Migration and Changes in Research Interest: A Life Cycle Analysis

Christian Chacua[*†] Richard B. Freeman[‡§]

February 27, 2022

Draft version. Please do not cite or distribute without permission of the authors.
Click here to download latest version

Abstract

Scientists' research topics change over time as research resolves some problems and raises new ones, as experience improves the skills and knowledge of the scientist, and as scientists move among locations that focus on different scientific problems. In this paper, we analyze changes in the geographic context of the papers of scientists from Colombia, a country with historically high rates of high-skilled emigration, when they emigrate to another country and when they return to Colombia. We measure geographic content by the geography terms in their titles and abstracts, and by the geography of their reference lists. By analyzing the geographic context of a paper, our analysis goes beyond traditional brain drain/brain circulation studies of the migration of researchers to how migration affects their actual work. We find that changes in geographical location are associated with changes in research topics in the social science and medical research fields where a geographic place is naturally salient and in natural sciences and engineering fields as well. Researchers who emigrate from Colombia write fewer papers with Colombia geographic context than comparable researchers who remain in the country, while migrants who return to the country publish more on topics relevant to Colombia than those who continue to work outside the country. As best we can tell, the primary factor for change in geographic context is the act of migration. In addition, we find a life cycle pattern in scientists referring to papers written in Colombia relative to papers with addresses outside the country: as a researcher ages, he/she shift from local sources of knowledge to global sources.

Key words: migration, research interest, academic careers, mobility of scientists, NLP.
JEL: O15, I23, J61, O30, C45

[*]Bordeaux School of Economics, UMR CNRS 6060 - University of Bordeaux, Bordeaux, France
[†]Growth Lab, Center for International Development - Harvard University, Cambridge, MA, United States
[‡]Department of Economics - Harvard University, Cambridge, MA, United States
[§]NBER, Cambridge, MA, United States

Corresponding author: Christian Chacua. Univ. Bordeaux, CNRS, BSE, UMR 6060, F-33600 Pessac, France. Email: ccdelgado@u-bordeaux.fr

We are grateful to Ernest Miguelez, Francesco Lissoni, Frank Neffke, Ricardo Hausmann, Christian Fons-Rosen and Aldo Geuna for their useful remarks. We also received valuable suggestions from participants at the Harvard Seminar in Economics of Science & Engineering (April 2021), the 6$^{\text{th}}$ Summer School on Data & Algorithms for ST&I studies (September 2021), the 5$^{\text{th}}$ RIE and 12$^{\text{th}}$ MEIDE conference (December 2021), the 4$^{\text{th}}$ RISE Workshop (December 2021) and the WICK#9 (December 2021). We acknowledge the financial support from the French National Research Agency under the program Investissements d'avenir (ANR-10-IDEX-03-02) and the TKC project (ANR-17-CE26-0016). Computer time for this paper was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the University of Bordeaux and of the University of Pau and the Adour Region.

0 Introduction

The brain-drain literature has traditionally looked at the emigration of highly skilled workers from lower income countries to higher income countries as a loss of human capital that limits the scientific /technical progress and economic growth of the source country; and viewed the return of emigrant scientists as desirable brain gain that can enhance local science and economic development. While work on brain circulation (Saxenian, 2005), ethnic networks (Kerr, 2008) and scientific collaborations (Xie & Freeman, 2021) directs attention at the ways emigrant and return migration benefits the source country by enhancing flows of information from the destination country to it, the existing literature focuses on the flows, with little attention to whether migration leads researchers to change their research to fit the country to which they move.

This paper assesses the effect of migration on the geographic context of research papers of scientists who move from Colombia, a middle income country that graduates relatively large numbers of scientists who emigrate to high income countries, of whom a substantial number return to the country. We focus on Colombia, in part because it has a unique detailed CV data set on scientists that allows us to measure career progression before and after spells of migration overseas and the geographic context of their papers. We measure geographic context via geographic terms in the titles, abstracts, and reference lists of published works. While we examine the movement of scientists between a low income and higher income countries at the heart of debates over the importance of brain-drain vs brain circulation, the methodology can be used to analyze any form of mobility in which research has a geographic dimension, such as moves from one area to another in a country or from one institution to another.

Our major finding is that migration is associated with changes in our titles and abstracts measure of Colombia's geographic context and in the addresses of authors of the papers that Colombian scientists refer to in their paper. Colombian scientists who go abroad shift from researching issues relating to Colombia to other issues. Emigrant scientists who come back to Colombia are, conversely, more likely to change interests towards Colombia. In both cases we contrast changes in the geographic context of papers of the migrants with that of otherwise similar scientists who did not change their location in the relevant period – for emigrants, the comparison group are Colombians who continued working in the country while for returnees the comparison group are emigrant scientists who stay abroad.

We also find that the distribution of references in the scientific papers of Colombian scientists changes over their careers regardless of where they locate. As scientists acquire more research experience, they reference more documents outside their country relative to local documents, even when their content is limited to national issues. Experience expands the base of knowledge from local to global.

The paper has four sections. Section one explains the value of looking at science through the lens of geography and of taking account of the scientist's career life cycle in the analysis. It gives the conceptual framework for our empirical work. Section two describes our data and how we use it to measure the geographic context of publications. Section three presents estimates of the impact of the location and life cycle of scientists on the geographic context of research. Section four comments on ways to extend our analysis and on its implications for migration policy.

1 Conceptual framework for analyzing geographic content of research paper

1.1 Research interests/topics associated with geography

The geographic context of a paper, defined as the extent to which a given paper refers to geographic aspects of the evidence it analyzes, is the natural lens to assess the effects of migration on research interests/topics. The key assumption on which the analysis builds is that a researcher in a given locality has more incentives to work on a topic relevant to that locality than to work on a similar topic in another location. In the case at hand, a researcher in Colombia and a researcher in another country with the same scientific interest could use the same scientific models and methodologies to examine problems but their subject matter would be country-specific. For example, a Colombian botanist interested in orchids would study the Cattleya trianae (an orchid unique to Colombia that makes it the country's national flower) rather than an orchid that grows in, say, France, while the French botanist would study the French orchid. A Colombian trade economist would focus on Colombia's international trade while the American trade expert would analyze the US trade balance. And so on.

