IMPRS-PHDS 2022 course (IDEM187) on Topics in Digital and Computational Demography – Day 4 (November 14th 2022)

BiblioDemography:

Using large-scale bibliometric data for demographic research; Advantages and pitfalls of using Scopus data to trace internal and international scholarly migration worldwide

Aliakbar Akbaritabar[1]

[1] Max Planck Institute for Demographic Research (MPIDR)
Akbaritabar@demogr.mpg.de

The main **goals** for this course (and day 4) are:

- To introduce students to **core demographic and social science methods** that are essential to interpret **digital trace data**;

- To introduce students to **core data science methods** that are key to advance our understanding of population processes in the context of the increasing heterogeneity of data sources useful for demographic research.

- To introduce students to **recent substantive advances** in the field of Digital and Computational Demography, with emphasis on fostering critical thinking about modern demographic analysis and (big) data-driven discovery.

- To help students **identify research questions** in their own area of substantive interest that could be addressed with innovative data sources, and support them in the process of devising an appropriate research plan.

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!
https://twitter.com/search?q=%23BiblioDemography&src=typed_query&f=live

AGENDA

1. Introduction (15 minutes, [video 1])

- Welcome and introduction
- What is bibliometric data?
- What type of questions can be studied using bibliometric data?
- What type of demographic questions can be studied using bibliometric data?
- Limitations and pitfalls of using bibliometric data.

2. Data Science skills to use bibliometric data (45 minutes, [videos: 2_1, 2_2])

- [video 2_1] Introduction to parallelised analysis of large-scale bibliometric, text and network data (with Dask in Python, DuckDB and DBeaver in SQL)
- [video 2_2] Hands-on example of parallelised analysis of bibliometric data
- [video 2_2] Hands-on example on use of text and network analysis

3. Example empirical study using bibliometric data (30 minutes, [video 3])

- Internal and international migration of scholars worldwide: Trends, patterns, and inter-relationships

KEY CONTRIBUTORS TO PROJECTS ON SCHOLARLY MIGRATION AT MPIDR

Emilio Zagheni
(MPIDR)

Aliakbar Akbaritabar
(MPIDR)

Andrea Miranda-Gonzalez
(UC Berkeley)

Samin Aref
(U. of Toronto)

Ebru Sanliturk
(MPIDR)

Maciej J. Dańko
(MPIDR)

Tom Theile
(MPIDR)

Xinyi Zhao
(MPIDR)

Study of scholarly migration at the MPIDR:

- Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2022). Return migration of German-affiliated researchers: Analyzing departure and return by gender, cohort, and discipline using Scopus bibliometric data 1996-2020. Scientometrics

- Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., Leasure, D. R., Lohmann, S., Negraia, D. V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M. D., Zagheni, E., & Zhao, X. (2022). *Digital and Computational Demography*. SocArXiv. https://doi.org/10.31235/osf.io/7bvpt

- Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2021a). International Migration in Academia and Citation Performance: An Analysis of German-Affiliated Researchers by Gender and Discipline Using Scopus Publications 1996-2020. ArXiv:2104.12380 [Cs]. http://arxiv.org/abs/2104.12380

- Subbotin, A., & Aref, S. (2021). Brain drain and brain gain in Russia: Analyzing international migration of researchers by discipline using Scopus bibliometric data 1996–2020. Scientometrics. https://doi.org/10.1007/s11192-021-04091-x

- Miranda-González, A., Aref, S., Theile, T., & Zagheni, E. (2020). Scholarly migration within Mexico: Analyzing internal migration among researchers using Scopus longitudinal bibliometric data. EPJ Data Science, 9(1), 34. https://doi.org/10.1140/epjds/s13688-020-00252-9

- Aref, S., Zagheni, E., & West, J. (2019). The Demography of the Peripatetic Researcher: Evidence on Highly Mobile Scholars from the Web of Science. In I. Weber, K. M. Darwish, C. Wagner, E. Zagheni, L. Nelson, S. Aref, & F. Flöck (Eds.), Social Informatics (pp. 50–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-34971-4_4

- Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., & Negraia, D. V. (2019). Demography in the digital era: New data sources for population research.

- Abel, G. J., Muttarak, R., Bordone, V., & Zagheni, E. (2019). Bowling Together: Scientific Collaboration Networks of Demographers at European Population Conferences. European Journal of Population, 35(3), 543–562. https://doi.org/10.1007/s10680-018-9493-1

- Akbaritabar, A., Zhao, X., & Zagheni, E. (2021). *Internal versus international scholarly mobility and migration worldwide*. International Union for the Scientific Study of Population (IUSSP), International Population Conference (IPC2021). https://ipc2021.popconf.org/abstracts/211016

- Emilio Zagheni, Tom Theile, Aliakbar Akbaritabar, Asli Ebru Sanliturk, and Maciej J. Dańko (2022) “Development and Global Migration of Scholars: Trends and Patterns revealed by Bibliometric Data”, under review

IN THE PIPELINE...

