in some cases to drop the time dimension altogether and explore, for example, how demographic patterns linked to internal migration (e.g. Bell, Charles-Edwards, Ueffing et al. 2015) or family forms vary as a function of generalized macro-level social changes, such as economic and human development (e.g. Ferrari and Macmillan 2019; Pesando 2019) or educational expansion (e.g. Esteve et al. 2016). Demographers have long been interested in characterizing population change over time and space in the pursuit of empirical regularities in population dynamics and have been drawn to ideas of convergence across countries or regions in demographic variables as encapsulated by the idea of the demographic transition. With the accumulation of these data sources, investigations of these theories of demographic convergence are now empirically informed by data with significant global coverage. Empirical projects that have leveraged the accumulation of comparative cross-national data sets have often found that significant diversity remains unexplained by either economic or social variables, that convergence in demographic indicators is far from

straightforward, and that heterogeneity, often between regions, is persistent (e.g. Esteve et al. 2012; Moultrie et al. 2012; Pesando 2019).

Although the harmonization of 'old' big data has been an exciting development and has injected new resources to sustain demographers' ambitions to describe empirical regularities in populations, the value of this harmonization has been questioned in some cases. The ability to harmonize across census or survey samples has emerged from a sustained push towards the standardization of census questionnaires within the international statistical system, as well as a wider push for cross-national comparability within the global sustainable development agenda. However, whether this standardization yields measures and comparisons that are socially meaningful and reflect lived experiences is not always clear-cut. This point is illustrated in debates surrounding the utility and meaning of the concept of the household as used in censuses and cross-national standardized surveys, especially in the context of sub-Saharan Africa, where family living arrangements and union statuses are much more diverse and nuanced than those captured in these instruments (Randall et al. 2011; Randall and Coast 2015; Hertrich et al. 2020). In contrast to these national population-wide data, longitudinal data collection endeavours in smaller geographical areas—as exemplified by the Health and Demographic Surveillance System (HDSS) sites spanning several LMICs within the International Network for the Demographic Evaluation of Populations and their Health (INDEPTH) (Sankoh and Byass 2012)—have more successfully incorporated culturally specific measures of household and family context while also capturing changes in households over time, for example due to migration (e.g. Townsend et al. 2002; Hosegood and Timæus 2006). However, the intensive resources and follow-up needed to run HDSS sites imply that these are generally limited to localized communities, and their wider generalizability to surrounding areas remains unclear.

The discussion so far has focused on the greater availability and harmonization of census microdata samples in the public domain. However, a separate question is whether the traditional full-enumeration decennial census model will continue as a method of administrative data collection and, relatedly, remain relevant as a data source for demographic research moving forward. Census-taking is a resource-intensive and logistically complex yet also political activity. Concerns about escalating costs, declining response rates, and privacy issues, as well as the need for more timely data and the excitement surrounding 'new' big data sources (such as mobile phones, satellite data, or web data) have regularly sparked political contestations and scientific discussions around the utility and sustainability of traditional censuses in recent decades, leading some demographers to speculate an imminent 'twilight' of the census (e.g. Coleman 2013). Nevertheless, based on censuses undertaken in the 2010 census round (censuses covering 2005–14), the traditional census model of full-field enumeration has so far persisted. Although several countries have postponed their census dates in response to Covid-19 (UNECE 2021), this model seems likely to remain the dominant one for the 2020 census round (covering 2015–24). While the 2000 census round saw 28 countries unable to undertake any census due to political instability or lack of human resources, this number declined to 11 in the 2010 census round, in which 227 out of 241 countries (94 per cent) undertook a census of some form (Kukutai et al. 2015).

This expansion of censuses across the world in the 2010 round has nonetheless occurred alongside a diversification of census methods towards the combination of full enumerations with administrative registers or sample surveys, or the replacement of census-taking altogether with population registers, as accomplished by the Nordic countries. Kukutai et al. (2015) classified 39 countries out of the 227 that undertook a population census in the 2010 round as having conducted one using an alternative method. This experimentation with alternative modes of population enumeration is most advanced in Europe, where 'a history of maintaining population registers and a broader public acceptance of personal data for statistical purposes' has made this possible (Kukutai et al. 2015, p. 5). This diversification highlights a stark inequality between high- and low-income countries in the tools available for population enumeration. In contexts such as the Nordic countries, the replacement of censuses with population registers has enabled the production of high-quality, regularly updated population data; however, in low-income countries (particularly in sub-Saharan Africa), accurate and complete censuses are crucially needed but 'likely to be least well-resourced or well-equipped' (Moultrie 2016, p. 259). As Moultrie noted, in addition to significant lags between censuses in sub-Saharan Africa, the estimated completeness of census enumeration (a key indicator of data quality) remains unknown, as few countries conduct post-enumeration surveys and, for those that do, results are not reported

transparently. This unreliability has ramifications for nationally representative surveys such as the DHS, for which the census serves as the sampling frame. Further, even as the core of census collection is affected by data quality concerns due to burdens facing national statistical offices (NSOs), paradoxically the demands on censuses to collect further information (e.g. on mortality) have only increased, due to continued deficiencies in vital registration systems.

To some extent, the opportunities presented by new technologies offer considerable possibilities to improve cost efficiency for census operations, although their adoption remains uneven and affected by available resources and technical capacity. The most significant development in the use of technology in recent census rounds has been the use of cartographic and geospatial tools (GPS, GIS, satellite imagery, and aerial photography) for planning and improved quality of mapping capacity (United Nations Statistics Division 2013). While this mapping has helped census planning operations, the use of remotely sensed data from satellite images in combination with administrative-unit based census data has also enabled the production of spatially disaggregated or gridded population estimates as outputs which more accurately represent the distribution of populations in space (Stevens et al. 2015; Wardrop et al. 2018). Ultimately the success of these 'top-down' techniques for producing high-resolution population estimates depends significantly on the quality of the census inputs and the overlap of census counts with mapped administrative units, which may be outdated in many low-income countries. In such settings, where full-enumeration census counts may be outdated or unavailable, 'bottom-up' approaches have been developed that use geospatial covariates extracted from satellite data in combination with census-like, population survey data available for a sample of areas to predict high-resolution population counts (Lloyd et al. 2017; Wardrop et al. 2018; Leasure et al. 2020).

In the context of enumeration, technological innovations adopted in the 2010 round included the use of handheld devices and internet-based questionnaires, as well as the use of mobile phones for monitoring of field operations. The use of handheld and mobile devices with inbuilt georeferencing capabilities and the incorporation of auxiliary information such as time and date provided opportunities for improved data quality and validation checks. Optical data capture and web-based data dissemination were among other types of technological innovations adopted by NSOs. Although face-to-face

interviews using paper questionnaires remained the most common mode of enumeration in censuses, internet-based enumeration became the second most common mode, although this trend varied strongly by region (United Nations Statistics Division 2013). While 44 per cent of countries in Europe and one-third of those in Asia offered internet-based enumeration, this option was not offered in South America or Africa. However, no country has so far used the internet as the sole mode of enumeration. The use of internet-based enumeration seems likely to be higher in the 2020 census round, and while this will have positive implications for the efficiency and timeliness of data collection, the use of multiple parallel modes of data collection also raises challenges. These include issues linked to the development of separate strategies, planning methods, and different skills and expertise for each mode, as well as data comparability and quality issues linked to mode effects. Although the intensification of the use of technological innovations in census-taking holds great promise, 'such use also requires additional efforts to ensure that the planning, development, testing and implementation of these different applications is successfully achieved' (United Nations Statistics Division 2020, p. 10).

While censuses, albeit in incrementally diversifying forms, remain the dominant mode of population enumeration in most of the world, population registers in Nordic countries and in other European settings, such as the Netherlands (e.g. Bakker et al. 2014), highlight the significant value of government administrative data. Through personal identification numbers unique to individuals, these longitudinal population registers allow for deterministic linkage of multiple administrative registers, thereby enabling analysis of multiple domains of the life course or linkage of individuals across generations. These data sources remain, in some sense, the gold standard for administrative data and one for other countries to aspire to, as the availability and use of individual-level administrative data continues to expand. The Nordic registers highlight the tremendous opportunities afforded by complete-population administrative registers, and they have facilitated the use of novel research designs including those involving multiple generations (e.g. Kolk 2014) or extended kin (e.g. Barclay et al. 2020), and the creation and linkage of unique kinds of contextual variables linked to neighbourhoods or workplaces (e.g. Lyngstad 2011; Holmlund et al. 2013). Even these gold-standard data, however, are not without their flaws, and over-coverage—arising from migration dynamics in which individuals who are not resident

remain registered—is an issue that can introduce bias (Monti et al. 2019).

Ultimately, administrative registers—while 'big' and 'deep', to borrow Ruggles' characterization of big population data, in terms of their coverage of people, events, or transactions—are shallow at capturing motivations, intentions, attitudes, and other subjective characteristics, something that surveys do better. A promising new direction to reinforce the complementarities between these two types of data and enhance the scope of administrative registers is through their linkage with survey data sets used in demographic research, especially in Europe: for example in the Netherlands (e.g. Bakker et al. 2014), in the UK, with the linkage of the Millennium Cohort Study (Tate et al. 2006) and the British Household Panel Study (Sala et al. 2012), and in the Nordic countries, through linkage with the Gender and Generations Survey (Gauthier et al. 2018). As this linkage requires consent from survey participants, non-consent is likely to be an important source of bias in this approach. This selection bias, however, is likely to be negligible in contexts with a longer history of open, research-accessible administrative data (e.g. the Nordic countries) and more salient in other contexts where these developments are still new and public acceptance weaker (e.g. the UK) (Sala et al. 2012; Mostafa 2016).

Surveys

By the 1990s sample surveys had become the most widely used method of demographic data collection in high-income countries as well as LMICs, especially in the context of fertility and family research. In Europe and North America, surveys had also moved to longitudinal data collection in the 1970s, with the use of prospective cohort studies and other instruments that shifted the focus of analyses from demographic status at one time point to changes over time through the collection of repeated measurements for the same individuals (Crimmins 1993). The kinds of surveys used by demographers have generally been broad, often multipurpose, high-quality probability samples with carefully designed sampling frames that are nationally representative, with detailed questionnaires and complex designs. This type of data collection is highly resource intensive, as in addition to the efforts needed for appropriate sampling and planning, the recruitment of respondents for such surveys requires repeated callbacks and refusal

conversion to ensure satisfactory response rates (Groves 2006). Standardized, nationally representative surveys widely used by demographers to study LMICs (e.g. the DHS) have also served multiple purposes, not just in terms of the topics that they cover but also in their use for both measurement and explanation of demographic outcomes. The DHS was already being used in the 1990s for estimating child mortality and fertility from retrospective birth histories, as well as to analyse theory-driven determinants of variation in these outcomes. The use of surveys for the measurement of demographic indicators was the consequence of weak or absent civil registration systems or inadequate census-based measures, along with the realization that indirect methods of demographic estimation applied to individual-level survey samples yielded reasonably good results (Cleland 1996).

Perusing the pages of *Population Studies*, it is evident that the use of sample surveys (both cross-sectional and longitudinal studies) has remained widespread in demographic research over the past 25 years. Nevertheless, concerns about declining response rates that are often non-random (especially in high-income countries), increasing costs associated with surveys due to additional recruitment efforts, worsening attrition in panel surveys, and lags in both data generation and reporting have raised challenges for the survey model of social research (Tolonen et al. 2006; Groves 2011; Mostafa and Wiggins 2015; Gauthier et al. 2018). Some of these discussions are quite similar to those described in the context of the relevance of censuses. Response rates in surveys in LMICs (e.g. the widely used DHS) remain high, but significant costs—and continued external technical assistance from the global North to the global South through aid agencies such as USAID—are needed to sustain them (Corsi et al. 2012; Short Fabic et al. 2012). Even in LMICs, urban residents may be less willing to respond, and panel attrition is often high and non-random, as seen in the longitudinal Study on Global Ageing and Adult Health (SAGE) (Kowal et al. 2012) and other household surveys (e.g. Alderman et al. 2001), although the impacts of attrition bias for coefficients in multivariate analyses are generally minimal (Alderman et al. 2001). Furthermore, longitudinal studies in LMICs, such as those collected by HDSS sites in the INDEPTH network, face significant challenges linked to the recruitment and retention of skilled personnel for data collection and management (Sankoh and Byass 2012).

Even as respondents have become less willing to respond, researchers are asking for and collecting

more types of information from them. Since the 1990s, surveys designed and used by demographers have increased in scope and become more varied in the kinds of data they collect, for example through the integration of biological or psychological measures at the individual level, richer contextual features such as geographical measures, or information from other household members or peers. The growing detail and complexity of survey instruments has meant that the coordination of these studies requires greater collaboration and cooperation in large interdisciplinary teams. The inclusion of new kinds of data in surveys, extending beyond self-reported indicators—for example geographical coordinates or biospecimens (e.g. blood, saliva samples, or genetic data)—has also made procedures linked to seeking informed consent and protecting confidentiality of participants more demanding.