Migration to a new locality would reduce the incentives of working on topics relevant to the new locality and potentially raise the incentives of working on topics relevant to the researchers' past location, producing a shift in the geographic component of their research. In trade theory lingo, we assume that someone in Colombia has a comparative advantage in researching issues relevant to Colombia, which they would lose if they go elsewhere, leading to a reduction in work on Colombian issues with emigration; while someone moving back to Colombia would shift her comparative advantage to Colombian topics.

There are at least three main reasons why a researcher could have more incentives to work on local topics rather than issues from far away geographies. First, as scientists' set preferences may be biased towards the places that have particular importance for them, it is natural to think that the place they live in has significant relevance for them. Moreover, if the place they live in is the place they have grown up or the place their family lives, they may have more personal incentives to care about it. Second, scientists' social context may create larger incentives to conduct research about their local environment. The institutions, funding environment, interactions with other colleagues co-located in the same geography (research networks), and the influence society exerts on scientists' research agendas can create larger incentives to study local geographies. Finally, local knowledge, local resources, and local information can make it less costly to work on a topic relevant to that locality than to work on a similar topic in another location. The costs of local research inputs and the cost of learning about local issues can be lower than for foreign ones, as scientists can have privileged access to information about their local context and to local know-how.

The geographic context of research varies widely among fields. Much empirical social science uses data or problems from a particular country – the scholar studies the school system, unionism, criminal justice, politics of a given country – making social science a natural place to expect migration to impact the geographic context of research topics. Similarly, much of medical science uses data or problems with diseases that have a geographic dimension – tropical diseases for tropical countries, studies of Alzheimer's disease in countries with an aging population, and so on, with implications that research topics would differ geographically. Analysis in more abstract fields such as mathematics or physics or astronomy (where the geography goes beyond any local area on earth) would appear to have less dependence on geography, but differences in local knowledge and orientation could readily affect research topics and show up in measures of geographic content. Russian mathematicians have historically pioneered some branches of math while Hungarian mathematicians helped create graph theory and combinatrics. The location of colliders and of telescopes of different types impact the work of physicists and astronomers, so that a move to a different location would likely change the astronomer or physicists focus of work. In these cases, measures of geographic context based on geographic terms in titles, abstracts, or references would work better if the collider or telescope was counted as part of the comparative advantage for being at particular locations.

Finally, there is another way in which location impacts the geographic context of research. This is through the geographic distribution of prestigious scientific journals, most of which are published in high income countries. To the extent that those journals focus largely on the problems of high income countries, and researchers want to publish in them, there will be bias in research interest toward high income country problems. Pursuing this logic, several studies claim the research topics are overly concentrated on the problems faced by rich developed countries, particularly the United States (Das et al., 2013; Porteous, 2020; Chavarro et al., 2017). Others report that the growing pressure to meet bibliometric indicators has amplified the incentive for researchers to shift their research interests from regional and contextualized problems in low-income countries towards high-income research agenda (Bianco et al., 2016; Martin & Whitley, 2010; Das et al., 2013).

1.2 Geography and academic fields

1.3 Life cycle and cohort effects

In addition to analyzing how migration affects the geographic content of research, we model the effect of two factors that might impact the researchers' choice of topic that could confound the effect of migration on that choice: life cycle changes in geographic concerns of scientists as they age and gain seniority; and cohort changes in geographic concerns due to which topics are "hot" when scientists begin their careers as students or post-docs. Life cycle changes are cogent because when a person migrates from one location to another, they also age and move further along their career life cycle. Cohort effects are cogent because they can alter the attractiveness of working on home country vs other topics, with potential impacts on migration behavior. Scientists in a cohort where many persons start their career working on an international hot topic are likely to differ in migration behavior and research interests than a cohort where many people start their career working on a local topic. Empirically, age/seniority and cohort dummy variables can isolate the average effect of migration but at the expense of hiding different patterns by age and cohort, suggesting the value of separate analyses of scientists at particular age/seniority levels and in particular cohorts.

How might the interest of researchers in their own country research change over the cycle? Current literature on the effect of age and seniority on research interests provides no clear answer. Some analysts stress that older scientists who have gotten tenure or received a prestigious prize will have more freedom to research unfamiliar topics and thus should be more likely than younger researchers to change the topics of their research (Borjas & Doran, 2015b; Messeri, 1988). Others note that path dependency in accumulation and creation of knowledge could produce the opposite pattern, as older/senior scientists have the most to lose by changing the content of their work (Hull et al., 1978; Planck, 2014). By contrast, younger scientists could be more likely to change the traditional rules of science and to develop revolutionary contributions that diverge from earlier work. (Kuhn, 1996). Looking at migration patterns, the fact that younger researchers are more likely to go overseas suggests that they are willing to move into situations where there is greater opportunity to work on topics in a different area. For example, Darwin traveled to the Galapagos at age 22. Wallace left the UK to go to Brazil at age 25 and later went to the Malay Archipelago, where he wrote his letter to Darwin on the theory of evolution. Of course, a senior scientist may change their interest without making such trips or as a result of some overseas visit.

As some topics may be more trendy than others when scientists start their academic careers, this will affect their initial interests. If there is path dependency, cohorts may have different evolution of research interests that should be accounted for when studying their changes. A cohort analysis also provides one way to control for changes in the situation of persons born in different periods, such as the higher educational attainment of more recent cohorts of scientists Zuckerman & Merton (1972), and their easier access to foreign research via advances in communication technologies per the internet, digital databases, etc. Hence, as the initial starting points affect the life-course of academic careers, it should also affect the likelihood of changing interests.

As shown next, our basic model assumes that the decision to migrate is exogenous. By general geography-time effects, we mean that the social demand for every type of interest is different at every time t and within the geographical boundaries of market g. That is, every society rewards differently the research conducted about a particular topic, according to their specific needs at a particular period.