- We will announce Scholarly Migration Database (SMD)’s beta website using participants’ emails
  - So please share your email address with us (write to: scholarlymigration@demogr.mpg.de and include “[sign-up for SMDatabase]” in subject line)

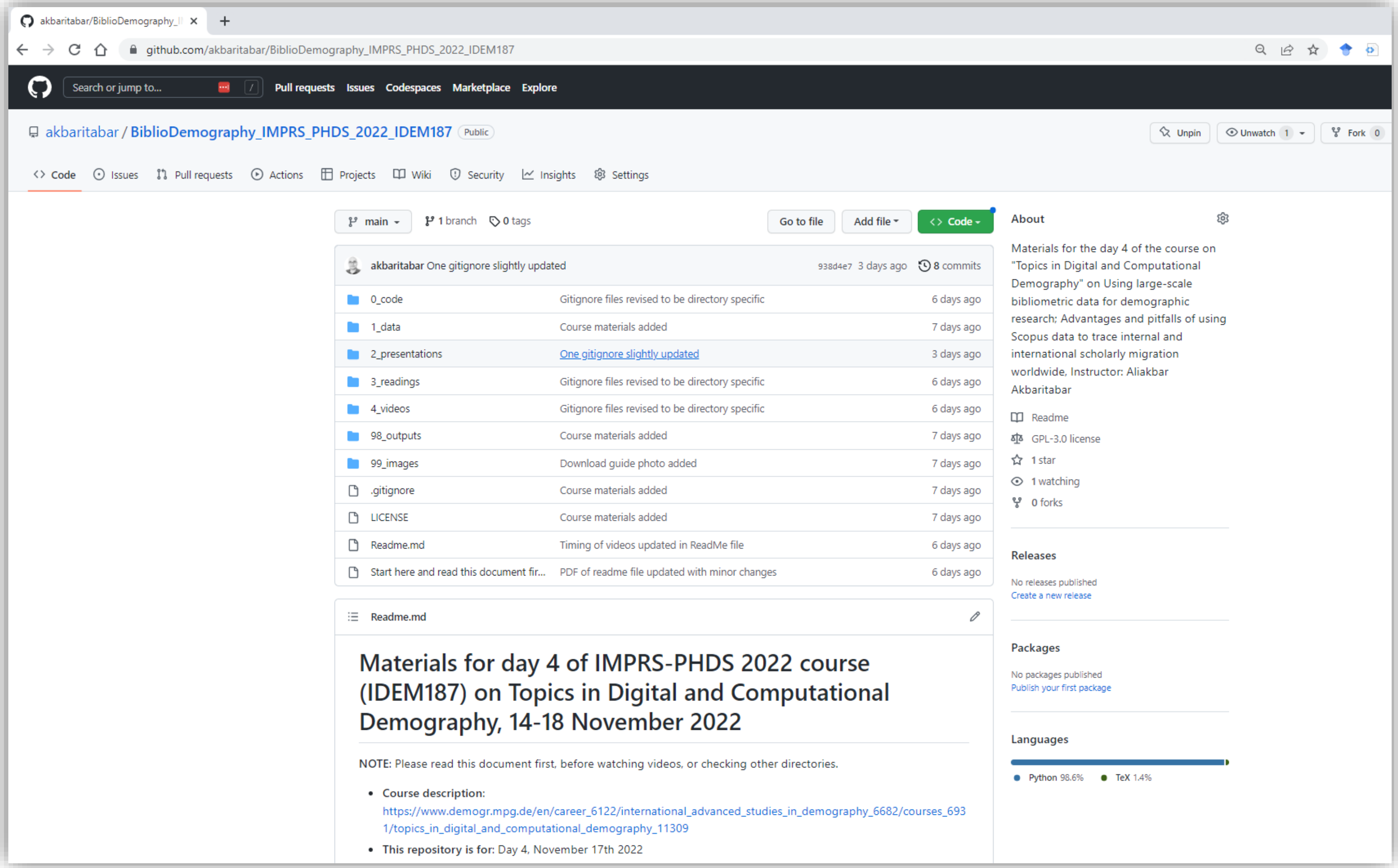
Scholarly Migration Database

Quantifying the Mobility of Scholars

Materials publicly available

Related links:

- Interview and references on using bibliometric data for demographic research: https://www.demogr.mpg.de/en/news_events_6123/news_press_releases_4630/news/how_to_use_bibliometric_data_for_demographic_research_10784



akbaritabar One gitignore slightly updated 938d4e7 3 days ago 8 commits

0_code	Gitignore files revised to be directory specific	6 days ago
1_data	Course materials added	7 days ago
2_presentations	One gitignore slightly updated	3 days ago
3_readings	Gitignore files revised to be directory specific	6 days ago
4_videos	Gitignore files revised to be directory specific	6 days ago
98_outputs	Course materials added	7 days ago
99_images	Download guide photo added	7 days ago
.gitignore	Course materials added	7 days ago
LICENSE	Course materials added	7 days ago
Readme.md	Timing of videos updated in ReadMe file	6 days ago
Start here and read this document fir...	PDF of readme file updated with minor changes	6 days ago

Readme.md

Materials for day 4 of IMPRS-PHDS 2022 course (IDEM187) on Topics in Digital and Computational Demography, 14-18 November 2022

NOTE: Please read this document first, before watching videos, or checking other directories.

