THÈSE EN COTUTELLE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ET UNIVERSITÀ DEGLI STUDI DELL'INSUBRIA

ÉCOLE DOCTORALE ENTREPRISE, ÉCONOMIE ET SOCIÉTÉ (ED 42)
DOCTORAL PROGRAM IN METHODS AND MODELS FOR ECONOMIC DECISIONS

SPÉCIALITÉ SCIENCES ÉCONOMIQUES

Par

Andriy Romanyuk

Foreign PhDs in Europe: Permanence and Research Careers

Doctorants étrangers en Europe : trajectoires et carrières scientifiques

Sous la direction de

Francesco LISSONI, Université de Bordeaux et Università Cattolica del Sacro Cuore
et de Andrea VEZZULLI, Università degli Studi dell'Insubria

Soutenue le 16 décembre 2025

Membres du jury :

M. STERZI, Valerio, Full Professor, Université de Bordeaux — **Président**
Mme MENDOLA, Mariapia, Full Professor, Università di Milano Bicocca — **Rapporteur**
M. PEZZONI, Michele, Associate Professor, Université Côte d'Azur — **Rapporteur**
Mme PISONI, Alessia, Associate Professor, Università degli Studi dell'Insubria — **Examinateur**

Abstract

This dissertation examines the mobility, retention, and career outcomes of foreign doctoral graduates in Europe, with a focus on France and the Netherlands. Using large-scale datasets that combine administrative records, theses, and scientific publications, it develops novel methodological tools and applies them to study how migration policies, academic careers, and student origins interact in shaping the trajectories of highly skilled migrants.

The first chapter introduces a name-based classification methodology employing Long Short-Term Memory (LSTM) models to infer the regional origin of doctoral graduates in France between 1990 and 2020. The approach distinguishes French-sounding from foreign names and further classifies non-French individuals into twelve regions. Results document major shifts in France's international doctoral population, with a decline in Arabic-origin students and an increase in East Asian ones, while also highlighting challenges in distinguishing second-generation migrants from native-born French individuals.

The second chapter investigates the impact of France's first dedicated visa scheme for foreign researchers, the carte de séjour "scientifique-chercheur" introduced in 1998, on the post-graduation stay rate of doctoral graduates. Using a new dataset of STEM graduates from 1992–2002 and applying a difference-in-differences approach, we find no evidence that the visa significantly affected retention patterns. While stay rates of non-EU graduates eventually converge with those of EU and French counterparts, this occurs only much later and appears unrelated to visa policy.

The third chapter analyzes the publishing activity and mobility of STEM doctoral graduates from Dutch universities between 2000 and 2020. While natives and foreign-born graduates publish at similar rates and volumes, foreign-born graduates are substantially less likely to publish from the Netherlands after graduation, particularly if coming from outside the European Economic Area. Concerning out-migration, the chapter investigates the extent of positive self-selection and finds it limited to Dutch natives, whose most productive graduates are more likely to move abroad, especially to the United States. The analysis also shows that the introduction of selective immigration policies, starting 2004, had no discernible effect on retention rates.

Together, these findings show that while migration policies play only a limited role in shaping doctoral graduates' retention in Europe, structural factors—such as funding schemes, origin composition, and academic career trajectories—remain central in understanding the mobility and contributions of international researchers.

Keywords: International migration; Migration policy; Graduate studies; Academic careers; Highly skilled migration; Machine learning

Résumé

Cette thèse analyse la mobilité, la rétention et les trajectoires professionnelles des docteurs étrangers en Europe, avec un accent particulier sur la France et les Pays-Bas. En mobilisant des ensembles de données de grande ampleur combinant dossiers administratifs, thèses et publications scientifiques, elle développe de nouveaux outils méthodologiques et les applique à l'étude des interactions entre politiques migratoires, carrières académiques et origines des étudiants.

Le premier chapitre présente une méthodologie de classification des origines régionales des docteurs fondée sur des modèles de réseaux de neurones récurrents (LSTM). Appliquée aux diplômés de doctorat en France entre 1990 et 2020, l'approche distingue les noms à consonance française des noms étrangers et répartit ces derniers en douze régions d'origine. Les résultats mettent en évidence d'importants changements dans la composition de la population doctorale internationale en France, avec un recul des étudiants originaires des pays arabophones et une progression de ceux venus d'Asie de l'Est, tout en soulignant la difficulté à distinguer les migrants de deuxième génération des natifs français.

Le deuxième chapitre étudie l'impact du premier dispositif français de visa destiné aux chercheurs étrangers, la carte de séjour "scientifique-chercheur" introduite en 1998, sur le taux de rétention post-doctorat. À partir d'un nouvel ensemble de données portant sur les docteurs en sciences, technologie, ingénierie et mathématiques (STEM) entre 1992 et 2002, et en mobilisant une approche de type différences-en-différences, nous ne trouvons aucun effet significatif de la mesure sur les comportements de rétention. Les taux de séjour des docteurs non-européens finissent par converger avec ceux des Européens et des Français, mais beaucoup plus tard et sans lien apparent avec la politique de visa.

Le troisième chapitre analyse l'activité de publication et la mobilité internationale des docteurs en STEM diplômés des universités néerlandaises entre 2000 et 2020. Alors que les diplômés natifs et étrangers publient avec des probabilités et des volumes similaires, les étrangers sont nettement moins susceptibles de publier depuis les Pays-Bas après leur soutenance, en particulier les non-Européens. Les réformes de visa de 2004 et 2008 n'ont eu aucun effet discernable sur leur rétention, et la sélection n'apparaît significative que pour les diplômés néerlandais : les plus productifs d'entre eux sont davantage enclins à poursuivre leur carrière à l'étranger, notamment aux États-Unis.

Dans l'ensemble, ces résultats montrent que les politiques migratoires jouent un rôle limité dans la rétention des docteurs en Europe, tandis que des facteurs structurels – tels que les dispositifs de financement, la composition par origine et les trajectoires académiques – demeurent centraux pour comprendre la mobilité et les contributions des chercheurs internationaux.

Mots-clés : Migration internationale; Politique migratoire; Études doctorales; Carrières académiques; Migration hautement qualifiée; Apprentissage automatique

Abstract

Questa tesi esamina la mobilità, la permanenza e gli esiti di carriera dei dottori di ricerca stranieri in Europa, con particolare attenzione alla Francia e ai Paesi Bassi. Utilizzando dataset che combinano archivi amministrativi, tesi e pubblicazioni scientifiche, essa sviluppa strumenti metodologici innovativi e li applica per studiare come le politiche migratorie, le carriere accademiche e l'origine degli studenti interagiscano nel determinare le traiettorie dei migranti altamente qualificati.

Il primo capitolo introduce una metodologia di classificazione basata sui nomi che impiega modelli Long Short-Term Memory (LSTM) per inferire la regione di origine dei dottori di ricerca in Francia tra il 1990 e il 2020. L'approccio distingue i nomi di "suono" francese da quelli stranieri e classifica ulteriormente gli individui non francesi in dodici regioni. I risultati documentano cambiamenti significativi nella composizione internazionale della popolazione di dottorandi in Francia, con una diminuzione degli studenti di origine araba e un aumento di quelli dell'Asia orientale, evidenziando al contempo le difficoltà nel distinguere i migranti di seconda generazione dai francesi nati nel Paese.

Il secondo capitolo analizza l'impatto del primo programma francese di visti dedicato ai ricercatori stranieri, la *carte de séjour "scientifique-chercheur"* introdotta nel 1998, sul tasso di permanenza post-laurea dei dottori di ricerca. Utilizzando un nuovo dataset di laureati in discipline STEM del periodo 1992–2002 e applicando un approccio di "Difference-in-Differences", non emergono evidenze di un effetto significativo del visto sui tassi di permanenza. Sebbene i tassi di permanenza dei laureati extra-UE tendano infine a convergere con quelli dei colleghi europei e francesi, ciò avviene solo molto più tardi e sembra non essere correlato alla politica dei visti.

Il terzo capitolo analizza l'attività di pubblicazione e la mobilità dei dottori di ricerca in discipline STEM provenienti dalle università olandesi tra il 2000 e il 2020. Sebbene i laureati nativi e quelli stranieri pubblichino a tassi e volumi simili, i laureati stranieri risultano molto meno propensi a pubblicare nei Paesi Bassi dopo la laurea, soprattutto se provenienti da Paesi al di fuori dello Spazio Economico Europeo. Per quanto riguarda la migrazione in uscita, il capitolo indaga la presenza di auto-selezione positiva, riscontrandola solo tra i laureati olandesi, i cui individui più produttivi tendono maggiormente a trasferirsi all'estero, in particolare negli Stati Uniti. L'analisi mostra inoltre che l'introduzione di politiche migratorie selettive, a partire dal 2004, non ha avuto effetti rilevabili sui tassi di permanenza.

