Global flows and rates of international migration of scholars

Aliakbar Akbaritabar  akbaritabar@demogr.mpg.de
Tom Theile  theile@demogr.mpg.de
Emilio Zagheni  office-zagheni@demogr.mpg.de
Global flows and rates of international migration of scholars

Aliakbar Akbaritabar1,*,†, Tom Theile1,†, and Emilio Zagheni1

1Max 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 countries. 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 countries1,2. While there is a wealth of literature on brain drain3, brain gain4 and brain circulation5,6, there is still a lack of reliable data on migration of scientists7,8. Here, we address this lack by leveraging two large-scale sources of bibliometric data, i.e., Elsevier’s proprietary Scopus9 and the openly available OpenAlex10 database. Building on previous experiences of the literature11–14, 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 demography15, and some of its sub-fields dealing with longevity (e.g., Human Mortality Database16), and fertility (e.g., Human Fertility Database17), 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 research18.

An exception to this data availability and reliability is the sub-field of demography dealing with migration19,20. It is difficult to find high-quality data on migration20. 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 Poland22.

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 migration23. 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 destination24,25. 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\textsuperscript{7,8}. With digitization, there are different sources of data used to provide estimates of migrant populations such as social media data\textsuperscript{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\textsuperscript{28–30}.

Bibliometric data has proven useful for demographic research\textsuperscript{28,29} and especially so in the case of scholars as a subset of the high-skilled population\textsuperscript{7,8,14}. Re-purposing these data and using academic affiliation addresses, allows constructing mobility trajectory of individual scholars\textsuperscript{11–14,31}. We use bibliometric data as a novel source of digital traces\textsuperscript{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\textsuperscript{9} and/or OpenAlex\textsuperscript{10}.

Here we introduce some illustrative examples of the type of migration research that is possible using bibliometric data\textsuperscript{28}.

Miranda-Gonzalez et al.\textsuperscript{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.\textsuperscript{31} investigated Scopus-published German scholars from 1996-2020, finding that fewer migrant women scholars return to Germany than men. Zhao et al.\textsuperscript{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\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.

In the migration data, the migration event identification is performed using the authors' and academic affiliations\textsuperscript{14–36} and migration events\textsuperscript{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\textsuperscript{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.\textsuperscript{9}, which includes more...
After the pre-processing of the constructed authorship records, we determined the country of residence for every author and used a two years preparation time for all publications to cover disciplinary differences in publication delay. 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. 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. 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) as follows:

\[ IMR_{i,t} = \frac{I_{i,t}}{N_{i,t}} \]  
\[ EMR_{i,t} = \frac{E_{i,t}}{N_{i,t}} \]  
\[ NM_{i,t} = I_{i,t} - E_{i,t} \]  
\[ NMR_{i,t} = \frac{I_{i,t} - E_{i,t}}{N_{i,t}} \]

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**


Acknowledgements

The original bibliometric data from Scopus is licensed by Elsevier and provided to us by the German Competence Network for Bibliometrics 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.

Author contributions statement


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

<table>
<thead>
<tr>
<th>Publication Title</th>
<th>Author Full Name</th>
<th>Affiliation</th>
<th>Publication Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper 1 title</td>
<td>John Doe</td>
<td>Affiliation 1, Country 1</td>
<td>2020</td>
</tr>
<tr>
<td>Paper 1 title</td>
<td>John Doe</td>
<td>Affiliation 2, Country 2</td>
<td>2020</td>
</tr>
<tr>
<td>Paper 1 title</td>
<td>Jane Doe</td>
<td>Affiliation 2, Country 2</td>
<td>2020</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>row_number</th>
<th>countrycode</th>
<th>year</th>
<th>countryname</th>
<th>world bank population</th>
<th>paddedpop_openalex</th>
<th>paddedpop_scopus</th>
<th>inmig_openalex</th>
<th>inmig_scopus</th>
<th>outmig_openalex</th>
<th>outmig_scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>USA</td>
<td>1997</td>
<td>United States</td>
<td>2.73E+08</td>
<td>621115</td>
<td>127289</td>
<td>14224</td>
<td>1675</td>
<td>15142</td>
<td>2143</td>
</tr>
<tr>
<td>2</td>
<td>USA</td>
<td>1998</td>
<td>United States</td>
<td>2.76E+08</td>
<td>699712</td>
<td>212982</td>
<td>18966</td>
<td>3071</td>
<td>18997</td>
<td>2997</td>
</tr>
<tr>
<td>3</td>
<td>USA</td>
<td>1999</td>
<td>United States</td>
<td>2.82E+08</td>
<td>716306</td>
<td>229842</td>
<td>21181</td>
<td>5862</td>
<td>20920</td>
<td>4838</td>
</tr>
<tr>
<td>4</td>
<td>USA</td>
<td>2000</td>
<td>United States</td>
<td>2.90E+08</td>
<td>740288</td>
<td>242324</td>
<td>23389</td>
<td>4116</td>
<td>25141</td>
<td>4610</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>migration_from</th>
<th>migration_to</th>
<th>migration_year_padding</th>
<th>n_migrations_Scopus</th>
<th>n_migrations_OpenAlex</th>
</tr>
</thead>
<tbody>
<tr>
<td>deu</td>
<td>usa</td>
<td>1998</td>
<td>756</td>
<td>800</td>
</tr>
<tr>
<td>deu</td>
<td>usa</td>
<td>1999</td>
<td>871</td>
<td>972</td>
</tr>
<tr>
<td>deu</td>
<td>usa</td>
<td>2000</td>
<td>909</td>
<td>950</td>
</tr>
<tr>
<td>deu</td>
<td>usa</td>
<td>2001</td>
<td>972</td>
<td>982</td>
</tr>
</tbody>
</table>
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.