- **Course description:** https://www.demogr.mpg.de/en/career_6122/international_advanced_studies_in_demography_6682/courses_6931/topics_in_digital_and_computational_demography_11309
- **This repository is for:** Day 4, November 17th 2022

About

Materials for the day 4 of the course on "Topics in Digital and Computational Demography" on Using large-scale bibliometric data for demographic research; Advantages and pitfalls of using Scopus data to trace internal and international scholarly migration worldwide, Instructor: Aliakbar Akbaritabar

Readme
GPL-3.0 license
1 star
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● Python 98.6% ● TeX 1.4%

Materials: https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187

Course assignments, live session and evaluation!

- Please read the instructions on the **ReadMe** file of the repository (the first page of the repository: https://github.com/akbaritabar/BiblioDemography_IMPRS_PHDS_2022_IDEM187) carefully for instructions.

- You can **choose the tool and language with which you are familiar**, and the **topic** which is most interesting to you and do a **minimum of 1 assignment**. Bring your responses to the live session.

- Live Q&A and discussion session (3 hours, 14:00-17:00 CET, hybrid format)

We will have a brief discussion (and interactive quiz) on the reading materials. Then we will go over the scripts to solve the assignments together. Follow the instructions to setup one of Python/R/SQL/Excel on your laptop and bring your responses to the assignments.

- Course evaluation

Evaluation of the course will be through a combination of a final exam (with two multiple choice questions), responses to assignments, and activity during the live session's discussions.

Please remember:

Limitations in use of bibliometric data (for demographic research)

- Data quality

- e.g., Scientific entity (e.g., authors, or institutions) **name disambiguation** (Tekles & Bornmann, 2020; Wu & Ding, 2013, Akbaritabar, 2021)**.**

- Higher level **epistemic questions** need be addressed while repurposing these data for demographic research (Laudel, 2003; Moed et al., 2013; Moed & Halevi, 2014)

- e.g., assigning the **country of affiliation in the first publication as the country of origin** for academic mobility is prone to error since that could simply be the country of graduation.

- First publication year as **academic birth**

- There is a **publication delay** that can hinder proper identification of the mobility period.

- These data are limited to only those scholars who have **actively published**

- In indexed scholarly journals, so **coverage** may be incomplete (and over-represented by **WEIRD** countries).