Nel complesso, questi risultati mostrano che, mentre le politiche migratorie giocano un ruolo limitato nel determinare la permanenza dei dottori di ricerca in Europa, fattori strutturali — come i sistemi di finanziamento, la composizione per origine e le traiettorie accademiche — restano centrali per comprendere la mobilità e il contributo dei ricercatori internazionali.

Parole chiave: Migrazione internazionale; Politiche migratorie; Studi di dottorato; Carriere accademiche; Migrazione altamente qualificata; Apprendimento automatico

Acknowledgements

I would first like to express my deepest gratitude to my supervisors, Ernest, Francesco and Andrea, for their invaluable guidance, support, and encouragement. Francesco has been both a supervisor and a mentor throughout this journey, while Andrea is undoubtedly one of the kindest people I have met in academia.

I am grateful to the Università Italo Francese (UIF/UFI), the Università degli Studi dell'Insubria, the ISI program, and the Pôle Doctorant of the Bordeaux School of Economics for their financial support, which made this work possible.

I thank all my DOC-TRACK colleagues, in particular Burcu, Catalina, Guido, Johannes and Michele. I also feel fortunate to have joined such a stimulating project, which not only allowed me to strengthen my technical skills but also to meet many wonderful academics. Special thanks go to my co-author Alberto, who has been a senior, a peer, and a friend during these years, making our collaboration on the project and my second chapter a truly rewarding experience.

I would also like to thank my peers from the Università degli Studi dell'Insubria and from my three years at the Bordeaux School of Economics. I am especially grateful to Catherine and Béka, whose company made my time in Bordeaux incredibly more enjoyable, as we shared the highs and lows of academic life together.

I am deeply thankful as well to my long-time friends in Benevento. While the list would be too long to name everyone, I especially want to thank Carmine, Cosimo, Eugenio, Francesco, Lorenzo, Marco, and Riccardo. We have supported each other for over a decade, and I hope we will continue to do so for many more decades to come. I also thank Alessandro and Francesco, who generously hosted me on countless visits to Milano and made every stay there a pleasure. Special mention goes to all my fellow MTG opponents, who have entertained me across several countries and helped keeping my mind sharp.

Finally, I owe heartfelt thanks to my family—my mother, my grandmother, and my stepfather. Above all, I am grateful for the opportunity they gave me to pursue the best education, for their constant support, and for their unwavering trust in my choices.

Contents

General Introduction 1

Chapter 1: Requiem: The Legacy of Names–Inferring the Regional Origin of Doctoral Graduates from Death Records 6
1 Introduction 7
2 Key Issues in Migrant Identification 9
2.1 Existing Methodologies for Identifying Country of Origin 10
2.1.1 Administrative Data Approaches 10
2.1.2 Databases Approaches 11
2.1.3 Machine Learning Based Approaches 11
2.2 Methodological Choice and Adaptation to the Context of Foreign Doctoral Graduates in France 11
3 Model Development and Data Processing 12
3.1 Overview of LSTM Classification 12
3.2 Model Design – Sequential Classification Model 14
3.3 Regions of Origin Taxonomy for Foreign Doctoral Graduates 15
3.4 Data Sources and Preparation 16
3.5 Evaluation of Model Performance 19
3.5.1 In-Sample Performance 19
3.6 Out-of-Sample Performance 24
4 The Regions of Origin of Doctoral Graduates in France 26
4.1 Requiem 1 – French vs Non-French 26
4.2 Requiem 2 – 12 Regions of Origin 29
5 Conclusions and Further Developments 31
References Chapter 1 33
Appendices Chapter 1 37

Chapter 2: The Chosen: Researcher Visas and Foreign PhD Graduates' Careers in France .. 48
1 Introduction 49
2 Migration policy and foreign PhDs in France 53
2.1 France's dual immigration policy 53

2.2 International PhD students in France 56
3 Identification strategy 57
4 Data 62
4.1 Region of origin 64
4.2 Gender 70
4.3 Publications 71
4.4 Descriptive statistics 73
5 Results 78
6 Conclusions 83
References Chapter 2 85
Appendices Chapter 2 90

Chapter 3: Stay rates and careers of foreign doctorate recipients from Dutch universities: A bibliometric analysis 117
1 Introduction 118
2 International STEM students and Innovation 121
3 The Netherlands as a case study 125
4 Data and methods 129
4.1 Dissertation data 130
4.2 Publication data 134
4.3 Estimation strategy 140
5 Results and discussion 145
6 Conclusions 162
References Chapter 3 167
Appendices Chapter 3 169

General Conclusion 242

General Introduction

The international mobility of students and skilled researchers has become an increasingly important phenomenon in the global knowledge economy. Over the past decades, the number of students crossing national borders for higher education has grown steadily, particularly in countries such as the United States, the United Kingdom, Canada, and Australia, which attract large shares of international enrollments (Aman, 2018; Gureyev et al., 2020). Large EU countries such as France and Germany also rank among the top destination countries, while several smaller ones - such as the Netherlands - host a disproportionate share of foreign students. Graduate students, particularly in science, technology, engineering, and mathematics (STEM) disciplines, constitute a significant portion of these flows. Their mobility is influenced not only by educational aspirations but also by labor market opportunities and immigration policies that shape their post-graduation stay intentions and decisions (Robinson-Garcia et al., 2019). Understanding the dynamics of international student mobility is crucial because it directly affects the global distribution of human capital, the supply of STEM workers, and patterns of innovation across countries.

International students are most overrepresented in STEM fields, which are critical for technological innovation and economic growth. In many advanced economies, foreign students earn a substantial share of doctoral degrees in key STEM disciplines, and a significant proportion remains in the host country to pursue research or professional careers (Aman, 2018; Gureyev et al., 2020). These observations indicate that international students serve as a key channel for host countries to acquire highly skilled labor, influencing both the quantity and quality of scientific output.

Immigration and visa policies play a central role in shaping these flows. The availability of post-graduation work permits and pathways to permanent residency acts as a strong incentive for international students to remain in the host country (Robinson-Garcia et al., 2019). For example, programs that extend temporary work options for STEM graduates can increase retention rates and influence the choice of fields among prospective students. Conversely, restrictions on visas or travel bans for specific nationalities can reduce enrollment from affected countries, leading to shifts toward alternative destinations (Aman, 2018).

The impact of international student mobility extends beyond labor supply into scientific productivity and innovation. Studies using publication and patent data have shown that international doc-

toral students contribute substantially to research output. Comparing affiliations and publication timing allows tracing of academic trajectories and mobility (Gureyev et al., 2020; Robinson-Garcia et al., 2019). For instance, the presence of foreign STEM graduates is positively correlated with both publications and patents, suggesting that retained talent contributes directly to technological innovation (Aman, 2018). However, individual productivity varies widely, and not all contributions are captured equally, highlighting the need for careful measurement (Lerchenmueller and Sorenson, 2016; Criscuolo et al., 2019).

Scientific publications provide a particularly rich source for analyzing mobility. Unlike patents, which are limited to certain technological domains and subject to legal and cost constraints, academic publications cover a broader range of disciplines and typically have lower barriers to production (O'Donoghue, 1998; Gureyev et al., 2020). They also occur much more frequently than patents, as the average scientific author publishes more often and over a longer time span than the average inventor patents. Thus, publication data allow tracing of institutional affiliations over time, offering insights into the movement of scholars across institutions and countries (Aman, 2018). Yet publications present challenges, such as author name disambiguation, multiple affiliations, and potential mismatches in timing or institution reporting (Lerchenmueller and Sorenson, 2016; Ge et al., 2016). Particularly, multiple affiliations complicate the identification of genuine mobility, as scholars may hold simultaneous appointments for collaborative or career advancement purposes (Moustafa, 2020; Hottenrott et al., 2021). Databases like Scopus, which assign unique identifiers to researchers and track career histories, represent a promising avenue to overcome these challenges (Aman, 2018).

The growing mobility of international students and skilled researchers raises both critical policy questions and methodological issues. To what extent do international students contribute to host-country scientific advancement and innovation? How do visa policies affect the retention of high-skilled labor? How can mobility and productivity be measured accurately given the limitations of bibliometric and administrative data? Addressing these questions is particularly relevant in contexts such as the European Union, where differences in post-PhD opportunities, visa regulations, and funding mechanisms affect career trajectories (Aman, 2018; Robinson-Garcia et al., 2019).

The three chapters of this dissertation integrate insights from international student migration, policy, and bibliometric data to examine the effects of visa regimes and stay rates on scientific

productivity. By leveraging publication-based measures of productivity alongside administrative information on student flows and policies, they provide a nuanced understanding of how human capital mobility shapes academic and technological outcomes. In doing so, it contributes to the broader literature on skilled migration, innovation, and policy evaluation (Gureyev et al., 2020; Lerchenmueller and Sorenson, 2016; Criscuolo et al., 2019).