Changes occurring to the DHS exemplify many of the trends affecting surveys used in demographic research more broadly. Since the mid-1990s, the DHS programme has added new questions into the standard and household questionnaires, introduced new modules on behaviour (e.g. gender-based violence, women's status, alcohol consumption), and integrated biomarkers covering a range of different health conditions (e.g. STIs, malaria, measles, and chronic conditions such as diabetes) across successive phases, moving beyond the collection of anthropometrics alone. The number of countries with sibling histories in the DHS for indirect measurement of maternal mortality and adult mortality has also grown, although concerns about data quality and mortality underestimation with these data remain (Bicego 1997; Timæus and Jasseh 2004; Masquelier 2013). In the first phase of the DHS, the household questionnaire consisted of 25 questions, and by the seventh phase it included 131 questions, after a peak of 226 in phase five. The increased length and complexity in the questionnaire has raised recurring discussions and concerns about deteriorating data quality (Pullum et al. 2013). Nevertheless, trends from more recent phases of data collection indicate that the general quality of the data in DHS surveys remains high, albeit with higher levels of missingness in some variables (e.g. use of antenatal healthcare) (Pullum 2019). While the quality of age and date information varies across regions, and issues of backward displacement of births or omission of recent births remain, these are relatively small and do not show systematic trends over time (Pullum and Becker 2014; Pullum and Staveteig 2017). Moving forward and looking

more broadly across different surveys, it is quite possible that questions of data quality may become more significant and maintenance of quality standards may require more intensive efforts in light of the continued complexity of survey questionnaires.

Reflecting these changes in DHS questionnaires, the topics of studies using the DHS in demographic journals have also diversified. Of 114 papers using the DHS published in four demographic journals (*Population Studies*, *Demography*, *Population and Development Review*, and *Demographic Research*) between 1998 and mid-2020, 62 (54.4 per cent) were in the area of fertility, with child health the next most common category. Gender (e.g. Kishor and Johnson 2006), reproductive health (e.g. Magadi et al. 2003), and HIV and sexual behaviours (e.g. Bongaarts 2007) have also emerged as significant themes of study. In 1996 the DHS began collecting GPS coordinates of cluster locations (primary sampling units) and in 2003, georeferenced data sets became available, which made these coordinates available to researchers (with random coordinate displacement added to the data sets to protect respondent privacy). This enabled the layering of contextual or geographical features to the individual characteristics collected in the surveys, for example those linked to climate and the environment, proximity to health infrastructure and institutions, and other aspects of the built environment (e.g. Hathi et al. 2017; Østby et al. 2018; Andriano and Behrman 2020; Grace et al. 2021). Some of this geospatial augmentation of survey data has led to the integration of surveys with new types of remote sensing data, such as the use of night-time lights observed via satellites (e.g. Dorélien et al. 2013; Rotondi et al. 2020).

The accumulation of large bodies of individual-level data with the potential for augmentation through different types of linkage has also bolstered the application of individual-level causal analysis approaches in demography, drawing from the potential outcomes framework (Neyman–Rubin–Holland model) that has seen widespread adoption in economics. Although the applicability and relevance of this model of causation—which relies extensively on ideals of experimentation for the interpretation of causal effects and is grounded at the individual level—has been challenged by demographers (e.g. Bhrolcháin and Dyson 2007), the pursuit of quasi-experimental or natural experiment approaches for exploiting exogenous variation has clearly witnessed an increase, on looking across demographic journals (e.g. Torche 2011; Andriano and Monden 2019; Polos and Fletcher 2019). The availability of data has

played an important part in this rise, as the detection of natural experiments or identification of specific subpopulations exposed to a particular type of treatment or intervention (e.g. school reform, health policy, economic shock) in otherwise broadly multipurpose data sets requires enough data to be able to implement these research designs. Research designs paying closer attention to issues of selection, such as within-family designs, have increased, with the advent of survey data sets where multiple observations are available and individuals can be linked to other family or household members (e.g. siblings) (e.g. Barclay and Myrskylä 2016; Rana et al. 2021). Similarly, the growing availability of prospective longitudinal studies, especially outside Europe and North America, has allowed for designs to enable analysis of changes in individual-level outcomes over time and for time-invariant fixed characteristics to be controlled while examining the effects of time-varying exogenous factors (e.g. Frankenberg et al. 2005; Song and Burgard 2008; Saha and van Soest 2011). Longitudinal data collections in LMICs— such as the 49 INDEPTH HDSS sites across 19 countries and also collections in China (e.g. China Family Panel Studies, China Health and Retirement Longitudinal Study), Indonesia (e.g. Indonesia Family Life Survey), and India (e.g. India Human Development Survey)—have provided new opportunities for understanding health, ageing, and family processes in settings where data have otherwise been sparse or limited to cross-sectional surveys.

The growing variety of data collected in surveys has also paved the way for new opportunities in interdisciplinary research, for example in the realm of biodemography. These developments are perhaps most prominently illustrated by research on health and ageing, where the integration of biomarkers linked to metabolic, cardiovascular, immune, and physical function in several population surveys has provided a deeper understanding of the physiological mechanisms and pathways by which social and demographic factors 'get under the skin' to affect health (Crimmins et al. 2010). Furthermore, these measures have provided an empirical basis for measuring and understanding biological risk and frailty, a key component of formal demographic models of mortality (Vaupel et al. 1979), and have helped to model the 'morbidity process' leading to mortality (e.g. Turra et al. 2005; Crimmins et al. 2010). A significant development for further strengthening the interface between demography and biology, with the potential to extend far beyond health research, has been the inclusion of

genome-wide genetic data in large-scale longitudinal population surveys, such as the US Health and Retirement Study (Hauser and Weir 2010) and British cohort studies (O'Neill et al. 2019). In contrast to earlier studies analysing genetic influences on demographic behaviours (which relied on specialized samples, such as twins), genome-wide molecular genetic data offer the potential to examine heritability of traits in larger samples of unrelated individuals and to incorporate genetic measures (e.g. polygenic scores) into research designs to explore fundamental questions about the interaction between genes and the social environment (Mills and Tropf 2020). While still incipient, papers using genetic measures and exploring gene–environment interactions are now visible on the pages of demographic journals (e.g. Gaydosh et al. 2018; Fletcher 2019), and demographers have also led the discovery of the genetic architecture of traits through genome-wide association studies (GWAS) of reproductive outcomes (e.g. age at first birth), contributing their findings directly to journals in genetics (Barban et al. 2016). Demographic insights, such as those linked to population heterogeneity across cohorts (Tropf et al. 2017) and the importance of population representativeness (Mills and Rahal 2019), have potential to make vital contributions to the further growth of this interdisciplinary enterprise.

Many features of survey data have been designed to fill some of the gaps identified by demographers in the past. For example, two types of concerns often noted with the use of survey data, especially in the literature on fertility and family, have been the focus on women and the analysis of determinants solely at the individual level without consideration of wider contextual features (e.g. Watkins 1993; Greene and Biddlecom 2000). Although men's questionnaires were incorporated into the DHS from 1987, the use of these surveys to study fertility and fertility decisions from men's perspectives was limited (e.g. Dodoo 1998; Schoumaker 2017). Standardized surveys, such as the DHS, do not collect data on other household members, although the importance of intergenerational influence is being recognized and information collected in some survey infrastructures, such as the Gender and Generations Programme (Dykstra et al. 2006; Gauthier et al. 2018).

Although theories on fertility behaviour and transitions have emphasized the importance of social networks for social learning and information diffusion (e.g. Bongaarts and Watkins 1996; Bernardi and Klärner 2014), the data available to study these processes empirically in existing multipurpose

surveys are still quite limited. Some promising attempts at analysing networks have been made in the context of more localized and specialized studies and samples, for example in Kenya (Kohler et al. 2001) and Malawi (Helleringer et al. 2009). Looking forward, digital technologies offer potential for more detailed and cost-efficient data collection on social networks and interactions within existing surveys (e.g. through measurement via mobile apps and sensors) and small-scale prototypes of such efforts already exist (see 'New big data and digital demography' section). In the meantime, in the absence of empirical data, efforts at exploring the effects of social networks on demographic processes have relied on agent-based simulation models, for example in the study of reproductive behaviours (e.g. Diaz et al. 2011; González-Bailón and Murphy 2013), marriage (e.g. Billari et al. 2007; Bijak et al. 2013), and migration (Entwisle et al. 2016; Klabunde et al. 2017). More broadly, agent-based modelling offers a novel opportunity to integrate social learning and macro–micro feedback mechanisms and to test the implications of micro-level theories at the macro level and integrate multiple data types (Grow and Bavel 2016; Kashyap and Villavicencio 2016; Willekens et al. 2017). Although the development of this kind of 'system-based' approach to population modelling (Courgeau et al. 2017) has seen momentum develop around it in the past decade, it is still not mainstream. An increasingly salient mechanism of diffusion is the use of technologies such as mobile phones and the internet. These are channels for information diffusion, social learning and feedback, and exposure to the life of others. Demographic surveys, in contrast, have been slow to incorporate information on the use of these technologies, but are now beginning to do so (e.g. Rotondi et al. 2020).

Civil registration and vital statistics systems

Civil registration and vital statistics (CRVS) systems remain a valuable source for demographic research on mortality and fertility in high-income countries where coverage of vital events in these systems is complete. For these countries, the launches of two online data collections—the Human Mortality Database (see HMD 2021) in 2002, and the Human Fertility Database (see HFD 2021) in 2009—have been important milestones for the provision of high-quality, internationally comparable data on vital events and demographic measures derived from them. The HMD, a collaboration between the University of California, Berkeley and the Max Planck Institute for Demographic Research, has focused on providing detailed data over long time periods, including the provision of both cohort and period measures and reliable data at advanced ages, an increasingly important consideration with continued improvements in longevity (Barbieri et al. 2015). Although HMD data are normally provided for annual counts, in response to the needs for monitoring the mortality impacts of the Covid-19 pandemic, the HMD team developed the Short-term Mortality Fluctuations (STMF) data series within the HMD in 2020, providing harmonized weekly mortality data for several countries. The HFD, a collaboration between the Max Planck Institute for Demographic Research and the Vienna Institute of Demography, followed on from the success of the HMD and emerged in a context of increasing interest in later childbearing and low fertility in industrialized countries (Jasilioniene et al. 2016). The analysis of these trends required a closer understanding of both period and cohort changes, along with changes in timing and parity-specific trends, all of which are areas where the HFD provides high-quality and cross-nationally comparable data to facilitate demographic research.

CRVS systems remain significantly underdeveloped in LMICs. A review by Mahapatra et al. (2007) estimated that in the period 1995–2004, only 26 per cent of the world's population lived in countries with complete registration of deaths and 30 per cent in countries with complete registration of births (defined as at least 90 per cent of events registered). While coverage was generally high for populations in Europe and the Americas, the worst coverage was in Africa and Asia. Improvements in civil registration systems have been slow or stagnant since the mid-1960s, although some countries with poor systems in the 1980s witnessed substantial progress in subsequent decades (e.g. Baltic states, South Korea, and Latin America including Brazil, Mexico, and El Salvador). Others more recently, in the 2000s, have shown improvements over time periods of less than a decade (e.g. Bahrain, Cyprus, Egypt, and Malaysia). These latter examples indicate that progress over shorter time spans is possible, with 'purposeful policies' and when 'new ICT technologies are applied' (Mikkelsen et al. 2015, p. 1405).

The deficiency of vital registration systems in LMICs has resulted in continued reliance on different strategies of 'interim substitutes' (Hill et al. 2007, p. 1726) for the estimation of mortality and fertility. The most widely used have been sample surveys such as the DHS and UNICEF's Multiple

Indicator Cluster Surveys (MICS), followed by censuses and, in some cases, sample registration systems (India, China) and HDSS sites (for small geographic areas in a handful of countries). While little change has occurred in the basic data used for the generation of fertility and child mortality indicators from retrospective birth histories, methodological refinements drawing on the use of more sophisticated statistical models for estimation have been made (Schoumaker and Hayford 2004; Schoumaker 2013; Alkema et al. 2014). Efforts have also been directed at addressing selection biases in the use of sibling histories for adult mortality estimation (Gakidou and King 2006; Obermeyer et al. 2010; Masquelier 2013). An important development in this area has been the shift towards Bayesian statistical approaches that take a more integrated and nuanced approach to the quantification of uncertainty associated with demographic indicators computed using one or multiple sources of imperfect data and modelling of trends from them (Alkema et al. 2012, 2014, 2016; Bijak and Bryant 2016; Wheldon et al. 2016; Alexander and Alkema 2018).