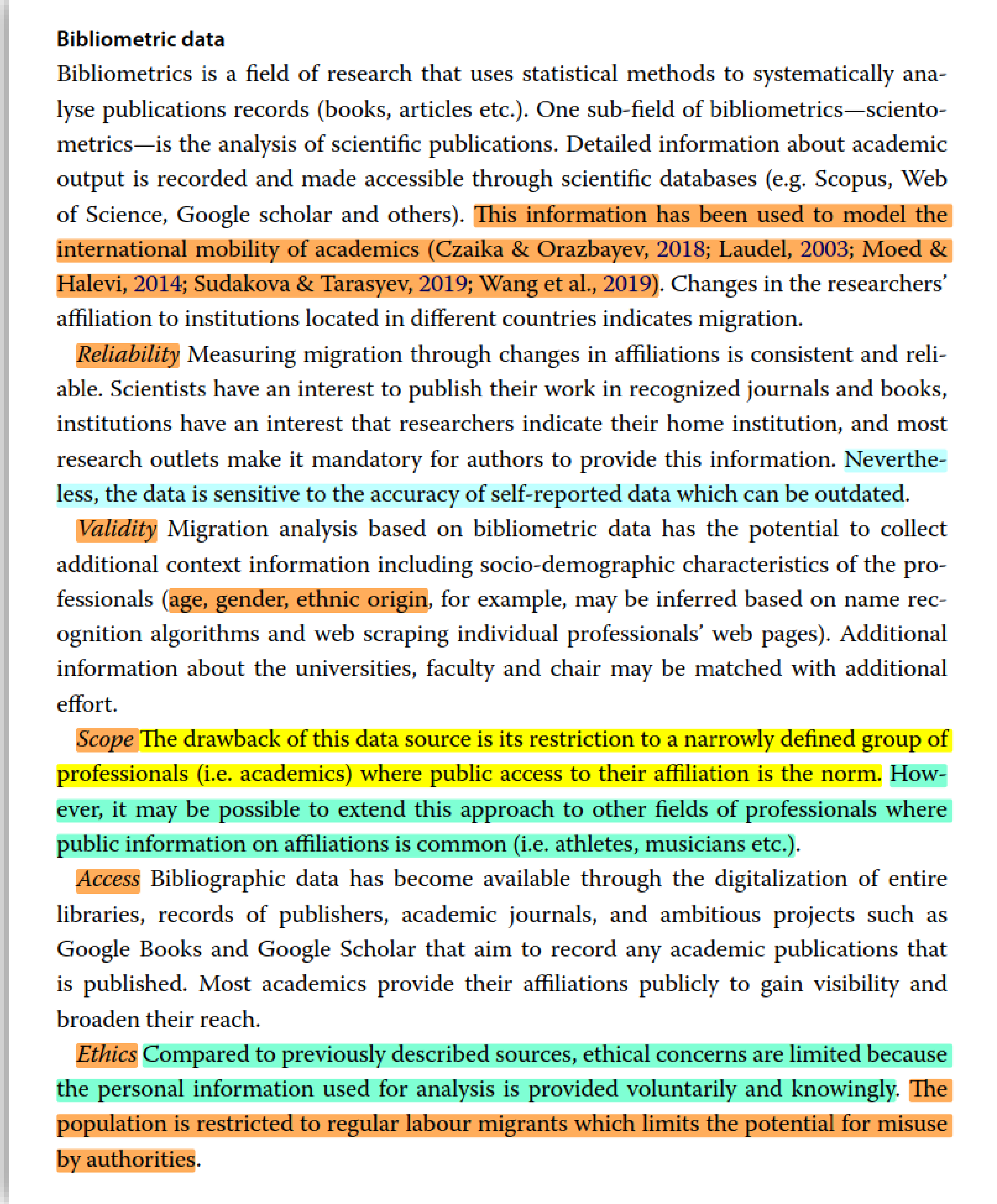
Tjaden, J. (2021). Measuring migration 2.0: A review of digital data sources. *Comparative Migration Studies*, *9*(1), 59. https://doi.org/10.1186/s40878-021-00273-x

Thanks to the online forum of the course and the user "MicolMorellini" for sharing this work which summarizes the use of bibliometric data for migration research in a nice way over these dimensions:

- Reliability
- Validity
- Scope
- Access
- Ethics

See also:

Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Alburez-Gutierrez, D., Gil-Clavel, S., Grow, A., Kim, J., Leasure, D. R., Lohmann, S., Negraia, D. V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M. D., Zagheni, E., & Zhao, X. (2022). *Digital and Computational Demography*. SocArXiv. https://doi.org/10.31235/osf.io/7bvpt



Bibliometric data

Bibliometrics is a field of research that uses statistical methods to systematically analyse publications records (books, articles etc.). One sub-field of bibliometrics—scientometrics—is the analysis of scientific publications. Detailed information about academic output is recorded and made accessible through scientific databases (e.g. Scopus, Web of Science, Google scholar and others). This information has been used to model the international mobility of academics (Czaika & Orazbayev, 2018; Laudel, 2003; Moed & Halevi, 2014; Sudakova & Tarasyev, 2019; Wang et al., 2019). Changes in the researchers' affiliation to institutions located in different countries indicates migration.

Reliability Measuring migration through changes in affiliations is consistent and reliable. Scientists have an interest to publish their work in recognized journals and books, institutions have an interest that researchers indicate their home institution, and most research outlets make it mandatory for authors to provide this information. Nevertheless, the data is sensitive to the accuracy of self-reported data which can be outdated.

Validity Migration analysis based on bibliometric data has the potential to collect additional context information including socio-demographic characteristics of the professionals (age, gender, ethnic origin, for example, may be inferred based on name recognition algorithms and web scraping individual professionals' web pages). Additional information about the universities, faculty and chair may be matched with additional effort.

Scope The drawback of this data source is its restriction to a narrowly defined group of professionals (i.e. academics) where public access to their affiliation is the norm. However, it may be possible to extend this approach to other fields of professionals where public information on affiliations is common (i.e. athletes, musicians etc.).

Access Bibliographic data has become available through the digitalization of entire libraries, records of publishers, academic journals, and ambitious projects such as Google Books and Google Scholar that aim to record any academic publications that is published. Most academics provide their affiliations publicly to gain visibility and broaden their reach.

Ethics Compared to previously described sources, ethical concerns are limited because the personal information used for analysis is provided voluntarily and knowingly. The population is restricted to regular labour migrants which limits the potential for misuse by authorities.

Future directions of research

- New services and methods to prepare **cleaner data**
- Increased availability through initiatives for **open access** to data.

