EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE   **Open Access**

Scholarly migration within Mexico: analyzing internal migration among researchers using Scopus longitudinal bibliometric data

Andrea Miranda-González[1,2], Samin Aref[2*], Tom Theile[2] and Emilio Zagheni[2]

*Correspondence:
aref@demogr.mpg.de
[2]Laboratory of Digital and
Computational Demography, Max
Planck Institute for Demographic
Research, Rostock, Germany
Full list of author information is
available at the end of the article

Abstract

The migration of scholars is a major driver of innovation and of diffusion of knowledge. Although large-scale bibliometric data have been used to measure international migration of scholars, our understanding of internal migration among researchers is very limited. This is partly due to a lack of data aggregated at a suitable sub-national level. In this study, we analyze internal migration in Mexico based on over 1.1 million authorship records from the Scopus database. We trace the movements of scholars between Mexican states, and provide key demographic measures of internal migration for the 1996–2018 period. From a methodological perspective, we develop a new framework for enhancing data quality, inferring states from affiliations, and detecting moves from modal states for the purposes of studying internal migration among researchers. Substantively, we combine demographic and network science techniques to improve our understanding of internal migration patterns within country boundaries. The migration patterns between states in Mexico appear to be heterogeneous in size and direction across regions. However, while many scholars remain in their regions, there seems to be a preference for Mexico City and the surrounding states as migration destinations. We observed that over the past two decades, there has been a general decreasing trend in the crude migration intensity. However, the migration network has become more dense and more diverse, and has included greater exchanges between states along the Gulf and the Pacific Coast. Our analysis, which is mostly empirical in nature, lays the foundations for testing and developing theories that can rely on the analytical framework developed by migration scholars, and the richness of appropriately processed bibliometric data.

Keywords: High-skilled migration; Internal migration; Computational demography; Science of science; Network science; Brain circulation

1 Introduction

The academic exchange of ideas goes beyond physical borders. As such, many scholars are highly mobile, and their work contributes to technological and economic advances in a number of locations over the course of their academic lives. There is a growing body of literature that focuses on the migration and mobility of scientists and their impacts at the international level. However, even though the geographic distribution of scholars is

Springer

both an outcome of regional disparities and a key source of inequality of opportunities for future generations, little is known about the drivers of these movements of researchers within country borders. Understanding these patterns can shed light on important regional deficits that identify areas of progress and opportunities for investment in human capital. From the public policy perspective, it is in the interests of regional governments to maintain a strong base of highly qualified scholars who can provide innovative and scientific solutions to public issues and collaborate with the private sector. In order to do so, regional governments should be aware of the underlying reasons for the migratory movements of researchers, and the associated sources of attraction at the national and the global levels. In order to identify these patterns, we propose an approach to study the internal migration of scholars using Scopus bibliometric data. We present our methods for measuring migratory movements, and discuss, as an illustrative case, the resulting network models of scholarly migration in Mexico.

There are a number of complementary theories of international and internal migration [1], and there is a continuous effort to create a bridge between disciplines in order to build a more cohesive migration framework. Some theoretical migration studies have suggested that migration that occurs within a country depends on the stages of its development as a society [2]. Similar to the Demographic Transition theory [3], Zelinsky's migration transition model identifies five phases in which spatial and time-specific characteristics (economic, social, and historical) determine mobility patterns [2]. Migration is studied by looking at different origins, destinations, and directions of migratory events. Migration estimates that provide a holistic view on the movements within a country would also facilitate the study of the temporal dynamics of a migration system from the perspective of a theoretical migration transition model.

The size and the impact of research in Mexico is perhaps not as well-known as those of other North American countries. According to SciVal 2010–2019 data,[a] Mexico has over 200,000 researchers (comparable to Switzerland) and over 217,000 pieces of scholarly output (comparable to Singapore), with an average number of citations per publication [4] of 9.4 (comparable to China, Brazil, and Poland) and a field-weighted citation impact [4] of 0.91 (comparable to Japan and Iran). Despite these features, research in Mexico has been an under-studied case in the scientometrics literature. Historically, Mexico has deployed policies to attract foreign researchers through scholarships and professionalization, which placed it at the forefront of Latin American countries [5]. Existing work on Mexico has focused on research production and collaboration [6]; on particular fields of scholarship, such as computer science [7], physics [8], and health sciences [9]; and on Mexican researchers abroad [10].

Mexico is also a particularly relevant case for conducting exploratory analysis of internal migration, because most existing discussions on migration and Mexico are about migration from Mexico to the United States (US), even though between 2005 and 2010, interstate and intra-state migration accounted for 3.5% and 3.1% of the population, respectively; whereas the rate of international migration was 1.1% [11]. The migration of Mexican researchers abroad, particularly to the US, seems to be tied to differences in favorable conditions of the labor market [12], and to changes in visa availability [13]. However, rather than focusing on a loss of talent, or *brain drain* [14], public policy can concentrate on harnessing the flows of researchers who come to (or return to) Mexico [15], and, similarly, on internal movers who can strengthen the domestic academic system.

It remains unclear whether scholarly migration in Mexico has increased or slowed down in the last two decades as a result of special socioeconomic conditions, such as government spending on public institutions, social inequality, and the availability of alternative jobs in the private sector. Domestically, there have been ongoing efforts by the government to promote scientific research and development. For instance, a National System of Researchers (SNI in Spanish) has been created in order to track and reward academic and teaching contributions. Despite such policies, limited data availability has made evaluations of current internal mobility among researchers difficult and problematic.

This study intends to contribute to the literature in two main ways: first, by re-purposing bibliometric data to analyze internal rather than international migration; and, second, by exploring the migratory movements of scholars in Mexico. Although our substantive focus is on a specific country, the proposed methodological framework of re-purposing bibliometric data for internal migration and the proposed method of analysis are directly applicable to a broader context. We proceed, first, by describing the current state of tertiary education in Mexico and the available data sources in Sect. 2. Then, in Sect. 3, we outline our methodology for adapting bibliometric data for migration research. We present our results on the demographic analysis in Sect. 4, and our results on network analysis in Sect. 5. Finally, we discuss and summarize our findings in Sect. 6.

2 Inferring migration patterns for scholars from census data

Mexico is composed of 32 states with specific natural endowments and economic infrastructures that enable them to accommodate over 126 million people. To better visualize the states where the movements occur, we include a map of Mexico in Fig. 1. In this study, we use two levels of aggregation: states and regions of Mexico. The latter are determined by the geographic, economic, and social similarities of the states. In total, we distinguish between five regions: (i) Northern states along the Mexico-US border, (ii) states in the Center with comparable economic status, (iii) states along the Pacific Coast, (iv) states surrounding Mexico City that share strong ties with the capital, and (v) states benefiting from the industry and tourism of the Gulf of Mexico and the Yucatán Peninsula.
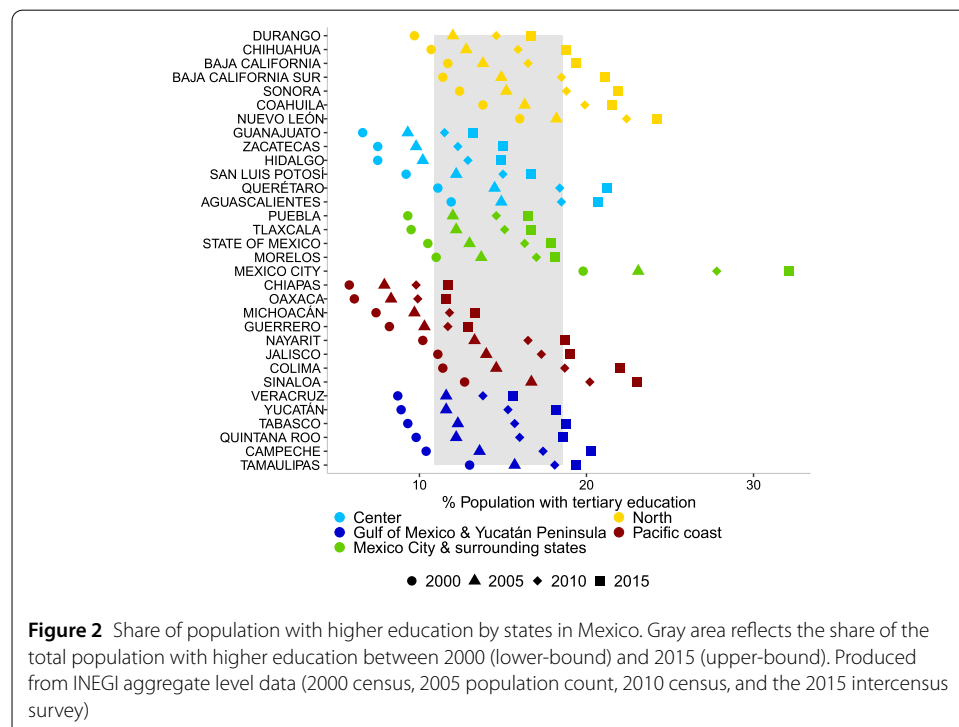
The data sources that can be used to study specific groups of migrants are particularly limited. In Mexico, the National Institute of Statistics, Geography and Informatics (IN-EGI) collects information on the socio-demographic and geographic characteristics of the

