**Digital Demography (Population Data and Science)**

**Course programme**

*Diego Alburez-Gutierrez*

Max Planck Institute for Demographic Research

Lab of Digital and Computational Demography

[alburezgutierrez@demogr.mpg.de](mailto:alburezgutierrez@demogr.mpg.de)

March 30 – April 3 2020

# Introduction

## Course description

Rapid increases in computational power and the growing access to the internet, social media and mobile phones use have radically changed our lives, the way we interact with each other and our behavior, including demographic choices. The digitalization of our lives has also led to the so-called “data revolution” that is transforming social sciences. In this course, participants will learn how traditional methods used in social sciences can help us make sense of new data sources, and how these new data sources may require new conceptual and methodological approaches.

## Goals

We will discuss a number of substantive topics related to the emergence of (big) data-driven discovery in the social sciences, with emphasis on population processes. By the end of the course, students will be familiar with relevant literature at the intersection of demographic research and digital social science. The main goals of the course are to introduce students

1. to recent substantive advances in the field of Digital and Computational Demography;
2. to some of the methods, approaches and tools of data science in the context of population research; and
3. to induce critical thinking about modern demographic analysis and (big) data-driven discovery

## Organisation and examination

This course consists of four lectures and one assignment. There are **required** and **optional** readings each week (all of them can be downloaded using [this link](https://github.com/alburezg/EDSD20_digital_demography/tree/master/Readings)). Only the former are obligatory but students will benefit from reading the latter before the class. There is no exam, only a final assignment which needs to be turned in at the end of the week (see section Assignment).

# Final assignment: “Big Data, Big Bias?”

In this assignment you will acquire hands-on experience in the use of crowd-sourced digital data. For this, you will use a sample of the *Familinx* database, which contains millions of genealogical records (i.e. family history data) scraped from [www.geni.com](http://www.geni.com). In this assignment, the goal is to use this data to compute simple demographic quantities and reflect on the limitations and possibilities of user-generated data.

A sample of the data, pertaining to profiles from Sweden, can be downloaded using [this link](https://github.com/alburezg/EDSD20_digital_demography/tree/master/Assignment), together with a codebook of the original data (note that not all columns are available in the sample dataset). In the sample dataset “sweden\_genealogy.csv”, each row represents a unique individual. The columns that identify the “father” and “mother” of a given row-person were added manually. The data looks like this:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **profileid** | **father** | **mother** | **birth\_year** | **death\_year** | **…** |
| 12589845 | 21489 | 3218 | 1912 | NA | … |
| 12648785 | 585202 | NA | 1658 | 1701 | … |
| 12698523 | 5545459 | 321681 | 1754 | 1755 | … |

This assignment consists of three exercises:

**Exercise 1**

Focusing on people born between 1750 and 1850, consider the following: how has lifespan developed historically in Sweden, according to the online genealogies?

* Compute the lifespan average by birth cohort and sex. For this exercise, I recommend you group birth cohorts by 15 years (e.g. 1750-1774; 1775-1799; etc.).
* Include a short description of your findings (max 200 words) and one figure that summarises them.

**Exercise 2**

What is the difference between lifespan and life expectancy? The two readings from Thursday seem to conflate both terms at points.

* Write a short paragraph (max 150 words) describing the connection between lifespan and life expectancy for a given birth cohort.
* Is it possible to evaluate this empirically using data from online genealogies? Write a short paragraph (max 250 words) indicating how you would do it
* For extra points, compute the cohort life expectancy for any given birth cohort using the genealogies

**Exercise 3**

What are potential sources of bias in the online genealogies? How can we evaluate these biases?

* Write a short paragraph (max 250 words) describing three potential sources in bias in online genealogies.
* Focus on one of the three biases identified above and provide evidence of its existence using empirical data
* Write a short paragraph with a potential solution to overcome this bias (max 200 words).

Please summarize the results of the exercises described above in a written report. Assignments are due Friday April 3 at midnight (CET). Send your assignment via email to [alburezgutierrez@demogr.mpg.de](mailto:alburezgutierrez@demogr.mpg.de) with the subject line “EDSD assignment”. Please make sure to include:

* A written report (Word document of pdf file). Call this file “[your\_surname]\_report.docx”
* The R scripts used to produce the empirical results presented in the written result. Clearly identify to which figure or calculation each part of the script refers. Call this file “[your\_surname]\_scripts.R”

# Lecture plan

## Monday March 30 2020, 11:30-13.00 - “Introduction to digital demography”

**Required readings**

Sections 2.1 to 2.3.10 of : Salganik, M. (n.d.). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press. <https://www.bitbybitbook.com/en/1st-ed/observing-behavior/observing-intro/>.

Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., and Negraia, D.V. (2019). *Demography in the Digital Era: New Data Sources for Population Research*. SocArXiv. doi:[10.31235/osf.io/24jp7](https://doi.org/10.31235/osf.io/24jp7).