Potential future questions:

- How much of the talent circulation has happened "**within**" national borders versus "**between**" nations?
- Are there **migration corridors** connecting specific regions globally, for example between two specific regions across countries or in the same country, or systems of circulation that involve several countries or subregions?
- Do **scientific collaborations** among scholars facilitate their future mobilities?
- Do scholars have **different probabilities of being mobile** based on the trajectory of their collaborations during their scientific career?
- Complex interactions between processes related to migration of scientists and scholarly collaborations as well as **institutional settings and policies**.
- Finding **migration hubs or regions** with high concentration of academic labour or high attractiveness for future mobility that can inform policy.
- Evaluate theories **explaining migration through network tie formation** (Massey et al., 1993).

Discussion on reading materials

- **Miranda-González et al. (2020)**

  1. What surprised you in this article?
  2. Do we have anyone from Mexico (perhaps 1 online, 1 in the room)?
     - **Please answer the same question, “What surprised you in this article?”**

- **If we compare this article with the results presented in the third video (internal versus international scholarly migration worldwide),**
  - Do you agree with Skeldon (2006)’s assumption that “internal” and “international” migration systems are interrelated?

- **Any points on the other suggested reading materials?**

Course quizzes!

Please go here:

https://app.sli.do/event/wAUmituoscEuk8VHZsKKzr/live/polls

Or open https://www.slido.com/

and enter: 4869 623

Quiz: Based on the shared videos and teaching materials, which one of the course goals were met (choose all that apply)?

1- To introduce students to **core demographic and social science methods** that are essential to interpret **digital trace data**;

2- To introduce students to **core data science methods** that are key to advance our understanding of population processes in the context of the increasing heterogeneity of data sources useful for demographic research.

3- To introduce students to **recent substantive advances** in the field of Digital and Computational Demography, with emphasis on fostering critical thinking about modern demographic analysis and (big) data-driven discovery.

4- To help students **identify research questions** in their own area of substantive interest that could be addressed with innovative data sources, and support them in the process of devising an appropriate research plan.

To answer, you can write 1, 2, 3, 4 in chat (multiple choices are allowed)

What do you think can "increase" an academic's propensity to migrate elsewhere? (what drives scholarly migration?)

Best if you give 1-2 word answers only (max 25 characters allowed). Please write in all small letters.

Those who did the “mobility trajectory” assignment, you could use your answers to the 3rd part for this quiz!

Quiz on assignments
Link:

- Which assignments did you work on, tried out, or resolved (choose all that apply)?
- For R: Mobility trajectory
- For R: Network analysis
- For R: Text analysis
- For Python: Mobility trajectory
- For Python: Network analysis
- For Python: Text analysis
- For Python: Parallelization
- For Excel: Mobility trajectory
- For SQL: SQL in DuckDB/DBeaver

Responses to assignments

- Which assignments did you work on, tried out, or resolved (choose all that apply)?
- **For R: Mobility trajectory**: 1) for viz, see on **right** and upload files (some will be shown after the break), 2) internal 3 moves, international 2 moves, 3) explanations for moves, speak up!
- **For R: Network analysis**: highest degree was 5
- **For R: Text analysis**: “Fixed effects” was the bigram with second highest use, 1866
- **For Python: Mobility trajectory**: same as above
- **For Python: Network analysis**: highest degree was 8
- **For Python: Text analysis**: 56 was the count of found noun-phrase-clauses in abstract 2 (or in python’s 0-indexed way abstract[1])
- **For Python: Parallelization**: out of 580 XML files, 500 were unique
- **For Excel: Mobility trajectory**: same as above
- **For SQL: SQL in DuckDB/Dbeaver**: out of 580 XML files, 500 were unique

- Based on the quiz votes (on Slido), we will work on the most popular assignments today!

Note: if you did the assignment on “**mobility trajectory**” and wish that your visualization to be shown after the break in the live session, and be included in the public repository (and Twitter) with your name, do the following:

1) Save a high-quality version of your visualization with your family name (in PDF or PNG format)

“**First_last_name.pdf**” or “**First_last_name.png**”

2) Upload it to **this** NextCloud folder (https://nextcloud.demogr.mpg.de/s/xYHmyx2YMDMHYCH)

using this **password**: “PHDS_visualization”

Q&A

Please raise any comments, [clarification or else] questions, points on the slides, recorded videos, and presented content!

Note: if you did the assignment on "**mobility trajectory**" and wish that your visualization to be shown after the break in the live session, and be included in the public repository (and Twitter) with your name, do the following:

1) Save a high-quality version of your visualization with your family name (in PDF or PNG format)

"**First_last_name.pdf**" or "**First_last_name.png**"

2) Upload it to **this** NextCloud folder (https://nextcloud.demogr.mpg.de/s/xYHmyx2YMDMHYCH)

using this **password**: "PHDS_visualization"

Hands-on session with examples

(live coding those assignments that were most popular)

Note: if you did the assignment on "**mobility trajectory**" and wish that your visualization to be shown after the break in the live session, and be included in the public repository (and Twitter) with your name, do the following:

1) Save a high-quality version of your visualization with your family name (in PDF or PNG format)

"**First_last_name.pdf**" or "**First_last_name.png**"

2) Upload it to **this** NextCloud folder (https://nextcloud.demogr.mpg.de/s/xYHmyx2YMDMHYCH)

using this **password**: "PHDS_visualization"

Break, 15 minutes

(then we will see the visualizations!)