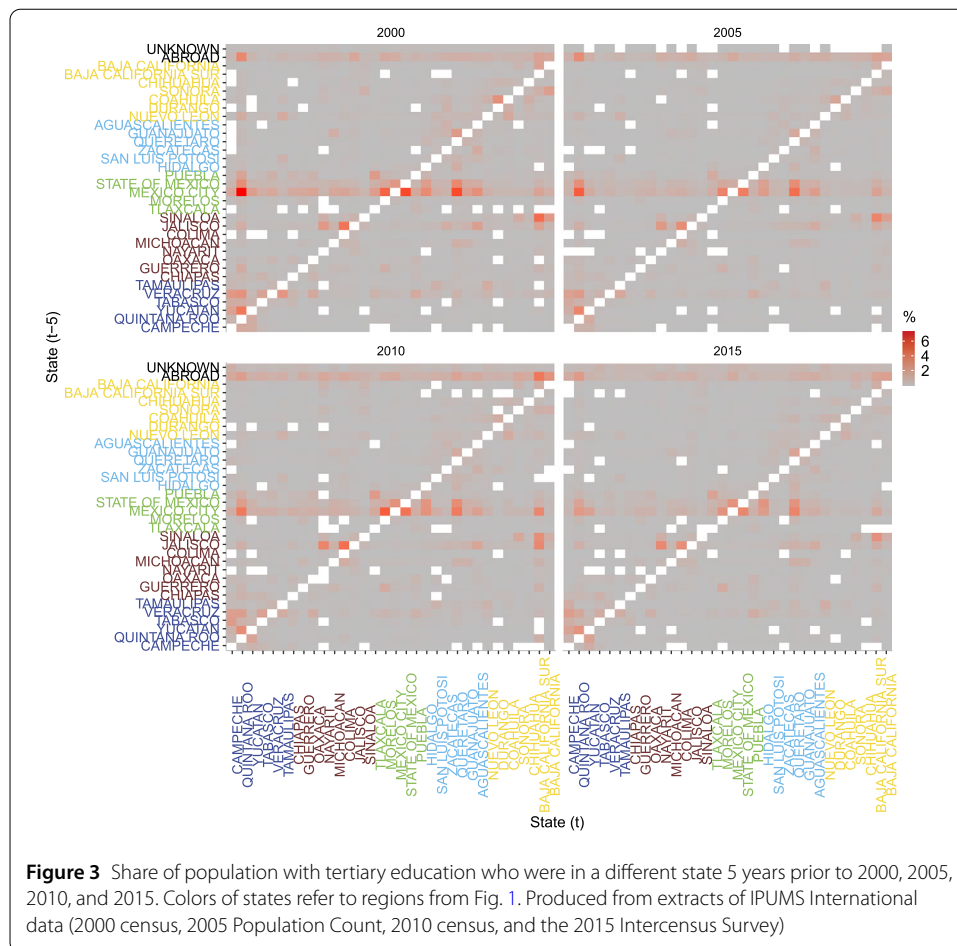**Figure 1** 32 states of Mexico grouped into five regions

complete population every 10 years in the census, and through a large representative sample between censuses (five-year intercensus or population count). At an aggregate level, the National Council for Population (CONAPO) of Mexico provides yearly estimates and projections of internal net migration rates for the general population. However, the census data have some important limitations. As the census information on mobility yields transitions from the past state to the current state, we are unable to count the number of mobility events that could have occurred in between. Therefore, we cannot obtain the necessary numerators for computing internal migration rates, nor can we infer multiple movements over short periods for the same person.

In the absence of bibliometric records, census data can be used to analyze the movement patterns of highly educated people, such as those with tertiary education. This is possible because census data contain information on an individual's current residence relative to their last place of residence or their location from five years before, as well as their last attained educational level. On the educational front, Fig. 2 shows the evolution of the share of population with tertiary education across different states in Mexico.[b]

The national shares, in gray, range from 10.9% in 2000 with an improvement up to 18.6% by 2015. Overall, there has been an increase in the share of tertiary-educated people in every state. While all states have experienced growth in the share of the population with tertiary education, many still lag far behind the national average. This is the case for Chiapas, Oaxaca, Michoacán, and Guerrero, all of which are states along the Pacific coast of Mexico. In several other states, such as in states in the North (in yellow), Querétaro, Aguascalientes, Colima, and Sinaloa, the share of tertiary-educated people has increased considerably. Unlike in all other states, in Mexico City – which, in addition to being the capital, is an important center for social and economic activities, and is the home of many

Figure 2 Share of population with higher education by states in Mexico. Gray area reflects the share of the total population with higher education between 2000 (lower-bound) and 2015 (upper-bound). Produced from INEGI aggregate level data (2000 census, 2005 population count, 2010 census, and the 2015 intercensus survey)

Figure 3 Share of population with tertiary education who were in a different state 5 years prior to 2000, 2005, 2010, and 2015. Colors of states refer to regions from Fig. 1. Produced from extracts of IPUMS International data (2000 census, 2005 Population Count, 2010 census, and the 2015 Intercensus Survey)

government offices – the share of the population with post-high school education has surpassed 30%, as of 2015.

We use the census microdata to focus on the movements among the population with tertiary education. Using IPUMS International extracts from 2000 to 2015 on Mexican census data [16], we exploit the information on the location of residence five years prior among people who have a higher level of education. Figure 3 shows the fraction of individuals with tertiary education who moved between two states (those who left the *y*-axis state five years prior to the census year to move to the *x*-axis state) over the tertiary-educated population of the destination state.

We can see that while there are movements from most of the states, Mexico City appears to play an especially important role, as it was a common origin of tertiary-educated movers. Moreover, the movements appear to be regional, in particular within the states of (i) the Gulf of Mexico and the Yucatán Peninsula (in blue), and (ii) Mexico City and its surrounding states (in green), which can be seen by looking at the concentrated red colors along the diagonal. However, using the census data to infer the migration rates of specific subgroups of the population proves to be difficult and severely limited by the five-year time span, even if the data are linked. On the other hand, the census provides information on important individual demographic characteristics, such as age and sex, which are not readily available from other sources like bibliometric data.

Based on these observations, we can state that using census data to analyze the movements of researchers would provide a general image of reduced movements of people with higher education that is not necessarily indicative of the movements of scholars. In other words, census data enabled us to observe the movements of tertiary-educated people, but not the movements of the research-active scholars who presumably contribute to the creation and dissemination of focused knowledge.

3 Methodology of re-purposing bibliometric data for internal migration

In order to analyze the international migration of researchers, many studies have relied on bibliometric databases. Recent studies offered proxies for place of residence [17]; provided information on bilateral international migration flows [18]; offered a methodological framework for dealing with multiple affiliations [19], analyzed the movements of highly mobile researchers and return migration flows [20], and studied the impact of migration on different fields of scholarship in a country [14]. In particular, Scopus has been widely used to analyze international mobility [14, 21, 22] due to its advantages compared to other bibliometric databases. For instance, Scopus provides a wider breadth of records for a number of disciplines [23] and offers a more reliable author ID, [24] which is considered to be more suitable for tracking the movements of individual researchers [25].

The availability of large-scale bibliometric data allows us to track the migration of researchers in a way that has not been possible with traditional sources of migration data, like censuses and surveys. Additionally, bibliometric data provide standardized data, which are suitable for comparative studies. The unit of the data is an *authorship record*, which is the linkage between an author and a publication. The data that we use in this paper involve 1.1 million authorship records of scholars who have published with Mexican affiliation addresses in sources covered by Scopus. Using these data, we analyze information for over 252,000 researchers across 32 states of Mexico and the changes in their affiliation addresses over the 1996–2018 period. In the next subsections, we explain the pre-processing steps, which are the key parts of repurposing bibliometric data for an analysis of internal migration among researchers.

3.1 Improving Scopus author IDs

We use the Scopus author ID [24, 25] to group the authorship records of individual researchers and to detect mobility events. While most author IDs correctly disambiguate individual researchers, and the Scopus author IDs are considered to be reliable for inferring mobility in the literature [25], the Scopus author identification system is not perfect, and there could be multiple distinct individuals sharing the same name (or similar names) who have been incorrectly allocated the same author ID. Author name disambiguation remains a challenge in scientometrics, although there have been many recent efforts to tackle it algorithmically [26, 27].

The first step of our pre-processing of the data involves treating the authorship records that are most likely to be impacted by the lack of accuracy in the Scopus author identification system. We extract all of the authorship records that are associated with more than 276[c] publications (an average of more than one publication per month during the 23 years of our analysis period). We implemented an approach inspired by the state-of-the-art methods in the literature [27] to further disambiguate the extracted set of authorship records through a conservative approach of allocating revised author IDs. We explain our

method briefly, while more information about the task of author name disambiguation can be found in [27].

Our algorithm takes the list of authorship records of a single suspicious author ID and compares every pair of authorship records on the list. For each pair, we check the similarity of the author name (by comparing characters), the number of shared co-authors (on the two authorship records), the article classification (discipline subjects of the publication venue), the funding organization or grant numbers, and the affiliation addresses. For each topic of pairwise comparison, we assign a score that is higher if the two authorship records have similar traits, and a score that is lower or negative if they are dissimilar. We sum the score into a single value, and calculate a distance matrix for all combinations of records in the list. We obtain clusters of authorship records using the *agglomerative clustering* algorithm in the *scikit-learn* software package [28], and assign a revised author ID to each cluster of authorship records.

Among more than 250,000 distinct author IDs in our data, only 27 author IDs were associated with more than 276 publications. However, these 27 author IDs accounted for 15,229 authorship records associated with 4207 distinct publications (note that some publications are shared between these authors). After applying our disambiguation algorithm, we identified 130 disjoint clusters of reasonably similar authorship records to which we assigned 130 revised author IDs, instead of the 27 IDs allocated by Scopus. There were two groups for which the Scopus author IDs were not changed. Upon manually investigating such cases, we found particularly prolific researchers who happened to have authored over 276 publications each. The majority of the 27 groups of authorship records were broken into smaller groups as multiple revised Author IDs were allocated to their records. We manually checked some of the cases, and found that they were indeed authorship records of multiple distinct individuals sharing the same name whose records were correctly disambiguated by our algorithm.

Our algorithm performs well for particularly suspicious outlier records selected from the end of the publication count distribution. However, running it on all of the data would not be helpful, because it was not designed to be a replacement for the Scopus author identification system. The conservative idea behind the algorithm is that by default, each pair of authorship records is from a distinct individual unless there is sufficient evidence (of similarity) to the contrary. Running such an algorithm on other groups of authorship records sharing the same author ID may incorrectly break the authorship records of a single researcher into smaller groups, especially if the person has visible points of change in their career; if, for instance, they have changed their academic name, discipline, and co-authors.

3.2 Inferring states from affiliations using a simple rule-based algorithm

A second and highly important step of our data pre-processing is aimed at extracting the state from the affiliation addresses of each authorship record. Affiliations are not standard, and they vary substantially. Scopus breaks down these affiliations into the following variables: name of an institutional unit or department, name of an institution, street address, city, postcode, and country. However, many of the fields may be missing. For instance, some authorship records may contain only the name of an institutional unit (e.g., a research group) without any address or postcode. Therefore, we first need to harmonize the text data by removing accents and special characters, and by standardizing common abbreviations. Then, we use a simple rule-based algorithm that sequentially looks for each

of the affiliation variables in dictionaries where the cities, postcodes, and names of institutions are mapped to their respective states. Specifically, we use as dictionaries data from the Mexican Post Office [29]; the ANUIES inventory of academic institutions [30]; and CADIIP [31], an inventory of public institutions in Mexico. For addresses, the algorithm looks for the name or abbreviation of a state within the string. As a result, for each authorship record, we obtain up to five potential states based on the previous matching process. Finally, we define the state to be the mode of the five possibilities, as long as it is detected at least twice. In doing so, we exploit all the available geographic information per record, rather than relying on a single variable. For 75% of the data, the states can be extracted reliably using the mode state outputs of this simple, rule-based algorithm. However, for the remaining 25% of the authorship records, reliable states cannot be extracted using the rule-based method, and a more sophisticated method is therefore required to identify the states. In Sect. 3.3, we explain how we developed a neural network approach to address this issue; while in Sect. 3.4, we discuss how this approach is applied to extract the remaining states.

