# Digital Demography

Ingmar Weber
Qatar Computing Research Institute, HBKU
P.O. 5825, Doha
Qatar
iweber@hbku.edu.qa

Bogdan State
Stanford University
Dept. of Computer Science
U.S.
bstate@stanford.edu

## ABSTRACT

Demography is the science of human populations and, at its most basic, focuses on the processes of (i) fertility, (ii) mortality and (iii) mobility. Whereas modern states are typically in a reasonable position to keep records on both fertility and mortality, through birth and death registrations, as well as through censuses, measuring the mobility of populations represents a particular challenge due to reasons ranging from inconsistencies in official definitions across countries, to the difficulty of quantifying illegal migration. At the same time, mere numbers, whether on births, deaths or migration events, shed little light on the underlying causes, hence providing insufficient information to policy makers.

The use of digital methods and data sources, ranging from social media data to web search logs, offers possibilities to address some of the challenges of traditional demography by (i) improving existing statistics or helping to create new ones, and (ii) enriching statistics by providing context related to the drivers of demographic changes. This tutorial will help to familiarize participants with research in this area.

First, we will give an overview of fundamental concepts in demographic research including the population equation. We also showcase traditional data collection and analysis methods such as census microdata, the construction of a basic life table, panel datasets and survival analysis.

In the second part, we present a number of studies that have tried to overcome limitations of traditional approaches by using innovative methods and data sources ranging from geo-tagged tweets [14, 42] to online genealogy. We will put particular emphasis on (i) methodological challenges such as issues related to bias, as well as on (ii) how to collect open data from the World Wide Web.

The slides and other material for this tutorial are available at `https://sites.google.com/site/digitaldemography/`.

## Keywords

Demography; Social Media; Digital Methods


*WWW'17 Companion,* April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051104

## 1. INTRODUCTION

Due to the increasing availability of large-scale data on human behavior collected on the social web, as well as advances in analyzing larger and larger data sets, interest in applying computational methods to study population dynamics continues to grow [7, 22]. Data scientists entering the interdisciplinary field of Computational Social Science (CSS) often lack background in theories and methods in the social sciences, whereas social scientists are often not aware of cutting edge advances in computational methods. This problem is felt particularly acutely in the field of demography, which has shown to be one of the most promising areas for the development of novel means of answering key scientific questions using digital data and computational methods.

Demography is the science of human populations, presenting a fundamental interest to policy makers and academics alike. At its most basic, demographic changes are driven by processes of fertility, mortality and mobility. Whereas modern states are typically in a reasonable position to keep records on both fertility and mortality (through birth and death registrations, as well as through censuses), measuring the mobility of populations represents a particular challenge due to reasons ranging from inconsistencies in official definitions across countries, to the difficulty of quantifying unauthorized migration. These difficulties have led to a particular interest in migration-related studies through the lens of digital data.

Organizations such as the United Nations Global Pulse[1] or the Data-Pop Alliance[2] have been founded around the idea of using "big data analytics" for projects related to global development and demographic research is often at the heart of those projects. Furthermore, "traditional" scientific societies such as the International Union for the Scientific Study of Population have started to offer training workshops in this domain[3] and have panels dedicated to "Big Data and Population Processes"[4]. As another example, the call for papers for the Annual Meeting of the Population Association of America[5] now explicitly includes topics such as "Big Data" and "Data Science". This creates a unique opportunity for

[1] `http://unglobalpulse.org/`
[2] `http://datapopalliance.org/`
[3] `http://iussp.org/en/training-workshop-social-media-and-demographic-methods-paa-2016`
[4] `http://iussp.org/en/panel/big-data-and-population-processes`
[5] `http://www.populationassociation.org/wp-content/uploads/PAA2017CallforPapers.pdf`

the WWW community to apply their skills to a new topic domain and to help shape the future of a new discipline.

This tutorial exposes computer scientists to digital demography by presenting (i) an overview of what demographic research is, (ii) data sets and methods traditionally used, (iii) novel data sets and computational approaches, and (iv) opportunities for future advances in this area. The goal of this tutorial is to give participants a rich repertoire of research questions, data sets, and methods that help to address challenges related to demographic changes.

The tutorial will run as a half-day event in the morning of April 4, 2017, and consists of two parts, explained in more detail in the following.

## 2. MOTIVATION

We will start with a general overview of what demography, as the science of human populations, entails. We will put particular emphasis on describing how the most fundamental processes of fertility, mortality and mobility are traditionally quantified. The first part of our tutorial will discuss the population equation as the fundamental demographic identity. We will also showcase census microdata, the construction of a basic life table, panel datasets and methods, survival analysis, as well as count models. Participants will be exposed to shortcomings of these traditional methods, particularly for issues related to mobility and generating estimates for subpopulations. We will also communicate to participants the ways in which digital data transgresses existing limitations, particularly for measurements related to mobility.