Note: if you did the assignment on "**mobility trajectory**" and wish that your visualization to be shown after the break in the live session, and be included in the public repository (and Twitter) with your name, do the following:

1) Save a high-quality version of your visualization with your family name (in PDF or PNG format)

"**First_last_name.pdf**" or "**First_last_name.png**"

2) Upload it to **this** NextCloud folder (https://nextcloud.demogr.mpg.de/s/xYHmyx2YMDMHYCH)

using this **password**: "PHDS_visualization"

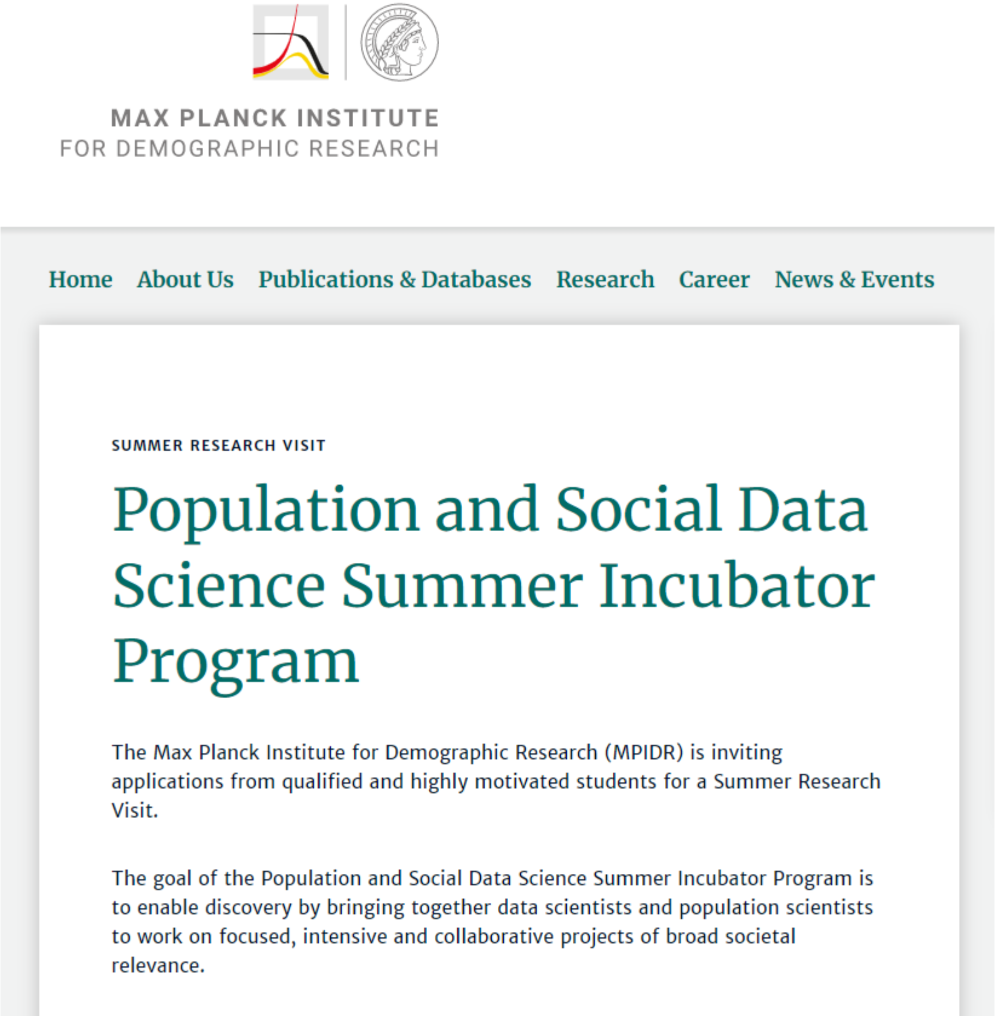
If you wish to do more

#BiblioDemography

Consider applying for this 3 months funded research visit at the MPIDR to spend the summer in Rostock!

