Digital Demography (Population Data and Science)

Course syllabus

2020-2021 European Doctoral School of Demography
Barcelona, March 01-05 2021
Zoom sessions: from Monday 01 to Thursday 04 March (14:00 – 15:30 CEST)

Instructor:

Diego Alburez-Gutierrez
Max Planck Institute for Demographic Research
Lab of Digital and Computational Demography
alburezgutierrez [at] demogr.mpg.de

A. 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 the 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 (online) data-driven discovery

Online lectures

Given the current Covid-19 situation, all lectures will be held online on Zoom. Students are expected to attend all sessions remotely. Please get in touch with the course conveyor in advance if you anticipate issues with this.

Organisation and examination

This course consists of four lectures and one assignment. There are **required** and **optional** readings each week (pdf files in the intranet). 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 next section).

Please make sure to download the course's GitHub repository to your computer:
https://github.com/alburezg/EDSD21_digital_demography

B. Final assignment: Mobility patterns during the first wave of the Covid-19 pandemic in England

In this assignment you will acquire hands-on experience in the use of digital trace data using three data sources. The first is the Google Community Mobility Reports (GCMR)[1]. This is the same dataset used by Basellini et al. (2020) in a paper we will discuss in the course[2]. The GCMR reports daily mobility data in six categories of location: residential, workplaces, supermarket and pharmacy (grocery), transit, retail, and parks. Data are provided as percentage variations in number of visits or time spent in each category, relative to apre-COVID-19 baseline period, defined from January 3 to February 6, 2020. In addition to Google, other services provide digital trace human mobility datasets, such as the Facebook Movement Range Maps[3] and the Apple COVID-19 Mobility Trends Reports[4].

This assignment consists of two exercises (one required and one optional):

Exercise 1 [required]

Fig A.8. in Basellini et al. (2020) shows a comparison of the mobility indicators provided by Apple, Facebook, and Google for England and Wales in weeks 1–33 of 2020. Your tasks are:

a. Download the raw mobility data from Google, Facebook, and Apple (links in footnotes).
b. Replicate Fig A.8. in Basellini et al. (2020). Don't worry about replicating the specific format of Fig A.8. (shapes, layout, colours, etc.)! The idea is to compare changes in mobility using three independent sources (Google, Apple, and Facebook). You can do this in a single plot or using three different plot (one per data source) if that is easier.
c. Discuss: Are the time-series from Google, Facebook, and Apple comparable? Argue with reference to the documentation of each data source. (max 250 words)

[1] https://www.google.com/covid19/mobility/
[2] Basellini, U., Alburez-Gutierrez, D., Del Fava, E., Perrotta, D., Bonetti, M., Camarda, C. and Zagheni, E. (2020). "Linking excess mortality to Google mobility data during the COVID-19 pandemic in England and Wales". SocArxiv. DOI: 10.31235/osf.io/75d6m
[3] https://data.humdata.org/dataset/movement-range-maps
[4] https://covid19.apple.com/mobility

Important things to keep in mind for this exercise:

- The sign of the residential category of Google is reversed in Fig A.8.
- The Google data gives estimates at the Local Authority level for the UK. You need to group these estimates by country (England and Wales). I have added two files in GitHub under "Assignment/Data" that can help you aggregate local authorities into regions or nations ("uk_regions.csv" and "uk_nations.csv").
- The estimates in Fig A.8. are weekly averages! If you plot the daily estimates, the lines will be much more wiggly. Remember to group the estimates by week.
    - The package `lubridate` can be useful for transforming text strings into dates in R.
    - The `cut` function (base R) can be useful for transforming days into weeks (type `?cut` in R).
- The Facebook data needs some restructuring. The "Movement Range Maps" data includes many columns that you will not use. If you have issues reading the (large) datafile to R, I suggest using the function `fread` from the `data.table` package.
- The Apple data also needs restructuring. The function ` pivot_longer` from the `tidyr` package may come in handy for transforming wide data into a long format.