**The population equation** represents the fundamental identity of demography:

$$\Delta P = (B + I) - (D + O)$$

Here $\Delta P$ represents the change in population for a territory over a period of time, $B$ and $D$ stand for the number of births and deaths in the territory, while $I$ and $O$ quantify the number of in- and out- migrants into and out of the territory. From this equation emerge the three main areas of demographic research: fertility, mortality and migration.

**Traditional data sources** for the study of demography are censuses and population registers. Population censuses, usually conducted every decade or every five years, have become a nearly-universal data collection procedure, with more than 190 countries conducting censuses during the 2000 round of censuses [36]. In their traditional formula, population censuses involve an interview-based approach, with census-takers going door-to-door and collecting population statistics from the entire country in a short amount of time. This approach translates into the recording of a "snapshot" of the state of population. In the language of data science, censuses yield "dimension tables" for the population of a country, recording the state of the population at a point in time.

A complementary approach to this traditional, state-based method uses event data, recorded in *population registers* [36]. Especially if countries implement national identification schemes that assign unique identifiers to individuals, registers can be used to update population estimates on a continuous basis. The population register approach is not universally feasible however, given that many governments lack either comprehensive registration schemes, or do not possess the data infrastructure required to maintain consistency between census data and register events. An alternative approach involves the use of periodical small-sample surveys (such as the American Community Survey) to update census estimates. Rolling censuses and modeling-based approaches offer other hybrid data collection methods for population statistics. Modeling-based approaches are particularly promising areas for the use of digital data. In particular, digital data can help improve or substitute census estimates for population density[12], economic growth [15] or poverty [9].

Beyond the simple tabulation of census estimates, there are a number of specialized methods demographers have refined for the analysis of population data. *Life tables* are used to compute age-specific quantities such as life expectancy. *Panel methods* such as fixed and random effects regression allow for the modeling of longitudinal relationships between variables. Beyond simple data mining, modern demographic methods are often tasked with establishing causal relationships. To assess causality in the context of longitudinal data researchers must often explicitly account for endogenous relationships between variables, case in which techniques such as *difference-in-difference* specifications are particularly useful.

Demographers have also designed methods to deal with the conceptual limitations of their data. Even the best life course data inherently suffers from a problem of *right-censoring*: in any longitudinal dataset it is possible to know how many individuals survived to the present date, but one cannot observe when individuals will exit the population of interest in the future. *Survival models*, such as proportional hazards regressions or accelerated failure time models represent canonical methods for dealing with these inherent issues. The confusion between effects due to age, period, and cohort are another issue demographers must often contend with. Because age, period and cohort are in a deterministic arithmetic relationship, standard regression models break down and different statistical techniques, under the general heading of Age-Period-Cohort models must be applied [10]. Because digital data often deals with life cycles – of users, content or entities – the interaction between demography and machine learning could also be particularly fruitful to industry practitioners looking for a sophisticated framework in which to investigate issues such as user churn, acquisition or resurrection.

Traces of digital activity are not the only instance of very large datasets from which demographers may extract valuable insight. "Big microdata," either from national censuses or from birth, death or marriage records is increasingly available to researchers [31]. The IPUMS project provides researchers from microdata from 250 censuses conducted in 79 countries [32]. Demographers are also already contending with the challenges of large-scale data management required to integrate large samples of microdata in the Terra Populus project [20]. Digital researchers may thus also benefit from learning about the complex management strategies adopted by demographers.

## 3. NEW OPPORTUNITIES

Studying entire populations at the individual level opens up new possibilities for the enrichment of demographic data. Digital datasets often offer rich individual-level attributes: gender, birth cohort, education, family and social relationships, or interests. All of these can be used to produce demographic estimates for very fine-grained populations, defined by any combination of attributes, which may help better describe more fine-grained demographic trends.

We will examine several case studies focused on new opportunities created by digital data in demographic applications. In addition to the measurement of migrations [17], we will also investigate unusual opportunities such as applying data mining to obituaries to obtain estimates of mortality, or the relationship between web searches and fertility and mortality regimes. Particular emphasis will be given to methods that attempt to address challenges related to biases in online data [44].