1.4 A toy model of choice of research interest

Consider a scientist i who must choose between researching interest *red* or interest *blue*, at three different stages in their academic careers $\forall t \in T = 1, 2, 3$ that capture the career life cycle. To model changes in cohort effects, we assume that the initial interests differ for every cohort ϕ, which we represent in a cost equation as some cohorts having lower costs in producing *red* at cost $C_{\phi,1,red}$, and other cohorts having lower costs in producing *blue* at cost $C_{\phi,1,blue}$. For simplicity, we suppose that every scientist chooses only one topic at every period and ignores the possible impact of expectations about changes in the importance of given topics over time.

We also assume that the social rewards for *red* and *blue* at period t and for every geography g, $P_{g,t,red}$ and $P_{g,t,blue}$, are exogenous, independent from each other, and not affected by the actions of i (as if i were price takers).[1]We can also suppose that society g has a preference for *red*, so rewards are on average historically larger, $\overline{P_{g,red}} > \overline{P_{g,blue}}$, and that g' has a preference for *blue*, such that $\overline{P_{g',blue}} > \overline{P_{g',red}}$.

For simplicity, we consider only two geographies g, where the scientist i was born, and g', a foreign country to i. Given these differences in the preferences, we can assume that the rate θ at which a scientist i learns about *red* in geography g is faster than the rate at which i can learn about *red* in geography g', so that $\theta_{red,g} < \theta_{red,g'}$. Analogically, i can learn about *blue* faster

[1] In more complex models, it is possible to analyze choices in which the actions of i influence the social demands for *red* and *blue*.

at geography g' than in geography g, so that $\theta_{blue,g'} > \theta_{blue,g}$. For simplicity, we assume there is no trade between g and g', so the only flow between g and g' is scientists. Therefore, i can only sell the knowledge they produce while being at the geography g at the prices determined by geography g.

For any i belonging to cohort ϕ, the costs of producing *red* at any period t are determined by (1). The total costs of producing *red* are determined by the initial costs attached to a cohort $C_{\phi,1,red}$ and the past accumulated know-how producing *red* in every geography the scientists have live in, $\theta_{red,g} \sum_{t>1}^{t-1} expe_{i,t,red,g} + \theta_{red,g'} \sum_{t>1}^{t-1} expe_{i,t,red,g'}$. We assume that as people get more experience in a particular topic, their costs of producing that knowledge decrease at a rate that is determined by $0 < \theta_{red,g} < 1$. For simplicity, we do not consider any rate of depreciation of acquired experience neither any forgetting rate.

$$C_{i,t,red} = C_{\phi,1,red} - \theta_{red,g} \sum_{t>1}^{t-1} expe_{i,t,red,g} - \theta_{red,g'} \sum_{t>1}^{t-1} expe_{i,t,red,g'} \tag{1}$$

In every period, i from cohort ϕ will research about *red* if the net rewards for researching *red* are larger than the net rewards of researching about *blue*, as expressed in equation (2). Every time i research about *red*, i will get more experience $expe_{i,t,red}$.

$$expe_{i,t,red,g} = \begin{cases} 1 \; if \; \pi_{i,t,red,g} > \pi_{i,t,blue,g} \\ 0 \; if \; \pi_{i,t,red,g} \leq \pi_{i,t,blue,g} \end{cases} \tag{2}$$

From equation (2) it should be noticed that $\pi_{i,t,red,g} = C_{i,t,red} - P_{g,t,red}$ and that $\pi_{i,t,blue,g} = C_{i,t,blue} - P_{g,t,blue}$. In this simple setup i does not have expectations about the future, so changes in interests at every time t are mostly explained by changes in the prices $P_{g,t,red}$.

When i moves from its national geography g to a foreign geography g', two main effects determine the changes in research interests: the new prices, $P_{g',t,red}$ and $P_{g',t,blue}$, and the new learning rates $\theta_{red,g'}$ and $\theta_{blue,g'}$. We assume that the move from g to g' is exogenous and creates a sharp shock for i, as its individual action is too small to affect the vector of prices or the learning rates at its new destination.

Migration from one country to another causes a sharp change in location that can be treated as a before-after "event study" to help isolate the effect of mobility on changes in research topics from reverse causality in which changes in interest produce changes in geographic content. To the extent that researchers can change the geographic context of their work relatively easily, say by switching from studying Colombian tomatoes to French tomatoes, while it is more difficult to change fields, we expect to see many more changes in the geographic content of papers than in fields. Given differences in the cost of shifting, changes in geographic context are more likely to reflect the effect of migration on the context of work than changes in academic fields, and thus more likely to be a cause of the change in context than an effect. For example, when a biologist changes from studying plants to animals after going to another country, it is more likely that he/she has chosen the best place to study animals before the decision to migrate. We try to differentiate the effect of migration on the geographic context of work from the reverse causation where changes in geographic interests cause migration by examining the pre-migration pattern of change in the geographic context of the migrant scientists' work. If past trends in interests were a good predictor of migration, our treatment of the migration decision as a cause rather than an effect of the change in the geographic context of papers would be an incorrect simplification and we would have to seek a more subtle way to isolate the change in location –> change in geographic content relation of interest.

Consider the example in which i is researching about interest *red* in geography g at time t. If i moves to geography g' for an exogenous shock at time $t+1$, i will change its interest to *blue* if the prices of producing *blue* at geography g' are considerable larger than the costs of producing at $t+1$. In the upcoming periods, particularly if i moves for educational reasons, changes of interests towards *blue* could be explained by the decrease in the costs of producing it, because of a large learning rate such that $\pi_{i,t+1,blue,g'} > \pi_{i,t+1,red,g'}$.

Finally, as we describe next, we seek to isolate the effects of geographical mobility on the geographic context of the research, we develop counterfactual matched samples and natural experiments to compare scientists with the same knowledge bases (see section 2.3).

