MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2023-018 I April 2023
https://doi.org/10.4054/MPIDR-WP-2023-018

Global flows and rates of international migration of scholars

Aliakbar Akbaritabar I akbaritabar@demogr.mpg.de
Tom Theile I theile@demogr.mpg.de
Emilio Zagheni I office-zagheni@demogr.mpg.de

Global flows and rates of international migration of scholars

Aliakbar Akbaritabar[1,*,†], **Tom Theile**[1,†], **and Emilio Zagheni**[1]

[1]Max Planck Institute for Demographic Research, Department of Digital and Computational Demography, Rostock, 18057, Germany
[*]Corresponding author: Aliakbar Akbaritabar (akbaritabar@demogr.mpg.de)
[†]These authors contributed equally to this work.

ABSTRACT

Lack of reliable and comprehensive migration data is one the major reasons that prevents advancements in our understanding of the causes and consequences of migration processes, including for specific groups like high-skilled migrants. We leverage large-scale bibliometric data from Scopus and OpenAlex to trace the global movements of a specific group of innovators: scholars. We developed pre-processing steps and offered best practices for the measurement and identification of migration events from bibliometric data. Our results show a high level of correlation between the count of scholars in Scopus and OpenAlex for most countries. While the magnitude of observed migration events in OpenAlex is larger than in Scopus, the bilateral flows among top pairs of origin and destination countries are consistent in the two databases. Even though OpenAlex has a higher coverage of non-Western countries, the highest correlations with Scopus are observed in Western cuntries. We share our aggregated estimates of international migration rates, and bilateral flows, at the country level, and expect that our estimates will enable researchers to improve our understanding of the causes and consequences of migration of scholars, and to forecast the future mobility of global academic talent.

Background & Summary

Scientists contribute to the research and development of countries[1,2]. While there is a wealth of literature on brain drain[3], brain gain[4] and brain circulation[5,6], there is still a lack of reliable data on migration of scientists[7,8]. Here, we address this lack by leveraging two large-scale sources of bibliometric data, i.e., Elsevier's proprietary Scopus[9] and the openly available OpenAlex[10] database. Building on previous experiences of the literature[11–14], we evaluated different measurements and analytical strategies. We provide best practices on how to re-purpose bibliometric information to prepare migration rates and flow estimates to study scholars worldwide. In addition to describing the pre-processing steps, and showcasing illustrative examples of migration measures and trends, we share aggregated estimates at the country level to enable more elaborated future studies on global academic talent circulation.

Preparing and providing public access to data with high quality is a well-established tradition in the scientific field of demography[15], and some of its sub-fields dealing with longevity (e.g., Human Mortality Database[16]), and fertility (e.g., Human Fertility Database[17]), to name a few. Following this practice, specific national contexts such as nordic nations have exemplary register data on the main life events of their whole population covering birth, death, marriage and divorce events enabling longitudinal research[18].

An exception to this data availability and reliability is the sub-field of demography dealing with migration[19,20]. It is difficult to find high-quality data on migration[15]. Some efforts on harmonizing migration data worldwide, e.g., Integrated Public Use Microdata Series (IPUMS)[21] have shown that migration, even between two neighboring countries, can lead to paradoxical data records due to differing definitions, methods of data collection, registration and digitization practices. One illustrative example is differing reported migration rates between Germany and Poland[22].

One of the usual approaches to remedy this lack of data on migration is to take birth, death, and growth rates of populations between two different time points and consider the difference between these rates as *unexplained factor*, *error term*, or *implied* and *estimated migration*[23]. But this is not a measurement of the actual migration events.

There have been efforts in the literature to use stock data, i.e., the count of the migrant population residing currently in a specific location in a given period, and by considering their country (or region) of origin and time of the move, it is possible to retrospectively estimate migration flows from that origin to this destination[24,25]. But these are estimates and prone to inaccuracies in case intermediary steps were taken between exiting the origin and arriving at the destination country.

This lack of reliable and longitudinal data on migration flows is more pronounced in the case of specific sub-populations, e.g., high-skilled migrants[7,8]. With digitization, there are different sources of data used to provide estimates of migrant populations such as social media data[20,26,27]. One such relatively under-explored data source is metadata of scientific publications accumulated by publishers or large companies, i.e., bibliometric data, which could provide a longitudinal *semi-census* information on scholars and their place of work over time[28–30].

Bibliometric data has proven useful for demographic research[28,29] and especially so in the case of scholars as a subset of the high-skilled population[7,8,14]. Re-purposing these data and using academic affiliation addresses, allows constructing mobility trajectory of individual scholars[11–14,31]. We use bibliometric data as a novel source of *digital traces*[28] and re-purpose them to answer questions regarding high-skilled and, specifically, scholarly migration flows and rates. The geographic scope of our data includes all countries worldwide for which the data is available in Scopus[9] and/or OpenAlex[10].

Here we introduce some illustrative examples of the type of migration research that is possible using bibliometric data[28]. Miranda-Gonzalez et al.[14] offered one of the few studies of internal scholarly migration using bibliometric data of all Scopus-published Mexican scholars from 1996-2018 and their mobility between regions of Mexico. They found that most of the scholars do not move and the capital, i.e., Mexico City, was the most preferred destination of emigrants. Zhao et al.[31] investigated Scopus-published German scholars from 1996-2020, finding that fewer migrant women scholars return to Germany than men. Zhao et al.[7] provided a gender perspective on the migration of scholars worldwide. They addressed a gap concerning the migration of scholars, whether male and female scholars participate equally in transnational mobility and how these patterns have shifted over time from a global perspective. They found that, while female researchers continued to be underrepresented among internationally mobile researchers and migrated over shorter distances, this gender gap was narrowing at a faster rate than the gender gap in the population of general active researchers. Subbotin and Aref[32] investigated Scopus-published Russian scholars from 1996-2020 finding that mobile scholars account for 5% of all scholars affiliated to Russia, and in recent years, the so-called *brain drain* from Russia is replaced with a more balanced *brain circulation*. Sanliturk et al.[33] studied the initial changes in the British academic environment after the Brexit referendum. The study shows evidence that after Brexit, scholars who started their academic careers in the EU countries have a higher probability to leave the UK, while scholars who started their academic careers in the UK have a higher probability to return to the UK. The results signal a compositional change rather than a brain drain in the British academic environment, in the years following the Brexit referendum. Sanliturk et al.[8] studied if the migration of scholars worldwide associates significantly with the economic development of countries (in terms of GDP per capita). Emigration propensity, on average, initially increases with economic development. They found the opposite pattern for the migration of scholars. Despite the reported inverse U-shape pattern of migration for the general population, in case of the academics, a U-shape pattern is observed. This means by increasing GDP, the migration of scholars first decreases and then it starts to increase in rich countries which could signal the return migration of graduates to their home countries.