Sample surveys have generally been more successful at estimating child mortality and fertility than at addressing data gaps in adult mortality. This data gap has meant that other population data collection exercises conducted by NSOs (e.g. censuses) have faced increased demands and burdens, for example through the inclusion of additional modules to monitor mortality. The collection of mortality information by age and sex in censuses for a reference period preceding the census has increased considerably, particularly after the UN working group for the 2010 round of population censuses recommended including these questions for those countries without alternative sources of adult mortality. By the 2010 census round, 72 countries incorporated mortality questions (a number that included countries across Africa, Asia, and Latin America), up from 53 in 2000 and 37 in the 1990 round (Hill et al. 2018). In addition to adult mortality, attempts have been made to estimate maternal mortality from census data through the inclusion of questions on the timing of deaths of women of reproductive age relative to pregnancies, thereby providing a way to measure pregnancy-related deaths. The number of countries with coverage of these questions expanded from 17 in the 2000 census round to 61 in the 2010 census round (Hill et al. 2018). Assessments of the effectiveness of census mortality questions have shown that estimates vary by type of method used with these sources (e.g. direct or indirect) (Odimegwu et al. 2018).

Ultimately while this approach may be cost effective, and methodological developments have improved insights gleaned from them, they are far from perfect substitutes for the development of civil registration systems and are best viewed 'as complementary with powerful synergistic potential' (Hill et al. 2007, p. 1733).

Compared with mortality and fertility, measurement of the third component of population change—migration—remains much more elusive, although progress has been made. The UN database on international migration, developed originally from research at the University of Sussex (Parsons et al. 2007) and subsequently extended and backdated by the UN and World Bank to 1990 (Skeldon 2018), provides biannual data available on international migrant stock by age, sex, and origin and destination countries and has been a significant achievement (UN Department of Economic and Social Affairs 2019). These tables of migrant stocks have been used to derive sequential flows (Abel and Sander 2014). Although the generation of data repositories is valuable, issues linked to data quality arising from different definitions of migration across states, different systems to enumerate migrants, and changing temporal dimensions of mobility all affect the interpretation of the numbers they provide (Skeldon 2018). The generation of cross-nationally comparable indicators of internal migration has also benefited from the improved availability of census microdata (Bell, Charles-Edwards, Kupiszewska et al. 2015; Bell, Charles-Edwards, Ueffing et al. 2015) and administrative registers linked to census data may further improve prospects for research in this area (Ernsten et al. 2018).

'New' big data and digital demography

The previous sections have outlined how, with the digital revolution, technological improvements in the past 25 years have helped to expand the capabilities of 'old' big population data sources, such as censuses and surveys. The spread and use of digital technologies, such as the internet, mobile phones, sensors, and cameras, as well as the increasing digitalization of different domains of social life, have themselves resulted in large volumes of new types of data on human activities, interactions, and behaviours—a category of data that has come to be known as 'digital trace' data.

The accumulation of digital traces has occurred as a result of two processes. First, the adoption of internet and mobile technologies implies that social life is

increasingly digitally mediated. For example, mobile phones and email are commonly used for communication, web search engines (e.g. Google) are used for information-seeking, and social media platforms, (e.g. Facebook, Twitter) are used for social interaction and exchange. These are essentially digital spaces where use of and engagement with these platforms and technologies generates digital records and data streams, which are regularly captured because these data are intrinsic to the business models of the private companies that provide these services, for example, to target advertisements. Second, the digitalization of data has resulted in the storage of diverse types of information—including about non-digital or offline life—as digital records of human activities. This development has clearly benefited the creation of repositories of long-standing population data sources, including previously paper-based records such as those from historical censuses. However, vast amounts of information linked to everyday activities (e.g. image and video recording of city life, consumer transactions) are also digitally stored and accessible, as are repositories of culturally or scientifically relevant materials, such as books and scientific journals.

Recent years have seen an emerging and increasing use of digital trace data sources for research on demographic topics or using demographic approaches, paving the way for the development of digital demography (Cesare et al. 2018). It is important to emphasize a key shift: that a lot of this work is now being done by demographers, often in collaboration with computer or information scientists, and is targeted to an audience of demographers, as well as others working broadly in the area of computational social science. Demographers are also joining discussions with international agencies and NSOs, who are increasingly interested in population measurement from non-traditional data sources (e.g. IUSSP 2015; Letouzé and Jutting 2015). The growth of this area is signalled by the regular presence of sessions on big data at international population conferences and the convening of two International Union for the Scientific Study of Population (IUSSP) panels, the first on 'Big Data and Population Processes' (2015–18) followed by the panel on 'Digital Demography' (2018–21) (see IUSSP 2021a). While this work has now begun to appear in demographic journals, a lot has been published in peer-reviewed proceedings of computer science conferences, where work in the areas of social computing, computational social science, and social informatics has grown rapidly and has a longer precedent.

Defining features of digital traces

What makes digital trace data a type of 'new' big data for population research that is different from 'old' big population data sources? The first pertains to the properties of the data and the second to the data-generating process or provenance of the data. Many definitions of digital big data emphasize their novelty in terms of their properties: volume, variety and velocity (Laney 2001; Sagiroglu and Sinanc 2013; Lazer and Radford 2017). While volume and variety are relevant, the more notable distinctions between population data sources (such as census and surveys) and digital trace big data lie in their velocity and the fact that these data are not systematically collected for research. Salganik has characterized these data sources as 'ready-mades' which, in contrast to 'custom-made' data such as censuses and surveys, are 'always on' and 'non-reactive' (Salganik 2019). The fact that these data are often by-products of the use of digital platforms provides opportunity for a more dynamic or continuous measurement in real time as events occur ('always on'), unlike that of data collection models such as decennial censuses or surveys, which involve asking questions at discrete points in time and require time for planning and fieldwork followed by additional lags for data production. Digital trace data also offer the potential to observe without asking ('non-reactive'), a feature that is promising but also raises important challenges for the use of data sources from an ethical standpoint. It is worth noting though that the non-reactive nature of data generation is not unique to digital traces and also holds for several (non-census) administrative data sources (including population registers, tax records and electronic health records) where the data are not systematically collected for research but generated as by-products of administrative transactions.

These differences in properties and provenance imply a number of features and issues that are unique to these data and relevant when considering their applications for demographic research. The first is their usability. Unlike rectangular data frames with rows and columns, many digital trace data are unstructured, messy, and come in formats unfamiliar to many demographers (e.g. JavaScript Object Notation (JSON)). They span media such as images, text, and time- or geographically stamped records of activity (e.g. metadata associated with calls, geographical location captured by apps, or geotags associated with specific tweets or posts). Constructs or covariates to enable statistical analysis

(e.g. regression) can be deduced from these data, but these constructs must first be operationalized based on what is observed, in contrast to data collection that occurs after the definition of a construct, as is usually the case in survey research (Cesare et al. 2018). The variety of formats, units of analyses, and also sizes of these data sets, which may contain millions of records, often require computational approaches for data management, retrieval, and analysis that are not yet a part of mainstream demographic training.

The second issue is that of bias and representativeness. Although the use of digital technologies has increased significantly, there are likely to be selection biases in who uses specific technologies, devices, or platforms, which limit the ability of these data sources to be population generalizable, a condition that has been sine qua non for the demographic enterprise. These biases are likely to be even more severe in LMICs, where despite data gaps being more significant, technological penetration remains uneven and digital divides are larger (ITU 2020). This limitation may help to explain why demographers have often been sceptical of digital data sources, although as discussed in the previous section, issues of selection bias or non-response bias are not absent from more traditional demographic data sources (e.g. incomplete vital registers or surveys). Indeed, issues of data quality and measurement are those that demographers have a long history of actively seeking to understand and address, and as I discuss later, emerging work by demographers working with digital traces indicates the extension of these insights to these new sources. Moreover, the lack of population representativeness of these data is in itself not a limitation per se and depends on how they are used in the context of a given research design. While selection bias may threaten the population generalizability of the estimation of levels of a quantity of interest, trend analysis is still feasible with biased data, using difference-in-difference type approaches (Zagheni and Weber 2015). Digital trace data, however, are affected by their own specific types of errors and bias. For example, they may be affected by algorithmic bias, whereby algorithms that are implemented on online platforms can shape behaviours, such that it may be difficult to assess whether an observed phenomenon is driven by human tendencies or algorithms (Lazer et al. 2014; Salganik 2019). A notable example of algorithmic bias is from one of the first and most famous applications of digital trace data for public health surveillance, Google Flu Trends, a platform that relied on aggregate web search queries for tracking flu outbreaks (Ginsberg et al. 2009). This platform was ultimately shut down in 2015, in response to the system beginning to overpredict flu prevalence systematically. A key factor underlying this was that flu-related searches increased significantly due to changes in Google Search's recommendation algorithm that promoted related search keywords to users (Lazer et al. 2014). A related issue to changing user behaviours is changing user composition—for example, given a landscape of changing service providers, migration between platforms may occur such that observed correlations from a given period may not persist.

Third, these data often come from and are owned by private companies, which has implications for their access, for the information available about them due to the often-restricted nature of proprietary algorithms that shape them, and by extension for issues such as research reproducibility. An example of this in the context of demographic research is provided by the Facebook advertisement platform, which consists essentially of marketing data about Facebook audience counts (users) provided to potential advertisers. These data have been repurposed for demographic studies of migration (e.g. Zagheni et al. 2017; Alexander et al. 2020; Rampazzo et al. 2021) based on the promising correlations observed between Facebook 'expats' and estimates of migrant stocks measured in large population surveys. While general descriptions of the expat category are provided by Facebook, the individual attributes determining whether a user is labelled an expat—whether based on geolocation or individual profile information, for example—remain unknown. Further, the numbers of expats as defined by Facebook also dropped significantly from March 2019 with little warning, likely due to changes in the underlying algorithm (Rampazzo et al. 2021). More broadly, the landscape of access to and use of these sources from a research perspective is unpredictable and uneven, and has arguably become even more restrictive in recent years in the face of legal changes, such as the introduction of the General Data Protection Regulation (GDPR) in Europe and privacy concerns (Freelon 2018; Bruns 2019). While some companies make their data (e.g. Twitter) or some form of aggregated data (e.g. Google Trends) available widely as per their terms of use, others have policies to restrict use to specific cases. Developing more democratic modes of access to these data sources for research—which focus on privacy preservation while also recognizing the data's scientific value, and which move beyond ad-hoc, private non-disclosure agreements—is

desirable, but large-scale prototypes of such mechanisms for access do not exist for the moment. While some prototypes are emerging (e.g. the Open Algorithms Project (OPAL 2021) that seeks to provide mobile phone data to researchers) and the Covid-19 pandemic accelerated the development of several initiatives for data sharing between private companies and university partners (e.g. Facebook's Covid-19 Symptom Survey and other Facebook 'Data for Good' outputs (Facebook 2021); also the COVID-19 Mobility Data Network (2021)), the development of frameworks for wider data sharing and even co-production is still much needed.

A fourth issue linked to the use of digital traces is that they are generated by users of specific services or technologies, who have not provided informed consent for the purposes of research. The absence of informed consent, a fundamental principle of survey research, in the generation of these data raises important ethical issues for researchers to consider when using them. On these grounds, it is argued that standards of privacy protection applied to them should be higher than those applied to data collected within the parameters of informed consent (Oberski and Kreuter 2020).

Demographic research with digital trace data

Two broad types of uses of digital trace data have emerged in the area of digital demography. The first strand of work has examined how digital trace data can be repurposed for measuring population indicators and processes, as well as understanding the contexts of demographic behaviours, particularly in domains where there are gaps in traditional sources of demographic data or measurement may be difficult. Given the challenges of migration measurement, considerable work in this vein has explored the potential of digital traces generated from the internet for estimating migration patterns. Nevertheless, examples of studies on mortality and fertility have also emerged. Crowdsourced online genealogies have been used for analysing longevity, including research on intergenerational correlations (Fire and Elovici 2015; Kaplanis et al. 2018). Aggregate Google Search queries conceptualized to proxy birth intentions have been used to predict fertility patterns (e.g. Billari et al. 2016; Wilde et al. 2020) and estimates of new parents on Facebook used to predict men's fertility (Rampazzo et al. 2018).

For migration, work was started by Zagheni and Weber, who used data on IP geolocation of a large population of Yahoo email users to estimate trends in international migration rates (Zagheni and Weber 2012). Other efforts have drawn on professional histories on networking sites—such as LinkedIn (State et al. 2014), bibliometric databases of scholarly publications such as Web of Science (Aref et al. 2019), air traffic flows (Gabrielli et al. 2019), social media such as Twitter (Fiorio et al. 2017; Yildiz et al. 2017), and Facebook's advertising platform (Zagheni et al. 2017; Alexander et al. 2019; Rampazzo et al. 2021)—to generate measures of flows and stocks of international or internal migrants and, in some cases, to integrate analyses of internal and international migration. While much of this work has been motivated by a need to fill data gaps in conventional sources, there have also been efforts to explore theoretical ideas by leveraging some of the flexible properties of digital traces. For example, Fiorio et al. 2021 explored how (re)defining temporal or geographical intervals of measurement can affect measured migration intensities. Aside from internet data, changes in the spatio-temporal distribution of mobile phone users detected using timestamped call detail records have been used to map population mobility, to capture seasonal internal mobility patterns that would otherwise be missed in conventional 'slower' data sources such as censuses (Blumenstock 2012; Deville et al. 2014; Lai et al. 2019), and to identify mobility changes in response to a crisis (Wilson et al. 2016). In contrast to the studies based on internet data sources, which largely focus on high-income settings, studies using mobile phones are more likely to cover LMICs, given the wider diffusion of mobile phones relative to the internet. The Covid-19 pandemic has further illustrated the value of using mobile phones for monitoring mobility changes as a broader tool for pandemic response strategies (Grantz et al. 2020; Oliver et al. 2020) and has seen a renewed call for more open access to aggregated mobility data from mobile phones (Buckee et al. 2020).