In particular, the first chapter develops a machine learning–based methodology to infer the regional origin of doctoral graduates in France. Its key contribution lies in generating an enriched dataset that goes beyond existing sources such as IdRef, which suffer from gaps and systematic underreporting of nationalities. By applying a sequential classification model to names and surnames, this chapter provides new evidence on the shifting composition of the international doctoral population in France from 1990 to 2020. It documents, in particular, a decline in Arabic-origin graduates and a steady rise of East Asian graduates. While the model achieves strong accuracy overall, it highlights the difficulty of distinguishing between French-born second-generation migrants and newly arrived foreign graduates, especially in recent cohorts. Beyond methodological innovation, the chapter opens new avenues for research on the link between migration and innovation by providing reliable origin information at the individual level.

The second chapter evaluates the effects of the 1998 Loi Reseda, which introduced a special residence permit dedicated to academic researchers (*carte de séjour scientifique-chercheur*) with the explicit aim of exempting foreign PhD graduates in France from the increasing restrictions to work immigration. Using a dynamic difference-in-differences framework, the chapter compares non-EU doctoral graduates—those directly affected by the policy—to French graduates, across the pre- and post-law cohorts. Contrary to expectations, the law did not significantly improve short-term or long-term stay rates overall. Some evidence of positive effects emerges for the 2001 cohort in terms of career permanence and supervisory roles, but these results are fragile and sensitive to model specifications. The analysis suggests that both institutional barriers, such as the delayed inclusion of Algerian nationals in the visa scheme, may explain the limited impact of the policy. Moreover, many foreign graduates had already accumulated enough years in France to qualify for citizenship, which reduced the marginal effect of the new residence permit. Overall, the chapter provides the first systematic evaluation of a European scientific visa policy and highlights its limited effectiveness.

The third chapter turns to the Netherlands and investigates the careers and mobility of a large subset of doctoral recipients in STEM fields between 2000 and 2020, whose country of birth could be obtained through webscraping. By combining dissertation data with publication records, the chapter traces the scientific trajectories of both native and foreign doctoral graduates, as well as their supervisors. It shows that foreign-born graduates are disproportionately concentrated in fields with broader non-academic career options, which helps explain their lower likelihood of pursuing academic publishing after graduation. Once field and supervisor sorting are taken into account, however, their productivity levels are comparable to natives. Consistent with prior literature, foreign graduates are significantly more likely to leave the Netherlands after graduation, though no strong evidence emerges of positive or negative selection on research productivity, except for some Chinese graduates bound by funding return requirements. Prior international affiliations stand out as the strongest predictor of emigration, underscoring the role of pre-existing mobility pathways. Notably, immigration policy changes in the 2000s, including the Knowledge Migrant Scheme and the EU Scientific Researcher visa, appear to have had little to no effect on the retention of foreign doctoral graduates.

References

Aman, V. (2018): "Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates," *Scientometrics*, 117, 705–720.

Criscuolo, P., O. Alexy, D. Sharapov, and A. Salter (2019): "Lifting the veil: Using a quasi-replication approach to assess sample selection bias in patent-based studies," *Strategic Management Journal*, 40, 230–252.

Ge, C., K.-W. Huang, and I. P. Png (2016): "Engineer/scientist careers: Patents, online profiles, and misclassification bias," *Strategic Management Journal*, 37, 232–253.

Gureyev, V. N., N. A. Mazov, D. V. Kosyakov, and A. E. Guskov (2020): "Review and analysis of publications on scientific mobility: Assessment of influence, motivation, and trends," *Scientometrics*, 124, 1599–1630.

Hottenrott, H., M. E. Rose, and C. Lawson (2021): "The rise of multiple institutional affiliations in academia," *Journal of the Association for Information Science and Technology*, 72, 1039–1058.

Lerchenmueller, M. J. and O. Sorenson (2016): "Author disambiguation in PubMed: evidence on the precision and recall of author-ity among NIH-funded scientists," *PLoS One*, 11, e0158731.

Moustafa, K. (2020): "Octopus affiliations," *Scientometrics*, 124, 2733–2735.

O'Donoghue, T. (1998): "A patentability requirement for sequential innovation," *The RAND Journal of Economics*, 654–679.

Robinson-Garcia, N., C. R. Sugimoto, D. Murray, A. Yegros-Yegros, V. Larivière, and R. Costas (2019): "The many faces of mobility: Using bibliometric data to measure the movement of scientists," *Journal of informetrics*, 13, 50–63.

Chapter 1

Requiem: The Legacy of Names—Inferring the Regional Origin of Doctoral Graduates from Death Records

Andriy Romanyuk[1,2]

[1]BSE UMR CNRS 6060 – Université de Bordeaux

[2]DiECO – Università dell'Insubria

Abstract

This paper introduces a name-based classification methodology using Long Short-Term Memory (LSTM) models to infer the regional origin of PhD students in France from 1990 to 2020. Employing a two-stage approach, the model first distinguishes doctoral graduates with French-sounding from foreign names and surnames, then classifies non-French individuals into twelve regions of origin. Results highlight significant shifts in the international doctoral student population, notably a decline in Arabic-origin students and an increase in East Asian students. While the model achieves high classification accuracy overall, it faces challenges distinguishing native-born French individuals from second-generation migrants, particularly post-2000.

JEL classification: O30, F22, C38

Keywords: Supervised Learning; International migration; Graduate Studies

I thank Clovis Sourisse, Valentine Petzold and Baptiste Comby for their valuable research assistance.

1 Introduction

The intersection of migration and innovation has gained increasing attention in recent years, as exemplified by the ongoing debate surrounding the H-1B visa in the United States.[1] Despite a substantial body of literature examining its economic and innovation-related effects (Kerr and Lincoln, 2010; Doran et al., 2014; Peri et al., 2015), questions regarding the role of migrants in research, invention, and entrepreneurship remain central to policy discussions.

Existing studies have consistently highlighted the significant contributions of foreign researchers, inventors, and entrepreneurs to host countries' innovation systems (Stephan and Levin, 2001; Chellaraj et al., 2008; Hunt et al., 2010). More recently, attention has shifted towards foreign doctoral graduates, given their critical role in knowledge production and technological progress. This increasing focus underscores the importance of distinguishing between foreign and native graduates. However, official records often lack comprehensive data on individuals' migration backgrounds, and when available, such information may be incomplete or misaligned with the specific needs of a study.

To address these gaps, various methodologies have emerged to infer individuals' countries of origin, nationality or ethnicity based on name analysis. Kerr (2008) was among the first to apply name analysis in innovation economics, using database-matching techniques and commercial datasets to infer individuals' ethnic backgrounds. Methodologies for inferring origin from names have grown steadily more sophisticated—moving from rule-based systems to machine-learning frameworks such as NamePrism(Ye et al., 2017). Building on this evolution, the present paper relies on the more recent Niggli (2023) model.

Just as migration and innovation research has been largely US-focused—raising concerns about the external validity of its findings for Europe—many name-based classification approaches rely on databases constructed primarily from US populations. However, European migration patterns differ significantly, necessitating a more context-specific approach in which training data are representative of the countries under study. In this paper, in order to identify foreign PhD students in France, I build on the methodology of Niggli (2023) by developing a tailored dataset for training

[1] For more information, see: https://www.nytimes.com/2025/01/24/business/h1b-visa-immigration.html, last accessed on January 29, 2025.

the model, ensuring it accurately reflects the migration dynamics relevant to France.

France has long been a key destination for international students. In 1998, it attracted 11% of all foreign students enrolled in OECD countries—ranking just behind the UK and Germany, and ahead of the US—with foreign students accounting for 7% of total tertiary enrollments (OECD, 2000). While its relative share declined to 4% by 2021, the domestic share of international students is now over 9% (OECD, 2022). The trend is even more pronounced at the doctoral level: in 2021, foreign PhD students made up 37% of France's doctoral population—with peaks of above 50% in fields like ICT and Engineering—and represented 5% of all international doctoral students in the OECD (Campus France, 2024).

The merit of my model lies in its ability to generate a dataset with individual-level information on the regional origin of doctoral graduates in France. When combined with existing data on graduation cohort, university, and field of study, this enriched dataset enables researchers to investigate a wide range of questions at the intersection of migration and innovation. It opens the door to empirical analyses that were previously limited by the lack of detailed origin information, offering new insights into the composition, mobility, and scientific contributions of the international doctoral population.