While the described studies were focused on specific national contexts, or the ones covering a global perspective used different measurement strategies and definitions for migration events, the data presented in this article covers all countries worldwide (which are covered in Scopus and OpenAlex), and leverages methods that we developed to enables comparative studies at a global scale. In addition, this paper presents the best practices adopted after testing different data pre-processing and analysis strategies such as the ones presented in the described literature.

Methods

We use a 2022 snapshot of Scopus and a 2023 snapshot of OpenAlex data. Because a lot of scholars do not publish in every year, the migration data suffers from left- and right-censoring. To prevent this, we limit the years of the migration data from 1998 to 2018. This limit stems from our license terms for Scopus data that spans from 1996 to 2022 and we maintain the same years in OpenAlex to be comparative. Scopus publications are limited to only *Article* and *Review* document types to ensure the highest quality of metadata. In addition, affiliation addresses delivered by Scopus are limited to only *author* affiliation addresses to exclude publisher and other types of addresses which are less relevant to trace scholarly migration. In OpenAlex the document type is limited to *journal articles* which has the highest share in their indexed publications.

Scientific entity name disambiguation

It is necessary to ensure that the bibliometric data used has sufficient quality. Lack of proper data or lower quality of metadata causes errors in identifying entities (e.g., authors or academic affiliations)[34–36] and migration events[14]. In other words, failing to identify scholars properly could cause a merger between different individuals' mobility trajectories. Organizational and academic addresses (i.e., affiliations) need to be correct for the migration event identification to work and to be reliable.

Author name. For author name disambiguation, we use identification numbers added to each unique author by Scopus[9]. Authors who do not have a Scopus author ID or are not indicated as *disambiguated* (i.e., active profiles) by Scopus are excluded from our analysis. The author_id identifies all publications of a single author in 94.4% of cases (recall) and has a precision of 98.1%, which means that records of two different authors could be merged by mistake under one author_id only in 1.9% of the cases. Precision and recall rates are quoted from Scopus and the study published by Baas et al.[9], which includes more detail on the disambiguation process and Scopus meta-data. For OpenAlex and at the moment, we use their provided author identification numbers but further studies on the equality of this identification number are needed.

Organization name. Organization names are disambiguated using the research organization registry (ROR) API and following steps outlined by Akbaritabar[35]. We use the full affiliation string from Scopus to geo-code it to different granular levels. For instance, "Max Planck Institute for Demographic Research (MPIDR), Rostock, Germany" is one affiliation address. But different authors who use this affiliation might write it with a different set of details e.g., to include or exclude city or country names or some might add department and laboratory names. Hence, different versions of this address need to be unified under a unique affiliation identification number to reduce the error in identifying a change in affiliation addresses which are used here as a proxy for a residential address change, i.e., a migration event. See Akbaritabar[35] for a more detailed description of the used methodology and a comparison of its performance with other organization name disambiguation methods.

Pre-processing steps

Figure 1 shows the steps needed in the data collection, processing and export that are described below. Bibliometric data as delivered by the database owner, i.e., Scopus, to the German Competence Network for Bibliometrics (Kompetenznetzwerk Bibliometrie, KB)[37] and through Max Planck Digital Library (MPDL) to us needs pre-processing to allow identification of migration events. KB prepares a relational database, hosts the Scopus data and provides access to us through PostgreSQL queries. We obtain the publication data from this database. For OpenAlex[10], we obtain the publicly available data and process it ourselves at the Max Planck Institute for Demographic Research (MPIDR).

Construction of authorship records. For each author, we obtained the list of all publications and we processed the metadata to assign a date to each author-publication-affiliation-triplet. We call this triplet an authorship record[38]. For example, a paper authored by John Doe and Jane Doe, where John Doe has the affiliations "1" and "2", while Jane Doe has only the affiliation "2", will result in three authorship records as shown in the illustrative example in Table 1. In the case of John who has two affiliations in the same year, we reduce the addresses using a mode-based method which is described further below.

Country of origin assignment. Bibliometric data does not include the country of origin of the authors, i.e., nationality or citizenship. It only indicates the authors' countries and addresses of affiliation at the time of publication that could be used as a proxy for their residential addresses[7,8,14,28]. The affiliation includes the address and the country and in some cases an affiliation_id, which identifies the same institutions, even if their names are spelt differently. The address information is available only in 87% of the Scopus records, but the country information is available in 99% of the records[9]. This step entails the harmonization of country codes and the treatment of missing values. For cases where the same affiliation address does not have a country assigned to them, we correct by filling in the missing countries.

In other words, instead of using the country code provided by Scopus or OpenAlex, we trained a model to predict the country based on the affiliation text. We then control the instances where the returned code by the model is different from the OpenAlex or Scopus country codes. In some cases, OpenAlex had a second affiliation mentioned in the same text. For instance, "Universität Stuttgart; Universidad Barcelona" are indicated as one affiliation and OpenAlex has declared a different country than the one returned by the model for this address. We defined a high threshold to select the country returned by the model's prediction (0.92) when it differs from OpenAlex. We excluded all records with lower confidences than the set threshold. This way, we are sure to exclude wrongly assigned country codes and keep the most of the correctly assigned ones.

Previous research has used the author's family name to find proxies for country of origin[39–43] and that could be an avenue for future development.

Identifying migration events

A migration event is identified based on a change in a scholars' affiliation addresses. This event could be identified in different ways. One can include all affiliations per author and any change in those affiliations as a proxy for *travel activity* and *mobility*[31,44], or consider only long-term *migration*. The migration counts and flows are based on the aggregation of all migration events that we could detect using changes in the affiliation addresses of authors.

Mode-based method

After the pre-processing of the constructed authorship records, we determined the country of residence for every author and year. Year is the lowest temporal level of analysis for migration events since bibliometric data does not have complete coverage on publication month and day. More specifically, we considered all the affiliation countries of an author_id in one year. If there is more than one country, we take the mode of all countries. If there is more than one mode, we check whether one of the modes was the previous country of residence and take that country as the new country of residence. If it was not the case, we would choose one of the mode countries randomly.

To determine migration events, an algorithm then goes through the years and logs a migration event when the country of residence changes. We assume a *two years* preparation time for all publications to cover disciplinary differences in publication delay[45]. If there are gaps in publication years (e.g., authors are not publishing continuously), we backward fill each publication year for two years and assume the author's residence to have changed two years earlier. If there is enough evidence (i.e., continuous publication activity), we consider the year when the modal affiliation changes as the migration year.