3.3 Developing a neural network and measuring its performance

We use the states obtained through the rule-based method as training data for developing a neural network for inferring the states of the remaining part of the data. We use the *Keras* library [32] with a *tensorflow* backend [33] in the Python programming language.

To assign a state to each row of the input consisting of a city, an institution, and an address, we adopted an approach commonly used in sentiment analysis literature [34] for predicting whether a sentence has a positive or a negative sentiment. Initially, we merge the city, the institution, and the address into one string for each row of the input. Then, we convert this string into a *feature vector* using a bag-of-words method with *term frequency inverse document frequency* [35] as the normalized frequency of a given word in a given row of the input relative to the frequency of that word in the whole data set. We use the words with the 3000 highest relative frequencies as the input layer of the neural network. The number of layers and the number of neurons of the neural network are depicted in Fig. 4.

The layers are densely connected, and have a dropout rate of 0.25. The neurons of the first layers use a rectified linear unit [36] as the activation function, while the output layer

Figure 4 A schematic representation of the neural network used for inferring states from affiliations
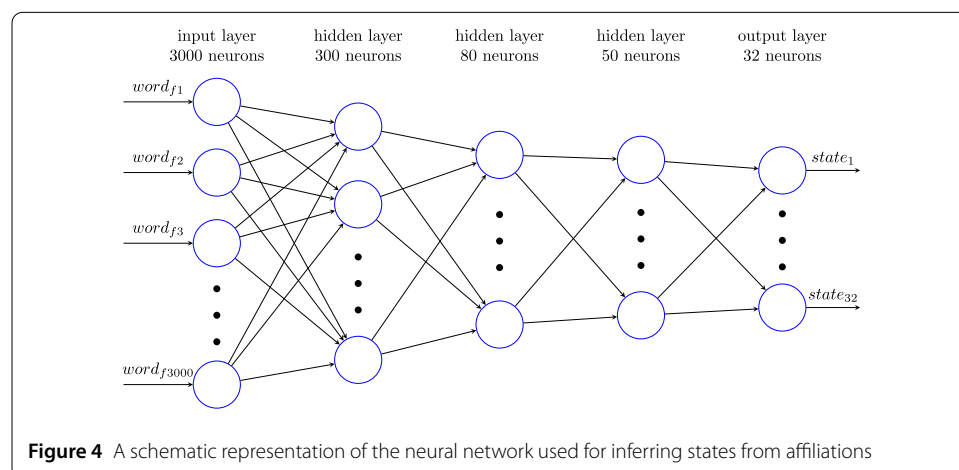
uses a *softmax* activation function [37]. The softmax activation function converts the activation of the output neurons into relative probabilities. The state that is assigned to the output neuron with the highest activation gets selected as the predicted state. We fine-tuned the number of layers, the number of neurons, and the dropout rate by trial and error to achieve a high degree of accuracy on the test data, while keeping the network as small as possible to avoid over-fitting.

We use the records for which the states were already obtained by the rule-based approach as training data for the neural network. Additionally, we manually labeled the states of 2448 authorship records that were not only unrecognizable by the rule-based approach, but were especially difficult to predict, as they failed on multiple rules. We added 1574 (70%) of those manually labeled records to the training set. Using a five-fold cross-validation, we obtained a median accuracy of 99.5% on this dataset. We then tested the accuracy of the method on the remaining 674 (30%) of the difficult-to-predict records, and obtained an accuracy rate of 96.2%. This could be considered a conservative evaluation for the accuracy of the neural network.

3.4 Inferring states from affiliations using the neural network

We used the neural network approach described in the previous subsection to predict the states of the records for which states could not be reliably obtained using our rule-based approach. We took additional steps to increase the overall reliability of the predictions. More specifically, we omitted a small fraction (less than 1%) of the total number of authorship records for which predicting a reliable state was particularly difficult even after using the neural network. Only the predictions with the lowest confidence score[d] were omitted from the dataset.

In summary, the states for 75% of the authorship records were reliably extracted using a simple rule-based method. The states for a remaining 24% of the authorship records were obtained using a neural network with a high degree of accuracy. Less than 1% of the authorship records were discarded because their states could not be extracted in a reliable way using either of our two methods.

3.5 Detecting moves based on changes in mode states across different years

After extracting states for the authorship records, we obtained the most frequent state for each researcher in each year. Note that, in some cases, a researcher may have multiple modal states in a year. We consider a move only if the changes in affiliations are such that the modal state of a researcher changes and the previous modal state disappears. For example, let us assume that the modal state for a researcher in years 2001, 2003, and 2006 are [Morelos], [Morelos, Hidalgo], and [Hidalgo], respectively. Our move detection algorithm iterates over the years in which the author has publications (and therefore modal states). When the algorithm reaches 2003, the modal state changes – it is a double mode – but the previous state has not disappeared; therefore, we do not consider that modal state change a move. Instead, we assume that the researcher is still affiliated with Morelos. When we reach 2006, the mode has changed, and the previous mode has disappeared. Thus, a move from Morelos to Hidalgo is recorded by the algorithm. To infer the time of the move, we consider the mid-point of the two years being processed by the algorithm when the move is detected. In the example above, the move-year is recorded as the average of 2003 and 2006. When the move-year is not an integer, we round it down to the closest integer.
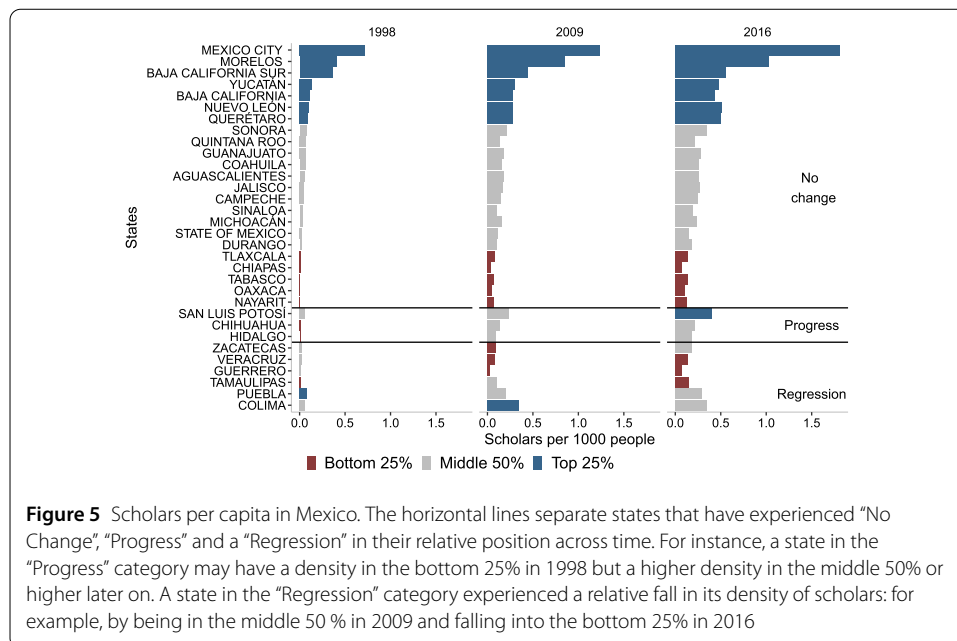
4 Results

4.1 General attributes of scholars in Mexico

One of the benefits of bibliometric data is that they enable us to conduct analyses at both the individual and the aggregate level. For example, we can obtain a profile of the median scholar in Mexico by looking at each scholar's mobility status, and complement this information with migration rates for each state. The data resulting from the pre-processing discussed in Sect. 3 allowed us to identify about 252,000 unique scholars in Mexico who were active during the 1996–2018 period. However, a large share (about 57%, or 146,000 authors) of these unique scholars are individuals who only have one authorship record, which prevents us from inferring their internal migration patterns. These one-time observations could be of scholars who moved abroad after publishing a single paper with a Mexican address, or of individuals who did not remain active in academia in Mexico. After removing these observations, about 96,000 authorship records are left, which cover 22,000 scholars who have moved at least once. That is, when we group individuals into moving and non-moving categories, the data show that only 22.8% of scholars who have published more than once also moved between states during the 1996–2018 period.

Three available characteristics of scholars are their academic age, number of publications, and total citation count. The publication dates allow us to compute the number of years each researcher has been active. *Academic age* is obtained by subtracting the year of an author's first publication from 2019, in the spirit of [20]. In doing so, we assume that scholars have not finished their academic trajectory, and that the first year of publication can be thought of as the birth year of a scholar. Overall, the median academic age of the scholars in our data is eight years, but the mobile scholars have a median age of 13 years, while their non-mobile counterparts have a median age of eight years. Indeed, the mobile scholars appear to be active for longer, which could be an artefact of having to build a more solid track record in order to be eligible for opportunities to transfer from one institute to another. In terms of a crude productivity measure, the median mobile scholar has five publications, whereas the median non-mobile scholar who has published more than once has three publications. Additionally, we can obtain a total citation count by author for the available publications, and normalize it by the academic age. In doing so, we control for the longer academic careers of some scholars. The median mobile scholar received 2.25 citations per year, while the median stationary scholar was cited 1.5 times per year.

A potential explanation for both of these productivity measures is that the typical mobile scholar has a longer career because it has taken them longer to accumulate networks to migrate to another state. At the same time, there could be a bias towards detecting mobility from researchers with more publications. While the Scopus database contains authorship records up to a certain date, the profiles of researchers are continuously evolving. Thus, the measures can only be used to infer the presence of differences between these two groups, rather than the magnitude of such differences.