To the best of my knowledge, the only dataset providing nationality information for doctoral graduates in France is IdRef, managed by ABES (*Agence bibliographique de l'enseignement supérieur*).[2] However, IdRef has limited historical coverage, with nationality data missing for approximately 30–40% of graduates from 1990 to 1999. Moreover, not all individuals with an IdRef record have nationality information included. Anecdotal evidence also suggests that when nationality is unknown during data entry, the default value is set to French. As a result, IdRef significantly underestimates the proportion of foreign graduates. According to my estimates, among the roughly 60% of graduates from the 1990–1999 period who have an IdRef record with nationality specified, only 5–7% are classified as foreign. This stands in stark contrast to the only official figure available for that period, which reports 27% foreign doctoral graduates in the 1999 cohort (Cohen, 2001).

The model's results indicate significant changes in the ethnic composition of international doc-

[2]IdRef is a French authority data platform that enables querying, creating, and editing standardized authority records for individuals and institutions. It is maintained by the academic documentation networks in France (e.g., SUDOC, Calames, STAR) and is primarily used by authorized professionals in higher education and research institutions. For more information on IdRef see: https://www.idref.fr, last accessed on: July 30, 2025

toral graduates in France between 1990 and 2020. The share of Arabic-origin doctoral graduates declined substantially, from about 45% to 27%, while East Asian representation steadily increased. Although the model demonstrates strong classification accuracy overall, distinguishing French-born individuals from second-generation migrants proved challenging, particularly after the year 2000 due to changing naming conventions linked to increased immigration.

The remainder of the paper is structured as follows. Section 2 provides an overview of the challenges involved in identifying migrants and reviews the methodologies proposed within the innovation and migration literature to address this issue. Section 3 presents the design of the proposed model, including standard performance metrics and confusion matrix analysis. Section 4 reports and discusses the empirical results. Finally, Section 5 concludes.

2 Key Issues in Migrant Identification

The definition of a migrant varies across institutions and contexts. UNESCO defines a migrant as any individual residing outside their usual place of residence for more than a year (UNESCO, 2023). In contrast, the French National Institute of Statistics and Economic Studies (INSEE) considers a migrant any person born abroad who resides in France, maintaining this status even after acquiring French nationality (INSEE, 2023). These different definitions highlight the complexities involved in classifying individuals based on migration background.

In migration studies, place of birth is the key variable for identifying migrants in data collection (Bilsborrow, 1997). Since this study focuses exclusively on international migration, I generalize this concept to *country of birth*, which has the advantage of being a stable characteristic over time, unlike nationality, which may change, particularly among internationally mobile individuals. However, depending on the study's objectives and data availability, nationality or ethnicity may also be used or preferred to identify migrants. Nationality is legally defined and often linked to one's place of upbringing, making it a practical classification criterion (Hambrick et al., 1998). Ethnicity, while offering an additional layer of identification—such as distinguishing second-generation migrants[3]—is more ambiguous, as it is closely tied to personal identity and self-perception (Phinney, 2013).

[3]For instance, Alba and Silberman (2002) distinguish second-generation migrants from North Africa as of French (*pied noir*), or Arabic ethnicity. This task is well suited to be performed by a machine learning model.

While country of birth is generally a reliable measure in migration research, it presents challenges in studies on innovation. When analyzing the career outcomes of foreign and native doctoral graduates—such as their publication records (Gaulé and Piacentini, 2013) or patent activity (Zheng and Ejermo, 2015; Buenstorf et al., 2022)—factors such as educational background and upbringing become crucial. For instance, a PhD student born in Morocco who migrated to France as a child or for undergraduate studies is likely to have a more similar academic trajectory to a French-born student than to a Moroccan-born student who moved to France solely for their PhD. However, capturing such nuances would require detailed individual-level data, including country of birth, nationality, and educational history—information that is rarely available at scale.

2.1 Existing Methodologies for Identifying Country of Origin

In innovation studies, several approaches exist to classify individuals' migration background, with some relying on direct administrative data and others inferring status from names and surnames. In this Section I overview these methodologies by looking at innovation studies on doctoral graduates when possible.

2.1.1 Administrative Data Approaches

Zheng and Ejermo (2015) study foreign doctoral graduates' inventive activity in Sweden by identifying migrants through Social Security Number, which covers the entire population and provides precise tracking. Such granularity and coverage of information is however quite rare. Even when the information is precise more often than not the coverage is not total. Chellaraj et al. (2008) for instance, to study the impact of international students on US innovation, as part of the study's dataset rely on National Science Foundation Survey of Earned Doctorates which has a response rate around 65%[4]. In contrast, Buenstorf et al. (2022) use a subset of German PhD dissertations with the nationality indicated for administrative reasons. However, both in both cases the sample may not fully represent the entire PhD student population.

[4]For more information see:https://ncses.nsf.gov/surveys/doctorate-recipients/2023#methodology, last accessed on February 03, 2025

2.1.2 Databases Approaches

When direct data is unavailable, researchers infer migration status through name and surname analysis. Kerr (2008) pioneered this approach in migration and innovation studies by matching names to a commercial database (*Melissa*), which assigns ethnicity to names and surnames. Subsequent studies in the context of doctoral graduates applied similar methods to identify ethnic groups. Gaulé and Piacentini (2013) create a database of names and surnames that are supposed to be Chinese and Mulders et al. (2024) match to a database with the historical distribution of Dutch surnames to identify ethnic minorities. However, these methods suffer from limited recall, as some names may not be present in reference databases, and the use of proprietary datasets raises concerns about result replicability.

2.1.3 Machine Learning Based Approaches

Advances in computational methods have led to more sophisticated approaches in name-based classification. To my knowledge there are no studies that use this methodology to infer the ethnicity or nationality of doctoral graduates. However, one tool used in innovation studies is NamePrism (Ye et al., 2017), which employs name embeddings-a way to convert names into numbers so that a model can understand patterns in them-from a large-scale database to classify individuals into 39 nationality groups covering over 90% of the world population.[5] More recently, Niggli (2023) developed an sequence-based model to classify inventors into ethnic groups tailored to innovation studies. It is worth mentioning the existence of deep learning approaches using transformers such as raceBERT (Parasurama, 2021).[6]

2.2 Methodological Choice and Adaptation to the Context of Foreign Doctoral Graduates in France

In this paper, I apply the classification algorithm developed by Niggli (2023) to distinguish between native and foreign doctoral graduates in France from 1990 to 2020. While Niggli (2023) focuses on

[5]For an example in innovation studies see: de Rassenfosse and Hosseini (2020)

[6]While deep learning models based on transformers (such as BERT-Bidirectional Encoder Representations from Transformers) are very powerful, they rely on unsupervised or self-supervised learning, making their decision-making process less interpretable. Additionally, these models often require significant computational resources, which can pose challenges for replicability and accessibility, especially for researchers or organizations with limited infrastructure. For more information on transformers please refer to Vaswani (2017).

the global inventor population, and therefore aims at an exhaustive classification of all international ethnic groups, my analysis requires an approach that better reflects France's migration dynamics, particularly the significant presence of first- and second-generation Northern and Sub-Saharan Africans in both the general population and among doctoral graduates (Campus France, 2017).

This algorithm classifies individuals based on their given name and surname, identifying phonetic and structural patterns that are often associated with ethnicity or country of birth rather than legal nationality. Given this characteristic, I classify foreign doctoral graduates in France according to their region of origin, following the association between language, ethnicity and their geographical distribution proposed by other existing taxonomies (Mateos et al., 2007; Ye et al., 2017).

This approach circumvents identification challenges inherent to name-based classification methods (e.g., the difficulty of distinguishing between a Spanish and a South American student based solely on name and surname). A detailed description of the methodology follows in Section 3.

For the purposes of this study, I define these classes as *regions of origin*, assuming that a student's language group and ethnic background coincides with their nationality. While this categorization is practical given data limitations and accounts for migration patterns relevant to the French academic system, it introduces potential biases. Specifically, individuals with foreign-sounding names who were born or educated in France may be misclassified as foreign. Additionally, distinguishing between first- and second-generation migrants remains difficult in the absence of administrative data.

3 Model Development and Data Processing

3.1 Overview of LSTM Classification

The sequence-based model Niggli (2023) introduces to classify name and surnames is the Long Short-Term Memory (LSTM) model. A LSTM model is a type of Recurrent Neural Netowrk (RRN). RRN is a specialized type of artificial intelligence designed to process and analyze sequential data, such as text, speech, or time-series information. Unlike traditional machine learning models that treat each data point independently, RRNs can remember previous inputs, allowing them to

detect patterns and dependencies in sequences. What makes LSTM models different from other RRNs is the use of *gates* (Hambrick et al., 1998), which allow the model to store only the relevant information and *forget* the irrelevant one (Goodfellow et al., 2016).

In the context of name and surname classification, the objective is to identify sequential structures and patterns that signal a likely origin. Certain linguistic markers, such as suffixes, provide valuable information. For example, the suffix "-enko" historically denotes "son of" in Ukrainian[7]. Similarly, phonetic characteristics can serve as indicators; for instance, the letter combinations "ou" and "eux" are commonly found in French surnames[8]. For a more detailed description of the LSTM model refer to Appendix D.