A distinct feature of demographic work using digital trace data has been its attention to examining the viability of these sources for population-generalizable measurement (e.g. Yildiz et al. 2017; Zagheni et al. 2017; Alexander et al. 2020). This has involved modelling and understanding biases in the quantities computed using different digital platforms by calibrating them against 'ground truth' measures, for example, derived from large, representative surveys or censuses. This type of approach, where parameters are estimated by linking faulty or imperfect data to more reliable data through a model, is similar to the logic of model life tables, which have been widely used to improve coverage of mortality

estimates in low-income countries with deficient vital registration (Coale and Trussell 1996). In a similar way, some studies have provided examples of how digital trace data could improve country coverage of indicators linked, for example, to men's fertility (Rampazzo et al. 2018) or gender inequality (Kashyap et al. 2020) in LMICs, by estimating models that link signals from social media to ground-truth indicators. Data on consumer transactions (Longley et al. 2018; Lansley et al. 2019) and satellite images (Lloyd et al. 2017; Leasure et al. 2020) can provide finer spatial resolution to subnational population estimates when combined with census and administrative data sets. More commonly, though, studies have articulated the value of digital traces in terms of their higher-frequency temporal resolution for nowcasting current patterns of demographic indicators before these appear in official statistics (e.g. Billari et al. 2016; Fiorio et al. 2017; Kashyap et al. 2017; Alexander et al. 2020; Wilde et al. 2020) or for monitoring changes during crises when traditional forms of data collection may be infeasible (e.g. Alexander et al. 2019; Palotti et al. 2020). A number of nowcasting efforts have found that the best predictions are generated by combining signals derived from digital traces with sources such as large-scale surveys and administrative data sets, emphasizing the former's value as complements, not substitutes. This has paved the way for efforts developing statistical frameworks that integrate both types of sources and quantify uncertainty, capitalizing on their relative strengths and weaknesses (e.g. Lansley et al. 2019; Alexander et al. 2020; Rampazzo et al. 2021).

Digital traces of different types can provide complementary measurement of contexts and environments, expressions of identities and sentiments, and information-seeking behaviours that are relevant for understanding demographic behaviours and outcomes. In cases where there may be social desirability biases, for example in the case of abortion (Reis and Brownstein 2010; Leone et al. 2021) or sex-selective abortion (Kashyap et al. 2017), data from aggregated Google searches have been shown to capture information-seeking behaviours that might not be readily measured in surveys. Analyses of sentiment around parenthood, as expressed by what people tweet on social media, might shed light on normative responses but also on concerns and contexts surrounding parenthood (Mencarini et al. 2019). Photos have provided a novel opportunity to analyse interracial friendships (Berry 2006) or, in other cases, an opportunity to improve age measurement in contexts where age reporting is missing or faulty (Helleringer et al. 2019). Images of cars in a neighbourhood captured by Google Street View have also been used to infer the socio-economic characteristics of US cities (Gebru et al. 2017).

While the examples so far refer to cases where digital traces already exist and require repurposing, a hybrid approach, where digital traces are actively collected by researchers within survey instruments, provides a promising strategy for linkage of different types of data (Stier et al. 2020) that could also be fruitfully applied for demographic research. Palmer et al. (2013) provided an example of how mobile phones can be used to collect survey data with location-based information, thereby examining individuals in their activity spaces in a more dynamic way that is more aligned with how they actually experience space, rather than in discrete census blocks. Other examples that could be applied to demographic research include the use of sensors to capture social interactions and networks by measuring spatial proximity (Cattuto et al. 2010; Kiti et al. 2016) and the use of metadata from call records (Kreuter et al. 2020). In addition to overcoming measurement challenges linked to recall bias in self-reports, these passive data collection approaches could also help ameliorate respondent burden—a growing issue with increasingly lengthy survey instruments—while providing avenues to obtain informed consent.

The second strand of work using digital trace data has sought to apply demographic approaches to study *who* is online or to understand demographic behaviours (e.g. dating and mate search) in digital spaces. This work is motivated by the fact that the digital revolution has itself created new *online* populations. Understanding the demographic composition and biases of digitally connected populations globally is likely to become increasingly important, given continued interest in using internet and mobile technologies for either passive observation or active recruitment (e.g. for survey data collection). While the use of online panels and social media recruitment for surveys in high-income countries has increased significantly due to their cost efficiency (Groves 2011; Schneider and Harknett 2019; Kalimeri et al. 2020), there are also emerging examples of data collection via mobile phones and social media in LMICs (e.g. Tamgno et al. 2013; Diamond-Smith et al. 2020; Coffey et al. 2021). This trend towards leveraging mobile, internet, and social media technologies for data collection significantly accelerated during the Covid-19 pandemic, when rapid data collection was clearly needed but traditional face-to-face approaches for

collection infeasible (e.g. Adjiwanou et al. 2020; Grow et al. 2020; Battiston et al. 2021; Feehan and Mahmud 2021), stimulating wider discussions among population scientists about such approaches (e.g. IUSSP 2021b). Population-generalizable uses of these approaches require a deeper understanding of who is using these technologies or specific platforms. For example, significant gender inequalities in internet and mobile phone access exist in South Asia and sub-Saharan Africa, but many social surveys provide limited data on information technology use by sex and other characteristics (Fatehkia et al. 2018). Understanding demographic differentials in the use of specific social media platforms may require us to turn to digital traces (Gil-Clavel and Zagheni 2019) or collect data from specific platforms (e.g. Facebook or Google) and then generalize to a wider population of internet users, using models (Fatehkia et al. 2018; Feehan and Cobb 2019; Kashyap et al. 2020). For example, Feehan and Cobb highlighted how network sampling approaches—similar to indirect mortality estimation approaches that involve asking others, such as with sibling histories—can also be used to deduce internet adoption from a limited online sample.

The increasing digitalization of different life domains implies that digital inequalities also have implications for demographic processes and social inequalities. Moreover, studies of online life can illuminate whether inequalities from the offline world are reinforced or transformed online. They may also provide a lens to study social interactions and behaviours that were previously difficult to study. An example of this is provided by the literature on internet dating. Although demographers have often written about the marriage market, they have rarely observed mate search dynamics. Preferences are often inferred from outcomes (e.g. marriage rates) but never directly observed. Studies of online dating provide a unique opportunity to understand these behaviours (e.g. Potârcă and Mills 2015; Bruch and Newman 2018).

Demography and the data revolution: Emerging opportunities and challenges

To conclude, I return to the question posed at the beginning: has demography witnessed a data revolution? Yes. Over the past 25 years, the data ecosystem for demographic research has been significantly enriched, as demonstrated by the augmented granularity and new features available for 'old' big population data sources (such as censuses,

administrative data, and surveys) and the growing use of 'new' big data sources. This is not to say that progress has been even, and a vision of the data revolution that strives to leave no one 'invisible' (IAEG 2014) must confront the striking inequalities that persist between the data-rich Global North and data-poor Global South. Bridging these gaps requires financial, political, and intellectual investments towards developing robust data infrastructures for different types of data (including censuses, administrative data, vital statistics, and new forms of data), as well as sustained efforts to strengthen scientific training and capacity in LMICs to sustain the development of long-term studies. While methodological ingenuity through the use of statistical, including Bayesian, models has helped to extract value from the data that are available, these approaches have starkly illuminated the high levels of uncertainty that affect population estimation and vital rates based on deficient data.

Although new data opportunities have helped to address some lingering concerns, they have also raised new ones. First, the expansion in volume, variety, and opportunity for linkage across multiple different types of data has occurred against a backdrop of growing concern about non-response, greater threat of respondent reidentification, and need for privacy protection, because with more detailed data also comes the potential for misuse. Moving forward, maintaining sensitivity and foresight towards the potential for data *misuse*, while mitigating the scientific loss incurred by *missed* use, will become an increasingly necessary balancing act, where population scientists will need to make an active contribution to public debates to articulate the trade-offs. These tensions have already come to the forefront in the charged debates surrounding the implementation of differential privacy (DP) as a method of disclosure control in the 2020 US Census (Ruggles et al. 2019). The introduction of statistical noise entailed by DP to prevent respondent reidentification threatens to have an adverse effect on block-level estimates and population denominators for the calculation of disaggregated rates derived from census microdata (Santos-Lozada et al. 2020; Hauer and Santos-Lozada 2021) and could significantly compromise the scientific and policy impacts that more granular data have achieved.

Second, the different kinds of data sources enabled by the data revolution have provided momentum for both the macro-level *discovery* stage of demographic research that aims to describe empirical regularities in populations and the

explanation stage that seeks to understand why and how they occur (Billari 2015), but more work is needed to integrate the two approaches. The availability of integrated cross-national databases for demographic research has sustained important work that focuses on good population-level description. There may indeed be even more opportunity to revitalize the discovery of empirical regularities in population-level dynamics by drawing on machine learning techniques applied to census microdata and large-scale surveys (e.g. De Maria et al. 2019; Hauer and Bohon 2020; Salganik et al. 2020) or to discover new regularities in demographic processes when observed at finer temporal or spatial resolutions, as provided through digital traces (e.g. Fiorio et al. 2021). This macro-level orientation has coexisted with research focused on explaining individual-level variations and understanding pathways, in which a greater emphasis and understanding of issues of selection, heterogeneity, and causal interpretation is also now visible. There is clearly a role for thinking more expansively about causality in demography and for considering methods of demographic accounting, including techniques such as standardization and decomposition focused on the macro level as types of causal analysis, as argued by Bhrolcháin and Dyson (2007). These approaches may yield new insights when applied to different online populations in the context of digital demography, for example to understand differences between platforms that are attributable to behaviours vs demographic compositions (e.g. Cesare et al. 2018). The deeper combination of different types of data including both the new and old, as well as the development of approaches such as empirically informed agent-based and microsimulation models that conceptualize populations as systems, offer promising opportunities for the integration of micro and macro levels of analysis.

Third, and linked to the second issue, we need to rethink which methods are used with the data opportunities we now have. With the increasing use of individual-level survey data over the past five decades, demographers have come to rely extensively on the tools of inferential statistics, which were developed with small, not large, data sets in mind. Bohon has described this in terms of a culture in which demographers have been collecting and analysing 'big data in a small way' (Bohon 2018, p. 323). Discussions on de-emphasizing *p*-values in scientific analyses within the broader research community are highly relevant for demographers, as our data sets are already large and have become larger. A recognition of these issues within the demographic community has emerged clearly, as shown, for example, by the stance adopted by the editorial board of *Demographic Research* on *p*-values (Bijak 2019). We need to move further towards emphasizing different aspects of our models, for example the size of effects, and towards a deeper assessment of the magnitude or meaning of these effects. It will also be increasingly important to recall that bigger data with richer features do not necessarily equate to better or unbiased data. While these issues have been discussed in relation to ongoing 'new' big data sources, for which demographic techniques for assessing data quality and examining biases have the potential to be reconfigured, these lessons should not be forgotten with more traditional data sources either. In any case, demographers are uniquely positioned to make a vital contribution to how 'new' big data sources can be used for careful population-generalizable measurement and to play a pivotal role in the broader development of the field of computational social science. Demographic insights linked to understanding and validating population representativeness are also salient for the careful integration of biological measures within surveys, for example polygenic scores drawing on genome-wide association studies (Mills and Tropf 2020). The data revolution, with its incumbent data opportunities, has forged the pathway for interdisciplinary research at the interface of demography with other disciplines, whether biology, economics, behavioural science, or computer/information science, but there is still much more room to grow.

Fourth, a key advantage of demographic research is that it often draws on data that are available in the public domain and accessible through web-based repositories. This is important from the perspectives of researcher access and reproducibility, both areas of increasing scientific discussion. This mode of access means that some of the concerns about reproducibility of findings (e.g. the 'replication crisis' in psychology)—due to data that are not shared or in the public domain, or, as with some emerging forms of big data, arising from restricted or protected data from companies—are less immediately pertinent within the field. However, as demographers come to rely more on more detailed or restricted-access data or to draw on data from commercial providers, we will need to generate solutions for maintaining open, transparent models of research, while aligning with regulations on confidentiality and licencing. The use of pre-analysis or preregistration plans, which have come to be used in experimental research in fields including psychology and political science, could offer a model for

greater transparency, although the extension of this model for the analysis of secondary observational data sets is unclear. Standards for research reproducibility are still evolving, however, and although positive signs of change are visible (e.g. inclusion of code with publications), open questions remain about how best to develop these approaches, given the changing data ecosystem and also in light of the non-trivial computational power (and technical skill) that replicating analyses increasingly demands.

Capitalizing on the opportunities presented by the data revolution requires that we retain some of the old, while adapting flexibly to incorporate the new. Some of the new will require (re)training in methods of data collection, management, and analyses, but also, importantly, data ethics. It will also require us to draw on theories and ideas from other disciplines. While the data revolution in demography has clearly started, it has only just begun.