Nominator and denominator populations

Once we have detected all migration events, we aggregate them by country and year into emigration and immigration counts that allow calculating other measures such as net migration rate. To generate measures of exposure (i.e., the denominators for migration rates, or the population size of researchers per country and year), we counted the number of active scholars for each year and country. Active scholars in a given year include those who publish at least an article or a review during that year.

To deal with missing observations, we assumed that an author who did not publish in a particular year was still part of the population of active scholars if he or she published one or two years before. Finally, we excluded authors who had *only* one indexed publication during their entire career from the denominator of the scholars' population. The reason for this exclusion is twofold: first, these scholars could be junior researchers who have graduated or those who leave academia. Since we do not have a live *census* of all academics globally, we cannot consider them as part of the pool of *active scholars*. Furthermore, in each given year, there is a fraction of scholars who enter the pool of active publishers (by having their first publication in the sample) and exit this pool of publishing scholars in the next years[46]. Counting them among active scholars would over-inflate the population of scholars and cause our measures to be artificially smaller. Second, because by definition of the mode country per year, these scholars who had publications only in one year could not have migrated (i.e., contributed to the nominator), hence, it is reasonable to exclude them from the denominator.

Bilateral migration flows

Each identified migration event, based on a change in mode country of affiliation, connects a pair of countries i.e., an origin (O) and destination (D) country. Using these OD pairs and the determined year of migration, we can construct yearly bilateral flows between these countries and origin-destination matrices[24,25]. These matrices are not based on estimates and they include actual migration events observed in the data as described above. This enables us to identify migration corridors where a large proportion of scholars move between specific pairs of countries.

Measures

To evaluate the exposure of populations to migration events, we calculate different measures. We calculate *in-migration* (equation 1), *out-migration* (equation 2), and *net migration* count (equation 3) and rates (equation 4)[14,47] as follows:

$$IMR_{i,t} = \frac{I_{i,t}}{N_{i,t}} \tag{1}$$

$$EMR_{i,t} = \frac{E_{i,t}}{N_{i,t}} \tag{2}$$

$$NM_{i,t} = I_{i,t} - E_{i,t} \tag{3}$$

$$NMR_{i,t} = \frac{I_{i,t} - E_{i,t}}{N_{i,t}} \tag{4}$$

where i is the country, t is the year, $I_{i,t}$ is the inflow of scholars entering a country and $E_{i,t}$ is the outflow of scholars exiting that country over the total number of scholars in the country in a given year, i.e., $N_{i,t}$.

Data Records

As a result of this research, two data-sets are prepared at the country level on global international migration rates and flows and shared publicly in form of CSV files. Examples of these datasets are presented here.

Table 2 shows four illustrative example rows of the *migration rates* data per country year combination. Each row in this table is a country and year combination and columns provide information from OpenAlex and Scopus on the count of scholars, padded population with the two-years backward filling method described before, and the number of incoming and outgoing scholars of this country. This table includes further columns which come from the World Bank data and other sources, e.g., general population of the country, GDP per capita, income level, etc.

Table 3 shows four illustrative example rows of the *flow* data per country pair and year combination. Each row in this table is a pair of two countries, i.e., origin (O) and destination (D), and the next columns give the count of scholars who have migrated from O to D in the given year based on Scopus and OpenAlex data.

Technical Validation

We have carried out different validation steps on the described methods. We controlled the effect of different backward and forward padding settings for publication years (e.g., for the years when a scholar does not publish) and how it affects the migration rates. In addition, we controlled the quality of the bibliometric metadata and considered limits on the subset of the data with sufficient quality to be included in our analysis. This entailed excluding specific starting and ending years, affiliation addresses that were not for authors (e.g., publishing houses and similar), and document types to ensure the reliability of the results.

Here we present 1) examples of illustrative results that can be obtained from the constructed dataset and 2) results of our comparison and validation between the size of the population of scholars and net migration rates based on Scopus and OpenAlex.

Figure 2 shows the international net migration rates (NMR) worldwide based on Scopus (top) and OpenAlex (bottom) per 1,000 scholars. It shows a consistent pattern for most countries worldwide. In some exceptional countries, the NMR calculated using Scopus and OpenAlex differ, e.g., see the cases of Canada, Guyana, Costa Rica, Honduras, Bolivia, Russia, China, India, Iran, Turkmenistan, Sudan, Angola, Mozambique, and Philippines, to name a few, as illustrative examples where the colors are different between maps on the top and bottom.

Figure 3 shows the example of the United States and the temporal trend of flows of scholars arriving in the US from other countries (left) and leaving the US to other countries (right) based on Scopus (top) and OpenAlex (bottom). The magnitude of flows based on OpenAlex is much larger than Scopus and while we have limited the publications in both databases to articles and reviews, this could indicate that the higher coverage of publications in OpenAlex might help discover some under-explored scholarly migration corridors worldwide. Nevertheless, as described in the methods section, the quality of the author name disambiguation and identifiers in OpenAlex needs further evaluation in future research.

Figure 4 shows the top 15 country pairs with the highest bilateral flows of scholars where the origin (Y-axis) and destination (X-axis) pairs based on Scopus (top) and OpenAlex (bottom) are presented. While in most of these country pairs colors which are normalized based on the size of population of scholars are consistent, the printed labels inside the cells, that show the actual count of scholars, have larger magnitudes in OpenAlex (bottom).

Figure 5 shows the correlation between the population of scholars (left) and net migration rates (right) compared over continental regions worldwide. It is clear that while the population of scholars between the two databases correlate to a high degree over years with a median correlation close to 1, but net migration rate fluctuates to a much higher degree. This could signal a large difference in coverage of individual migration trajectories between these two databases and can also stem from the small net migration rates which fluctuate with small differences in measurement rather than population counts which are larger and small changes do not cause them to fluctuate.

Usage Notes

Please note that in our join operation on Scopus and OpenAlex data and in order to be inclusive, we keep all country-year pairs where one of these databases have counts. While in our visualizations, we exclude the rows where one does not have measurement for a country-year pair. Please consider to filter the rows according to your goals while using the dataset.

Code availability

All scripts to replicate the presented analysis and figures are publicly accessible alongside the aggregated datasets based on Scopus and OpenAlex on GitHub at "https://github.com/MPIDR/Global-flows-and-rates-of-international-migration-of-scholars".