Scopus data allow us to obtain complete scholar counts per year once we aggregate the data per state. However, the publications may not occur in consecutive years, and the bibliometric data can only offer screenshots at specific times. In the absence of an observation, we must assume that each scholar remained in the same state over the last and following two years of a publication.[e] After all, obtaining positions in other states and mov-

Figure 5 Scholars per capita in Mexico. The horizontal lines separate states that have experienced "No Change", "Progress" and a "Regression" in their relative position across time. For instance, a state in the "Progress" category may have a density in the bottom 25% in 1998 but a higher density in the middle 50% or higher later on. A state in the "Regression" category experienced a relative fall in its density of scholars: for example, by being in the middle 50 % in 2009 and falling into the bottom 25% in 2016

ing is costly and difficult in academia. After these initial analyses, we aggregate counts of scholars, and use separate CONAPO midyear population estimates [38] from each state to obtain the density of scholars, as seen in Fig. 5.

For every year, each bar represents the number of scholars per 1000 people in a given state, and its color indicates the state's relative position in a given year, which allows for comparisons in the presence of population growth trends. Over time, there is an increase in the density of scholars per state that reaches a maximum of 1.8 per 1000 residents of Mexico City. While many states remain in the same percentile, some states – San Luis Potosí, Chihuahua, Hidalgo – experience an increase in the number of scholars, and fall in the "Progress" band. There are still five states that do not fare as well as the rest of the states, even if their scholar density increases, and thus fall into the "Regression" category. Within this group, only Colima and Tamaulipas have shares of the population with tertiary education that are higher than the national value in the last available year (Fig. 2). When we look at the states with the highest densities (in blue) in 2016, we see Yucatán which, coincidentally, has a smaller share of the tertiary-educated population than the national average in Fig. 2. Therefore, the finding that a state has many scholars does not necessarily imply that the state has high levels of education. Moreover, the differences in the factors that attract scholars seem to persist. For example, in 2016, the density of scholars in Chiapas continued to be lower than that of Mexico City.

4.2 Measures of internal migration

Our data set provides the history of movements over years, from which we can determine whether a scholar has relocated to a different state. As the time that it takes to conduct a research project and to publish the results varies by discipline, for many scholars, there is a gap of years between each publication – and, therefore, between each potential recorded movement. As such, the time span used to define the entries into and the exits from a state may penalize scholars in areas with a typically lengthy publication process. Therefore,

using the mid-point between the year of change in the state of affiliation, as mentioned in Sect. 3.5, incorporates the assumption that movements to new affiliations are costly, and that the publication process is lengthy. We use a one-year net migration rate, $NMR_{i,t}$, of movements of scholars from state i in year t:

$$NMR_{i,t} = \frac{I_{i,t} - E_{i,t}}{N_{i,t}} \times 1000. \tag{1}$$

Overall, the net migration rates are calculated as the difference between the entering ($I_{i,t}$) and the exiting ($E_{i,t}$) populations in state i between time t and $t-1$, as a share of the total population of scholars ($N_{i,t}$) in the state at the given time t. Scopus data allow for the measurement of the numerators and denominators for the net migration rates of scholars. In the absence of Scopus data, assuming that the scholars' movements were similar to those of the general population would be a serious limitation.

We included several additional modifications for the components of this rate. As we mentioned previously, the numerator t is the year corresponding to the mid-point between the change in states of a scholar's movement history. We removed observations for which the total number of movements (entries and exits) fall within the lowest 15% of observations, which removed cases that may not be representative relative to the scholar population size. The denominator is the aggregate count of scholars used for Fig. 5, which uses a two-year backward and forward filled technique. As a result, the choice of filling locations limits the period for which migration rates can be computed to 1998–2016. Finally, from this point onward, all rates are expressed per 1000 people.

Figure 6 illustrates the net migration rates of researchers in 32 states of Mexico. The net migration rates vary across states and within regions to the extent that we can start considering that there are underlying characteristics beyond geography that cause these disparities. Intuitively, a positive rate means that the entering researchers surpass those who left, while the opposite is the case for a negative rate. Rates close to zero correspond to situations in which there is no apparent loss or gain of scholars. That is, the number of scholars who leave is mostly offset by the number of scholars who enter. As academic institutions are bound by physical and budgetary constraints, there is a limit to the number of scholars who can be attracted and expelled. Unlike population growth rates, a consistently positive net migration rate of scholars may not be sustainable for some states.

We can observe three patterns experienced by the states: (i) steady oscillations around zero, (ii) downward or upward period-specific trends, or (iii) large variation. When we consider the states of the Center, we see that in most of them, the rates fluctuate around zero (Guanajuato and San Luis Potosí); but that in others, such as Zacatecas and Hidalgo, the rates vary substantially. Consider the specific case of Aguascalientes. In recent decades, the state has become an industrial and technology hub that has attracted many universities and research centers. The initial downward but positive trend we observe in Aguascalientes may be the result of this industrial dynamism. However, with time, the growth in research opportunities may have decreased as the migration rate now approaches zero.

When we turn to the states in the Gulf of Mexico and the Yucatán Peninsula, we see that there are periods of substantial variation. Specifically, there are rates close to −100

Figure 6 Net migration rates for scholars by regions
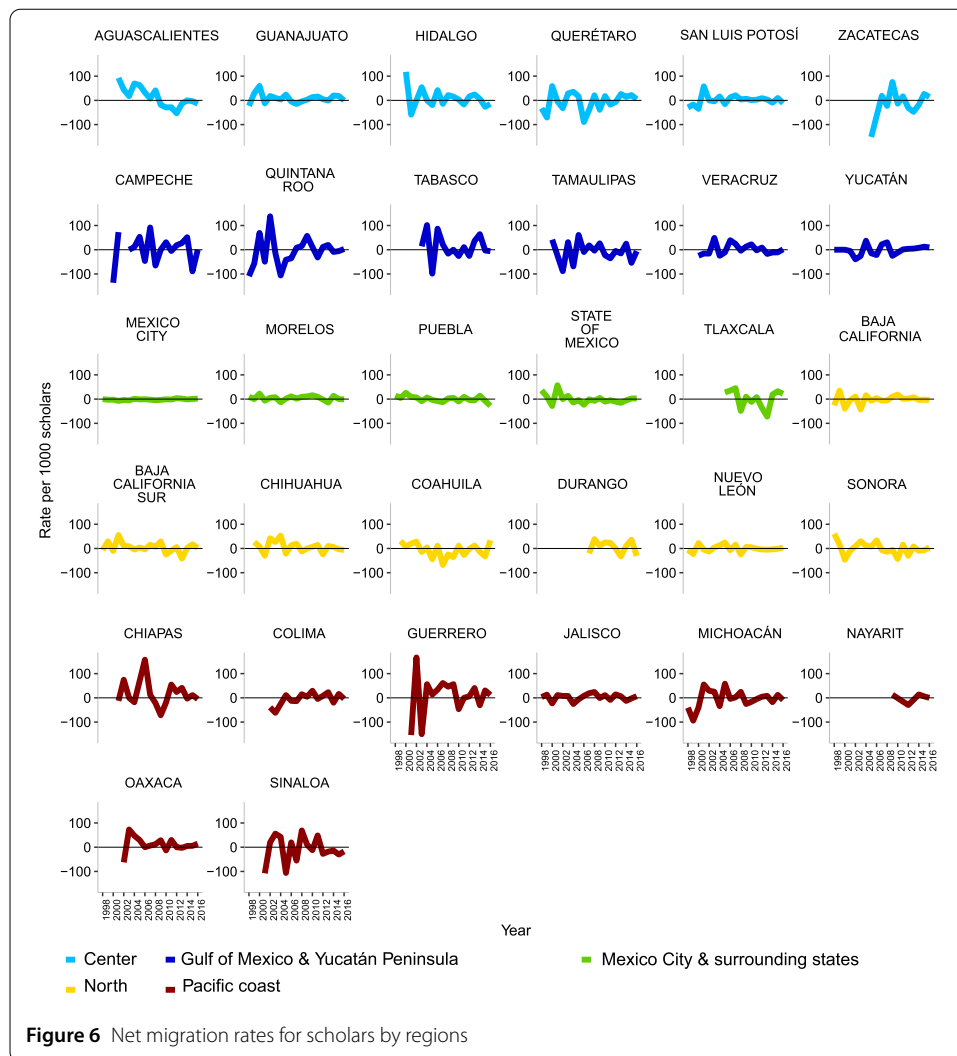
scholars per 1000 people. This means that there was a large-scale departure of scholars. These substantial variations coincide with the first few years of the Scopus data, which may be subject to a left-truncation such that our count of mobile and total scholars might be lower than the actual number. Therefore, the variation we observe may be an artefact of the lack of observed movements, so the results for the first few years of the analysis period should be interpreted with caution.

On the opposite end of the spectrum, Yucatán shows steady rates of around zero for the last years of the study period. We can focus on the states with the highest density of scholars as of 2016 from Fig. 5: Mexico City, Morelos, Baja California Sur, Yucatán, Baja California, Nuevo León, and Querétaro. Overall, their net migration rates fluctuate smoothly at around zero. Some of the lowest migration rates in absolute terms are found in Mexico City. It is possible that these states have already reached their capacity, but their institutions are still able to offer academic placements to scholars. Therefore, the zero net migration rates could imply that there is no brain drain in the states with the most scholars per capita.