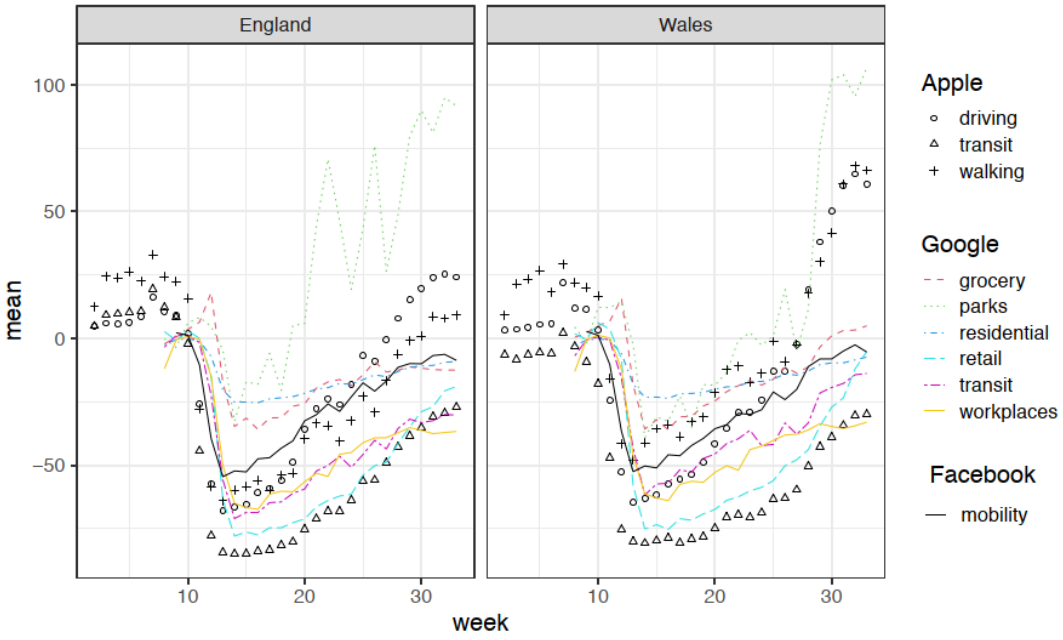

Fig A.8. Comparison of mobility indicators provided by Apple (categories driving, transit and walking), Facebook (category mobility) and Google (categories residential, workplaces, grocery, transit, retail and parks) for England and Wales in weeks 1--33 of 2020. See Basellini et al (2021) for details.

Exercise 2 [OPTIONAL!!]

Fig A.9. in Basellini et al. (2020) shows the share of missing population in the Google data (GCMR) by region, week and category (residential, workplaces, grocery, transit, retail and parks) for ten regions in England and Wales in weeks 8–33 of 2020. Your tasks are:

d. Replicate Fig A.9. from Basellini et al. (2020) using data from the GCMR exclusively (no need to use Facebook or Apple data)
e. Discuss: Why is it important to visualize missing values in the GCMR data? What do we learn from this visual analysis? (max 250 words)