2 Empirical strategy

2.1 Data

Our main data comes from curriculum data on Colombian scientists (CvLAC), complemented with data from ORCID, matched with publication data from Microsoft Academic Graph (MAG). CvLAC is an online platform that records standardized curriculum data of people researching in Colombia, managed by the Colombian Ministry of Science and Innovation (Minciencias, formerly known as Colciencias). Profiles are mostly filled by users and by their academic institutions. Information is updated for active scientists who work in Colombian universities and research institutions, but it can be outdated for people who are not based anymore in Colombia. As CvLAC data is used by the Minciencias for evaluation purposes and when attributing grants and research funds, academic and scientific institutions have strong incentives to keep the profiles of their researchers updated. However, when a scientist moves abroad and does not apply to grants/funding in Colombia and is not affiliated to a Colombian institution, the incentives to keep a CvLAC profile updated decrease. For these reasons, to complement the information from CvLAC, we rely on ORCID and MAG. ORCID is the global Open Researcher and Contributor ID that uniquely identifies authors and contributors of scholarly communication. As of 18 November 2021, the number of live accounts reported by ORCID was 12,742, 475 https://orcid.org/statistics. For all the scientists with a CvLAC profile or whose names are mentioned by their thesis advisors, we look for an ORCID profile. We then complete the missing information in CvLAC with the information available in ORCID. To track published documents by the same authors that are not included in the CvLAC profile, we also match CvLAC with MAG. MAG is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals, conferences, and fields of study. By adding information from ORCID and MAG we obtain coverage of scientists with missing or outdated information in CVLAC, in particular, those who moved to other countries.

We restrict our analysis to scientists who had a public CvLAC profile in August 2021 or whose names are mentioned by their advisors in the list of directed bachelor's and master's theses.[2] Furthermore, we focus on the scientists who are nationals from Colombia and who

[2]For those scientists, we retrieve information on career paths from ORCID or infer it from their publication

obtained a bachelor's degree in a Colombian university. We further limit our analysis to the profiles that contain information on the thesis titles (i.e. bachelor's, master's, or Ph.D.'s) that give us information on the geographic locus of the researcher.

2.2 Classifying the geographic locus of research

We measure the geographic locus of scientists' research interests by the geographic terms in the title, abstract, and references in their science papers. We develop a taxonomy to classify any type of academic document (e.g. thesis, articles, books, conference proceedings, among others) and aggregate the classified documents by time period, to obtain dynamic measures of research interests at the scientist level.

To identify the geographic locus of research interests of published works, we consider the content and the cognitive scope. On the one hand, the content is a direct and time-invariant measure of a document's main interest according to its textual content. As our focus is on geography, we extract the geographical locations that are mentioned and studied in titles and abstracts and examine if these places fall inside or outside a limited geographical area. On the other hand, cognitive scope is an indirect and time-variant measure of the geographical origin of the knowledge that supports a work. It relies on the space of ideas and focuses on the geography of the documents an article cites. We measure how local or global is the knowledge used to create a work, according to how much a document cites works from its same geographical area via addresses of researchers on those papers.

We set Colombia's administrative boundaries, as defined by the World Bank, to define the relevant area of study. By content, we classify a work as "Colombia" if its title or its abstract mentions at least one place that is located inside Colombia's administrative boundaries and as "Foreign" if all the geographical mentions are located outside Colombia. If a work does not include any geographical mentions in their title or in their abstract, it is classified as "Non-geographic." By cognitive scope, we label works as "Local" if they cite disproportionately works written by scientists with Colombian affiliations relative to the average number of citations to works of Colombian scientists in the whole field, and "Global" otherwise. Combining both the content and the cognitive scope, we then obtain a taxonomy made up of six classes, presented in table 1. the appendix gives more details about the way we compute these two dimensions, A.

Table 1: A taxonomy for studying research interest

		Content (G. in Titles and Abstracts)		
		Colombia	Foreign	Non-geographic
Cognitive scope	Local	LC	LF	LN
(G. in References)	Global	GC	GF	CN

Source: Own proposition. The content measures the degree of geography in the titles and abstracts while cognitive scope measures the degree of geography in the reference lists.

To determine scientists' research interests for a particular period, we aggregate the interests expressed in their published works. For a scientist i, we determine its research interest for the class r in time t, as G_{itr}. It is the simple count of the N documents that were published at time t and that were classified in the class r. We also define g_{itr} as the average interest of i for class

records in MAG.

r during time t. g_{itr} is the share of documents referring to research interest r, with respect to all the published documents written by i at time t.

$$g_{itr} = \frac{G_{itr}}{|N_{it}|} = \frac{|N_{itr}|}{|N_{it}|} \tag{3}$$

2.3 Linking Migration and changes in geographic research interest

We use a Difference-in-Difference approach, to estimate the probability that research interests change after a geographical move compared to a pre-mobility period. The first difference is the change in research interests of the scientist who moved. The second difference is the change in research interests over the same period for scientists who did not move, but who were similar in the baseline period and at the same career stage as determined by a synthetic counterfactual sample. As presented in figure 1, we focus on two types of geographical moves: emigration and return migration. In the first case, for every emigrant scientist i, we find a non-mobile scientist j who was as likely as i to go abroad in the baseline period, but who did not move. In the second case, for every returnee scientist i, we find a non-mobile emigrant scientist j who was as likely as i to come back home in the baseline period, but who stayed abroad. In addition, we also distinguish between two reasons to move: education (getting a degree) and work (getting a position).

The non-mobile scientist j has to have the same initial baseline research interest; cohort of graduation; highest degree/seniority at the time of traveling; gender; and academic field as the mobile scientist for whom they are the control. We perform some robustness checks in which we also impose that i and j were graduated from the same university or had the same bachelor's/master's thesis advisor.

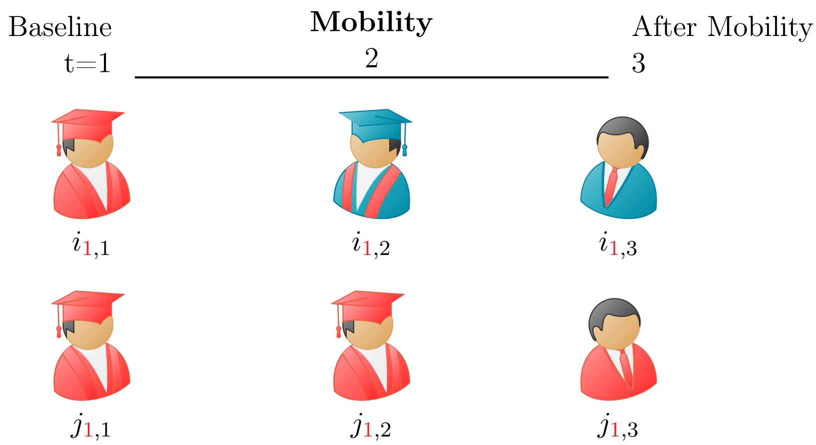
Following Borjas & Doran (2015a), we estimate p_{int}, the probability of deviating from a baseline research interest, at the nth paper written by scientist i, at the period t. The starting model, as shown in equation (4), includes a vector of scientists fixed effects δ_i; a vector of calendar year fixed effects δ_t; a dummy variable A that takes the value of one if a scientist went abroad or zero if the scientist is part of the control sample; a dummy T that takes the value of one, if the nth paper was published in the post-mobility period or zero otherwise; and a set of control variables at the scientist-time level Z_i. As our sample considers paper-scientists pairs, we cluster standard errors at the level of articles.