**Geolocation data** has already shown a great deal of promise for the development of new demographic measures and for the improvement of existing ones. IP geolocation has been successfully used to measure international migrations [43, 34], while call-record data has been used in the development of socio-spatial measures of segregation [3] or intra-country migrations [28]. Other digital sources of location information that have been explored for the study of migrations include Twitter [42, 1], LinkedIn [33], Skype [19]. At a more fine-grained level, mobility patterns have also been studied using data from services such as Foursquare [27] or (now-defunct) Gowalla and BrightKite [8].

Mobility data in particular has already proven its use to epidemiologists by helping obtain more fine-grained estimates of the spread of infections diseases [41]. But digital traces may also be useful in improving estimates of poverty (e.g. by combining data collected through the Internet with estimates derived from satellite data, as per [9]). Google Street View presents another innovative dataset, having been used successfully to estimate US demographics [11].

**Life course data** In addition to digitized birth, death and marriage records, and to the increasing availability of census micro-data, other digital datasets are promising to provide even more detailed and timely information into fertility and mortality patterns. Web search data was an early data source used to gauge both abortion [30] and fertility intentions [2]. More recently, public health researchers have used online obituaries to improve estimates of cancer mortality [35]. Information about major life events has also been successfully extracted from Twitter data [23]. The mining of family ties is yet another area in which digital data has shown promise for the study of demography [6].

**Evaluation of digital data** Digital demographic datasets are vulnerable to both measurement error and selection bias. Measurement error is a reflection of the limitations of both digital sensors and of data exchange and collection protocols. Location sensors – whether based on GPS or cell-tower triangulation – have inherent reliability issues, which are distributed in a non-random fashion across the world, with higher-income individuals having access to newer-generation smartphones with improved geolocation functions. IP geolocation databases display their own reliability issues, due to the extremely complex process through which IP addresses are allocated on the Internet [29]. And self-reports from users are vulnerable to misrepresentation, both intentional and absent-minded. The large size of many digital datasets mitigates this problem, but it does not do so entirely. Measurement error becomes a concern particularly in the study of populations with rare characteristics, and researchers must be mindful of the noisy process through which digital data is generated.

Selection bias poses an even more intractable problem for the study of demography with digital data. At the core of the problem is the “digital divide” [26], a fundamental inequality of the 21st century, with individuals across the world having widely different levels of access to the Internet [4, 40, 7]. The digital divide is not binary – it operates on a continuum of access and sophistication [13], which dictates that the initial user populations of any Internet service in a country will typically be composed of its technological elite, usually a widely different group of individuals in demographic and socio-economic characteristics from the rest of the population. For instance, taking an Internet service’s Indian user-base to be representative of India’s population is a dangerous thing to do when most of the country has poor or no access to the Internet.

For circumspect demographers issues of representativity produced by selection bias may disqualify most digital data from scientific relevance, save for data with near-universal coverage, such as satellite images. This need not be the case. While heterogeneity abounds, the world’s Internet and mobile phone usage is increasing rapidly. According to the ITU [18], as of mid-2016 just under 3.5b individuals, 47% of the world’s population, were online, compared with just over 1b nine years before. Moreover, researchers are also developing methods for bias correction based on the validation of digital data against pre-existing datasets collected through traditional means, as well as against other digital datasets, when traditional data is unavailable [44].

**Combining traditional and digital datasets** may be useful not only to reduce the bias of estimates derived from digital data sources, but also to improve the reliability of statistics using traditional methods [7]. Researchers have combined call records and field surveys [5], and have shown that combining a traditional epidemiological dataset provided by the U.S. Center for Disease Control with Google Flu Trends predictions may have helped avoid the now-infamous misfiring of that previously well-regarded algorithm [21]. Digital data will fully realize its promise only if researchers are mindful of the very important limitations of these novel datasets and devise appropriate strategies for the mitigation of both noise and selection bias.

# 4. TARGET AUDIENCE

This tutorial is aimed at participants with a basic level of data mining and data processing skills. The content covered in the tutorial is designed to introduce both PhD students and researchers interested in learning or advancing their current knowledge of digital methods for demographic research.

# 5. PREREQUISITES

Participants should have the basic skills of data harvesting, processing, and analysis. The workshop will cover various forms of statistical and probabilistic analysis, and other forms of quantitative methods for interrogating social data.