Niggli (2023) develops a machine learning model to classify worldwide inventors into 18 ethnic origin categories (Figure A.3). These categories are selected to capture both major inventor nationalities and significant diasporas in key patenting countries, drawing from established taxonomies (Mateos et al., 2007; Miguelez et al., 2013; Ye et al., 2017). To construct a training dataset for this task, he combines two sources: the Olympic Athletes Dataset[9] (1896–2016), which links athletes to ethnic origins via their national team affiliations, and a sample of inventors drawn from USPTO and EPO patent records, selected based on addresses in countries corresponding to the 18 target ethnic origins. The inventors are then classified using NamePrism (Ye et al., 2017), yielding a labeled dataset suitable for training and validating an LSTM model.

A key limitation of the athlete-based data is that it is unbalanced—some national teams are overrepresented, and post-1980 immigration complicates the link between nationality and ethnic origin in countries like the UK and France. To reduce bias, athletes from such countries after 1980 are excluded. Still, disparities remain, which pose challenges for machine learning classification, as unbalanced training data can lead to models biased toward majority classes or underperforming on underrepresented ones. To address this, Niggli (2023) supplements the dataset with sampled inventors from each ethnic group, aiming to create a more balanced and representative training set.

[7]Examples include "Shevchenko," "Marchenko," "Bondarenko," and "Tymchenko."

[8]Examples include "Dufour," "Boucher," "Rousseau," and "Fabreaux."

[9]Thanks to Randi H. Griffin, the data can be accessed at: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results, last accessed on February 05, 2025

3.2 Model Design - Sequential Classification Model

To address the specific distributional characteristics of the PhD population in France, I implement a two-stage classification strategy illustrated in Figure 1. Official statistics indicate that the share of foreign doctoral graduates has fluctuated between 30% and 40% over the past three decades (see Figure C.8 in Appendix C). Within foreign doctoral graduates then, some regions of origin are significantly more represented than others (see Table C.2 in Appendix C).

Given this context, a single-step multi-class classifier would require training on a highly imbalanced dataset, with the French as the predominant region of origin comprising 60-70% of the sample, potentially leading to poor performance, especially on underrepresented classes. In the context of machine learning classification, the term *class* refers to the category or label assigned to each observation based on the model's predictions. In what follows, I will use the terms *class* and *regions of origin* interchangeably, as each predicted class corresponds to an inferred region of origin.

Instead, I propose a sequential classification architecture. In the first stage—**Requiem_1**—the model performs a binary classification to distinguish between individuals with *French*-sounding and *non_French*-sounding names and surnames. Since the proportion of *non_French* graduates is relatively close to that of *French* graduates, this step can be trained on a balanced dataset, enhancing model stability and reducing the need for imbalance correction.

Individuals classified as *non_French* are then passed to the second stage—**Requiem_2**—which assigns them to one of 12 pre-defined regions of origin. This stage operates under significant class imbalance, as some regions of origin are far less represented than others. To mitigate this, I adopt an algorithm-level class weighting approach (López et al., 2013; Chen et al., 2024), where each class receives a weight inversely proportional to its relative frequency in the training data:

$$\text{class_weight}_j = \frac{n_{\text{total_samples}}}{n_{\text{samples}_j} \times n_{\text{classes}}}$$

This strategy ensures that the model places greater emphasis on underrepresented classes, helping to improve classification accuracy and reduce majority-class bias. The overall architecture thus allows for a robust classification pipeline tailored to the demographic realities of the dataset.

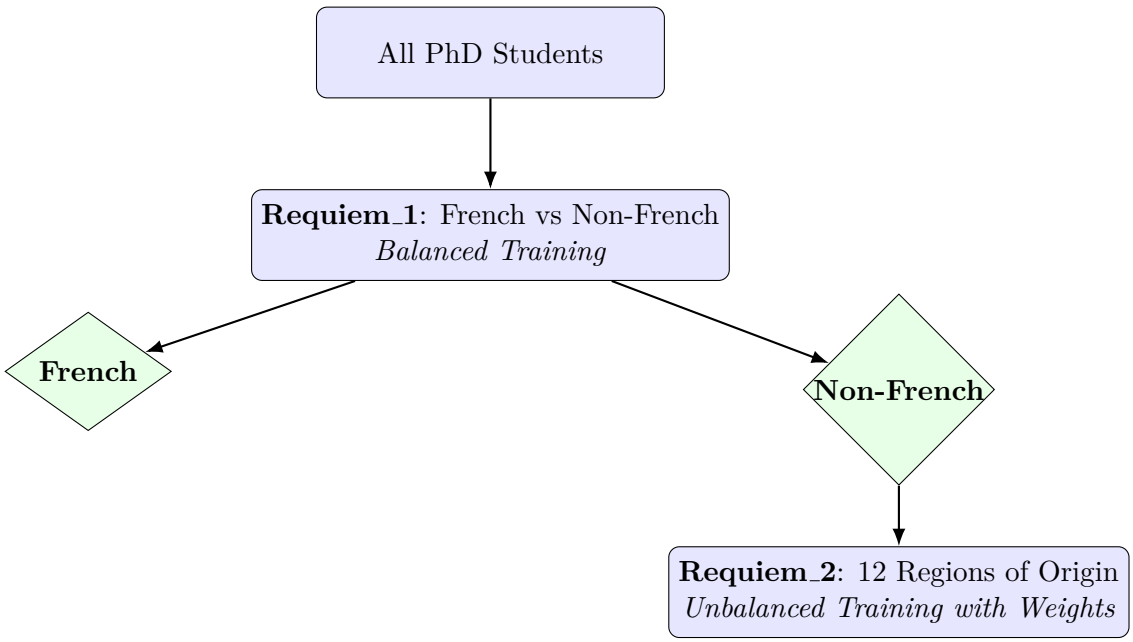


Figure 1: LSTM-Based Two-Stage Name Classification Model

3.3 Regions of Origin Taxonomy for Foreign Doctoral Graduates

I construct my taxonomy by building on the classification frameworks proposed by Niggli (2023) and Ye et al. (2017). However, my taxonomy is primarily based on linguistic classifications (Ruhlen, 1994; Gray and Atkinson, 2003; Vossen and Dimmendaal, 2020), drawing on the geographical distribution of languages across countries as well as the relationship between language and ethnicity (Mateos et al., 2007; Mu, 2015).

I adapt and expand the existing taxonomies to better reflect the specific context of foreign doctoral graduates in France. In particular, I refer to the earliest available statistics that report the country of origin of international doctoral candidates in France-namely the 2011–2012 cohort documented by Campus France (2017)-and assign to each country of origin a region. To account for the different migration patterns characteristic of the 1990s, I further extend the taxonomy by introducing two additional regions of origin: *Turkic* and *Indo_Iranian*. These regions were notably

represented among international university students in France during that period (MESR, 1994). Table 1 presents the resulting mapping between regions of origin and countries of origin for the population of foreign doctoral and university students.

Table 1: Taxonomy of Regions of origin and countries of origin of foreign PhD (2011-2012) and University (1994-1995) Students

Regions of origin	Countries	Population-Cohort
Anglo_Saxon	United States	PhD Stud.s 2011-2012
Arabic	Tunisia, Algeria, Lebanon, Morocco	PhD Stud.s 2011-2012
Balto_Slavic	Russia	PhD Stud.s 2011-2012
East_Asian	China, Vietnam	PhD Stud.s 2011-2012
Italian	Italy	PhD Stud.s 2011-2012
Niger_Congolese	Senegal, Ivory Coast, Cameroon, Gabon, Madagascar	PhD Stud.s 2011-2012
Portuguese	Brazil, Portugal	PhD Stud.s 2011-2012
Romanian	Romania	PhD Stud.s 2011-2012
Spanish	Spain	PhD Stud.s 2011-2012
West_North_Germanic	Germany, Belgium	PhD Stud.s 2011-2012
Indo_Iranian	Iran	University Stud.s 1994-1995
Turkic	Turkey	University Stud.s 1994-1995

3.4 Data Sources and Preparation

The doctoral graduates that I classify are retrieved from **Theses.fr**. **Theses.fr** is a platform that serves as France's official repository for Electronic Theses and Dissertations (ETD). Managed by the French Bibliographic Agency for Higher Education (ABES) under the oversight of the Ministry for Higher Education and Research since 1985, it archives PhD dissertations from as early as the 1900s, with more detailed records available in recent decades. The database provides key details such as the author, abstract, graduation year, supervisor, field of study, and awarding university.[10] For the purpose of the classification the relevant information are the name, surname and graduation year of each PhD student.