$$p_{int} = \delta_i + \delta_t + \beta_1 T + \beta_2 (T * A) + Z_i \gamma + \varepsilon \tag{4}$$

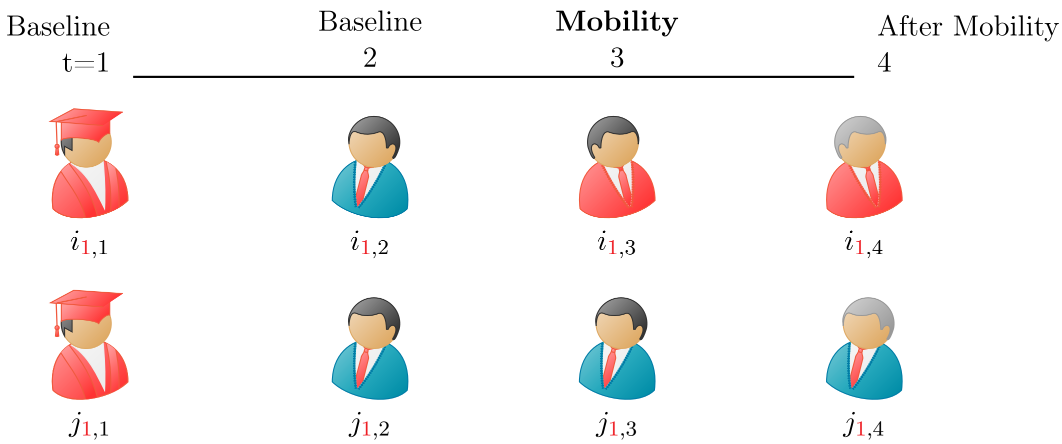
The key coefficient is β_2, which measures the probability of changing interests, for the scientists who moved and published papers in the after-mobility period compared to the non-mobile comparable scientist. We expect that scientists who move are more likely to change their research interests.

To assess whether the main causal direction in the data is from geographical mobility to changes in geographic research interest rather than from changes in interest leading to migration, we conduct some additional experimental designs. We consider samples of scientists who applied to mobility grants offered by Minciencias, to study Ph.D. programs abroad (Convocatoria Doctorados en el Exterior) and to repatriate emigrant scientists (Es Tiempo de Volver grant). In particular, we compare people who got the grant and move against people who were short-

Figure 1: Counterfactual samples of scientists for the baseline model



(a) Emigrants.



(b) Returnees.

Source: Own construction. Note: i is a scientist who moves and j is a counterfactual scientist who does not move but who was likely as i to migrate in the pre-mobility period. Panel (a) shows the case of emigrants for educational reasons and (b) the case of returnees for working reasons.

listed and did not move. Furthermore, we exploit the structure of the different calls, according to the type of mobility, to perform additional robustness checks.

3 Results

3.1 The Life Cycle of Research Interest

Using the taxonomy presented in section 2.2, we examine whether research interests change as the person's career unfolds irrespective of emigrating from or returning to Colombia. Per our conceptual model 1.4, interests change in the normal progress of academic careers. When scientists get more experience in a particular area, they accumulate know-how on it, which reduces the cost of future research on that topic. Therefore, changes in research interest should be less likely as people advance in their academic careers and get older.

Here we conduct baseline measures of research interests for every cohort of scientists who got their bachelor's degree in a year t. First, we apply our taxonomy to the full sample of articles found in CvLAC, MAG, and ORCID. Next, we conduct an analysis of differences in research interest by field.

Table 2 shows the classification of the published documents in our total sample. As our empirical model focuses on the changes of scientists at the nth published document, we count scientists-publication pairs. We find that around 35% of the total publication-scientist pairs have geographical information on their titles and abstracts. Among those with geographical information, around 54% have at least one reference to Colombia. Concerning the cognitive scope, more than 60% of the classified documents, cite disproportionately works written by authors with Colombian affiliation. However, the analysis of the cells shows interesting variations, as most of the works about Colombia cite disproportionately works from authors outside Colombia, and most of the works about foreign places cite disproportionately works written by authors based in Colombia. Note, however, that the cognitive scope dimension has not been measured yet for a considerable share of works, as they may not be available in MAG, or if they were matched, information on the reference list or the geocoded address of their authors is not available.

Table 2: Classification of published documents

		Content		
		Colombia	Foreign	Non-geographic
Cognitive scope	Local	14545	18538	97352
	Global	17060	12960	52068
	NA	77426	63016	222617
	TOTAL	109031	94514	372037

Source: Own estimations, using data from CvLAC, MAG and ORCID. Note: Counting scientist-publication pairs. NA includes non-matched with MAG and matched with MAG, but without information on the reference list or on the geocoded affiliations.

Figure 2 shows the evolution of research interest for three different cohorts of scientists, according to the year of graduation of their bachelor's degree. It can be seen that as scientists acquire more research experience, research works refer less to local documents, even when they write about Colombia. In particular, when looking at the average share of g_{itr} over time, with respect to the total number of scientists from the cohort, it can be observed that the share of global documents is always increasing. Moreover, there are some important differences in the share of documents by cohorts. For example, while the cohort of scientists who got their bachelor's degree between 1990-1994 was initially more interested in foreign topics, the 2000-2004 cohort was initially more interested in Colombian topics. Moreover, the more recent cohorts have a larger initial share of global documents, which may be the result of the new communication technologies that facilitate access to foreign literature.