Notes and acknowledgements

ORCID

Ridhi Kashyap http://orcid.org/0000-0003-0615-2868

References

Abel, Guy J., and Nikola Sander. 2014. Quantifying global international migration flows, *Science* 343(6178): 1520–1522. doi:10.1126/science.1248676

Adjiwanou, Vissého, Nurul Alam, Leontine Alkema, Gershim Asiki, Ayaga Bawah, Donatien Béguy, Valeria Cetorelli et al. 2020. Measuring excess mortality during the COVID-19 pandemic in low- and lower-middle income countries: The need for mobile phone surveys, *SocArXiv*. doi:10.31235/osf.io/4bu3q

Alburez-Gutierrez, Diego, Emilio Zagheni, Samin Aref, Sofia Gil-Clavel, André Grow, and Daniela Veronica Negraia. 2019. Demography in the digital era: New data sources for population research, *Preprint. SocArXiv*. doi:10.31235/osf.io/24jp7

Alderman, Harold, Jere R. Behrman, Hans-Peter Kohler, John A. Maluccio, and Susan Cotts Watkins. 2001. Attrition in longitudinal household survey data: Some tests for three developing-country samples, *Demographic Research* 5: 79–124. doi:10.4054/DemRes.2001.5.4

Alexander, Monica, and Leontine Alkema. 2018. Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model, *Demographic Research* 38: 335–372. doi:10.4054/DemRes.2018.38.15

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2019. The impact of hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook data, *Population and Development Review* 45(3): 617–630. doi:10.1111/padr.12289

Alexander, Monica, Kivan Polimis, and Emilio Zagheni. 2020. Combining social media and survey data to nowcast migrant stocks in the United States, *Population Research and Policy Review*. doi:10.1007/s11113-020-09599-3

Alkema, Leontine, Doris Chou, Daniel Hogan, Sanqian Zhang, Ann-Beth Moller, Alison Gemmill, Doris Ma Fat et al. 2016. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group, *The Lancet* 387 (10017): 462–474. doi:10.1016/S0140-6736(15)00838-7

Alkema, Leontine, Jin Rou New, Jon Pedersen, Danzhen You, and all members of the UN Inter-agency Group for Child Mortality Estimation and its Technical Advisory Group. 2014. Child mortality estimation 2013: An overview of updates in estimation methods by the United Nations Inter-Agency Group for Child Mortality Estimation, *PLOS ONE* 9(7): e101112. doi:10.1371/journal.pone.0101112

Alkema, Leontine, Adrian E. Raftery, Patrick Gerland, Samuel J. Clark, and Francois Pelletier. 2012. Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa, *Demographic Research* 26(15): 331–362. doi:10.4054/DemRes.2012.26.15

Andriano, Liliana, and Julia Behrman. 2020. The effects of growing-season drought on young women's life course transitions in a Sub-Saharan context, *Population Studies* 74(3): 331–350. doi:10.1080/00324728.2020.1819551

Andriano, Liliana, and Christiaan W. S. Monden. 2019. The causal effect of maternal education on child mortality: Evidence from a quasi-experiment in Malawi and Uganda, *Demography* 56(5): 1765–1790. doi:10.1007/s13524-019-00812-3

Aref Samin, Emilio Zagheni, and Jevin West. 2019. The demography of the peripatetic researcher: Evidence on highly mobile scholars from the Web of Science, in Ingmar Weber, Kareem M. Darwish, Claudia Wagner, Emilio Zagheni, Laura Nelson, Samin Aref, and Fabian Flöck (eds), *Social Informatics*. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 50–65. doi:10.1007/978-3-030-34971-4_4

Bakker, Bart F. M., Johan van Rooijen, and Leo van Toor. 2014. The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics, *Statistical Journal of the IAOS* 30(4): 411–424. doi:10.3233/SJI-140803

Barban, Nicola, Rick Jansen, Ronald de Vlaming, Ahmad Vaez, Jornt J. Mandemakers, Felix C. Tropf, Xia Shen et al. 2016. Genome-wide analysis identifies 12 loci influencing human reproductive behavior, *Nature Genetics* 48(12): 1462–1472. doi:10.1038/ng.3698

Barbieri, Magali, John R. Wilmoth, Vladimir M. Shkolnikov, Dana Glei, Domantas Jasilionis, Dmitri Jdanov, Carl Boe et al. 2015. Data resource profile: The Human Mortality Database (HMD), *International Journal of Epidemiology* 44(5): 1549–1556. doi:10.1093/ije/dyv105

Barclay, Kieron, Anna Baranowska-Rataj, Martin Kolk, and Anneli Ivarsson. 2020. Interpregnancy intervals and perinatal and child health in Sweden: A comparison within families and across social groups, *Population Studies* 74(3): 363–378. doi:10.1080/00324728.2020.1714701

Barclay, Kieron, and Mikko Myrskylä. 2016. Advanced maternal age and offspring outcomes: Reproductive aging and counterbalancing period trends, *Population and Development Review* 42(1): 69–94. doi:10.1111/j.1728-4457.2016.00105.x

Battiston, Pietro, Ridhi Kashyap, and Valentina Rotondi. 2021. Reliance on scientists and experts during an epidemic: Evidence from the COVID-19 outbreak in Italy, *SSM – Population Health* 13(March): 100721. doi:10.1016/j.ssmph.2020.100721

Bell, Martin, Elin Charles-Edwards, Dorota Kupiszewska, Marek Kupiszewski, John Stillwell, and Yu Zhu. 2015. Internal migration data around the world: Assessing contemporary practice, *Population, Space and Place* 21(1): 1–17. doi:10.1002/psp.1848

Bell, Martin, Elin Charles-Edwards, Philipp Ueffing, John Stillwell, Marek Kupiszewski, and Dorota Kupiszewska. 2015. Internal migration and development: Comparing migration intensities around the world, *Population and Development Review* 41(1): 33–58. doi:10.1111/j.1728-4457.2015.00025.x

Bernardi, Laura, and Andreas Klärner. 2014. Social networks and fertility, *Demographic Research* 30: 641–670. doi:10.4054/DemRes.2014.30.22

Berry, Brent. 2006. Friends for better or for worse: Interracial friendship in the United States as seen through wedding party photos, *Demography* 43(3): 491–510. doi:10.1353/dem.2006.0020

Bhrolcháin, Máire Ní, and Tim Dyson. 2007. On causation in demography: Issues and illustrations, *Population and Development Review* 33(1): 1–36. doi:10.1111/j.1728-4457.2007.00157.x

Bicego, George. 1997. Estimating adult mortality rates in the context of the AIDS epidemic in Sub-Saharan Africa: Analysis of DHS sibling histories, *Health Transition Review* 7: 7–22. Available: https://www.jstor.org/stable/40652323

Bijak, Jakub. 2019. Editorial: P-values, theory, replicability, and rigour, *Demographic Research* 41(32): 949–952. doi:10.4054/DemRes.2019.41.32

Bijak, Jakub, and John Bryant. 2016. Bayesian demography 250 years after Bayes, *Population Studies* 70(1): 1–19. doi:10.1080/00324728.2015.1122826

Bijak, Jakub, Jason Hilton, Eric Silverman, and Viet Dung Cao. 2013. Reforging the wedding ring: Exploring a semi-artificial model of population for the United Kingdom with Gaussian process emulators, *Demographic Research* 29(27): 729–766. doi:10.4054/DemRes.2013.29.27

Billari, Francesco C. 2015. Integrating macro- and micro-level approaches in the explanation of population change, *Population Studies* 69(S1): S11–S20. doi:10.1080/00324728.2015.1009712

Billari Francesco C., Francesco D'Amuri, and Juri Marcucci. 2016. Forecasting births using Google, in *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*. Valencia: Editorial Universitat Politècnica de València, pp. 119–119. doi:10.4995/CARMA2016.2015.4301

Billari, Francesco C., Alexia Prskawetz, Belinda Aparicio Diaz, and Thomas Fent. 2007. The "wedding-ring": An agent-based marriage model based on social interaction, *Demographic Research* 17: 59–82. doi:10.4054/DemRes.2007.17.3

Billari, Francesco C., and Emilio Zagheni. 2017. Big Data and population processes: A revolution? doi:10.31235/osf.io/f9vzp.

Blumenstock, Joshua E. 2012. Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda, *Information Technology for Development* 18(2): 107–125. doi:10.1080/02681102.2011.643209

Bohon, Stephanie A. 2018. Demography in the big data revolution: Changing the culture to forge new frontiers, *Population Research and Policy Review* 37(3): 323–341. doi:10.1007/s11113-018-9464-6

Bongaarts, John. 2007. Late marriage and the HIV epidemic in Sub-Saharan Africa, *Population Studies* 61 (1): 73–83. doi:10.1080/00324720601048343

Bongaarts, John, and Susan Cotts Watkins. 1996. Social interactions and contemporary fertility transitions, *Population and Development Review* 22(4): 639–682. doi:10.2307/2137804

Brass, William. 1996. Demographic data analysis in less developed countries: 1946–1996, *Population Studies* 50 (3): 451–467. doi:10.1080/0032472031000149566

Bruch, Elizabeth E., and M. E. J. Newman. 2018. Aspirational pursuit of mates in online dating markets, *Science Advances* 4(8): eaap9815. doi:10.1126/sciadv.aap9815

Bruns, Axel. 2019. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research, *Information, Communication & Society* 22 (11): 1544–1566. doi:10.1080/1369118X.2019.1637447

Buckee, Caroline O., Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, Urs Gasser, Yonatan H. Grad et al. 2020. Aggregated mobility data could help fight COVID-19, *Science* 368(6487): 145–146. doi:10.1126/science.abb8021

Cattuto, Ciro, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks, *PLOS ONE* 5(7): e11596. doi:10.1371/journal.pone.0011596

Cesare, Nina, Hedwig Lee, Tyler McCormick, Emma Spiro, and Emilio Zagheni. 2018. Promises and pitfalls of using digital traces for demographic research, *Demography* 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2

Cleland, John. 1996. Demographic data collection in less developed countries 1946–1996, *Population Studies* 50 (3): 433–450. doi:10.1080/0032472031000149556

Coale, Ansley, and James Trussell. 1996. The development and use of demographic models, *Population Studies* 50 (3): 469–484. doi:10.1080/0032472031000149576

Coffey, Diane, Payal Hathi, Nazar Khalid, and Amit Thorat. 2021. Measurement of population mental health: Evidence from a mobile phone survey in India, *Health Policy and Planning* 36(5): 606–619. doi:10.1093/heapol/czab023

Coleman, David. 2013. The twilight of the census, *Population and Development Review* 38: 334–351. doi:10.1111/j.1728-4457.2013.00568.x

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. The role of administrative data in the big data revolution in social science research, *Social Science Research* 59 (September): 1–12. doi:10.1016/j.ssresearch.2016.04.015

Corsi, Daniel J., Melissa Neuman, Jocelyn E. Finlay, and S. V. Subramanian. 2012. Demographic and Health Surveys: A profile, *International Journal of Epidemiology* 41(6): 1602–1613. doi:10.1093/ije/dys184

Courgeau Daniel, Jakub Bijak, Robert Franck, and Eric Silverman. 2017. Model-based demography: Towards a research agenda, in André Grow and Jan Van Bavel (eds), *Agent-based Modelling in Population Studies: Concepts, Methods, and Applications*. The Springer Series on Demographic Methods and Population Analysis. Cham: Springer International Publishing, pp. 29–51. doi:10.1007/978-3-319-32283-4_2

Courgeau, Daniel, Robert Franck, and Julia Key. 2007. Demography, a fully formed science or a science in the making? An outline programme, *Population* 62 (1): 39–45. doi:10.3917/popu.701.0039

COVID-19 Mobility Data Network. 2021. Available: https://www.covid19mobility.org/partners/ (accessed: 30 July 2021).