For the classification tasks, I construct a novel dataset representative of the migration flows in France. This is based on the **Register of Deceased Persons in France**(*Fichier des personnes décédées*). Managed by INSEE (*Institut national de la statistique et des études économiques*), this register contains information on individuals deceased in France, including their name, surname, country of birth, date of birth and date of death. For the purpose of the study I am interested

[10]For more information, see: https://documentation.abes.fr/sudoc/regles/CodesUnivEtab.htm, last accessed on: January 29, 2025

in the years and countries of birth and their frequency of association to each name and surname mentioned in the Register.[11]

I begin by collecting records of deceased individuals in France from the Register of Deceased Persons in France for the period 1990 to 2023. Considering an average graduation age of 30 (Vourc'h, 2010; Pommier et al., 2022), and considering that graduation years in my dataset range from 1990 to 2020, I select individuals born between 1960 and 1990 to ensure representativeness. To increase the number of observations, I include an additional decade on the left tail of the distribution (i.e., individuals born between 1950 and 1960).

After cleaning the data and removing homonyms, the resulting dataset comprises 2,567,043 observations, of which 2,230,125 individuals were born in France and 336,918 abroad. Each entry includes the individual's first name, surname, and country of birth.

Among the foreign-born individuals, I first identify those who have French-sounding names and surnames as they may be French individuals born abroad. To identify and exclude these cases, I apply the original algorithm developed by Niggli (2023) to the foreign-born subset. This results in 92,313 individuals (27.40%) being classified as *French* and subsequently removed from the sample. The remaining sample consists of 244,605 individuals classified as *non_French*. I then select individuals with countries of birth representative of the population of foreign doctoral graduates in France based on the taxonomy presented in Section 3.3. Table B.1 in the Appendix B, shows the mapping between regions of origin and country of birth for the foreign deceased individuals selected for training. This filtering process yields 228,328 individuals, as those born in countries not included in the taxonomy are excluded.

To avoid imbalances in model training (López et al., 2013; Chen et al., 2024), I construct a balanced dataset for training **Requiem_1**, which is a binary classifier distinguishing between *French* and *non_French* individuals. From the 2.2 million French-born individuals, I randomly sample 228,328 individuals born before 1970. This cutoff is motivated by evidence that the share of French-sounding names begins to decline steadily after 1970, coinciding with a sharp increase in immigration from Morocco and Sub-Saharan Africa (Coulmont and Simon, 2019), as illustrated in Figure A.1 in the Appendix A. For **Requiem_2**, I classify all individuals in the 12 regions of

[11]For more information on the Register see: https://www.insee.fr/fr/information/4190491, last accessed on: January 29, 2025

origin displayed in Table 1. Certain regions of origin relevant to the broader population of foreign doctoral graduates in France are underrepresented in the sample. For instance, Italians accounted for 11.57% of foreign doctoral graduates in France in the 2011–2012 cohort (Campus France, 2017), yet are represented by only 10,045 individuals in the dataset (see Table 2). To construct a balanced training set, one possible strategy would be to downsample overrepresented regions of origin or to augment underrepresented classes using an alternative dataset, as in Niggli (2023), by incorporating individuals classified using NamePrism (Ye et al., 2017).

However, the *Arabic* region of origin—comprising countries that together represent 44.54% of foreign doctoral graduates in the 2011–2012 cohort (Campus France, 2017)—is the most represented class. Given its empirical relevance, I refrain from downsampling this class. Yet, the substantial disparity between this and smaller classes (e.g., Romanians, as shown in Table 2) would require an unfeasibly large number of synthetic observations to achieve class balance through data augmentation, which would compromise the validity of the classification exercise.

Instead, to address class imbalance, I adopt an algorithmic solution (López et al., 2013; Chen et al., 2024), applying class weighting based on inverse class frequency. The resulting weights, which ensure that each class contributes proportionately to the model training, are reported in the "Weight" column of Table 2.

Table 2: Regions of origin sizes and computed class weights

Regions of origin	**Count**	**Weight**
Arabic	83156	0.2278
Portuguese	30715	0.6168
Niger_Congolese	29884	0.6339
Balto_Slavic	16268	1.1645
Spanish	13093	1.4469
West_North_Germanic	10637	1.7810
East_Asian	10448	1.8132
Italian	10045	1.8859
Anglo_Saxon	8241	2.2988
Turkic	7964	2.3787
Indo_Iranian	3695	5.1269
Romanian	3182	5.9535
Total	228328	

3.5 Evaluation of Model Performance

Ideally, the performance of a classification model should be validated using external data—that is, data entirely independent of the model development process—to ensure its generalizability. However, such external datasets are not always available or sufficiently large to support robust analysis. Therefore, I conduct diagnostic checks on in-sample data, specifically using the portion of the dataset set aside for validation (i.e., not used in training), which represents approximately 10% of the full sample—around 45,000 individuals in **Requiem_1** and almost 23,000 in **Requiem_2**. This internal validation is particularly valuable given the large sample size, which allows for more precise assessment of model behavior. In addition, I replicate the same diagnostics on a much smaller out-of-sample dataset for which I have partial geographic information on doctoral graduates. This information, available only for the 1992–2001 cohorts, refers to various indicators of origin—such as country of birth, or the location where individuals completed their secondary education, bachelor's degree, or master's degree. While limited in both temporal and sample coverage, this external validation provides an important benchmark for evaluating the model's robustness across different contexts.

3.5.1 In-Sample Performance

The classification report for **Requiem_1** in Table 3 demonstrates excellent model performance in distinguishing between *French* and *non_French* classes. Both classes exhibit high precision (all at or above 0.95), indicating that the model is both accurate and consistent across the two categories. Notably, the precision for the *French* class (0.968) is slightly higher than for the *non_French* class (0.958), whereas the *non_French* class has a marginally higher recall (0.957 vs. 0.968). This balance results in an almost identical F1-score of 0.963 and 0.962 for the two classes. The overall accuracy, macro average, and weighted average scores all stand at 0.963, underscoring the robustness and fairness of the model across a balanced dataset.[12]

[12]Model performance is evaluated using standard classification metrics. Precision measures the accuracy of positive predictions: Precision $= \frac{\text{TP}}{\text{TP+FP}}$, where TP (True Positives) are correctly predicted positive cases, and FP (False Positives) are incorrectly predicted positives. Recall assesses the model's ability to identify all actual positives: Recall $= \frac{\text{TP}}{\text{TP+FN}}$, where FN (False Negatives) are actual positives missed by the model. F1-score is the harmonic mean of Precision and Recall: F1 $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$, balancing both concerns. Accuracy captures overall correctness: Accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, with TN (True Negatives) being correctly predicted negatives.

Table 3: Classification Report - **Requiem_1**

Class	Precision	Recall	F1-Score	Support
non_French	0.958	0.968	0.963	22733
French	0.968	0.957	0.962	22733
Accuracy			0.963	45466
Macro Avg	0.963	0.963	0.963	45466
Weighted Avg	0.963	0.963	0.963	45466

Table 4 presents the classification performance of **Requiem_2** across the twelve regions of origin. The overall accuracy of the model is high at 0.898, with a strong weighted F1-score of 0.900, indicating solid general performance across a diverse set of classes. The macro average F1-score of 0.866 suggests that the model maintains relatively balanced performance even across less represented classes.

Arabic, *Portuguese*, and *East_Asian* achieve the highest F1-scores (above 0.92), indicating highly reliable classification for these classes. In contrast, performance is weakest for the *Indo_Iranian*, with a notably lower F1-score of 0.692, suggesting difficulty in distinguishing this class—likely due to a combination of lower support and higher class overlap. Similarly, *Anglo-Saxon* and *West_North_Germanic* classes have comparatively modest scores, with F1-scores of 0.844 and 0.802 respectively.

The model performs particularly well on classes with larger support, such as *Arabic* (8316) and *Portuguese* (3072), suggesting that the amount of training data positively influences classification performance. The *Turkic* and *Italian* classes, despite having fewer examples, still achieve F1-scores close to or above 0.85, highlighting good model generalization.

Table 4: Classification Report - **Requiem_2**

Class	Precision	Recall	F1-Score	Support
Anglo_Saxon	0.792	0.903	0.844	824
Arabic	0.969	0.907	0.937	8316
Balto_Slavic	0.943	0.903	0.923	1627
East_Asian	0.932	0.919	0.925	1045
Italian	0.777	0.960	0.859	1005
Niger_Congolese	0.867	0.819	0.842	2988
Portuguese	0.936	0.955	0.945	3072
Romanian	0.853	0.874	0.863	318
Spanish	0.857	0.859	0.858	1309
West_North_Germanic	0.765	0.843	0.802	1064
Indo_Iranian	0.596	0.824	0.692	369
Turkic	0.844	0.960	0.898	796
Accuracy			0.898	22733
Macro Avg	0.844	0.894	0.866	22733
Weighted Avg	0.905	0.898	0.900	22733

Identifying Foreign Doctoral Graduates: French vs. Non-French

Table 5 presents a breakdown of *False Positives—non_French* individuals classified as *French*—and *False Negatives—French* individuals classified as *non_French*—based on predictions from **Requiem_1**. The table displays the misclassification patterns with respect to the regions of origin defined in **Requiem_2**.