Mexico City and its surrounding states, with the exception of Tlaxcala, together represent a notable case, as their rates are all close to zero. Beyond the researcher population,
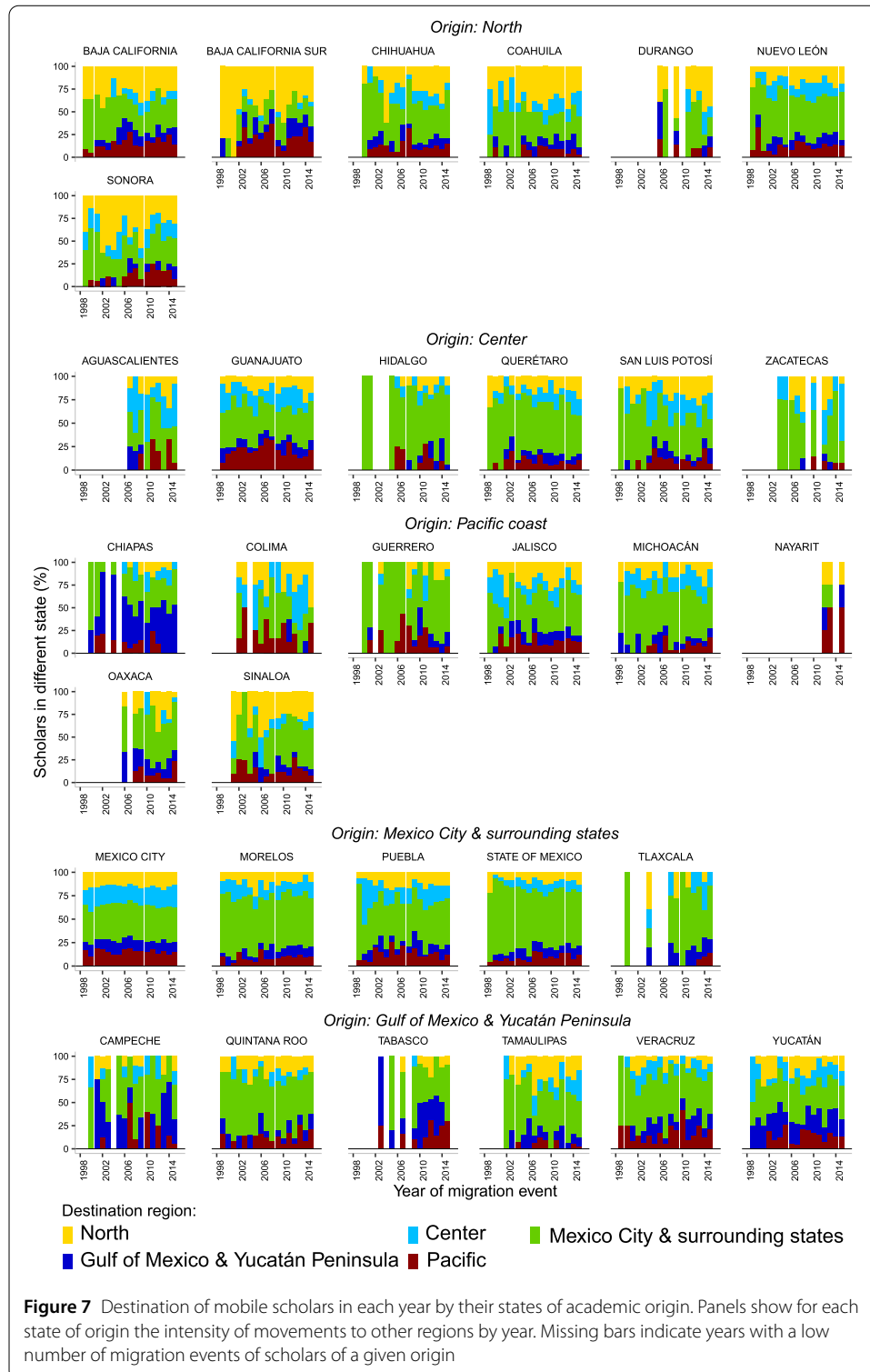
these states act as a large metropolitan area in which the economic and human flows are substantial. Unlike other regions, this area may be subject to overlapping characteristics that drive the similar migration rates. Conversely, the Pacific coast is an example of a region with large differences across states. Jalisco, Michoacán, and Guerrero are neighboring states, but each has distinct migration rates.

Overall, our analysis of the net migration rates is an example of how bibliometric data can be used to detect details of the profiles of states that would not be visible when using alternative data, such as the census. Another perspective on the movements of scholars is to focus on changes relative to the state of the first publication. Figure 7 contains plots of the shares of mobile scholars who have moved to other states, by state of academic origin. Each row groups the states of academic origin within a given region. For instance, the first plot of Fig. 7 shows that scholars who started off in Baja California mostly migrated to states within the Northern region, and to Mexico City and its surrounding states. There are cases in which the bars in the plot are missing; this occurs when there are few migration movements. Specifically, the lowest 15% of the number of movements within a state per year are dropped from the chart. That is, the empty space is significant, as it indicates the states in which the migration of scholars is an extremely rare event. In states such as Durango, Aguascalientes, Zacatecas, Nayarit, Oaxaca, and Tlaxcala, the processed data can detect substantive movements for the last decade only.

Figure 7 suggests that there is a degree of selectivity in the inter-state movements in a given state of origin. Scholars whose academic origins are in the Northern states tend to move within the region, with the exception of Nuevo León. Although the Center states are surrounded by all of the remaining groups, it appears that scholars in the Center states are strongly attracted to Mexico City and the surrounding areas, as well as to the Pacific Coast (for the case of Guanajuato). Most of the mobile scholars who come from one of the states along the Pacific Coast or the Gulf of Mexico also prefer Mexico City and the surrounding states. However, since these regions have many neighboring states from other regions, there is some variation of preferences within regions. For instance, scholars from Chiapas seem to prefer going to the Gulf of Mexico region which is in line with the fact that it is the closest region other than Mexico City and the surrounding states. Similarly, Sinaloa is next to Durango, which belongs to the Northern states, and many of its originating scholars constitute the yellow bars representing the Northern states.

The least diverse outcomes are observed for the mobile scholars who began their career in Mexico City and surrounding states, as most move within the region. Once more, a potential explanation for this finding is that economic and political activities are centralized in Mexico City, and these developments have caused spillover connections to the neighboring states. Unlike for all other states, the shares of scholars moving to each region who began their career in Mexico City are similar. That is, even if scholars originated in Mexico City, they tend to also move to other regions, which implies that the destinations of those with their origins in Mexico City are not concentrated within one region. This origin-destination analysis suggests that regardless of the scholars' state of origin, Mexico City and its surrounding states are preferred destinations.

The flexibility of the bibliometric data allows us to calculate additional migration rates to infer changes in the intensity and the redistribution effects of internal migration. First,

Figure 7 Destination of mobile scholars in each year by their states of academic origin. Panels show for each state of origin the intensity of movements to other regions by year. Missing bars indicate years with a low number of migration events of scholars of a given origin

the Crude Migration Intensity (CMI) measures the share of migration events relative to the population size [39]. Then, the Migration Effectiveness Index (MEI) and the Aggregate Net Migration Rate (ANMR) [39], measure the effect of migration on the redistribution

of a population within the country. Formally,

$$CMI_t = 100 \times \frac{M_t}{\sum_i N_{i,t}}, \tag{2}$$

$$MEI_t = 100 \times \frac{\sum_i |I_{i,t} - E_{i,t}|}{\sum_i (I_{i,t} + E_{i,t})}, \tag{3}$$

$$ANMR_t = 100 \times \frac{0.5 \sum_i |I_{i,t} - E_{i,t}|}{\sum_i N_{i,t}} \tag{4}$$

$$= \frac{CMI_t \times MEI_t}{100}, \tag{5}$$

where i denotes a region and t is the year of the movement.[f] The numerator of the MEI contains the total sum of differences between the number of scholars entering ($I_{i,t}$) and exiting ($E_{i,t}$) a given region i. The MEI measures the net migration balance in an area as a share of the total number of individuals who moved either from, or away from, the zone. When the MEI is large, for every movement, there is a large contribution of the net migration rate, which means that migration is efficient in redistributing the population of scholars. While the ANMR has the same numerator as the MEI, it is weighted by the total population of scholars ($N_{i,t}$), and it can be expressed in terms of the CMI and the MEI. By changing the denominator to the population of scholars, the ANMR becomes an indicator of the general effect of population redistribution of moving scholars on the population, as shown in its second specification. Finally, due to our data format, the CMI is computed from the ANMR and CMI by rearranging equation (5). Otherwise it would be obtained by the number of migration events (M_t), including those that enter and leave, over the total population at risk.

Figure 8 shows values of the three indicators for mobile scholars for the 1998–2016 period. The results of the CMI measure displayed in Fig. 8(a) shows that the country intensity is between five and nine, and has a decreasing trend. However, when we look at the intensity within regions, we see that the Center and Gulf regions experience the largest intensities. In other words, these regions seem to be more dynamic than the rest of Mexico as they experience more migration events after standardizing by their population size. This finding is an alternative reading of the results displayed in Fig. 7, which focused only on the shares relative to the origin state, rather than movements within the region regardless of the origin.[g] Mexico City and the surrounding states have the lowest regional CMI, which suggests that controlling for scholar population size reduces the regional importance of mobile scholars. While Fig. 7 shows that many scholars prefer Mexico City and the surrounding states, their contribution seems to diffuse relative to all scholars within the region.

In general, the country MEI in Fig. 8(b) shows a downward trend, which suggests that the movements of scholars are less efficient over time. When we examine the within-region migration efficiency of scholars, we observe that the MEI measures are highly volatile, except those for Mexico City and its surrounding states. In comparison to the rest of the states, the low MEI from the Mexico City area suggests that there is a less efficient movement of scholars within this region.

As expected, the results for the ANMR measurements shown in Fig. 8(c) are similar to those in the MEI. However, the ANMR values suggest that the turnover of scholars as a