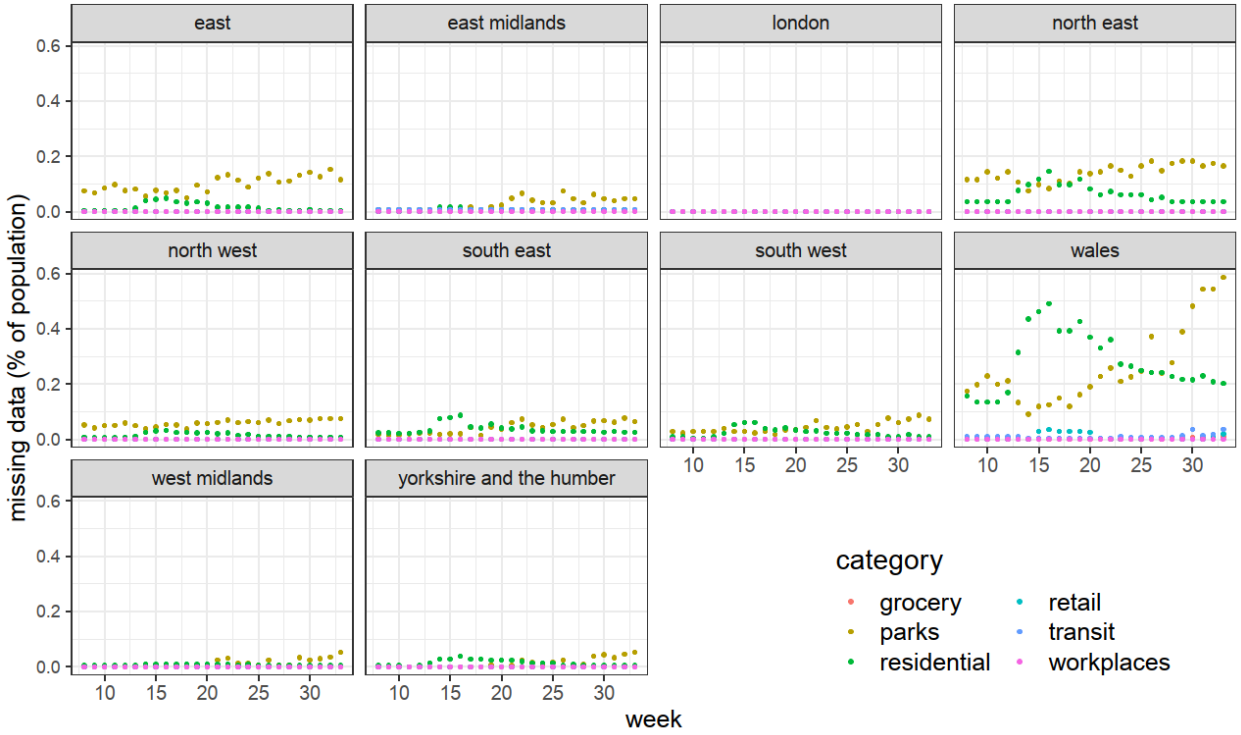

Fig A.9. Share of missing population in the GCMR by region, week and category (residential, workplaces, grocery, transit, retail and parks) for ten regions in England and Wales in weeks 8--33 of 2020. See Basellini et al (2021) for details.

Submitting the assignment

Summarise the results of the exercises in a written report using Rmarkdown using the template provided under "Assignment/R" in the course's GitHub repository (see above). When you submit, please make sure to include the .Rmd and the. pdf files, writing your surname in the file names (e.g., "lopez.Rmd"). Each file must include (1) the R code used to produce the empirical results as in-line chunks (2) the written text required for each exercise. You can read more about Rmarkdown here: https://rmarkdown.rstudio.com/lesson-1.html.

Assignments are due Friday March 05 at midnight (Anywhere on Earth). Send your assignment via email to alburezgutierrez [at] demogr.mpg.de with the subject line "EDSD 2021 assignment".

C. Lecture plan

Monday March 01, 14:00 – 15:30 CEST - "Introduction to digital demography"

Required readings

Edelmann, A., Wolff, T., Montagne, D., and Bail, C.A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology* 46(1):61–81. doi:10.1146/annurev-soc-121919-054621.

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

Optional 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/.

Lazer, D.M.J., Pentland, A., Watts, D.J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M.J., Strohmaier, M., Vespignani, A., and Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science* 369(6507):1060–1062. doi:10.1126/science.aaz8170.

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.

Tuesday March 02, 14:00 – 15:30 CEST - "Digital trace data (1)"

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

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.

Website: https://www.digitalgendergaps.org/data/?report=2020-03-02

Optional readings

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

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.

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.

Wednesday March 03, 14:00 – 15:30 CEST - "Digital trace data (2)"

Required readings

Basellini, U., Alburez-Gutierrez, D., Del Fava, E., Perrotta, D., Bonetti, M., Camarda, C. and Zagheni, E. (2020). "Linking excess mortality to Google mobility data during the COVID-19 pandemic in England and Wales". SocArxiv. DOI:10.31235/osf.io/75d6m.

Optional readings

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinger, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., Vispute, S., Vu, M., Wellenius, G., Williams, B., and Wilson, R.J. (2020). *Google COVID-19 Community Mobility Reports: Anonymization Process Description*. https://arxiv.org/abs/2004.04145.

Drake, T.M., Docherty, A.B., Weiser, T.G., Yule, S., Sheikh, A., and Harrison, E.M. (2020). The effects of physical distancing on population mobility during the COVID-19 pandemic in the UK. *The Lancet Digital Health* 2(8):e385–e387. doi:10.1016/S2589-7500(20)30134-5.

Thursday March 04, 14:00 – 15:30 CEST – "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.

Optional readings

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., and Zagheni, E. (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.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* 31(3):980–991. doi:10.1016/j.ijforecast.2014.06.001.

Friday March 05 – No class

Students are encouraged to work on their assignments (see assignment description above).