Crimmins, Eileen M. 1993. Demography: The past 30 years, the present, and the future, *Demography* 30(4): 579–591. doi:10.2307/2061807

Crimmins, Eileen, Jung Ki, and Sarinnapha Vasunilashorn. 2010. Biodemography: New approaches to understanding trends and differences in population health and mortality, *Demography* 47(1): S41–S64. doi:10.1353/dem.2010.0005

De Maria, Pier Francesco, Leonardo Tomazeli Duarte, Álvaro de Oliveira D'Antona, and Cristiano Torezzan. 2019. Digital humanities and big microdata: New approaches for demographic research, in Leonardo Bacelar Lima Santos, Rogério Galante Negri, and Tiago José de Carvalho (eds), *Towards Mathematics, Computers and Environment: A Disasters Perspective*. Cham: Springer International Publishing, pp. 217–231. doi:10.1007/978-3-030-21205-6_11

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel et al. 2014. Dynamic population mapping using mobile phone data, *Proceedings of the National Academy of Sciences* 111(45): 15888–15893. doi:10.1073/pnas.1408439111

Diamond-Smith, Nadia, Avery E. Holton, Sarah Francis, and Drew Bernard. 2020. Addressing anemia among women in India—an informed intervention using Facebook Ad manager, *MHealth* 6: 39. doi:10.21037/mhealth-19-237a

Diaz, Belinda Aparicio, Thomas Fent, Alexia Prskawetz, and Laura Bernardi. 2011. Transition to parenthood: The role of social interaction and endogenous

networks, *Demography* 48(2): 559–579. doi:10.1007/s13524-011-0023-6

Dodoo, F. N. 1998. Men matter: Additive and interactive gendered preferences and reproductive behavior in Kenya, *Demography* 35(2): 229–242. doi:10.2307/3004054

Dorélien, Audrey, Deborah Balk, and Megan Todd. 2013. What Is urban? Comparing a satellite view with the Demographic and Health Surveys, *Population and Development Review* 39(3): 413–439. doi:10.1111/j.1728-4457.2013.00610.x

Dribe, Martin, J. David Hacker, and Francesco Scalone. 2014. The impact of socio-economic status on net fertility during the historical fertility decline: A comparative analysis of Canada, Iceland, Sweden, Norway, and the USA, *Population Studies* 68(2): 135–149. doi:10.1080/00324728.2014.889741

Dykstra, Pearl, Matthijs Kalmijn, Trudie Knijn, Aafke Komter, Aart Liefbroer, and Clara Mulder. 2006. *Family Solidarity in the Netherlands.* Available: https://repub.eur.nl/pub/18191/.

Einav, Liran, and Jonathan Levin. 2014. The data revolution and economic analysis, *Innovation Policy and the Economy* 14(January): 1–24. doi:10.1086/674019

Entwisle, Barbara, Nathalie E. Williams, Ashton M. Verdery, Ronald R. Rindfuss, Stephen J. Walsh, George P. Malanson, Peter J. Mucha et al. 2016. Climate shocks and migration: An agent-based modeling approach, *Population and Environment* 38(1): 47–71. doi:10.1007/s11111-016-0254-y

Ernsten, Annemarie, David McCollum, Zhiqiang Feng, Dawn Everington, and Zengyi Huang. 2018. Using linked administrative and census data for migration research, *Population Studies* 72(3): 357–367. doi:10.1080/00324728.2018.1502463

Esteve, Albert, Ron Lesthaeghe, and Antonio López-Gay. 2012. The Latin American cohabitation boom, 1970–2007, *Population and Development Review* 38(1): 55–81. doi:10.1111/j.1728-4457.2012.00472.x

Esteve, Albert, Christine R. Schwartz, Jan Van Bavel, Iñaki Permanyer, Martin Klesment, and Joan Garcia. 2016. The end of hypergamy: Global trends and implications, *Population and Development Review* 42(4): 615–625. doi:10.1111/padr.12012

Facebook. 2021. Available: https://dataforgood.fb.com/tools/symptomsurvey/ (accessed: 30 July 2021).

Fatehkia, Masoomali, Ridhi Kashyap, and Ingmar Weber. 2018. Using Facebook ad data to track the global digital gender gap, *World Development* 107(July): 189–209. doi:10.1016/j.worlddev.2018.03.007

Feehan, Dennis M., and Curtiss Cobb. 2019. Using an online sample to estimate the size of an offline population, *Demography* 56(6): 2377–2392. doi:10.1007/s13524-019-00840-z

Feehan, Dennis M., and Ayesha S. Mahmud. 2021. Quantifying population contact patterns in the United States during the COVID-19 pandemic, *Nature Communications* 12(1): 893. doi:10.1038/s41467-021-20990-2

Ferrari, Giulia, and Ross Macmillan. 2019. Until work do us part: Labour migration and occupational stratification in non-cohabiting marriage, *Population Studies* 73(2): 197–216. doi:10.1080/00324728.2019.1583359

Festy, Patrick. 2007. Numbering same-sex couples in censuses and population registers, *Demographic Research* 17: 339–368. doi:10.4054/DemRes.2007.17.12

Fiorio, Lee, Emilio Zagheni, Guy Abel, Johnathan Hill, Gabriel Pestre, Emmanuel Letouzé, and Jixuan Cai. 2021. Analyzing the effect of time in migration measurement using georeferenced digital trace data. *Demography* 58(1): 51–74. doi:10.1215/00703370-8917630

Fiorio, Lee, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué. 2017. Using Twitter data to estimate the relationship between short-term mobility and long-term migration, in *Proceedings of the 2017 ACM on Web Science Conference.* WebSci'17. Troy, NY: Association for Computing Machinery, pp. 103–110. doi:10.1145/3091478.3091496

Fire, Michael, and Yuval Elovici. 2015. Data mining of online genealogy datasets for revealing lifespan patterns in human population, *ACM Transactions on Intelligent Systems and Technology* 6(2): 28:1–28:22. doi:10.1145/2700464

Fletcher, Jason M. 2019. Environmental bottlenecks in children's genetic potential for adult socio-economic attainments: Evidence from a health shock, *Population Studies* 73(1): 139–148. doi:10.1080/00324728.2018.1498533

Frankenberg, Elizabeth, Wayan Suriastini, and Duncan Thomas. 2005. Can expanding access to basic healthcare improve children's health status? Lessons from Indonesia's "Midwife in the Village" programme, *Population Studies* 59(1): 5–19. doi:10.1080/0032472052000332674

Freelon, Deen. 2018. Computational research in the post-API age, *Political Communication* 35(4): 665–668. doi:10.1080/10584609.2018.1477506

Gabrielli, Lorenzo, Emanuel Deutschmann, Fabrizio Natale, Ettore Recchi, and Michele Vespe. 2019. Dissecting global air traffic data to discern different types and trends of transnational human mobility, *EPJ Data Science* 8(1): 26. doi:10.1140/epjds/s13688-019-0204-x

Gakidou, Emmanuela, and Gary King. 2006. Death by survey: Estimating adult mortality without selection bias from sibling survival data, *Demography* 43(3): 569–585. doi:10.1353/dem.2006.0024

Gauthier, Anne H., Susana Laia Farinha Cabaço, and Tom Emery. 2018. Generations and Gender Survey study profile, *Longitudinal and Life Course Studies* 9(4): 456–465. doi:10.14301/llcs.v9i4.500

Gaydosh, Lauren, Daniel W. Belsky, Benjamin W. Domingue, Jason D. Boardman, and Kathleen Mullan Harris. 2018. Father absence and accelerated reproductive development in non-Hispanic white women in the United States, *Demography* 55(4): 1245–1267. doi:10.1007/s13524-018-0696-1

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, *Proceedings of the National Academy of Sciences* 114(50): 13108–13113. doi:10.1073/pnas.1700035114

Gil-Clavel, Sofia, and Emilio Zagheni. 2019. Demographic differentials in Facebook usage around the world, *Proceedings of the International AAAI Conference on Web and Social Media* 13(July): 647–650. https://ojs.aaai.org/index.php/ICWSM/article/view/3263

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data, *Nature* 457(7232): 1012–1014. doi:10.1038/nature07634

González-Bailón, Sandra, and Tommy E. Murphy. 2013. The effects of social interactions on fertility decline in nineteenth-century France: An agent-based simulation experiment, *Population Studies* 67(2): 135–155. doi:10.1080/00324728.2013.774435

Google Ngram viewer. 2021. Available: https://books.google.com/ngrams (accessed: 30 July 2021).

Grace, Kathryn, Andrew Verdin, Audrey Dorélien, Frank Davenport, Chris Funk, and Greg Husak. 2021. Exploring strategies for investigating the mechanisms linking climate and individual-level child health outcomes: An analysis of birth weight in Mali, *Demography* 58(2): 499–526. doi:10.1215/00703370-8977484

Grantz, Kyra H., Hannah R. Meredith, Derek A. T. Cummings, C. Jessica E. Metcalf, Bryan T. Grenfell, John R. Giles, Shruti Mehta et al. 2020. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology, *Nature Communications* 11(1): 4961. doi:10.1038/s41467-020-18190-5

Greene, Margaret E., and Ann E. Biddlecom. 2000. Absent and problematic men: Demographic accounts of male reproductive roles, *Population and Development Review* 26(1): 81–115. doi:10.1111/j.1728-4457.2000.00081.x

Groves, Robert M. 2006. Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly* 70(5): 646–675. doi:10.1093/poq/nfl033

Groves, Robert M. 2011. Three eras of survey research, *Public Opinion Quarterly* 75(5): 861–871. doi:10.1093/poq/nfr057

Grow, André, and Jan Van Bavel. 2016. *Agent-Based Modelling in Population Studies: Concepts, Methods, and Applications.* Cham: Springer.

Grow, André, Daniela Perrotta, Emanuele Del Fava, Jorge Cimentada, Francesco Rampazzo, Sofia Gil-Clavel, and Emilio Zagheni. 2020. Addressing public health emergencies via Facebook surveys: Advantages, challenges, and practical considerations, *Journal of Medical Internet Research* 22(12): e20653. doi:10.2196/20653

Guilmoto, Christophe Z. 2017. Gender bias in reproductive behaviour in Georgia, Indonesia, and Vietnam: An application of the own-children method, *Population Studies* 71(3): 265–279. doi:10.1080/00324728.2017.1330489

Hathi, Payal, Sabrina Haque, Lovey Pant, Diane Coffey, and Dean Spears. 2017. Place and child health: The interaction of population density and sanitation in developing countries, *Demography* 54(1): 337–360. doi:10.1007/s13524-016-0538-y

Hauer, Mathew, and Stephanie Bohon. 2020. Causal inference in population trends: Searching for demographic anomalies in big data, *SocArXiv*. doi:10.31235/osf.io/xn2v9

Hauer, Mathew E., and Alexis R. Santos-Lozada. 2021. Differential privacy in the 2020 census will distort COVID-19 rates, *Socius: Sociological Research for a Dynamic World* 7(January): 237802312199401. doi:10.1177/2378023121994014

Hauser, Robert M., and David Weir. 2010. Recent developments in longitudinal studies of aging in the United States, *Demography* 47(1): S111–S130. doi:10.1353/dem.2010.0012

Helleringer, Stephane, Hans-Peter Kohler, Agnes Chimbiri, Praise Chatonda, and James Mkandawire. 2009. The Likoma Network Study: Context, data collection and initial results, *Demographic Research* 21(15): 427–468. doi:10.4054/DemRes.2009.21.15

Helleringer, Stephane, Chong You, Laurence Fleury, Laetitia Douillot, Insa Diouf, Cheikh Tidiane Ndiaye, Valerie Delaunay et al. 2019. Improving age measurement in low- and middle-income countries through computer vision: A test in Senegal, *Demographic Research* 40(9): 219–260. doi:10.4054/DemRes.2019.40.9

Henry, Louis. 1967. *Manuel de Démographie Historique.*

Hertrich, Véronique, Pascaline Feuillet, Olivia Samuel, Assa Doumbia Gakou, and Aurélien Dasré. 2020. Can we study the family environment through census data? A comparison of households, dwellings, and domestic units in rural Mali, *Population Studies* 74(1): 119–138. doi:10.1080/00324728.2019.1694166

HFD. 2021. Available: www.humanfertility.org (accessed: 30 July 2021).

Hilbert, Martin, and Priscila López. 2011. The world's technological capacity to store, communicate, and compute information, *Science* 332(6025): 60–65. doi:10.1126/science.1200970

Hill, Kenneth, Peter Johnson, Kavita Singh, Anthony Amuzu-Pharin, and Yagya Kharki. 2018. Using census data to measure maternal mortality: A review of recent experience, *Demographic Research* 39 (December): 337–364. doi:10.4054/DemRes.2018.39.11

Hill, Kenneth, Alan D. Lopez, Kenji Shibuya, and Prabhat Jha. 2007. Interim measures for meeting needs for health sector data: Births, deaths, and causes of death, *The Lancet* 370(9600): 1726–1735. doi:10.1016/S0140-6736(07)61309-9

HMD. 2021. Available: www.mortality.org (accessed: 30 July 2021).

Holmlund, Helena, Helmut Rainer, and Thomas Siedler. 2013. Meet the parents? Family size and the geographic proximity between adult children and older mothers in Sweden, *Demography* 50(3): 903–931. doi:10.1007/s13524-012-0181-1

Hosegood, Victoria, and Ian Timæus. 2006. Household composition and dynamics in KwaZulu Natal, South Africa: Mirroring social reality in longitudinal data collection, in Etienne Van De Walle (ed), *African Households: Censuses and Surveys*. New York: M.E. Sharpe, pp. 98–117.

IAEG. 2014. *A World That Counts: Mobilising the Data Revolution for Sustainable Development.* Available: https://www.undatarevolution.org/report/.

ITU. 2020. *Measuring Digital Development: Facts and Figures 2020.* Available: https://www.itu.int/en/ITU-D/Statistics/Dashboards/Pages/IFF.aspx.

IUSSP. 2015. The IUSSP on a data revolution for development, *Population and Development Review* 41(1): 172–177. doi:10.1111/j.1728-4457.2015.00041.x

IUSSP. 2021a. Available: https://iussp.org/en/panel/digital-demography (accessed: 30 July 2021).