References

1. Hunt, J. Immigrant patents boost growth. *Science* **356**, 697–697, 10.1126/science.aan2468 (2017). Publisher: American Association for the Advancement of Science.
2. Bound, J., Khanna, G. & Morales, N. Reservoir of foreign talent. *Science* **356**, 697–697, 10.1126/science.aan2956 (2017). Publisher: American Association for the Advancement of Science.
3. Docquier, F., Lohest, O. & Marfouk, A. Brain Drain in Developing Countries. *The World Bank Econ. Rev.* **21**, 193–218, 10.1093/wber/lhm008 (2007).
4. Kapur, D. & McHale, J. *Give Us Your Best and Brightest: The Global Hunt for Talent and Its Impact on the Developing World* (Center for Global Development, 2005).

5. Wiesel, T. Fellowships: Turning brain drain into brain circulation. *Nature* **510**, 213–214, 10.1038/510213a (2014). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7504 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Developing world;Institutions;Research management Subject_term_id: developing-world;institutions;research-management.

6. Appelt, S., van Beuzekom, B., Galindo-Rueda, F. & de Pinho, R. Chapter 7 - Which Factors Influence the International Mobility of Research Scientists? In Geuna, A. (ed.) *Global Mobility of Research Scientists*, 177–213, 10.1016/B978-0-12-801396-0.00007-7 (Academic Press, San Diego, 2015).

7. Zhao, X., Akbaritabar, A., Kashyap, R. & Zagheni, E. A gender perspective on the global migration of scholars. *Proc. Natl. Acad. Sci.* **120**, e2214664120, 10.1073/pnas.2214664120 (2023).

8. Sanliturk, E., Zagheni, E., Dańko, M. J., Theile, T. & Akbaritabar, A. Global patterns of migration of scholars with economic development. *Proc. Natl. Acad. Sci.* **120**, e2217937120, 10.1073/pnas.2217937120 (2023). Publisher: Proceedings of the National Academy of Sciences.

9. Baas, J., Schotten, M., Plume, A., Côté, G. & Karimi, R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quant. Sci. Stud.* **1**, 377–386 (2020).

10. Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 10.48550/arXiv.2205.01833 (2022). ArXiv:2205.01833 [cs].

11. Laudel, G. Studying the brain drain: Can bibliometric methods help? *Scientometrics* **57**, 215–237 (2003).

12. Moed, H. F., Aisati, M. & Plume, A. Studying scientific migration in scopus. *Scientometrics* **94**, 929–942 (2013).

13. Moed, H. F. & Halevi, G. A bibliometric approach to tracking international scientific migration. *Scientometrics* **101**, 1987–2001 (2014).

14. Miranda-González, A., Aref, S., Theile, T. & Zagheni, E. Scholarly migration within Mexico: analyzing internal migration among researchers using Scopus longitudinal bibliometric data. *EPJ Data Sci.* **9**, 34, 10.1140/epjds/s13688-020-00252-9 (2020).

15. Skeldon, R. Interlinkages between internal and international migration and development in the Asian region. *Population, Space Place* **12**, 15–30, https://doi.org/10.1002/psp.385 (2006). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/psp.385.

16. Jdanov, D. A., Jasilionis, D., Shkolnikov, V. M. & Barbieri, M. Human mortality database. *Encycl. gerontology population aging/editors Danan Gu, Matthew E. Dupre. Cham: Springer Int. Publ. 2020* (2019).

17. Jasilioniene, A. *et al.* Data resource profile: the human fertility database. *Int. J. Epidemiol.* **45**, 1077–1078e (2016).

18. Thygesen, L. C. & Ersbøll, A. K. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur. journal epidemiology* **29**, 551–558 (2014).

19. King, R. & Skeldon, R. 'Mind the Gap!' Integrating Approaches to Internal and International Migration. *J. Ethn. Migr. Stud.* **36**, 1619–1646, 10.1080/1369183X.2010.489380 (2010).

20. Alexander, M., Polimis, K. & Zagheni, E. Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. *Popul. Res. Policy Rev.* 10.1007/s11113-020-09599-3 (2020).

21. Center, M. P. Integrated Public Use Microdata Series, International: Version 7.3, 10.18128/D020.V7.3 (2020). Version Number: 7.3 Type: dataset.

22. Mooyaart, J., Dańko, M. J., Costa, R. & Boissonneault, M. Quality assessment of european migration data. *Changes* **1**, 11 (2021).

23. Canudas-Romo, V., Shen, T. & Payne, C. F. The Components of Change in Population Growth Rates. *Demography* **59**, 417–431, 10.1215/00703370-9765067 (2022).

24. Abel, G. J. & Cohen, J. E. Bilateral international migration flow estimates for 200 countries. *Sci. Data* **6**, 82, 10.1038/s41597-019-0089-3 (2019). Bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Geography;Social sciences Subject_term_id: geography;social-sciences.

25. Abel, G. J. & Cohen, J. E. Bilateral international migration flow estimates updated and refined by sex. *Sci. Data* **9**, 1–11, 10.1038/s41597-022-01271-z (2022). Number: 1 Publisher: Nature Publishing Group.

26. Alexander, M., Polimis, K. & Zagheni, E. The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. *Popul. Dev. Rev.* **45**, 617–630, 10.1111/padr.12289 (2019).

27. Vieira, C. C. *et al.* The interplay of migration and cultural similarity between countries: Evidence from Facebook data on food and drink interests. *PLOS ONE* **17**, e0262947, 10.1371/journal.pone.0262947 (2022). Publisher: Public Library of Science.

28. Kashyap, R. *et al.* *Digital and computational demography* (Edward Elgar Publishing, 2023). Pages: 48-86 Publication Title: Research Handbook on Digital Sociology Section: Research Handbook on Digital Sociology.

29. Alburez-Gutierrez, D. *et al.* Demography in the digital era: New data sources for population research (2019).

30. Sugimoto, C. R. *et al.* Scientists have most impact when they're free to move. *Nature* **550**, 29–31, 10.1038/550029a (2017).

31. Zhao, X., Aref, S., Zagheni, E. & Stecklov, G. Return migration of german-affiliated researchers: analyzing departure and return by gender, cohort, and discipline using scopus bibliometric data 1996–2020. *Scientometrics* 10.1007/s11192-022-04351-4 (2022).

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

33. Şanlıtürk, A. E., Aref, S., Zagheni, E. & Billari, F. C. Homecoming after Brexit: evidence on academic migration from bibliometric data. Tech. Rep. WP-2022-019, Max Planck Institute for Demographic Research, Rostock (2022). 10.4054/MPIDR-WP-2022-019. Edition: 0.

34. Tekles, A. & Bornmann, L. Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quant. Sci. Stud.* 1–38, 10.1162/qss_a_00081 (2020). Publisher: MIT Press.