When disaggregating data by academic fields, as shown in figure 3, some important differences emerge. As expected, published documents in engineering have fewer geographical references in their content and if they do, they are mostly about foreign places. For social sciences, most of the documents with geographical mentions in their content refer to Colombia, and the average shares are relatively more stable than from medicine. Moreover, when looking at the cognitive scope, the reliance on global documents is also increasing over time.

3.2 Emigrant scientists

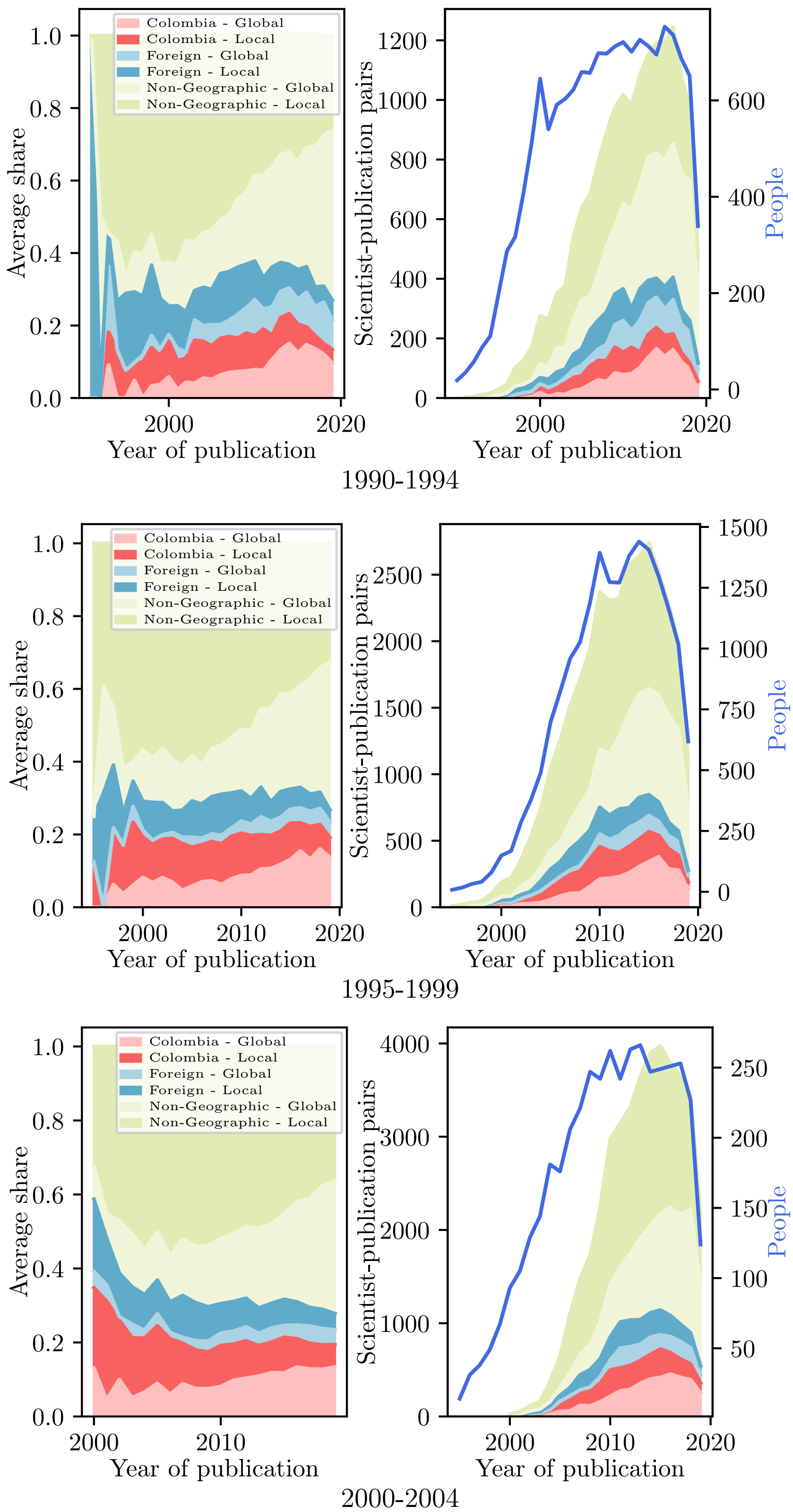
We study the effect of emigration on research interest. In particular, we try to answer three main questions: How do research interests change when scientists move to another country? What is the effect of moving for educational reasons? What is the effect of moving for professional work? To answer them, we follow the procedure described in section 2.3, to measure the probability of deviating from a baseline research interest after going abroad. We distinguish between two reasons to go abroad: to get a Ph.D. degree and to get a new job. In our basic estimations, we consider the whole sample of emigrant Colombian scientists, to have a first measure of changes in research interests that are due to emigration. Next, we design an identification strategy that tries to determine the causality link of this relationship, using a small sample of scientists who applied for Minciencias Ph.D. grants.

Table 3 shows the results of the basic estimations of changes in research interests, considering the whole sample of emigrant scientists. We present the results for emigrants that move for pursuing a degree abroad.[3] The coefficient of post-mobility measures the natural change of interest for both emigrant and non-emigrant scientists after the mobility period. It is positive and significant, suggesting that changes in research interest occur after the mobility period for educational reasons. This is expected, as the interest for research in geographical places can move according to along the life cycle of a scientist (see previous section, 3.1). Moreover, the coefficient of the interaction term "Post-mobility * Went Abroad" is positive and significant at a $p - value < 0.1$, suggesting that people who went abroad changed their interest after the mobility period. Hence, they are more likely to change interests, than the people who did not move.