IUSSP. 2021b. Available: https://iussp.org/fr/data-collection-during-pandemic-global-south-challenges-and-opportunities (accessed: 30 July 2021).

Jasilioniene, Aiva, Tomáš Sobotka, Dmitri A. Jdanov, Kryštof Zeman, Dora Kostova, Evgeny M. Andreev, Pavel Grigoriev et al. 2016. Data resource profile: The Human Fertility Database, *International Journal of Epidemiology* 45(4): 1077–1078. doi:10.1093/ije/dyw135

Kalimeri, Kyriaki, Mariano G. Beiro, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2020. Traditional versus Facebook-based surveys: Evaluation of biases in self-reported demographic and psychometric information, *Demographic Research* 42: 133–148. doi:10.4054/DemRes.2020.42.5

Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits et al. 2018. Quantitative analysis of population-scale family trees with millions of relatives, *Science* 360(6385): 171–175. doi:10.1126/science.aam9309

Kashyap, Ridhi, Francesco C. Billari, Nicolo Cavalli, Eric Qian, and Ingmar Weber. 2017. *Ultrasound Technology and "Missing Women" in India: Analyses and now-casts based on Google searches.* IUSSP Conference Paper, Cape Town. Available: https://iussp.confex.com/iussp/ipc2017/mediafile/Presentation/Paper4760/IUSSP.pdf.

Kashyap, Ridhi, Masoomali Fatehkia, Reham Al Tamime, and Ingmar Weber. 2020. Monitoring global digital gender inequality using the online populations of Facebook and Google, *Demographic Research* 43: 779–816. doi:10.4054/DemRes.2020.43.27

Kashyap, Ridhi, and Francisco Villavicencio. 2016. The dynamics of son preference, technology diffusion, and fertility decline underlying distorted sex ratios at birth: A simulation approach, *Demography* 53(5): 1261–1281. doi:10.1007/s13524-016-0500-z

Kishor, Sunita, and Kiersten Johnson. 2006. Reproductive health and domestic violence: Are the poorest women uniquely disadvantaged? *Demography* 43(2): 293–307. doi:10.1353/dem.2006.0014

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* Thousand Oaks, CA: Sage.

Kiti, Moses C., Michele Tizzoni, Timothy M. Kinyanjui, Dorothy C. Koech, Patrick K. Munywoki, Milosch Meriac, Luca Cappa et al. 2016. Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors, *EPJ Data Science* 5(1): 1–21. doi:10.1140/epjds/s13688-015-0062-0

Klabunde, Anna, Sabine Zinn, Frans Willekens, and Matthias Leuchter. 2017. Multistate modelling extended by behavioural rules: An application to migration, *Population Studies* 71(S1): 51–67. doi:10.1080/00324728.2017.1350281

Klüsener, Sebastian, Martin Dribe, and Francesco Scalone. 2019. Spatial and social distance at the onset of the fertility transition: Sweden, 1880–1900, *Demography* 56(1): 169–199. doi:10.1007/s13524-018-0737-9

Kohler, Hans-Peter, Jere R. Behrman, and Susan C. Watkins. 2001. The density of social networks and fertility decisions: Evidence from South Nyanza District, Kenya, *Demography* 38(1): 43–58. doi:10.1353/dem.2001.0005

Kolk, Martin. 2014. Multigenerational transmission of family size in contemporary Sweden, *Population Studies* 68(1): 111–129. doi:10.1080/00324728.2013.819112

Kowal, Paul, Somnath Chatterji, Nirmala Naidoo, Richard Biritwum, Wu Fan, Ruy Lopez Ridaura, Tamara Maximova et al. 2012. Data resource profile: The World Health Organization Study on Global AGEing and Adult Health (SAGE), *International Journal of Epidemiology* 41(6): 1639–1649. doi:10.1093/ije/dys210

Kreuter, Frauke, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. 2020. Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent, *Social Science Computer Review* 38 (5): 533–549. doi:10.1177/0894439318816389

Kukutai, Tahu, Victor Thompson, and Rachael McMillan. 2015. Whither the census? Continuity and change in census methodologies worldwide, 1985–2014, *Journal of Population Research* 32(1): 3–22. doi:10.1007/s12546-014-9139-z

Lai, Shengjie, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Nick W. Ruktanonchai, Alessandro Sorichetta, Jessica Steele, Tracey Li et al. 2019. Exploring the use of mobile phone data for national migration statistics, *Palgrave Communications* 5(1): 1–10. doi:10.1057/s41599-019-0242-9

Laney, Douglas. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note 6.

Lansley, Guy, Wen Li, and Paul A. Longley. 2019. Creating a linked consumer register for granular demographic analysis, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(4): 1587–1605. doi:10.1111/rssa.12476

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: Traps in big data analysis, *Science* 343(6176): 1203–1205. doi:10.1126/science.1248506

Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis et al. 2009. Life in the network: The coming age of computational social science, *Science (New York, N.Y.)* 323(5915): 721–723. doi:10.1126/science.1167742

Lazer, David, and Jason Radford. 2017. Data ex machina: Introduction to big data, *Annual Review of Sociology* 43(1): 19–39. doi:10.1146/annurev-soc-060116-053457

Leasure, Douglas R., Warren C. Jochem, Eric M. Weber, Vincent Seaman, and Andrew J. Tatem. 2020. National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty, *Proceedings of the National Academy of Sciences* 117(39): 24173–24179. doi:10.1073/pnas.1913050117

Leone, Tiziana, Ernestina Coast, Sonia Correa, and Clare Wenham. 2021. Web-based searching for abortion information during health emergencies: A case study of Brazil during the 2015/2016 Zika outbreak, *Sexual and Reproductive Health Matters* 29(1): 1883804. doi:10.1080/26410397.2021.1883804

Letouzé, Emmanuel, and Johannes Jutting. 2015. *Official statistics, Big Data and Human Development*. Data-Pop Alliance White Paper Series. New York.

Lloyd, Christopher T., Alessandro Sorichetta, and Andrew J. Tatem. 2017. High resolution global gridded data for use in population studies, *Scientific Data* 4(1): 170001. doi:10.1038/sdata.2017.1

Longley, Paul, James Cheshire, and Alexander Singleton. 2018. *Consumer Data Research*. UCL Press.

Lyngstad, Torkild Hovde. 2011. Does community context have an important impact on divorce risk? A fixed-effects study of twenty Norwegian first-marriage cohorts, *European Journal of Population / Revue Européenne de Démographie* 27(1): 57–77. doi:10.1007/s10680-010-9226-6

Magadi, Monica Akinyi, Eliya Msiyaphazi Zulu, and Martin Brockerhoff. 2003. The inequality of maternal health care in urban sub-Saharan Africa in the 1990s, *Population Studies* 57(3): 347–366. doi:10.1080/0032472032000137853

Mahapatra, Prasanta, Kenji Shibuya, Alan D. Lopez, Francesca Coullare, Francis C. Notzon, Chalapati Rao, and Simon Szreter. 2007. Civil registration systems and vital statistics: Successes and missed opportunities, *The Lancet* 370(9599): 1653–1663. doi:10.1016/S0140-6736(07)61308-7

Masquelier, Bruno. 2013. Adult mortality from sibling survival data: A reappraisal of selection biases, *Demography* 50(1): 207–228. doi:10.1007/s13524-012-0149-1

Mencarini, Letizia, Delia Irazú Hernández-Farías, Mirko Lai, Viviana Patti, Emilio Sulis, and Daniele Vignoli. 2019. Happy parents' tweets: An exploration of Italian Twitter data using sentiment analysis, *Demographic Research* 40: 693–724. doi:10.4054/DemRes.2019.40.25

Mikkelsen, Lene, David E. Phillips, Carla AbouZahr, Philip W. Setel, Don de Savigny, Rafael Lozano, and Alan D. Lopez. 2015. A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress, *The Lancet* 386(10001): 1395–1406. doi:10.1016/S0140-6736(15)60171-4

Mills, Melinda C., and Charles Rahal. 2019. A scientometric review of genome-wide association studies, *Communications Biology* 2(1): 1–11. doi:10.1038/s42003-018-0242-0

Mills, Melinda C., and Felix C. Tropf. 2020. Sociology, genetics, and the coming of age of sociogenomics, *Annual Review of Sociology* 46(1): 553–581. doi:10.1146/annurev-soc-110619-031647

Monti, Andrea, Sven Drefahl, Eleonora Mussino, and Juho Härkönen. 2019. Over-coverage in population

registers leads to bias in demographic estimates, *Population Studies* 74(3): 451–469. doi:10.1080/00324728.2019.1683219

Mostafa, Tarek. 2016. Variation within households in consent to link survey data to administrative records: Evidence from the UK Millennium Cohort Study, *International Journal of Social Research Methodology* 19(3): 355–375. doi:10.1080/13645579.2015.1019264

Mostafa, Tarek, and Richard Wiggins. 2015. The impact of attrition and non-response in birth cohort studies: A need to incorporate missingness strategies, *Longitudinal and Life Course Studies* 6(2): 131–146. doi:10.14301/llcs.v6i2.312

Moultrie, Thomas A. 2016. Demography, demographers and the "data revolution" in Africa, *Afrique Contemporaine No 258* 2: 25–239. doi:10.3917/afco.258.0025

Moultrie, Tom A., Takudzwa S. Sayi, and Ian M. Timæus. 2012. Birth intervals, postponement, and fertility decline in Africa: A new type of transition? *Population Studies* 66(3): 241–258. doi:10.1080/00324728.2012.701660

Obermeyer, Ziad, Julie Knoll Rajaratnam, Chang H. Park, Emmanuela Gakidou, Margaret C. Hogan, Alan D. Lopez, and Christopher J. L. Murray. 2010. Measuring adult mortality using sibling survival: A new analytical method and new results for 44 countries, 1974–2006, *PLOS Medicine* 7(4): e1000260. doi:10.1371/journal.pmed.1000260

Oberski, Daniel L., and Frauke Kreuter. 2020. Differential privacy and social science: An urgent puzzle, *Harvard Data Science Review* 2: 1. doi:10.1162/99608f92.63a22079

Odimegwu, Clifford, Vesper H. Chisumpa, and Oluwaseyi Dolapo Somefun. 2018. Adult mortality in sub-Saharan Africa using 2001‒2009 census data: Does estimation method matter? *Genus* 74(1): 10. doi:10.1186/s41118-017-0025-3

OECD. 2013. *New Data for Understanding the Human Condition – OECD*. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. Available: https://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf.

Oliver, Nuria, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle, *Science Advances* 6(23): eabc0764. doi:10.1126/sciadv.abc0764.

O'Neill, Dara, Michaela Benzeval, Andy Boyd, Lisa Calderwood, Cyrus Cooper, Louise Corti, Elaine Dennison et al. 2019. Data resource profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER), *International Journal of Epidemiology* 48 (3): 675–676. doi:10.1093/ije/dyz004

OPAL. 2021. Available: https://www.opalproject.org/home-en (accessed: 30 July 2021).

Østby, Gudrun, Henrik Urdal, Andreas Forø Tollefsen, Andreas Kotsadam, Ragnhild Belbo, and Christin Ormhaug. 2018. Organized violence and institutional child delivery: Micro-level evidence from Sub-Saharan Africa, 1989–2014, *Demography* 55(4): 1295–1316. doi:10.1007/s13524-018-0685-4.

Palmer, John R. B., Thomas J. Espenshade, Frederic Bartumeus, Chang Y. Chung, Necati Ercan Ozgencil, and Kathleen Li. 2013. New approaches to human mobility: Using mobile phones for demographic research, *Demography* 50(3): 1105–1128. doi:10.1007/s13524-012-0175-z

Palotti, Joao, Natalia Adler, Alfredo Morales-Guzman, Jeffrey Villaveces, Vedran Sekara, Manuel Garcia Herranz, Musa Al-Asad et al. 2020. Monitoring of the Venezuelan exodus through Facebook's advertising platform, *PLOS ONE* 15(2): e0229175. doi:10.1371/journal.pone.0229175

Parsons, Christopher R., Ronald Skeldon, Terrie L. Walmsley, and L. Alan Winters. 2007. Quantifying international migration, a database of bilateral stocks, in Çağlar Özden and Maurice Schiff (eds), *International Migration, Economic Development and Policy*, Washington, The World Bank, pp. 17–58.

Permanyer, Iñaki. 2013. Using census data to explore the spatial distribution of human development, *World Development* 46(June): 1–13. doi:10.1016/j.worlddev.2012.11.015

Pesando, Luca Maria. 2019. Global family change: Persistent diversity with development, *Population and Development Review* 45(1): 133–168. doi:10.1111/padr.12209

Polos, Jessica, and Jason Fletcher. 2019. Caesarean section and children's health: A quasi-experimental design, *Population Studies* 73(3): 353–368. doi:10.1080/00324728.2019.1624810

Potârcă, Gina, and Melinda Mills. 2015. Racial preferences in online dating across European countries, *European Sociological Review* 31(3): 326–341. doi:10.1093/esr/jcu093

Preston, Samuel H. 1996. Population studies of mortality, *Population Studies* 50(3): 525–536. doi:10.1080/0032472031000149596

Pullum, Thomas W. 2019. *Strategies to assess the quality of DHS data*, September. Available: https://dhsprogram.com/publications/publication-MR26-Methodological-Reports.cfm.