35. Akbaritabar, A. A quantitative view of the structure of institutional scientific collaborations using the example of Berlin. *Quant. Sci. Stud.* **2**, 753–777, 10.1162/qss_a_00131 (2021).

36. Donner, P., Rimmert, C. & van Eck, N. J. Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quant. Sci. Stud.* **1**, 150–170, 10.1162/qss_a_00013 (2019).

37. Kompetenzzentrum bibliometrie (competence centre for bibliometrics). https://bibliometrie.info/ (2021). Accessed: 2021-09-24.

38. Aref, S., Zagheni, E. & West, J. The demography of the peripatetic researcher: Evidence on highly mobile scholars from the web of science. In *International Conference on Social Informatics*, 50–65 (Springer, 2019).

39. Basu, A. Some differences in research publications of indian scientists in india and the diaspora, 1986–2010. *Scientometrics* **94**, 1007–1019, 10.1007/s11192-012-0800-z (2013).

40. Begum, M., Roe, P., Webber, R. & Lewison, G. UK ethnic minority cancer researchers: their origins, destinations and sex. In *Proceedings of ISSI 2017*, 568–579 (University of Wuhan, China, 2017).

41. Grilli, J. & Allesina, S. Last name analysis of mobility, gender imbalance, and nepotism across academic systems. *Proc. Natl. Acad. Sci.* **114**, 7600–7605, 10.1073/pnas.1703513114 (2017).

42. Ioannidis, J. P. *et al.* Comprehensive mapping of local and diaspora scientists: a database and analysis of 63951 greek scientists. *Quant. Sci. Stud.* 1–28, 10.1162/qss_a_00136 (2021).

43. Lewison, G., Kumar, S., Wong, C.-Y., Roe, P. & Webber, R. The contribution of ethnic groups to malaysian scientific output, 1982–2014, and the effects of the new economic policy. *Scientometrics* **109**, 1877–1893, 10.1007/s11192-016-2139-3 (2016).

44. Robinson-Garcia, N. *et al.* The many faces of mobility: Using bibliometric data to measure the movement of scientists. *J. Informetrics* **13**, 50–63, 10.1016/j.joi.2018.11.002 (2019).

45. Björk, B.-C. & Solomon, D. The publishing delay in scholarly peer-reviewed journals. *J. Informetrics* **7**, 914–923, 10.1016/j.joi.2013.09.001 (2013).

46. Akbaritabar, A., Traag, V. A., Caimo, A. & Squazzoni, F. Italian sociologists: a community of disconnected groups. *Scientometrics* 10.1007/s11192-020-03555-w (2020).

47. Bell, M. *et al.* Cross-national comparison of internal migration: issues and measures. *J. Royal Stat. Soc. Ser. A (Statistics Soc.* **165**, 435–464, 10.1111/1467-985X.t01-1-00247 (2002). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-985X.t01-1-00247.

Acknowledgements

The original bibliometric data from Scopus is licensed by Elsevier and provided to us by the German Competence Network for Bibliometrics[37] through the project "Kompetenzzentrum Bibliometrie" grant number 16WIK2101A via the Max Planck Digital Library (MPDL). Under the original license terms, only aggregated results based on the raw metadata which accompany our scientific publications as replication materials can be made publicly available. No individual data from Elsevier Scopus is shared. OpenAlex data is publicly available[10].

Author contributions statement

Conceptualization: A.A., T.T., E.Z. Methodology: A.A., T.T., E.Z. Software: T.T. Validation: A.A. Formal analysis: A.A., T.T. Investigation: A.A., T.T., E.Z. Resources: E.Z. Data Curation: T.T. Writing - Original Draft: A.A. Writing - Review & Editing: A.A., T.T., E.Z. Visualization: A.A., T.T. Supervision: E.Z. Project administration: A.A. Funding acquisition: E.Z.

Competing interests

The authors declare that they have no competing interests.

Figures & Tables

Table 1. An illustrative example of one publication and its respective authorship records

Publication Title	Author Full Name	Affiliation	Publication Year
Paper 1 title	John Doe	Affiliation 1, Country 1	2020
Paper 1 title	John Doe	Affiliation 2, Country 2	2020
Paper 1 title	Jane Doe	Affiliation 2, Country 2	2020

Table 2. Illustrative example of international migration rates data per country and year combination. Each row includes one country and year and the next columns provide information from OpenAlex and Scopus (note, these are illustrative example rows to show the structure of the CSV file and the numbers might differ from the data).

row_number	**countrycode**	**year**	**countryname**	**world bank population**	paddedpop_openalex	paddedpop_scopus	inmig_openalex	inmig_scopus	outmig_openalex	outmig_scopus
1	USA	1997	United States	2.73E+08	621115	127289	14224	1675	15142	2143
2	USA	1998	United States	2.76E+08	699712	212982	18966	3071	18997	2997
3	USA	1999	United States	2.79E+08	716306	229842	21181	5862	20920	4838
4	USA	2000	United States	2.82E+08	740288	242323	23389	5116	23141	3610

Table 3. Illustrative example of international migration flow data per country pair and year combination. Each row includes a pair country and year and the next columns provide information from OpenAlex and Scopus (note, these are illustrative example rows to show the structure of the CSV file and the numbers might differ from the data).

migration_from	**migration_to**	**migration_year_padding**	**n_migrations_Scopus**	**n_migrations_OpenAlex**
deu	usa	1998	756	800
deu	usa	1999	871	972
deu	usa	2000	909	950
deu	usa	2001	972	982