Figure 8 Migration intensity indices, over time. Dashed line represent country indices
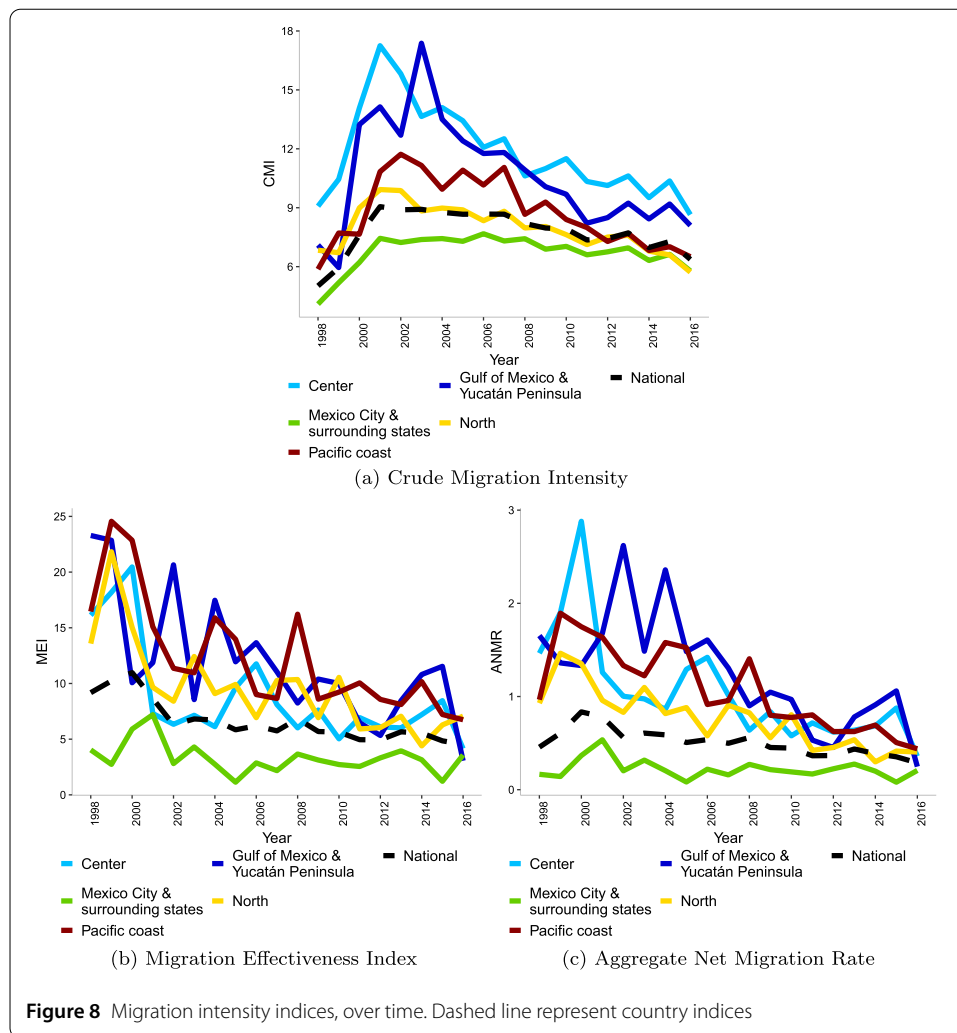
share of the population is much lower and stable for Mexico City and the surrounding states. Overall, these results jointly suggest that the net migration level to the Mexico City area is lower than the total volume of mobile scholars in that region, and, therefore, that the level of population redistribution is low.

The results presented throughout the section highlight the versatility of bibliometric data for demographic analyses of scholars. By re-purposing the data, we can create profiles for states based on their scholar population and their movements. Even when there is heterogeneity within the regions, states with higher scholar density tend to have migration rates around zero, which suggests that there is a brain *balance*, rather than gain or drain. However, even when we use intensity measures, we cannot analyze between-region patterns from a demographic standpoint. The following section contains a network analysis that provides a complementary perspective to help us better understand the mobility of scholars.

5 Internal migration as a network

The data from scholars who have moved between different states can be used to create a migration network by representing each state as a node, and creating a directed edge (i,j)

for each mobility event from state i to state j. Creating such a migration network enables us to analyze the system of scholarly internal migration in Mexico as a whole. We have made the data publicly available in a *FigShare* data repository [40].

5.1 Networks and their temporal dynamics

Figure 9(a) shows the direction and the magnitude of migratory movements of scholars in Mexico between 1996 and 2018. The five states that send the most scholars include Mexico City, the State of Mexico, Morelos, Nuevo León, and Guanajuato, respectively (these are also the states that receive the most scholars, and in the same order, except that Guanajuato outranks Nuevo León). These five states represent the capital city and its surrounding states, as well as the states that contribute the most to national GDP. As it can be seen in panel (a) of Fig. 9, the two flows between every pair of states are mostly reciprocal, with similar intensities (which means that the adjacency matrix of the directed network is very close to being a symmetric matrix). Aggregating the flows for each node, we see a similar pattern in node degrees. The in- and the out-degree of the same nodes are strongly correlated (the Pearson correlation coefficient is 0.999).

Mexico City has the highest incoming and outgoing flows of mobile scholars; i.e., its flows are almost three times as large as the flows in the State of Mexico, which is the second degree central node of the network. This is likely due to Mexico City's political and economic importance, and because it is home to many large national universities and research institutes. There is substantial heterogeneity in the degrees of the nodes. The states with the lowest in- and out-degrees are Nayarit, Tlaxcala, Colima, Durango, and Zacatecas (in increasing order of degrees); their in- and out-flows are less than 0.02 of those of Mexico City.
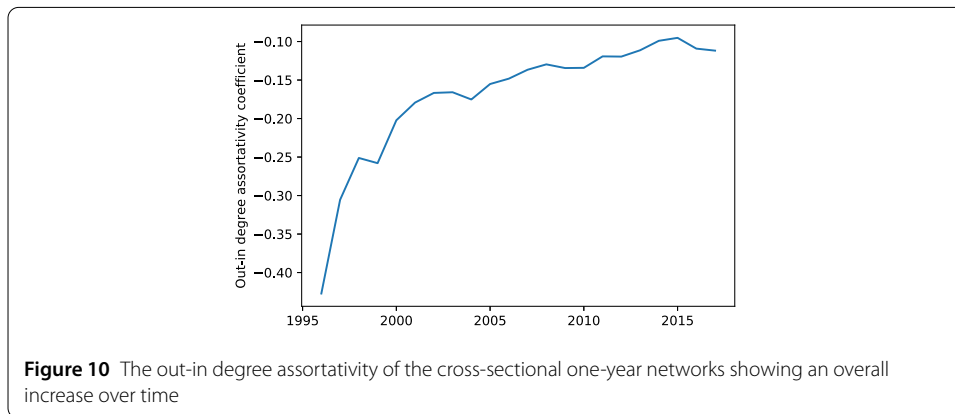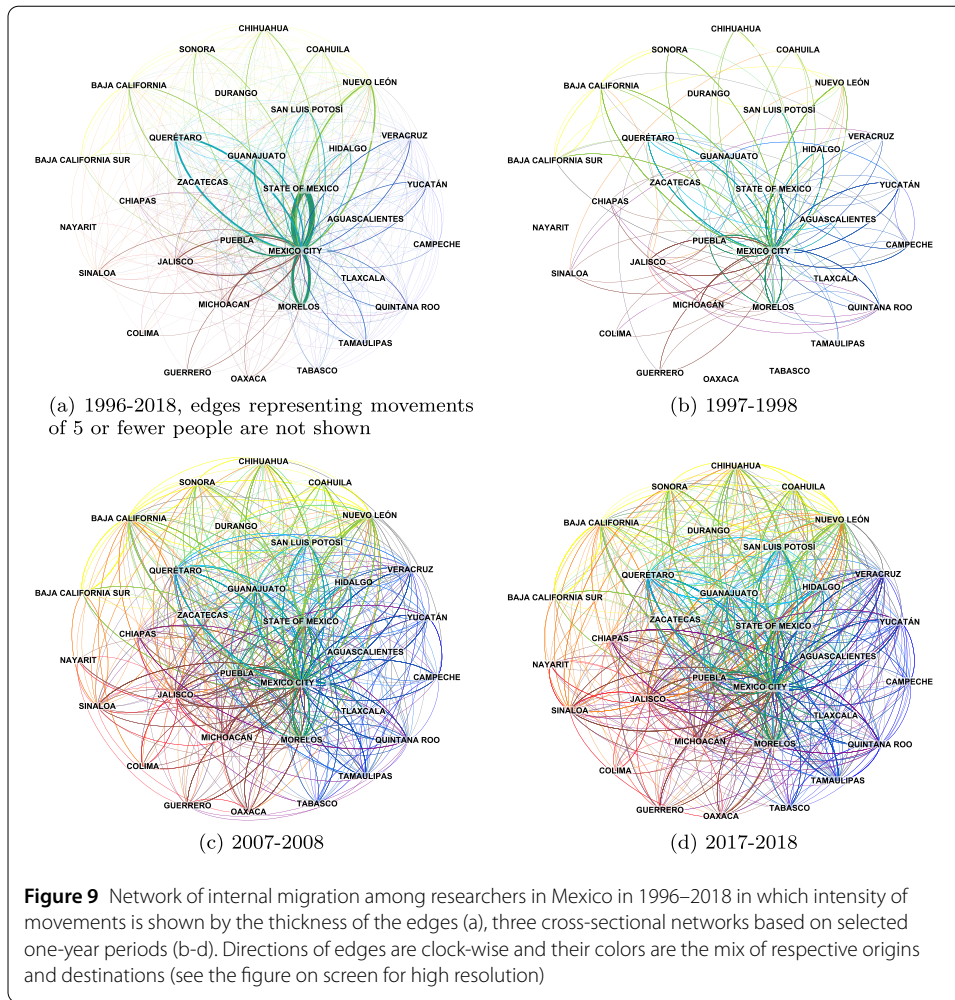
Subpanels (b)–(d) of Fig. 9 highlight the period movements of scholars between states. Overall, the migration network of researchers has not only become more dense (from a density of 0.08 in 1996–1997 to 0.5 in 2017–2018), but its edges have also become more diverse over the past two decades. For instance, in more recent years, the states along the Pacific coast (red) have had a greater exchange (purple edges) with states along the Gulf of Mexico and the Yucatán Peninsula (blue).

5.2 Assortativity of networks over time

The degree assortativity of a network captures the correlations between the degrees of adjacent nodes [41]. In many social networks, there is a tendency between nodes of a similar degree to connect (assortative mixing by degree); whereas in many technological and biological networks, high degree nodes tend to connect with low degree nodes (dissortative mixing by degree). Moreover, in some other networks, there is no correlation between degrees of adjacent notes (random mixing pattern) [42]. In directed networks, the correlation can be measured in four different ways by using either the in- or the out-degree for the source and target nodes (in-in, in-out, out-in, and out-out). We measure the out-in assortativity of directed graph G using Eq. (6) [41].

$$r(G) = \frac{\sum_{j,k}(e_{j,k} - q_j^{\text{in}} q_k^{\text{out}})}{\sigma_{\text{in}}\sigma_{\text{out}}}. \tag{6}$$

In Eq. (6), $r(G)$ is the out-in degree assortativity coefficient for directed graph G. $e_{j,k}$ is the probability that a randomly chosen directed edge leads into a vertex of in-degree j and

Figure 9 Network of internal migration among researchers in Mexico in 1996–2018 in which intensity of movements is shown by the thickness of the edges (a), three cross-sectional networks based on selected one-year periods (b-d). Directions of edges are clock-wise and their colors are the mix of respective origins and destinations (see the figure on screen for high resolution)

Figure 10 The out-in degree assortativity of the cross-sectional one-year networks showing an overall increase over time

out of a vertex of out-degree k. The term q_j^{in} (q_k^{out}) represents the excess in-degree (out-degree) distribution [41] of directed graph G and σ_{in} (σ_{out}) is the standard deviation of the distribution q_j^{in} (q_k^{out}).