Pullum, Thomas W., and Stan Becker. 2014. *Evidence of omission and displacement in DHS birth histories*, September. Available: https://www.dhsprogram.com/publications/publication-mr11-methodological-reports.cfm.

Pullum, Thomas W., Bruno Schoumaker, Stan Becker, and Sarah E. K. Bradley. 2013. *An assessment of DHS estimates of fertility and under-five mortality.* IUSSP Conference Paper. Available: https://iussp.org/sites/default/files/event_call_for_papers/Pullum_Schoumaker_Becker_Bradley_IUSSP_2013.pdf.

Pullum, Thomas W., and Sarah Staveteig. 2017. *An assessment of the quality and consistency of age and date reporting in DHS surveys, 2000–2015*, August. Available: https://www.dhsprogram.com/publications/publication-MR19-Methodological-Reports.cfm.

Rampazzo, Francesco, Jakub Bijak, Agnese Vitali, Ingmar Weber, and Emilio Zagheni. 2021. A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom, *Demography*.

Rampazzo, Francesco, Emilio Zagheni, Ingmar Weber, Maria Rita Testa, and Francesco Billari. 2018. *Mater certa est, pater numquam: What can Facebook advertising data tell us about male fertility rates?* Twelfth International AAAI Conference on Web and Social Media. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17891.

Rana, Md Juel, John Cleland, T. V. Sekher, and Sabu S. Padmadas. 2021. Disentangling the effects of reproductive behaviours and fertility preferences on child growth in India, *Population Studies* 75(1): 37–50. doi:10.1080/00324728.2020.1826564

Randall, Sara, and Ernestina Coast. 2015. Poverty in African households: The limits of survey and census representations, *The Journal of Development Studies* 51(2): 162–177. doi:10.1080/00220388.2014.968135

Randall, Sara, Ernestina Coast, and Tiziana Leone. 2011. Cultural constructions of the concept of household in sample surveys, *Population Studies* 65(2): 217–229. doi:10.1080/00324728.2011.576768

Reid, Alice, Hannaliis Jaadla, Eilidh Garrett, and Kevin Schürer. 2020. Adapting the own children method to allow comparison of fertility between populations with different marriage regimes, *Population Studies* 74(2): 197–218. doi:10.1080/00324728.2019.1630563

Reis, Ben Y., and John S. Brownstein. 2010. Measuring the impact of health policies using internet search patterns: The case of abortion, *BMC Public Health* 10(1): 514. doi:10.1186/1471-2458-10-514

Rodríguez-Vignoli, Jorge, and Francisco Rowe. 2018. How is internal migration reshaping metropolitan populations in Latin America? A new method and new evidence, *Population Studies* 72(2): 253–273. doi:10.1080/00324728.2017.1416155

Rosenfeld, Michael J. 2010. Nontraditional families and childhood progress through school, *Demography* 47(3): 755–775. doi:10.1353/dem.0.0112

Rotondi, Valentina, Ridhi Kashyap, Luca Maria Pesando, Simone Spinelli, and Francesco C. Billari. 2020. Leveraging mobile phones to attain sustainable development, *Proceedings of the National Academy of Sciences* 117(24): 13413–13420. doi:10.1073/pnas.1909326117

Ruggles, Steven. 2014. Big microdata for population research, *Demography* 51(1): 287–297. doi:10.1007/s13524-013-0240-2

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. 2019. Differential privacy and census data: Implications for social and economic research, *AEA Papers and Proceedings* 109(May): 403–408. doi:10.1257/pandp.20191107

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. Historical census record linkage, *Annual Review of Sociology* 44(1): 19–37. doi:10.1146/annurev-soc-073117-041447

Ruggles, Steven, Evan Roberts, Sula Sarkar, and Matthew Sobek. 2011. The North Atlantic Population Project: Progress and prospects, *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 44(1): 1–6. doi:10.1080/01615440.2010.515377

Sagiroglu, Seref, and Duygu Sinanc. 2013. Big data: A review, *2013 International Conference on Collaboration Technologies and Systems (CTS)* 42–47. doi:10.1109/CTS.2013.6567202

Saha, Unnati Rani, and Arthur van Soest. 2011. Infant death clustering in families: Magnitude, causes, and the influence of better health services, Bangladesh: 1982–2005, *Population Studies* 65(3): 273–287. doi:10.1080/00324728.2011.602100

Sala, Emanuela, Jonathan Burton, and Gundi Knies. 2012. Correlates of obtaining informed consent to data linkage: Respondent, interview, and interviewer characteristics, *Sociological Methods & Research* 41(3): 414–439. doi:10.1177/0049124112457330

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age.* Princeton: Princeton University Press.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration, *Proceedings of the National Academy of Sciences* 117(15): 8398–8403. doi:10.1073/pnas.1915006117

Sankoh, Osman, and Peter Byass. 2012. The INDEPTH network: Filling vital gaps in global epidemiology, *International Journal of Epidemiology* 41(3): 579–588. doi:10.1093/ije/dys081

Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Verdery. 2020. How differential privacy will affect our understanding of health disparities in the United States, *Proceedings of the National Academy of Sciences* 117(24): 13405–13412. doi:10.1073/pnas.2003714117

Schmertmann, Carl P., Suzana M. Cavenaghi, Renato M. Assunção, and Joseph E. Potter. 2013. Bayes plus

Brass: Estimating total fertility for many small areas from sparse census data, *Population Studies* 67(3): 255–273. doi:10.1080/00324728.2013.795602

Schneider, Daniel, and Kristen Harknett. 2019. What's to like? Facebook as a tool for survey data collection, *Sociological Methods & Research*. doi:10.1177/0049124119882477

Schoumaker, Bruno. 2013. A Stata module for computing fertility rates and TFRs from birth histories: TFR2, *Demographic Research* 28: 1093–1144. doi:10.4054/DemRes.2013.28.38

Schoumaker, Bruno. 2017. Measuring male fertility rates in developing countries with Demographic and Health Surveys: An assessment of three methods, *Demographic Research* 36: 803–850. doi:10.4054/DemRes.2017.36.28

Schoumaker, Bruno, and Sarah R. Hayford. 2004. A person-period approach to analysing birth histories, *Population (English Edition)* 59(5): 689–702. doi:10.3917/pope.405.0689

Schürer, K., and E. Higgs. 2014. *Integrated Census Microdata (I-CeM)*, *1851-1911*. [Data Collection]. UK Data Service. SN: 7481.

Short Fabic, Madeleine, YoonJoung Choi, and Sandra Bird. 2012. A systematic review of Demographic and Health Surveys: Data availability and utilization for research, *Bulletin of the World Health Organization* 90 (8): 604–612. doi:10.2471/BLT.11.095513

Skeldon, Ronald. 2018. International migration, internal migration, mobility and urbanization: Towards more integrated approaches, *UN*. doi:10.18356/a11581d8-en

Song, Shige, and Sarah A. Burgard. 2008. Does son preference influence children's growth in height? A comparative study of Chinese and Filipino children, *Population Studies* 62(3): 305–320. doi:10.1080/00324720802313553

State, Bogdan, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. 2014. Migration of professionals to the U.S., in Luca Maria Aiello and Daniel McFarland (eds), *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings.* Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 531–543. doi:10.1007/978-3-319-13734-6_37

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data, *PLOS ONE* 10(2): e0107042. doi:10.1371/journal.pone.0107042

Stier, Sebastian, Johannes Breuer, Pascal Siegers, and Kjerstin Thorson. 2020. Integrating survey data and digital trace data: Key issues in developing an emerging field, *Social Science Computer Review* 38(5): 503–516. doi:10.1177/0894439319843669

Tamgno, James K., Roger M. Faye, and Claude Lishou. 2013. *Verbal autopsies, mobile data collection for monitoring and warning causes of deaths.* 2013 15th International Conference on Advanced Communications Technology (ICACT), pp. 495–501.

Tate, A. Rosemary, Lisa Calderwood, Carol Dezateux, and Heather Joshi. 2006. Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study, *International Journal of Epidemiology* 35(2): 294–298. doi:10.1093/ije/dyi287

Timæus, Ian M., and Momodou Jasseh. 2004. Adult mortality in Sub-Saharan Africa: Evidence from Demographic and Health Surveys, *Demography* 41(4): 757–772. doi:10.1353/dem.2004.0037

Tolonen, Hanna, Satu Helakorpi, Kirsi Talala, Ville Helasoja, Tuija Martelin, and Ritva Prättälä. 2006. 25-year trends and socio-demographic differences in response rates: Finnish Adult Health Behaviour Survey, *European Journal of Epidemiology* 21(6): 409–415. doi:10.1007/s10654-006-9019-8

Torche, Florencia. 2011. The effect of maternal stress on birth outcomes: Exploiting a natural experiment, *Demography* 48(4): 1473–1491. doi:10.1007/s13524-011-0054-z

Townsend, Nicholas, Sangeetha Madhavan, Stephen Tollman, Michel Garenne, and Kathleen Kahn. 2002. Children's residence patterns and educational attainment in rural South Africa 1997, *Population Studies* 56(2): 215–225. doi:10.1080/00324720215925

Tropf, Felix C., S. Hong Lee, Renske M. Verweij, Gert Stulp, Peter J. van der Most, Ronald de Vlaming, Andrew Bakshi et al. 2017. Hidden heritability due to heterogeneity across seven populations, *Nature Human Behaviour* 1(10): 757–765. doi:10.1038/s41562-017-0195-1

Turra, Cassio M., Noreen Goldman, Christopher L. Seplaki, Dana A. Glei, Yu-Hsuan Lin, and Maxine Weinstein. 2005. Determinants of mortality at older ages: The role of biological markers of chronic disease, *Population and Development Review* 31(4): 675–698. doi:10.1111/j.1728-4457.2005.00093.x

UN Department of Economic and Social Affairs. 2019. *International Migrant Stock 2019*. Available: https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp.

UNECE. 2021. *Censuses of the 2020 Round.* Available: https://statswiki.unece.org/display/censuses/Censuses+of+the+2020+round (accessed: 30 July 2021).

United Nations Statistics Division. 2013. *Overview of National Experiences for Population and Housing Censuses of the 2010 Round.* Available: https://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10706.aspx.

United Nations Statistics Division. 2020. *Report on the Results of the UNSD Survey on 2020 Round Population and Housing Censuses.*

Vaupel, James W., Kenneth G. Manton, and Eric Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* 16(3): 439–454. doi:10.2307/2061224

Wardrop, N. A., W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson et al. 2018. Spatially disaggregated population estimates in the absence of national population and housing census data, *Proceedings of the National Academy of Sciences* 115(14): 3529–3537. doi:10.1073/pnas.1715305115

Watkins, Susan Cotts. 1993. If all we knew about women was what we read in *Demography*, what would we know? *Demography* 30(4): 551–577. doi:10.2307/2061806

Wheldon, Mark C., Adrian E. Raftery, Samuel J. Clark, and Patrick Gerland. 2016. Bayesian population reconstruction of female populations for less developed and more developed countries, *Population Studies* 70(1): 21–37. doi:10.1080/00324728.2016.1139164

Wilde, Joshua, Wei Chen, and Sophie Lohmann. 2020. *COVID-19 and the future of US fertility: What can we learn from Google?* Working Paper 13776. IZA Discussion Papers. Available: https://www.econstor.eu/handle/10419/227303.

Willekens, Frans, Jakub Bijak, Anna Klabunde, and Alexia Prskawetz. 2017. The science of choice: An introduction, *Population Studies* 71(S1): 1–13. doi:10.1080/00324728.2017.1376921

Wilson, Robin, Elisabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power, Simon Tudge, Miguel Gonzalez, Sam Guthrie et al. 2016. Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake, *PLoS Currents*. doi:10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c

Wrigley, E. A., and R. S. Schofield. 1983. English population history from family reconstitution: Summary results 1600–1799, *Population Studies* 37(2): 157–184. doi:10.1080/00324728.1983.10408745

Yildiz, Dilek, Jo Munson, Agnese Vitali, Ramine Tinati, and Jennifer A. Holland. 2017. Using Twitter data for demographic research, *Demographic Research* 37(46): 1477–1514. doi:10.4054/DemRes.2017.37.46

Zagheni Emilio, and Ingmar Weber. 2012. You are where you e-mail: Using e-mail data to estimate international migration rates, in *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci'12. Evanston, Illinois: Association for Computing Machinery, pp. 348–351. doi:10.1145/2380718.2380764

Zagheni, Emilio, and Ingmar Weber. 2015. Demographic research with non-representative internet data, *International Journal of Manpower* 36(1): 13–25. doi:10.1108/IJM-12-2014-0261

Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging Facebook's advertising platform to monitor stocks of migrants, *Population and Development Review* 43(4): 721–734. doi:10.1111/padr.12102