# 6. RELEVANCE TO WWW

Digital Demography has been a growing area of interests for the last several years with publications related to inferring demographic attributes, migration, gender changes and related topics appearing in data mining venues. See the following link for 20 example publications with “demographi*” in the title that have been published in WWW, WSDM, CIKM, HT, ICDM or ICWSM https://goo.gl/Sozczv. Though many of these papers focus on the inference of demographic attributes, there is also a growing body of work using online data sources for studying population-level statistics, such as migration. In particular, there's been work on this topic published by the presenters at computer science conferences on using email data (WebSci'12 [43], WSDM'13 [34]), Twitter data (WWW'14 [42]), LinkedIn data (SocInfo'14 [33]), Google+ data (ASONAM'16 [25]) and Facebook data (WebSci'16 [16]). At WWW'17, the "Social Network Analysis and Computational Social Science" track is closely related to the topic of Digital Demography. Digital Demography requires an understanding of the Web and the algorithms that operate on the Web and therefore WWW is a perfect venue to attract the right people who could become interested in this area.

# 7. PREVIOUS EDITIONS

To the best of our knowledge, this is the first tutorial of its kind. However, the tutorial on Computational Social Science at WWW'16 [39][6], co-hosted by Ingmar Weber, is related, as is the workshop on Social Media and Demographic Research at ICWSM'16[7], co-hosted by Bogdan State.

# 8. PRESENTERS

*Bogdan State* is an MS candidate in Computer Science at Stanford University, where he also recently completed a PhD in Sociology. He is interested in using Internet data to decipher the basic mechanisms of human social interaction. His experience includes four years working as a data scientist in Silicon Valley. Bogdan works on the Facebook Core Data Science team, where his contributions have ranged from developing large-scale business intelligence systems to improving the performance of ranking models. Bogdan has published 10 peer-reviewed articles.

*Ingmar Weber* is a principal scientist in the Social Computing group at the Qatar Computing Research Institute. His uses large amounts of online data to study offline phenomena including international migration, societal fragmentation and lifestyle diseases. He has published over 100 peer-reviewed articles and is an ACM Distinguished Speaker[8]. He has extensive experience with tutorials including editions at WSDM'13 ("Data-driven Political Science", [38]), at CIKM'13 ("Twitter and the Real World"[9]), at WWW'16 ("Computational Social Science"[10]) as well as at summer schools ("Web Science" at RuSSIR'12[11] and "Computational Social Science" at RuSSIR'16[12]). With Yelena Mejova and Michael Macy he has edited a CUP book on "Twitter: A Digital Socioscope" [24].


# 9. ACKNOWLEDGMENTS

We would like to acknowledge our friend and long-term collaborator Emilio Zagheni. Emilio introduced Ingmar to the world of demographic research in 2011 in reaction to prior work related to web search demographics [37]. The resulting publication [43], in turn, caught Bogdan's attention and led to a follow-up study [34] and, eventually, a whole line of research. Without Emilio's initiative and guidance this would not have happened.


[6]Material at https://sites.google.com/site/csswwwtutorial/

[7]https://sites.google.com/site/smdrworkshop/2016

[8]http://www.dsp.acm.org/view_lecturer.cfm?lecturer_id=7123

[9]https://sites.google.com/site/twitterandtherealworld/home

[10]https://sites.google.com/site/csswwwtutorial/

[11]http://romip.ru/russir2012/section.php?id=122#WebSci

[12]http://romip.ru/russir2016/russir-2016-program/

# 10. REFERENCES


[1] Geo-located Twitter as proxy for global mobility patterns, author=Hawelka, Bartosz and Sitko, Izabela and Beinat, Euro and Sobolevsky, Stanislav and Kazakopoulos, Pavlos and Ratti, Carlo, journal=Cartography and Geographic Information Science, volume=41, number=3, pages=260–271, year=2014, publisher=Taylor & Francis.

[2] F. Billari, F. D'Amuri, and J. Marcucci. Forecasting births using Google. Annual Meeting of the Population Association of America, http://www.carmaconf.org/wp-content/uploads/pdfs/4301.pdf, 2013.

[3] V. Blondel, G. Krings, I. Thomas, et al. Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *Brussels Studies*, 42(4):1–12, 2010.

[4] J. Blumenstock and N. Eagle. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proc. ICTD*, 2010.

[5] J. E. Blumenstock and N. Eagle. Divided we call: disparities in access and use of mobile phones in rwanda. *Information Technologies & International Development*, 8(2), 2012.

[6] M. Burke, L. A. Adamic, and K. Marciniak. Families on facebook. In *Proc. ICWSM*, 2013.

[7] N. Cesare, H. Lee, T. McCormick, E. S. Spiro, and E. Zagheni. Promises and pitfalls of using digital traces for demographic research. 2016. https://ssrn.com/abstract=2839585.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. SIGKDD*, 2011.

[9] C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, and E. Bright. A global poverty map derived from satellite data. *Computers & Geosciences*, 35(8):1652–1660, 2009.