Figure 10 illustrates $r(G)$ for the cross-sectional one-year migration networks. As it can be seen in Fig. 10, the degree assortativity coefficient of the networks has generally increased over time during the 1996–2018 period. This increase in assortativity can be ex-

plained as the gradual transformation of the network from a dissortative mixing pattern ($r(G)$ = −0.45) to a random (less dissortative) mixing pattern ($r(G)$ = −0.1).

In the first few years of the period under study, movements were mostly from low out-degree nodes (states with small outgoing flow) to high in-degree nodes (states with large incoming flow), and from high out-degree nodes to low in-degree nodes. Given the strong correlation between the in- and the out-degrees of the same nodes, all four types of degree assortativity measures show similar increases, albeit with minor differences (in-in, in-out, and out-out plots are not shown to avoid redundancy). Therefore, we can say that the majority of flows in the first few years were from states with low flows to states with high flows, and vice versa.

However, over the last few years, the degrees of adjacent nodes are hardly correlated. The mixing patterns of the networks in more recent years display less dissortativity, and instead indicate that researchers are circulating between the states, irrespective of their degrees. This observation suggests that the mixing pattern of the network is affected by a relative increase in mobility between states with similar flow intensities (low-flow to low-flow and high-flow to high-flow) in the past two decades.

5.3 Community structure induced by the dynamics of flows

To detect communities in networks, different approaches (and algorithms) can be used. According to Fortunato and Hric [43], these approaches belong to five categories: optimization, statistical inference, dynamics, consensus clustering, and spectral methods. As the edges in our migration networks represent flows between states, we expect that detecting and analyzing communities based on the dynamics of the flows may reveal a more meaningful structure. *InfoMap* is a popular algorithm for detecting communities from the dynamics category [44], which relies on random walk dynamics in the network, and on the intuition that a hypothetical random walker stays in the dense regions of the network for a longer period of time. For running InfoMap, we use the *MapEquation* software package [45, 46] to detect how the flows between the states lead to the formation of network communities between them. The network communities could be thought of as different components of the Mexican scholarly migration system, which are revealed due to the heterogeneity of the flows.

For this purpose, we have created an alluvial map [45] of the network flows and communities over time, which is illustrated in Fig. 11 (99% of the total flow is shown). The height of each community is proportional to its total flow. The structure of these communities clearly shows a dense core, with relatively few nodes accounting for the majority of the total flows, and an outlying, loosely connected periphery made up of a relatively large number of nodes accounting for a small portion of the flows. Such a network structure is referred to as a core-periphery structure [47].

As it can be seen in Fig. 11, there were eight communities in the migration network in 1996. The core was made up of 14 states listed in the figure. The other communities in 1996 were smaller, and were made up of at most three states (as shown in Fig. 11). Most of the communities of 1996 have a dynamically changing flow over different years, and therefore break into smaller groups over time, which join back together in some years. For instance, the community shown in dark blue is made up of Baja California, Baja California Sur, and Coahuila in 1996. This community breaks into different groups the next year, and then reforms to its previous state one year later. By 2018, Baja California and Baja California Sur are together in a two-node community.

Figure 11 Communities of the states over time featuring a core-periphery structure
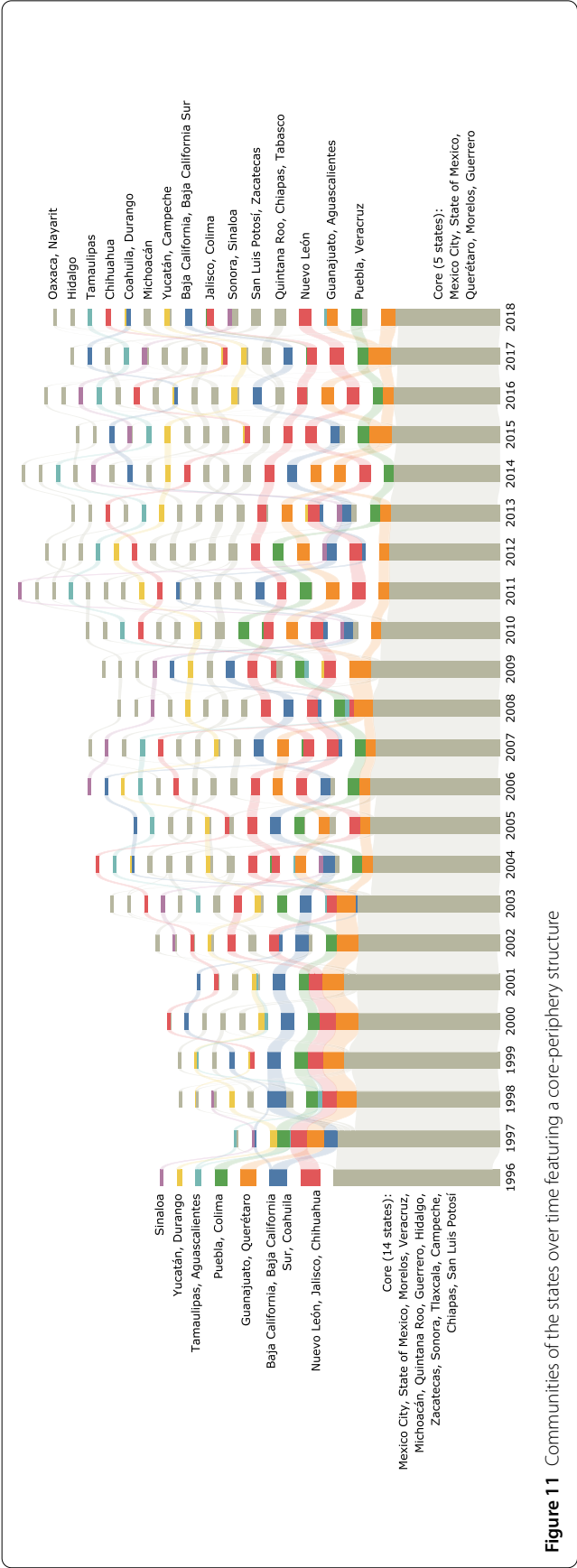
As the dynamics of flows in some other communities are more volatile, their nodes may break off at some point, and never again form a community. For instance, Nuevo León, Jalisco, and Chihuahua formed a three-node community in 1996, which is shown in dark red. The three nodes then diverge into different groups. Eventually, in 2018, each of them is a part of a different community, possibly with other states. In 2018, the number of communities increases to 16, and the core community becomes smaller, both in terms of the relative share of the annual flow (represented by height) and the number of nodes, which is reduced to five (states are listed in the figure).

6 Discussion and summary

This project was undertaken to analyze the internal migration of researchers within a country, and to enhance and evaluate the potential suitability of bibliometric data for such an analysis. After investigating what is achievable within the limits of classic sources of migration data, such as surveys and censuses, we outlined a detailed methodological framework for re-purposing bibliometric data for studying internal migration.

Our framework involves an author name disambiguation method that enhances the quality of author identification offered by the provider of bibliometric data, and ascertains its accuracy in identifying the publications of the same author by treating the outliers. An initial step of our proposed method involves a parsimonious rule-based algorithm that infers states from affiliation addresses. The output of this step, combined with manually tagged data, is then used to train a neural network for predicting the states for the rest of the data. The produced state variable is used, as the sub-national level of data aggregation on which the locations of researchers are inferred over a number of years. The final step of our methodological framework deals with detecting migration events from changes in modal states of affiliations. While we illustrated this methodological framework for the case of Mexico, this approach is applicable to a broader context.

In an effort to bridge the gap between demographic methods and the literature on the migration of scholars, we estimated key internal migration measures over time, such as the net migration rate, the migration effectiveness index, and the aggregate net migration rate. To the best of our knowledge, these types of estimates have not previously been produced using bibliometric data. Substantively, we provided a detailed picture of scholarly migration within Mexico, we showed patterns of movements by origin and destination over time, and we offered comparisons of the states and regions with respect to scholarly migration.

We also modeled the movements of researchers as a network, which enabled us to analyze them holistically, and to evaluate their progression through time. Using the degree of the nodes, we identified the symmetry of pairwise flows of opposite direction, and the heterogeneity that exists at the level of flows between the states. Using degree assortativity, we explained the temporal changes in the correlations of the in- and the outflows between the states, and how the movements between states with similar total flows have increased relatively over time. We documented a core-periphery structure in the network, and explored its dynamics through detecting network communities induced by the changes in the flows between states. The analysis revealed a system whose core community has decreased in both size and total flow over time, while smaller peripheral nodes gained a larger share of the network flow and formed their own communities.

The changes observed in patterns of scholarly migration between states can be viewed from the perspective of a theoretical model like Zelinsky's migration transition model [2]

introduced in Sect. 1. Although Zelinsky's model is based on a general population of migrants, it may have a bearing on mobile sub-populations such as scholars, as we argue that their mobility could change with respect to the developmental stages of the society. The initial patterns of migration between rural and urbanized states for the first few years of the period under study (Fig. 9 (a)) was followed by an increase in assortativity, under which flows between nodes of similar degrees (including high-degree nodes) grew in relative terms (Fig. 10). This suggests that mobility among scholars in Mexico might have been experiencing a phase similar to Zelinsky's *late transitional society* [2]. The assertion of the theory is that at this stage, there is a relative increase in migration between urban centers. This, in turn, results in circular migration within a single metropolitan region of the network. With our current analysis, we cannot test whether broad mobility theories, like the migration transition model, also apply to scholars. Similarly, providing conclusive evidence on how the trajectory of scholarly migration is similar to or different from the experience of the general population is beyond the scope of this article. However, we believe that our analysis, which is mostly empirical in nature, lays the foundations for testing and developing theories that can rely on the analytical framework developed by migration scholars, and the richness of appropriately processed bibliometric data.