For the *False Positives*, the true country of birth is known, allowing us to identify, for each region of origin, the proportion of individuals incorrectly classified as French. The highest proportion is found among individuals classified as *West_North_Germanic* (9.82%), likely driven by individuals from Belgium. This is followed by *Anglo_Saxon* individuals and, notably, by regions of origin associated with countries with a strong history of migration to France—first from Africa and the Middle East (e.g., *Niger_Congolese*, *Arabic*, *Indo_Iranian*), and then from Europe (e.g., *Spanish*, *Italian*).

In the case of *False Negatives*, since French individuals do not have an assigned region of origin by default, I infer their region of origin by reclassifying them using **Requiem_2**.[13] I then compute, for each region of origin, its proportion over the total number of *False Negatives*. The majority (42.89%) are classified as *Arabic*, which likely reflects the presence of second-generation immigrants—individuals born in France but with Arabic parents. Similar reasoning may apply to

[13]Each individual is assigned to the region of origin with the highest probability output by **Requiem_2**.

individuals classified as *Italian* (11.34%), *Niger_Congolese* (9.48%), and *Spanish* (9.48%). Finally, the *West_North_Germanic* class (8.45%) again likely includes individuals from Belgium.

Table 5: Breakdown of False Positives and False Negatives by Regions of Origin for **Requiem_1**

Regions of origin	**Region X Misclassified as French**	**French Misclassified as Region X**
Anglo_Saxon	5.68	4.23
Arabic	3.98	42.89
Balto_Slavic	0.84	4.12
East_Asian	2.17	2.16
Italian	1.74	11.34
Niger_Congolese	4.39	9.69
Portuguese	0.47	3.40
Romanian	0.34	1.65
Spanish	2.20	9.48
West_North_Germanic	9.82	8.45
Indo_Iranian	2.98	1.96
Turkic	0.37	0.62
Formula	$\frac{\text{False Positives for Region X}}{\text{Total Region X}} \times \%$	$\frac{\text{False Negatives classified as Region X}}{\text{Total False Negatives}} \times \%$

Note: False Positives are non-French individuals classified as French. False Negatives are French individuals classified as non-French. To know where they would be confounded in the second stage, I classify them with **Requiem_2** (selecting the class with the highest value for each individual).

Another way to interpret the Classification Report (Table 4) is through the visualization of the corresponding confusion matrix, a table that summarizes a classification model's performance by showing the number of correct and incorrect predictions for each class. This representation is particularly useful for understanding how misclassified individuals are distributed across classes. Figure 2 presents a row-normalized confusion matrix, effectively illustrating the *recall* for each class as reported in the classification report. Diagonal values correspond to the recall scores for each class, while off-diagonal entries indicate the proportion of instances from a given class that were misclassified into others. Overall, recall values are high across all classes, never falling below 0.82. Notable patterns include modest confusion between *Anglo_Saxon* and *West_North_Germanic* (around 0.03–0.04), as well as misclassification of *Niger_Congolese* individuals primarily as *Portuguese* and *Arabic* (both around 0.04). Similarly, *Spanish* individuals are often confused with *Portuguese* and *Italian* (approximately 0.04), while *Indo_Iranian* individuals are frequently misclassified as *Arabic* (0.06) or *Niger_Congolese* (0.04).

Figure 3 provides a complementary view by normalizing the confusion matrix column-wise,

thereby representing the *precision* for each class. The lowest precision is observed for the *Indo_Iranian* class, with 10% of its true instances misclassified as *Niger_Congolese* and 19% as *Arabic*. The *Arabic* class, in particular, appears to attract the highest number of misclassifications from other classes. This may be attributed to its disproportionately large size during training (see Table 2), potentially biasing the classifier toward this region of origin.

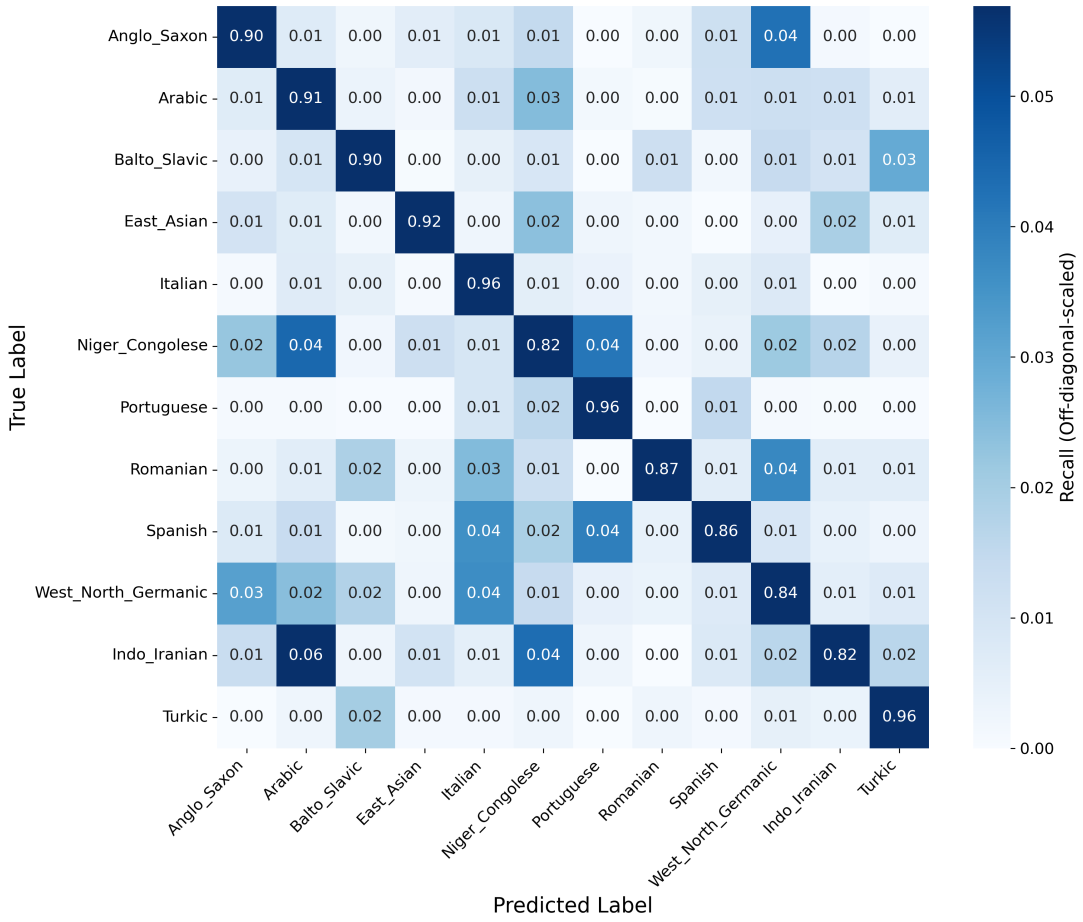


Figure 2: Requiem_2 - Confusion Matrix - Recall

Note: The Confusion Matrix is normalized row-wise in order to represent Recall. The heatmap is scaled on the off-diagonal values.