[10] S. E. Fienberg and W. M. Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological methodology*, 10:1–67, 1979.

[11] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. Lieberman Aiden, and F.-F. Li. Using deep learning and google street view to estimate the demographic makeup of the us. https://arxiv.org/abs/1702.06683, 2017.

[12] A. Gros and T. Tiecke. Connecting the world with better maps, Feb. 2016. https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/.

[13] E. Hargittai. Second-level digital divide: Differences in people's online skills. *First Monday*, 7(4), 2002.

[14] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[15] J. V. Henderson, A. Storeygard, and D. N. Weil. Measuring economic growth from outer space. *The American Economic Review*, 102(2):994–1028, 2012.

[16] A. Herdagdelen, B. State, L. A. Adamic, and W. Mason. The social ties of immigrant communities in the united states. In *Proc. WebSci*, 2016.

[17] C. Hughes, E. Zagheni, G. J. Abel, A. Wisniowski, A. Sorichetta, I. Weber, and A. J. Tatem. Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data. *European Commission project reports*, 2016. http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7910&type=2&furtherPubs=yes.

[18] ITU. ICT Facts and Figures, July 2016. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf.

[19] R. Kikas, M. Dumas, and A. Saabas. Explaining international migration in the skype network: The role of social network features. In *Proc. SIdEWayS*, pages 17–22, 2015.

[20] T. A. Kugler, D. C. Van Riper, S. M. Manson, D. A. Haynes II, J. Donato, and K. Stinebaugh. Terra populus: Workflows for integrating and harmonizing geospatial population and environmental data. *Journal of Map & Geography Libraries*, 11(2):180–206, 2015.

[21] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[22] E. Letouze and J. Jutting. Official statistics, big data and human development. *Data-Pop Alliance White Paper Series*, 2015. https://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf.

[23] J. Li, A. Ritter, C. Cardie, and E. H. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. 2014.

[24] Y. Mejova, I. Weber, and M. W. Macy. *Twitter: a digital socioscope*. Cambridge University Press, 2015.

[25] J. Messias, F. Benevenuto, I. Weber, and E. Zagheni. From migration corridors to clusters: The value of Google+ data for migration studies. In *Proc. ASONAM*, 2016.

[26] P. Norris. *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press, 2001.

[27] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[28] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one*, 7(6):e39253, 2012.

[29] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. Ip geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, 41(2):53–56, 2011.

[30] B. Y. Reis and J. S. Brownstein. Measuring the impact of health policies using internet search patterns: the case of abortion. *BMC Public Health*, 10(1):514, 2010.

[31] S. Ruggles. Big microdata for population research. *Demography*, 51(1):287–297, 2014.

[32] S. Ruggles, R. McCaa, M. Sobek, and L. Cleveland. The IPUMS collaboration: Integrating and disseminating the world's population microdata. *Journal of demographic economics*, 81(2):203, 2015.

[33] B. State, M. Rodriguez, D. Helbing, and E. Zagheni. Migration of professionals to the U.S. - evidence from linkedin data. In *SocInfo*, pages 531–543, 2014.

[34] B. State, I. Weber, and E. Zagheni. Studying Inter-national Mobility Through IP Geolocation. In *WSDM*, pages 265–274, 2013.

[35] G. Tourassi, H.-J. Yoon, and S. Xu. A novel web informatics approach for automated surveillance of cancer mortality trends. *Journal of biomedical informatics*, 61:110–118, 2016.

[36] UNSTATS. Population and housing censuses, Feb. 2017. https://unstats.un.org/unsd/demographic/sources/census/alternativeCensusDesigns.htm.

[37] I. Weber and C. Castillo. The demographics of web search. In *Proc. SIGIR*, 2010.

[38] I. Weber, A.-M. Popescu, and M. Pennacchiotti. Data-driven Political Science. In *Proc. WSDM*, pages 777–778, 2013.

[39] I. Weber, C. Wagner, M. Strohmaier, and L. M. Aiello. Computational social science for the world wide web (CSSW3). In *Proc. WWW*, pages 1037–1038, 2016.

[40] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee. Heterogeneous mobile phone ownership and usage patterns in kenya. *PloS one*, 7(4):e35319, 2012.

[41] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[42] E. Zagheni, V. R. K. Garimella, I. Weber, et al. Inferring international and internal migration patterns from twitter data. In *Proc. WWW*. ACM, 2014.

[43] E. Zagheni and I. Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proc. WebSci*, pages 348–351. ACM, 2012.

[44] E. Zagheni and I. Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, 2015.