Info and application link:

https://www.demogr.mpg.de/en/career_6122/jobs_fellowships_1910/population_and_social_data_science_summer_incubator_program_11474



SUMMER RESEARCH VISIT

Population and Social Data Science Summer Incubator Program

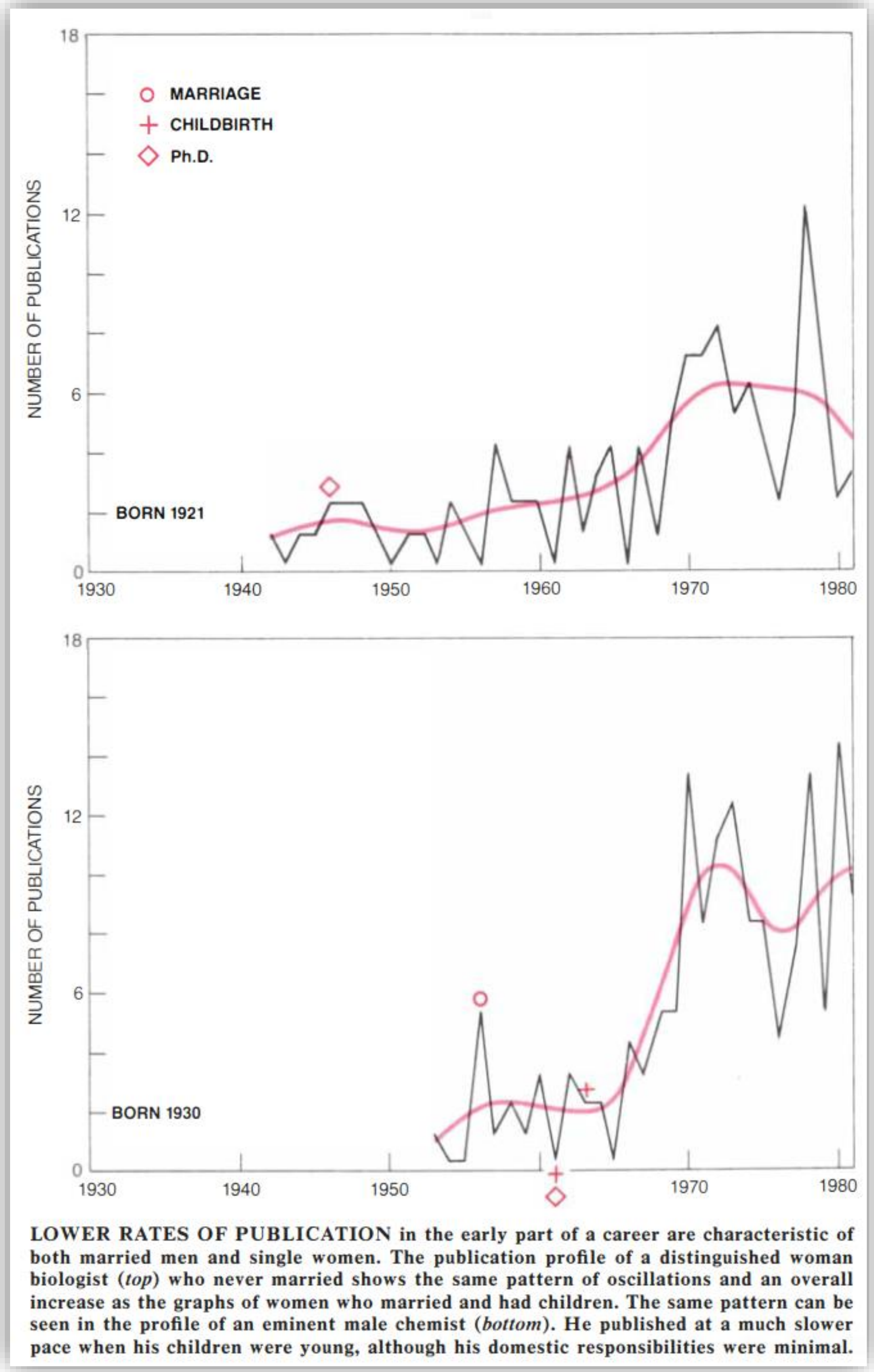
The Max Planck Institute for Demographic Research (MPIDR) is inviting applications from qualified and highly motivated students for a Summer Research Visit.

The goal of the Population and Social Data Science Summer Incubator Program is to enable discovery by bringing together data scientists and population scientists to work on focused, intensive and collaborative projects of broad societal relevance.

A very innovative interview set-up and my favourite visualization (for 1987!)