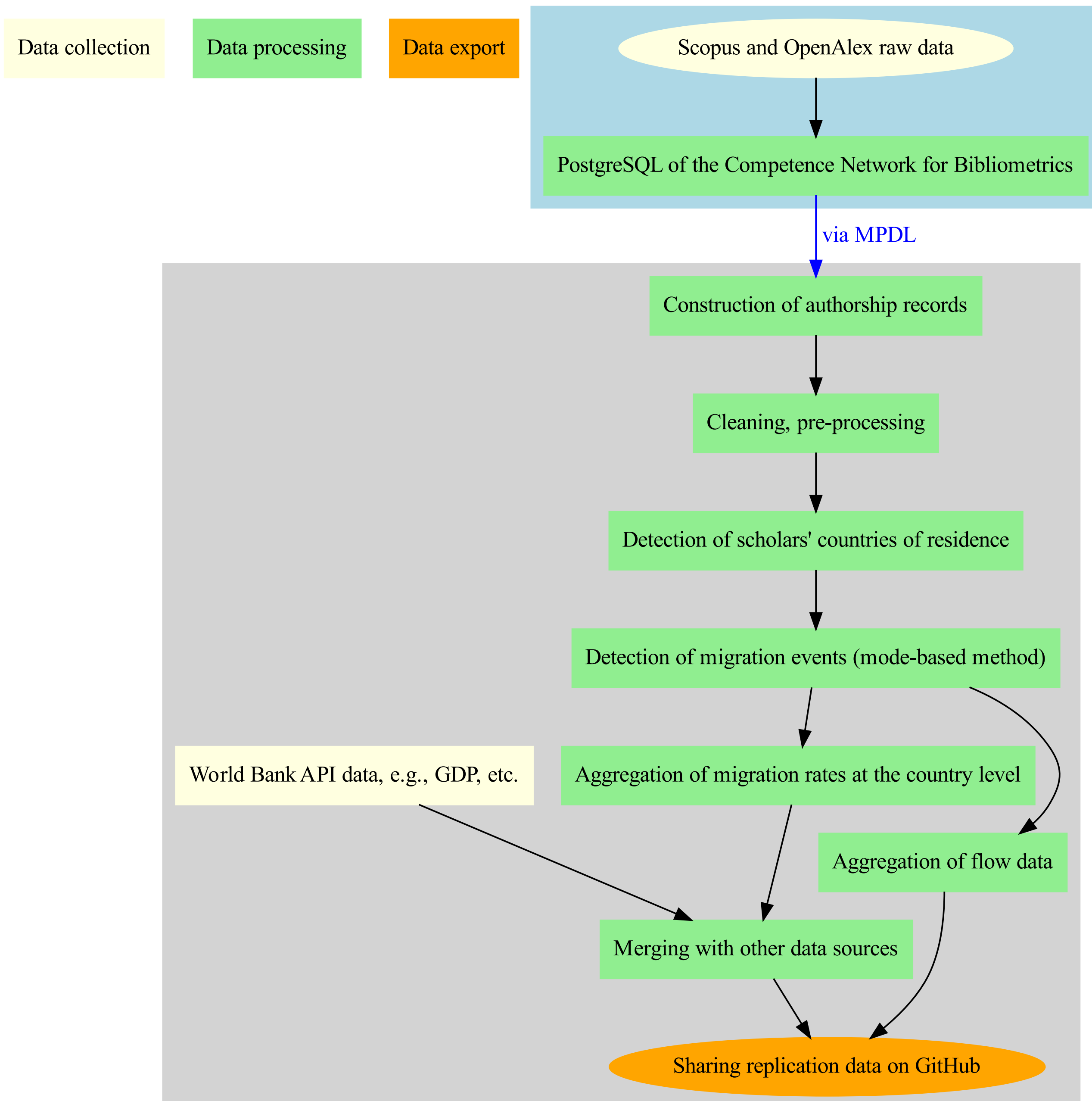


Figure 1. Data collection (light yellow steps), processing (green) and export (orange) pipeline to prepare migration of scholars dataset with one part carried out at the Competence Network for Bibliometrics (light blue bounding-box on the top) and shared with us via the Max Planck Digital Library (MPDL), and the rest of the steps are carried out at the Max Planck Institute for Demographic Research (MPIDR, gray bounding-box on the bottom).

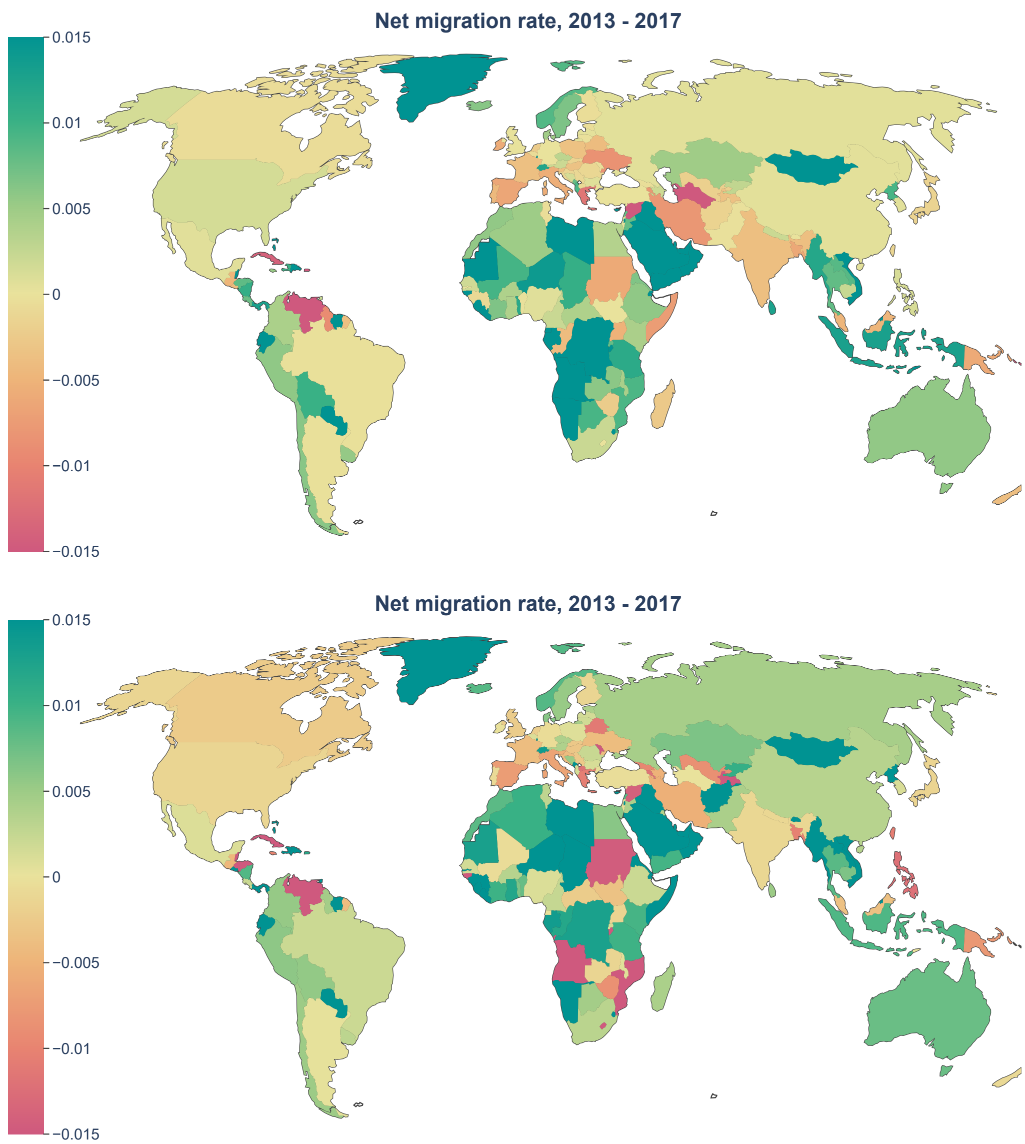


Figure 2. Net-migration rates of scholars worldwide from 2013-2017 based on Scopus (top) and OpenAlex (bottom) in terms of rate per 1,000 active scholars.