By offering novel methodological contributions, our work demonstrates that large-scale longitudinal bibliometric data could be harnessed to obtain valuable insights into the internal migration patterns of scholars when coupled with an algorithmic method for producing a sub-national level for data aggregation. Among other contributions, we documented (i) that mobile scholars, on average, publish more and have higher citation counts than non-mobile scholars; (ii) that there is substantial heterogeneity in net migration rates across states, but that Mexico City and the surrounding states maintain a position of centrality as key areas that attract scholars from all regions of Mexico, and send scholars to all states; (iii) that, overall, the crude migration intensity and the migration effectiveness index for scholars have declined over time, which could indicate that the redistribution of scholars in Mexico has decreased in the past two decades.

More broadly, we leveraged demographic techniques to provide extensive results on scholarly migration in Mexico, a topic that has been understudied despite the considerable size and impact of the country's research. The methodological framework proposed in this study aims to facilitate organizing data and information about the migration of scholars that can be used to evaluate the likelihood of alternative hypotheses, and to build the foundations of a theory of scholarly migration.

While standard demographic measures, such as net migration rates, are essential for quantifying migratory movements, a more comprehensive picture of scholarly migration is obtained by additionally drawing from network approaches. Demographic and network approaches complement each other in providing a more comprehensive view on the dynamics of scholarly migration that is consistent with the transitional nature of migration systems. For example, while we observed a general decreasing trend in the crude migration intensity over the past two decades, we also found that the migration network has become more dense and more diverse, and has been characterized by increased exchanges between the states along the Gulf and the Pacific Coast. The combination of methods and data that we present opens up new opportunities for developing a theoretical framework for understanding scholarly migration within country boundaries.

Availability of data and materials
The bibliometric data used in this study is proprietary and cannot be released. Scopus data is owned and maintained by Elsevier. Our network dataset is made publicly available in a *FigShare* data repository [40].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AMG and SA did the majority of the work related to the research design, the data processing, the analysis, and the writing and editing of the manuscript. TT contributed to the development of the neural network and the author disambiguation algorithm. TT and EZ contributed to the writing and revision of the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Demography, University of California Berkeley, Berkeley, CA, USA. [2]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research, Rostock, Germany.

Endnotes
   [a] SciVal is a research profiling system and a web-based analytics solution provided by Elsevier www.scival.com (accessed on 25/06/2020)
   [b] Aggregate data were obtained from INEGI [48], which defines higher education as technical or commercial studies with concluded high school studies, professional studies (bachelor's degree or equivalent), specialties, master's degree, or PhD.
   [c] While this number is not chosen through a rigorous scientific process, we remind the reader that the aim here is to consider a threshold for particularly suspicious authorship records whose size could vary depending on the quality of the author disambiguation method used by the provider of the bibliometric data. Given the very small fraction of such records in the Scopus data, this process would not change our substantive results and discussion. Nonetheless, we apply a healthy degree of skepticism to ensure a sound data analysis that starts with a treatment of outliers.
   [d] The confidence score is the activation of the neuron with the highest activation in the softmax output layer.
   [e] Essentially, we fill backwards and forwards the two neighboring missing values around a publication year. Due to this padding, the effective sample when using scholar counts is reduced to 1998–2016.
   [f] We extend the time-invariant measure in [39] into a time series.
   [g] For example, the North CMI is composed of the yellow bars of states that belong in the Northern region, and of the yellow bars from researchers who did not start off in this region but eventually moved within it.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
 1. Massey DS, Arango J, Hugo G, Kouaouci A, Pellegrino A, Taylor JE (1993) Theories of international migration: a review and appraisal. Popul Dev Rev 19(3):431–466
 2. Zelinsky W (1971) The hypothesis of the mobility transition. Geogr Rev 61(2):219–249
 3. Coale AJ (1989) Demographic transition. In: Eatwell J, Newman P, Milgate M (eds) Social economics. Palgrave Macmillan, London, pp 16–23
 4. Colledge L (2014) Snowball metrics recipe book. Snowball Metrics Program Partners, Amsterdam
 5. Aupetit SD (2006) The brain drain in Mexico—a subject for research… or agenda? Glob Soc Educ 4(1):103–120
 6. Lancho-Barrantes BS, Cantú-Ortiz FJ (2019) Science in Mexico: a bibliometric analysis. Scientometrics 118(2):499–517
 7. Uddin A, Singh VK, Pinto D, Olmos I (2015) Scientometric mapping of computer science research in Mexico. Scientometrics 105(1):97–114
 8. Reyes-Gonzalez L, Gonzalez-Brambila CN, Veloso F (2016) Using co-authorship and citation analysis to identify research groups: a new way to assess performance. Scientometrics 108(3):1171–1191
 9. Macías-Chapula CA (2013) Comparative analysis of health public policy research results among Mexico. Chile Argent Scientometrics 95(2):615–628
 10. Marmolejo-Leyva R, Perez-Angon MA, Russell JM (2015) Mobility and international collaboration: case of the Mexican scientific diaspora. PLoS ONE 10(6):0126720

11. Téllez Vázquez Y, López Ramírez J, Romo Viramontes R (2014) Prontuario de migración interna. Technical report, Consejo Nacional de Población (CONAPO), Mexico City, Mexico

12. Lozano Ascencio F, Gandini L, Ramírez-García T (2015) Devaluación del trabajo de posgraduados en México y migración internacional: los profesionistas en ciencia y tecnología. Migr desarro 13(25):61–89

13. Rodríguez Gómez R (2009) Migración de personal altamente calificado de México a estados unidos: una exploración del fenómeno. Rev Electrón Investig Educ 11(2):1–34

14. Subbotin A, Aref S (2020) Brain drain and brain gain in Russia: Analyzing international mobility of researchers by discipline using Scopus bibliometric data 1996–2020. https://arxiv.org/pdf/2008.03129

15. Tuirán R, Ávila JL (2013) Migración calificada entre México-estados unidos: desafíos y opciones de política. Migr desarro 11(21):43–63

16. Minnesota Population Center (2019) Integrated Public Use Microdata Series, International: Version 7.2 [dataset] Minneapolis. https://doi.org/10.18128/D020.V7.2

17. Czaika M (2018) High-skilled migration: drivers and policies. Oxford University Press, New York. https://doi.org/10.1093/oso/9780198815273.001.0001

18. Czaika M, Orazbayev S (2018) The globalisation of scientific mobility, 1970–2014. Appl Geogr 96:1–10. https://doi.org/10.1016/j.apgeog.2018.04.017

19. Robinson-García N, Sugimoto CR, Murray D, Yegros-Yegros A, Larivière V, Costas R (2019) The many faces of mobility: using bibliometric data to measure the movement of scientists. J Informetr 13(1):50–63. https://doi.org/10.1016/j.joi.2018.11.002

20. Aref S, Zagheni E, West J (2019) The demography of the peripatetic researcher: evidence on highly mobile scholars from the Web of Science. In: Proceedings of the 11th international conference on social informatics. Lecture notes in computer science

21. Moed HF, Plume A et al (2013) Studying scientific migration in Scopus. Scientometrics 94(3):929–942

22. Moed HF, Halevi G (2014) A bibliometric approach to tracking international scientific migration. Scientometrics 101(3):1987–2001

23. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G (2008) Comparison of PubMed, Scopus, Web of Science, and Google scholar: Strengths and weaknesses. FASEB J 22(2):338–342

24. Kawashima H, Tomizawa H (2015) Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. Scientometrics 103(3):1061–1071

25. Aman V (2018) Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates. Scientometrics 117(2):705–720. https://doi.org/10.1007/s11192-018-2895-3

26. Tekles A, Bornmann L (2019) Author name disambiguation of bibliometric data: a comparison of several unsupervised approaches. arXiv:1904.12746

27. D'Angelo CA, van Eck NJ (2020) Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. Scientometrics 123:883–907

28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

29. Servicio Postal Mexicano (2019) Descarga de Códigos Postales. Data retrieved from Servicio Postal Mexicano. https://www.correosdemexico.gob.mx/SSLServicios/ConsultaCP/CodigoPostal_Exportar.aspx

30. Asociación Nacional de Universidades e Instituciones de Educación Superior (2018). Anuarios Estadísticos de Educación Superior. Data retrieved from ANUIES http://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

31. Instituto Nacional de Estadística, Geografía e Informática (2012) Catálogo de dependencias e instituciones de interés público 2012. Technical report, Mexico City, Mexico

32. Chollet F, et al (2015) Keras: the Python Deep Learning library. https://keras.io

33. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. Software available from https://tensorflow.org/

34. Mäntylä MV, Graziotin D, Kuutila M (2018) The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. Comput Sci Rev 27:16–32

35. Tokunaga T, Makoto I (1994) Text categorization based on weighted inverse document frequency. In: Special Interest Groups and Information Process Society of Japan (SIG-IPSJ). Citeseer

36. Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for lvcsr using rectified linear units and dropout. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE Comput. Soc., Los Alamitos, pp 8609–8613

37. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

38. Consejo Nacional de Población (2018) Proyecciones de la Población de México y de las Entidades Federativas, 2016–2050. Data retrieved from the CONAPO, https://datos.gob.mx/busca/dataset/proyecciones-de-la-poblacion-de-mexico-y-de-las-entidades-federativas-2016-2050/resource/c9ae0945-20c0-4a21-b040-fc0679c64d7a

39. Bell M, Blake M, Boyle P, Duke-Williams O, Rees P, Stillwell J, Hugo G (2002) Cross-national comparison of internal migration: issues and measures. J R Stat Soc, Ser A, Stat Soc 165(3):435–464

40. Miranda-González A, Aref S, Theile T, Zagheni E (2020) Dataset of internal migration among researchers between states in Mexico over 1996–2018. FigShare. https://doi.org/10.6084/m9.figshare.12619016

41. Newman ME (2003) Mixing patterns in networks. Phys Rev E 67(2):026126

42. Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89:208701. https://doi.org/10.1103/PhysRevLett.89.208701

43. Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659:1–44. https://doi.org/10.1016/j.physrep.2016.09.002

44. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123. https://www.pnas.org/content/105/4/1118.full.pdf. https://doi.org/10.1073/pnas.0706851105
45. Rosvall M, Bergstrom CT (2010) Mapping change in large networks. PLoS ONE 5(1):1–7. https://doi.org/10.1371/journal.pone.0008694
46. Edler D, Eriksson A, Rosvall M (2013) The MapEquation software package. https://mapequation.org
47. Borgatti SP, Everett MG (2000) Models of core/periphery structures. Soc Netw 21(4):375–395
48. Instituto Nacional de Estadística, Geografía e Informática (2015) Características educativas de la población. Data retrieved from INEGI. https://www.inegi.org.mx/