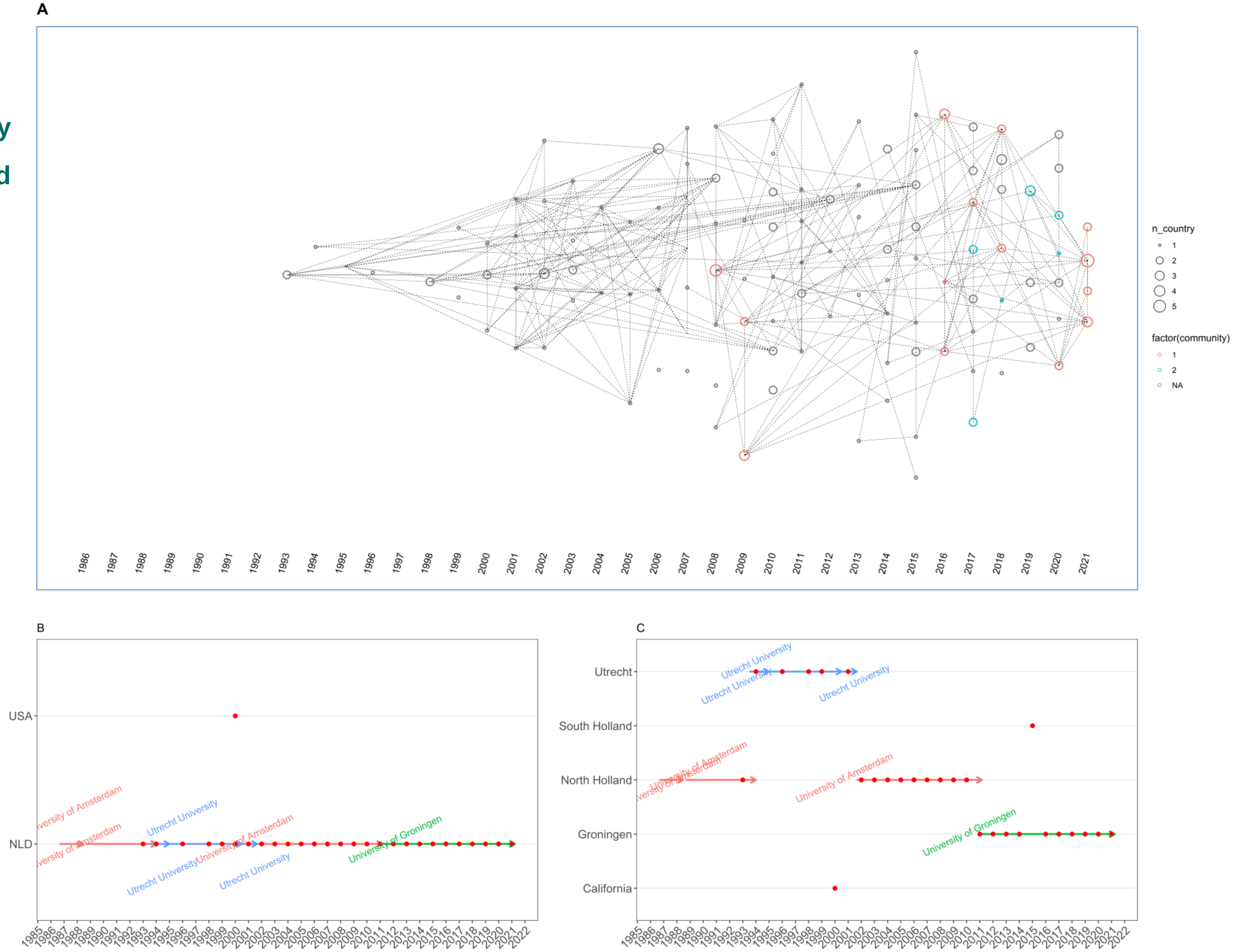
- Using bibliometric data alongside demographic life events

Cole, J. R., & Zuckerman, H. (1987). Marriage, Motherhood and Research Performance in Science. Scientific American, 256(2), 119–125. JSTOR. https://doi.org/10.1038/scientificamerican0287-119



LOWER RATES OF PUBLICATION in the early part of a career are characteristic of both married men and single women. The publication profile of a distinguished woman biologist (*top*) who never married shows the same pattern of oscillations and an overall increase as the graphs of women who married and had children. The same pattern can be seen in the profile of an eminent male chemist (*bottom*). He published at a much slower pace when his children were young, although his domestic responsibilities were minimal.

My own attempt at visualizing mobility trajectory at country and region level and overlay it with the network of collaboration



But assignment data was slightly modified from this picture, what is the difference?

Selected visualizations from students?

Did anyone upload?!

Note (last call): if you did the assignment on “**mobility trajectory**” and wish that your visualization to be shown after the break in the live session, and be included in the public repository (and Twitter) with your name, do the following:

1) Save a high-quality version of your visualization with your family name (in PDF or PNG format)

“**First_last_name.pdf**” or “**First_last_name.png**”

2) Upload it to **this** NextCloud folder (https://nextcloud.demogr.mpg.de/s/xYHmyx2YMDMHYCH)

using this **password**: “PHDS_visualization”

Please tweet with hashtag
#BiblioDemography,
a tribute to James W. Vaupel (1945-2022).
Thanks Ilya and Jonas for bringing up Jim's
labeling idea!
https://twitter.com/search?q=%23BiblioDemography&src=typed_query&f=live

Aliakbar Akbaritabar

Research Scientist

Akbaritabar@demogr.mpg.de

https://akbaritabar.github.io/

https://twitter.com/Akbaritabar

THANK YOU!