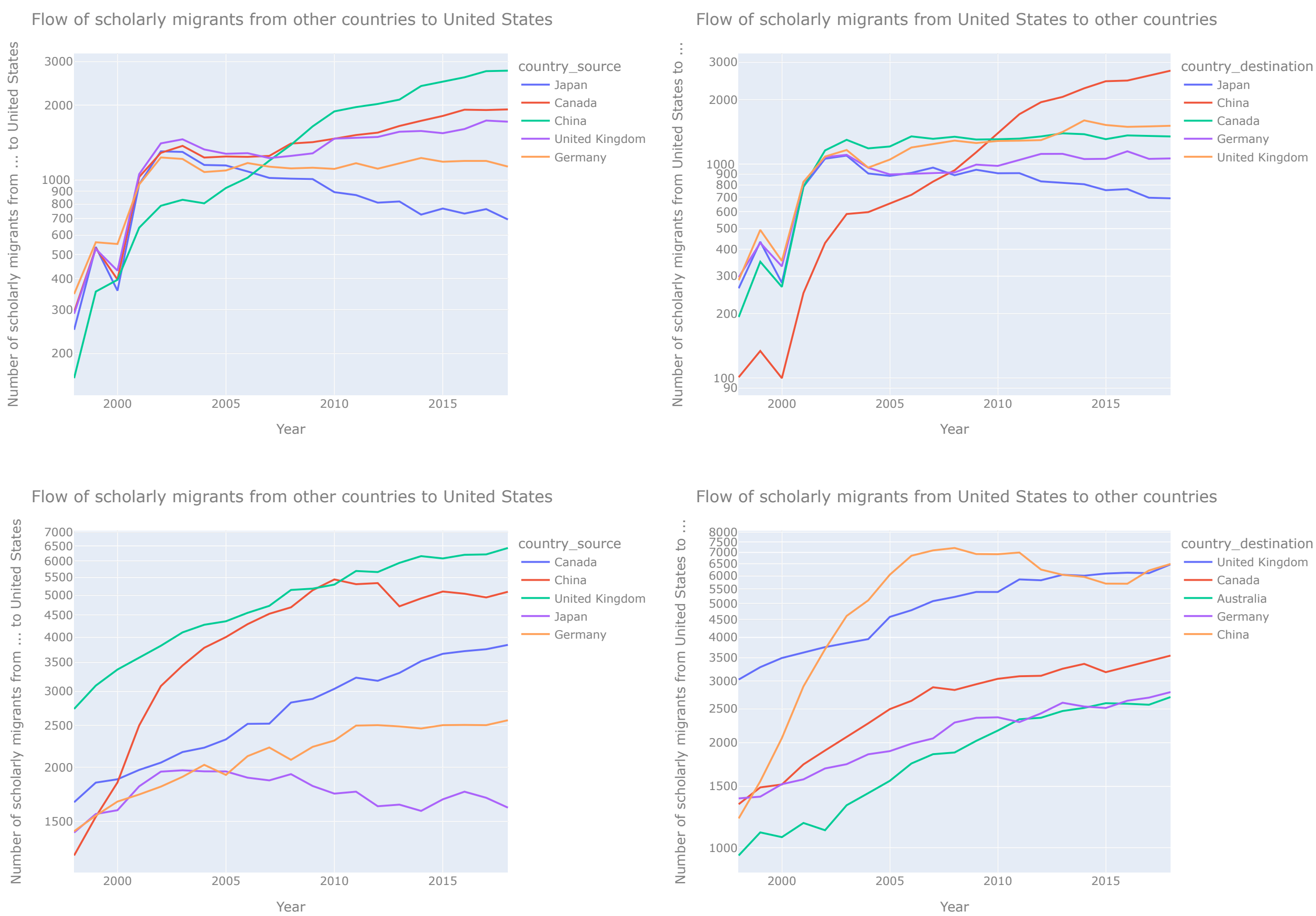
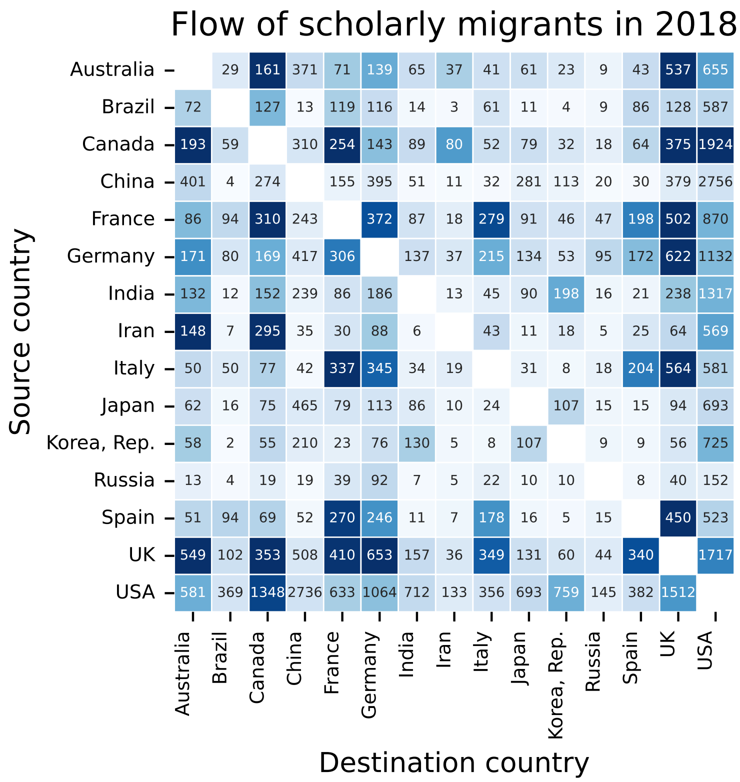


Figure 3. Temporal change in migration flows of scholars *to* (left) and *from* (right) the United States based on Scopus (top) and OpenAlex (bottom). The figure is limited to the top 5 origins and destinations with the highest flows. Flows are presented as actual count of scholars sent or received and the 2018 counts could be seen in Figure 4 as printed labels in cells.