**Optional readings**

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55(5):1979–1999. doi:[10.1007/s13524-018-0715-2](https://doi.org/10.1007/s13524-018-0715-2).

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE* 14(10):e0224134. doi:[10.1371/journal.pone.0224134](https://doi.org/10.1371/journal.pone.0224134).

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30(1):75–89. doi:[10.1057/jit.2015.5](https://doi.org/10.1057/jit.2015.5).

Zagheni, E. and Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. Paper presented at the 3rd Annual ACM Web Science Conference, Evanston, Illinois, 2012. doi:[10.1145/2380718.2380764](https://doi.org/10.1145/2380718.2380764).

Zagheni, E., Garimella, V.R.K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from Twitter data. Paper presented at the 23rd International Conference, Seoul, Korea, 2014. doi:[10.1145/2567948.2576930](https://doi.org/10.1145/2567948.2576930).

**Homework**

## Tuesday March 31 2020, 11:30-13.00 - “Crowd-sourced online data”

**Required readings**

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D.G., Price, A.L., and Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360(6385):171–175. doi:[10.1126/science.aam9309](https://doi.org/10.1126/science.aam9309).

Fire, M. and Elovici, Y. (2013). Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population. *arXiv:1311.4276 [cs, q-bio, stat]*. <http://arxiv.org/abs/1311.4276>.

**Optional readings**

Malmi, E., Gionis, A., and Solin, A. (2018). Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating. *arXiv:1802.06055 [physics, q-bio]*. <http://arxiv.org/abs/1802.06055>.

Antoun, C., Zhang, C., Conrad, F.G., and Schober, M.F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* 28(3):231–246.

## Wednesday April 1 2020, 11:30-13.00 - “Digital trace data”

**Required readings**

Alexander, M., Polimis, K. and Zagheni, E. (2019), The Impact of Hurricane Maria on Out‐migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45: 617-630. doi:[10.1111/padr.12289](https://doi.org/10.1111/padr.12289)

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development* 107:189–209. doi:[10.1016/j.worlddev.2018.03.007](https://doi.org/10.1016/j.worlddev.2018.03.007).

**Optional readings**

Sofia Gil’s tutorial on using the Facebook Marketing API: <https://github.com/SofiaG1l/Using_Facebook_API>

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook’s advertising platform to monitor stocks of migrants. *Population and Development Review* 43(4):721–734. doi:[10.1111/padr.12102](https://doi.org/10.1111/padr.12102).

Rampazzo, F., Zagheni, E., Weber, I., Testa, M.R., and Billari, F. (2018). Mater certa est, pater numquam: What can Facebook Advertising Data Tell Us about Male Fertility Rates? *arXiv:1804.04632 [cs]*. <http://arxiv.org/abs/1804.04632>.

**Homework**

1. Download the sample online genealogy from [this link](https://github.com/alburezg/EDSD20_digital_demography/tree/master/Assignment) and load it in R
2. Create graphs showing the distribution of years of births and years of death of the population

## Thursday April 2 2020, 15:00-17.00 - “Computational approaches”

**Required readings**

Verdery, A.M. and Margolis, R. (2017). Projections of white and black older adults without living kin in the United States, 2015 to 2060. *Proceedings of the National Academy of Sciences* 114(42):11109–11114. doi:[10.1073/pnas.1710341114](https://doi.org/10.1073/pnas.1710341114).

Zagheni, E. (2011). The Impact of the HIV/AIDS Epidemic on Kinship Resources for Orphans in Zimbabwe. *Population and Development Review* 37(4):761–783. doi:[10.1111/j.1728-4457.2011.00456.x](https://doi.org/10.1111/j.1728-4457.2011.00456.x).

**Optional readings**

Grow, A and Van Bavel, J. 2018. Agent-Based Modeling of Family Formation and Dissolution. In R. Schoen (Ed.), Analytical Family Demography (pp. 125-156). Springer Series on Demographic Methods and Population Analysis, (Vol. 47), Cham: Springer International Publishing.

Grow, A. and Van Bavel, J. (2020). The Gender Cliff in the Relative Contribution to the Household Income: Insights from Modelling Marriage Markets in 27 European Countries. *European Journal of Population*. doi:[10.1007/s10680-019-09547-8](https://doi.org/10.1007/s10680-019-09547-8).

Margolis, R. and Verdery, A.M. (2019). A Cohort Perspective on the Demography of Grandparenthood: Past, Present, and Future Changes in Race and Sex Disparities in the United States. *Demography* 56(4):1495–1518. doi:[10.1007/s13524-019-00795-1](https://doi.org/10.1007/s13524-019-00795-1).

## Friday April 3 2020 – No class

Students are encouraged to work on their assignments (see assignment description in pages 1-2).