In the experimental design, we only consider scientists who applied for the Minciencias Ph.D. Grants between 2009 and 2017. Those are grants given by the Colombian government to pursue a Ph.D. degree. There are two different calls every year: a call for Ph.D. programs in Colombia and a call for Ph.D. programs abroad. People can only apply to one call per year, so they reveal their preferences for staying in Colombia or going abroad at the moment of application. For every call, we have information on the shortlisted applicants and the grantees. We conduct two types of analysis: comparing shortlisted vs grantees at the call for Ph.D. programs abroad and

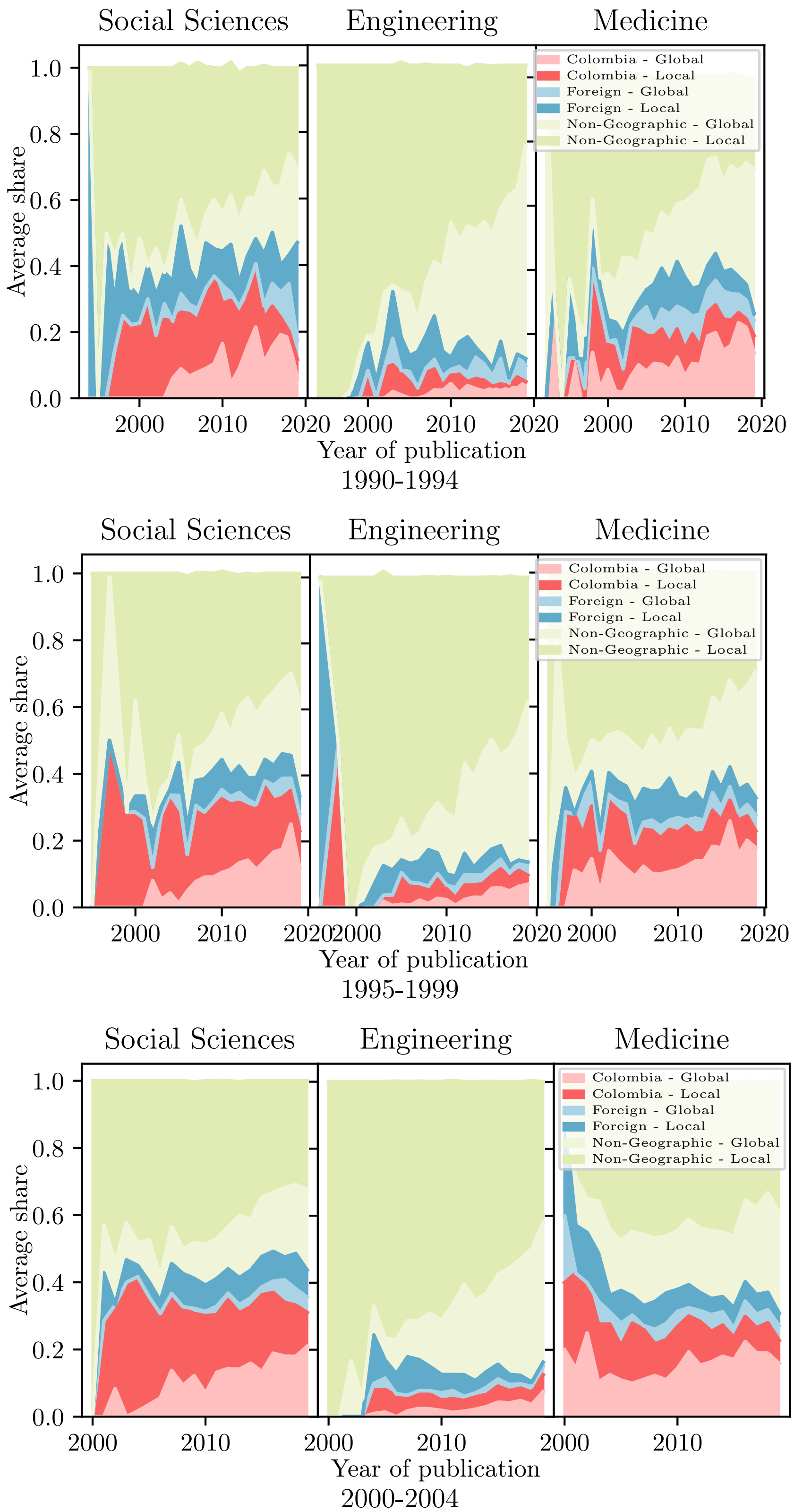
[3]The results for the case of mobility for work reasons are still in progress, so they are not included in this version of the draft paper.

Figure 2: Evolution of Research Interests for three different cohorts



Source: Own estimations, using data from CvLAC, MAG, and ORCID. Note: The left panel shows the average share of g_{itr} for every class, at the level of scientists. The right panel shows the number of scientist publication papers on the left axis and the number of people on the right axis. We only focus on scientists who have published at least one document every three years.

Figure 3: Evolution of Research Interests by field, for three different cohorts



Source: Own estimations, using data from CvLAC, MAG and ORCID. Note: Average share of g_{itr} for every class, by field, at the level of scientists.

Table 3: Full sample of emigrants

	Deviation of Interest after PhD
Post-mobility	0.1033**
	(0.0420)
Post-mobility * Went Abroad	0.0787*
	(0.0426)
Scientists fixed effects	Yes
Calendar year fixed effects	Yes
Obs	4233
Scientists	1058
R-squared Within	0.0426
R-squared Between	0.2535
R-squared overall	0.2046
Log-likelihood	-1711.1

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Source: Own estimations, using data from CvLAC, MAG and ORCID. Using the full sample of scientists with a CvLAC profile.

comparing people who applied to both grants and just got one of them.[4]

The results presented in table 4 suggest that changes in research interests are more common for people who study abroad. When comparing shortlisted candidates who stayed in Colombia and grantees of the Minciencias call for Ph.D. programs abroad, we find that people who went abroad are more likely to deviate from their baseline interest[5]. The coefficient of the interaction variable "Post-mobility * Went Abroad " is positive and significant and its magnitude is larger than the coefficient of the variable "Post-mobility." The coefficient for "Post-mobility" is also positive and the R-Squared within is low, suggesting that people change of interest after the mobility period, for other reasons than geographical mobility.

3.3 Returnees scientists

In progress.

[4] The second type of analysis is still in progress.

[5] In this model, we only consider documents that refer to geographical places by content, to estimate the probability of deviating interest form one location to another.