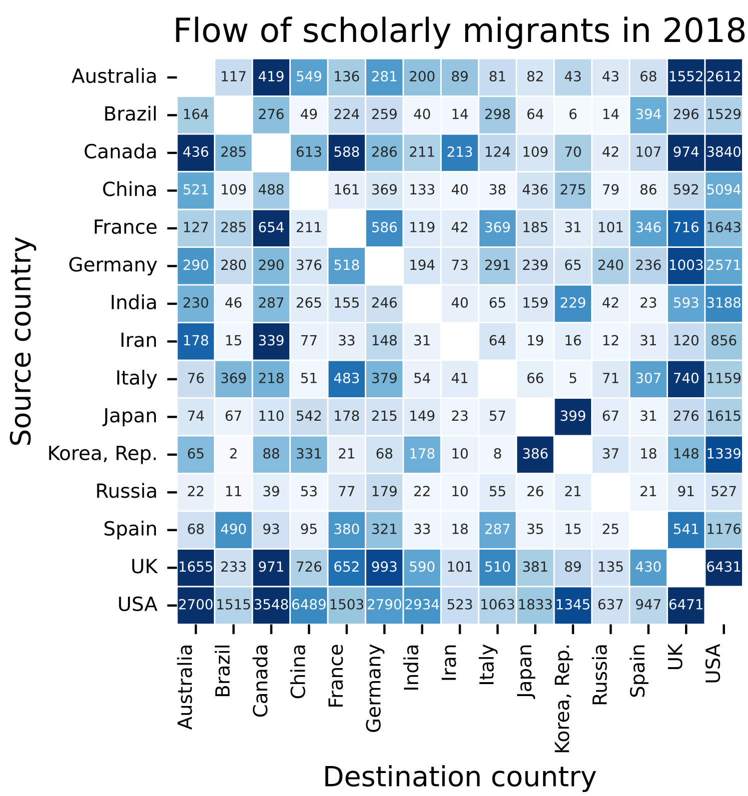


Figure 4. Bilateral flows of scholarly migration between top 15 pairs of countries with the highest exchanges based on Scopus (top) and OpenAlex (bottom). Numbers printed in cells are the actual count of scholars moved from a source country to the destination. Colors are based on the normalized flow of migrants. Normalization was done by dividing the total flow of scholars between each country pair by the total outflow from the source country times the total inflow of the destination country.

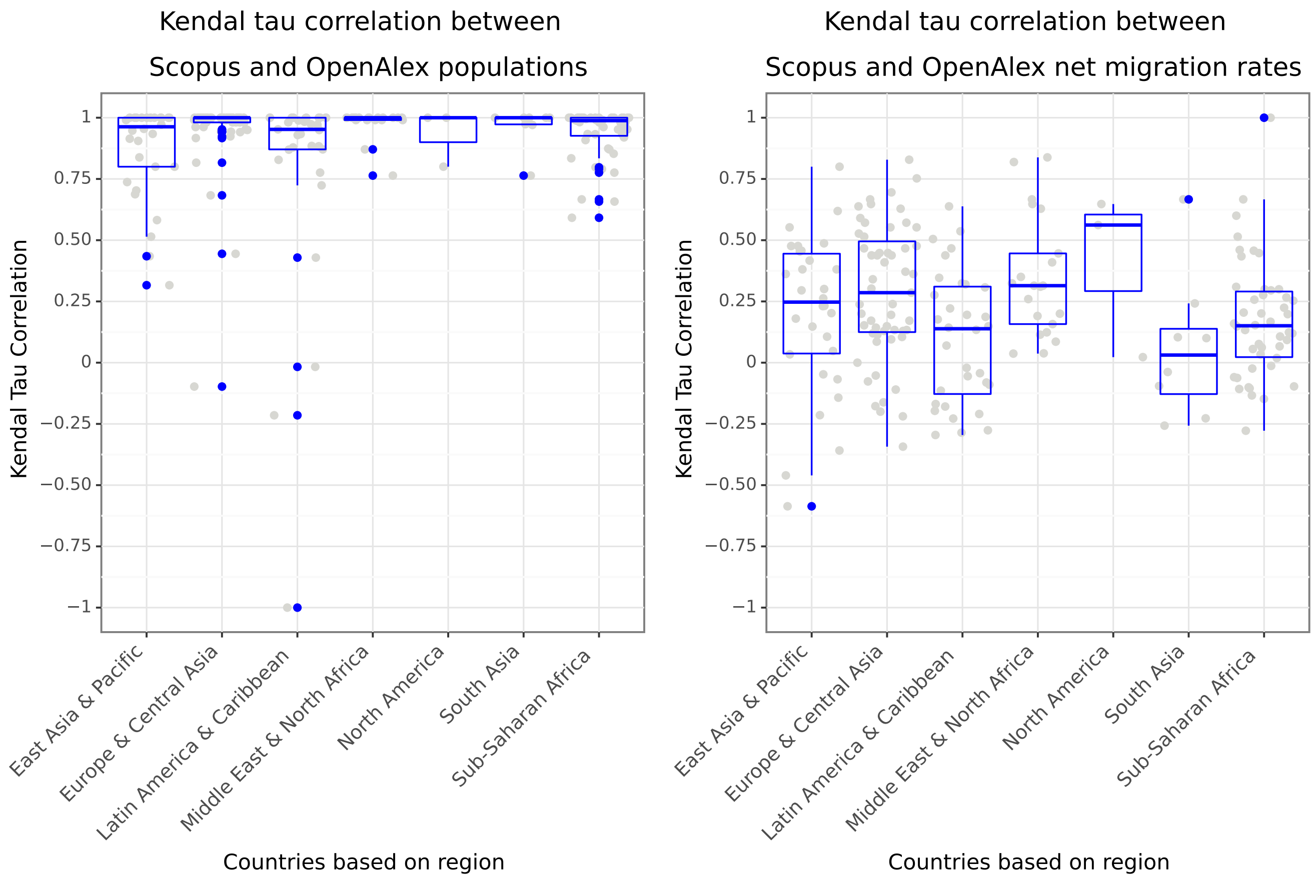


Figure 5. Kendal tau correlation between population (left) and net migration rates (right) from 1998-2018 based on Scopus and OpenAlex divided over different continental regions (X-axis). Each gray dot is one country's correlation measure and a jitter is added to the X-axis positioning of dots to reduce their overlap without substantive meaning. Blue boxplots and dots show the trend and median (thick line) of the same data.