Table 4: Emigrants who applied for the Minciencias Ph.D. grants

	Deviation of Interest After PhD
Post-mobility	0.0690***
	(0.0309)
Post-mobility * Went Abroad	0.0728**
	(0.0314)
Scientists fixed effects	Yes
Calendar year fixed effects	Yes
Obs	1604
Scientists	380
R-squared Within	0.0375
R-squared Between	0.1761
R-squared overall	0.1377
Log-likelihood	-1534.2

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Source: Own estimations, using data from CvLAC, MAG and ORCID. Using data from the Ph.D. grants to study abroad, given by Minciencias.

4 Conclusions and discussion

Our results show that measures of brain drain and brain gain can go beyond counting people flows to examine the actual research topics and career paths of emigrant and returnee scientists compared to others. Scientists who move abroad shift the focus of their research from their home country but many continue to work on problems that affect their own countries.

Moreover, they may acquire new skills and tools that could be used directly, if they come back home, or indirectly, through diaspora networks, to produce research outcomes of higher quality in their home country. On the other side of migration, when scientists come back home, they do not invariably work on topics connected to that country. Some continue to write more about topics relevant to the international research agenda, which is invariably set by scientists in high income countries rather than local issues. At the same time, the migrant scientists can be more proficient than persons like themselves. The net effect of migration on knowledge production useful to the source country depends on the changes in skills / human capital and the changes in research interests. Hat led them to look at a similar but different problem might give them a broader perspective. It is difficult to imagine Darwin or Wallace conceiving the theory of evolution without their trips out of England – to the Galapagos Island and to Borneo, respectively.

References

Bianco, M., Gras, N., & Sutz, J. (2016). Academic evaluation: Universal instrument?: Tool for development? *Minerva (London)*, *54*(4), 399–421.

Borjas, G. J., & Doran, K. B. (2015a). Cognitive mobility: Labor market responses to supply shocks in the space of ideas. *Journal of Labor Economics*, *33*(S1), S109-S145. Retrieved from https://doi.org/10.1086/676659 doi: 10.1086/676659

Borjas, G. J., & Doran, K. B. (2015b). Prizes and productivity how winning the fields medal affects scientific output. *Journal of human resources*, *50*(3), 728–758.

Chavarro, D., Tang, P., & Ràfols, I. (2017). Why researchers publish in non-mainstream journals: Training, knowledge bridging, and gap filling. *Research policy*, *46*(9), 1666–1680.

Das, J., Do, Q.-T., Shaines, K., & Srikant, S. (2013). US and them: The geography of academic research. *Journal of Development Economics*, *105*, 112–130.

Hull, D. L., Tessner, P. D., & Diamond, A. M. (1978). Planck's principle. *Science*, *202*(4369), 717–723.

Kerr, W. R. (2008). Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics*, *90*(3), 518–537.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed. ed.). Chicago, IL: University of Chicago Press.

Martin, B., & Whitley, R. (2010). The UK research assessment exercise: A case of regulatory capture? In R. Whitley, J. Gläser, & L. Engwall (Eds.), *Reconfiguring knowledge production: Changing authority relationships in the sciences and their consequences for intellectual innovation* (pp. 51–80). Oxford University Press.

Messeri, P. (1988). Age differences in the reception of new scientific theories: The case of plate tectonics theory. *Social Studies of Science*, *18*(1), 91-112. Retrieved from https://doi.org/10.1177/030631288018001004 doi: 10.1177/030631288018001004

Planck, M. (2014). *Scientific autobiography: And other papers*. Open Road Media.

Porteous, O. (2020). Research deserts and oases: Evidence from 27 thousand economics journal articles on africa. *Unpublished Working paper*.

Saxenian, A. (2005). From brain drain to brain circulation: Transnational communities and regional upgrading in India and China. *Studies in comparative international development*, *40*(2), 35–61.

Xie, Q., & Freeman, R. B. (2021, February). *The contribution of chinese diaspora researchers to global science and china's catching up in scientific research* (Working Paper No. 27169). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w27169 doi: 10.3386/w27169

Zuckerman, H., & Merton, R. K. (1972). Age, aging, and age structure in science. *Higher Education*, *4*(2), 1–4.

A Content and Cognitive Scope

A.1 Content

- We extracted from the titles and abstracts, all the mentions to geographical places using a set of BERT-based NLP models. Our focus is on titles and abstracts, as if a place is the focus of attention of a document, it is usually mentioned over there.
- We looked for four types of entities:
  - GPE: Geopolitical entities: Countries, cities, states
  - LOC: Non-Geopolitical entities: locations, mountain ranges, bodies of water
  - NORP: Nationalities or religious or political groups
  - FAC: Buildings, airports, highways, bridges, etc.
- Next, we geocoded all these entities, using Geonames, Google Maps Platform API, and some lists on nationalities.

Colombia:

Precise Locations for Intermediate-Depth Earthquakes in [the Cauca Cluster LOC], [Colombia GPE]

Foreign:

Comprehensive analysis of earthquake source spectra in southern [California GPE]

Non-Geographic:

Earthquake nests as natural laboratories for the study of intermediate-depth earthquake mechanics

A.2 Cognitive Scope

- For every published work w in field f, we compute the share of references to works written by Colombian authors $|Ref_{w,f,c}|$ with respect to the total number of documents listed in its reference list $|Ref_{w,f}|$.

- We then compute the Relative Local Cognitive Specialization index (RLCS), which compares the share of references to Colombian authors in w with the average share of references to Colombian authors of other works w', in the same field f:

$$RLCS_{w,f} = \frac{\frac{|Ref_{w,f,c}|}{|Ref_{w,f}|}}{\frac{1}{|w'_f|}\sum_{w'}\frac{|Ref_{w',f,c}|}{|Ref_{w',f}|}}$$

- If $RLCS_{w,f} > 1$, then the work is classified as local. Otherwise, it is classified as global.

B CvLAC profiles

A CvLAC profile is made up of eight main sections, that contains detailed information on scientists' academic careers. Among others, the most salient groups of information by section are:

- General data: Name, Nationality, education, professional experience, languages
- Training activities: mentoring, consulting, thesis directed
- Evaluation activities: Peer reviewing, evaluation/thesis committees
- Social appropriation: Conferences, scientific events, scientific societies/networks
- Bibliographic production: Articles, books, book chapters, etc.
- Technical Production: Software, consultancy/technical reports, patents, companies
- Artwork: Pieces of art, art events/workshops
- More information: Projects, other works