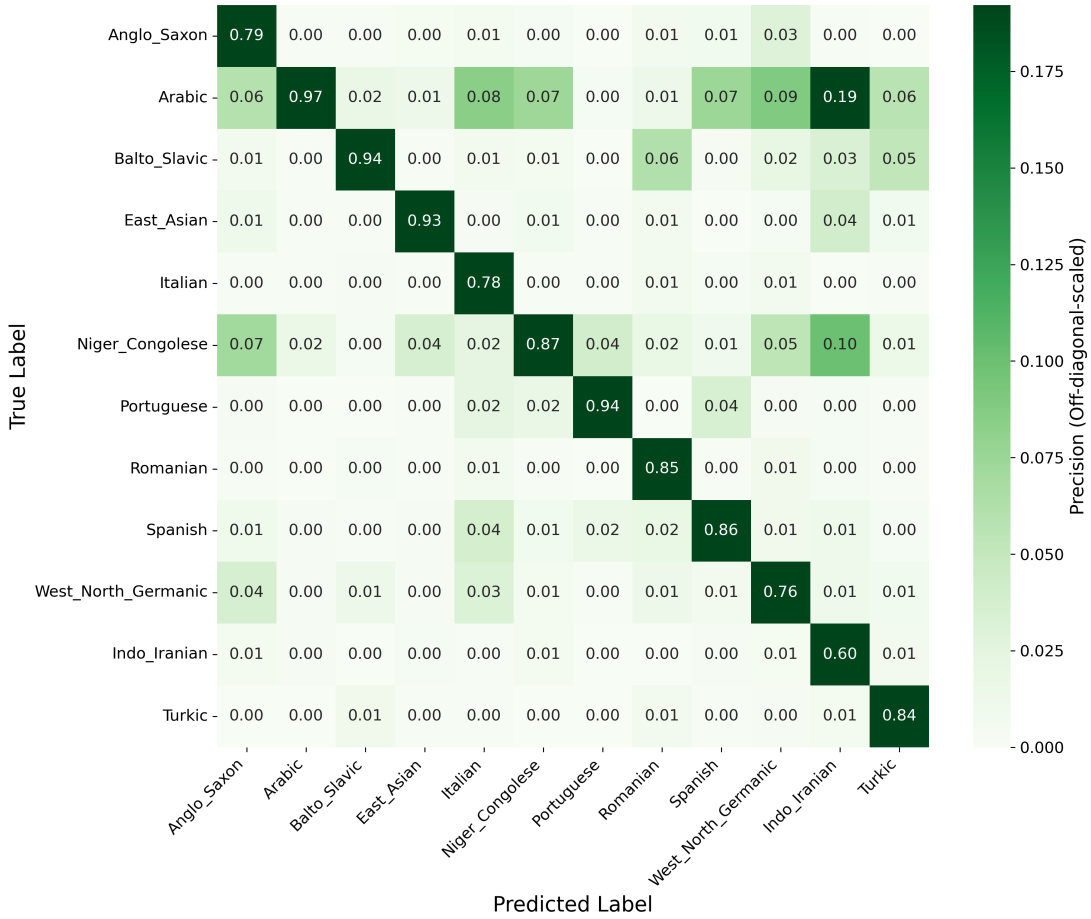


Figure 3: Requiem_2 - Confusion Matrix - Precision

Note: The Confusion Matrix is normalized column-wise in order to represent Precision. The heatmap is scaled on the off-diagonal values.

3.6 Out-of-Sample Performance

For a subset of doctoral graduates from the 1992–2001 cohorts, I have manually retrieved location-related information corresponding to one of the following: country of birth, high school, bachelor's degree, or master's degree. Overall, detailed location information is available for 786 doctoral graduates, with the share of non-French individuals averaging approximately 20% across defence year cohorts. For further details on the construction of this sample, see Appendix E. According to the only official statistic available for the period, the proportion of foreign doctoral graduates in the 1999 cohort was around 27% (Cohen, 2001). This comparison suggests that the hand-curated sample underestimates the true share of foreign graduates.

Table 6 presents the classification performance by type of available information. The model achieves consistently high precision for French doctoral graduates, exceeding 0.95 across all information sources. Recall is similarly strong, reaching 0.88 when country of birth is known, 0.90 for high school, and declining slightly to 0.78 and 0.74 for bachelor and master data, respectively.

For non-French graduates, the model performs well only when country of birth is available, with a precision of 0.77. Precision drops to 0.61 when based on high school location, 0.59 for bachelor, and as low as 0.13 for master's degree location. However, recall remains relatively stable around 0.88–0.93 for all sources except the master level, where it decreases to 0.78. This decline likely reflects the known fact that many international doctoral graduates complete their master's studies in France, making this variable a poor indicator of origin (CJC, 2010).

From the model's perspective, the most relevant and consistent signal is the country of birth, which aligns closely with the region of origin classification based on name and surname analysis. However, if the focus is on post-graduation careers, high school location may be more informative, as individuals who attended high school in France—regardless of their country of birth—are likely to have undergone similar educational socialization within the French system.

Unfortunately, given that non-French individuals represent only around 20% of the sample, it is not possible to replicate the full classification report across all 12 non-French classes. The limited sample size lacks the statistical power and class-specific numerosity required for such disaggregated analysis.

Table 6: Classification Report - **Requiem_1**: Using Observed Location Information, 1992-2001

Info Type	**Precision**		**Recall**		**Accuracy**	**Support**
	Non-French	French	Non-French	French		
Country of Birth	0.77	0.95	0.89	0.88	0.88	146
High School	0.61	0.98	0.88	0.90	0.90	164
Bachelor	0.59	0.97	0.93	0.78	0.82	282
Master	0.13	0.99	0.78	0.74	0.74	191

4 The Regions of Origin of Doctoral Graduates in France

4.1 Requiem_1 - French vs non_French

I apply the **Requiem_1** model, introduced in Section 3.2, to all doctoral graduates who defended their thesis in France between 1990 and 2020, using data retrieved from **Theses.fr**. The dataset comprises a total of 374,358 theses.

Figure 4 presents the temporal distribution of the estimated share of foreign doctoral graduates in France. The red squares represent the estimates derived from the **Requiem_1** model. The orange line (2005–2020) corresponds to official statistics on foreign doctoral graduates, while the green line (1990–2005) shows inferred proportions based on the share of foreign students in French universities overall, using available administrative data for those years. For further details on the construction of these official and inferred statistics, see Appendix C.

According to the only official statistics available prior to 2005, the share of foreign doctoral graduates in the 1999 cohort was approximately 27% (Cohen, 2001). In comparison, the estimate produced by **Requiem_1** for the same cohort is 33.6%, suggesting a discrepancy of roughly 7 percentage points. From 2000 onwards, the model increasingly overestimates the share of foreign doctoral graduates—by approximately 10% around 2007 and up to 20% by 2015.

The overestimation observed in the model's predictions may be attributed to several factors. First, the model is trained only on French-born individuals before 1970. While this rule ensures high precision, it comes at the cost of low recall: some individuals who are actually French are misclassified as foreign. Second, this limitation is likely exacerbated by increasing migration flows to France over recent decades, which have made the distinction between French and foreign-sounding names and surnames progressively more complex.

As shown in Figure 5, a study by Coulmont and Simon (2019) reveals that the proportion of newborns with French-sounding names fell from nearly 100% in 1945 to approximately 55% in 2005[14]. Considering that PhD graduates in France are, on average, around 30 years old (Vourc'h, 2010; Pommier et al., 2022), we can estimate the share of newborns with French names corresponding to the birth years of each PhD cohort.

[14]Defined as names with consonances characteristic of the French language (Coulmont and Simon, 2019).

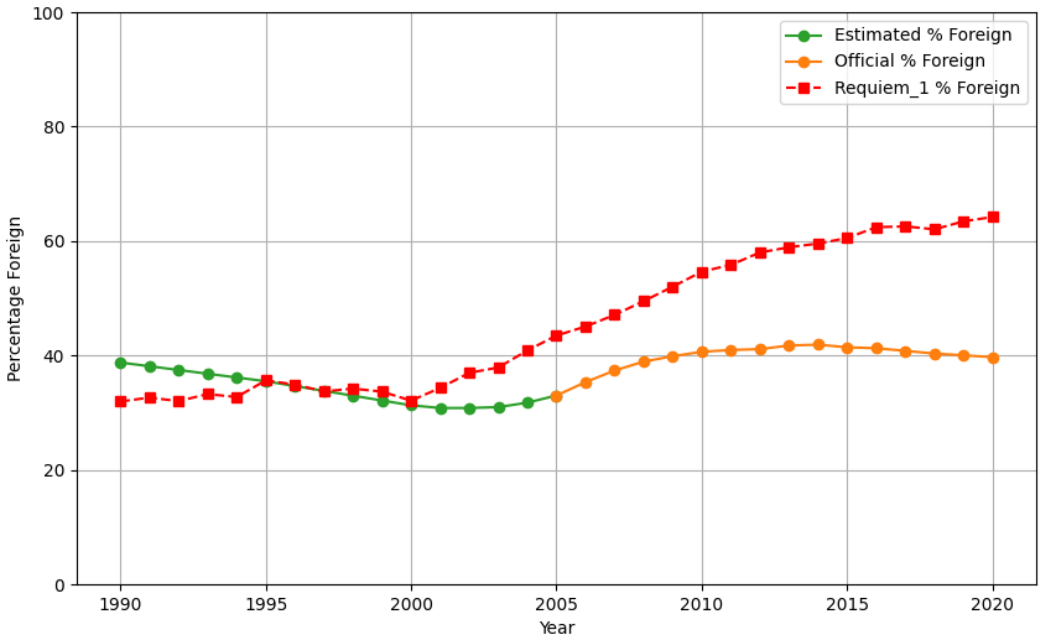


Figure 4: Percentage of foreign doctoral graduates in France from 1990 to 2020

Note: The percentage of foreign doctoral graduates between 1990 and 2005 is estimated using the statistics of foreign students in Franch universities overall (see Appendix C)

For PhD graduates between 1990 and 2000, the estimated proportion of newborns with French-sounding names during their birth years (1960–1970) ranged from 95% to 90%. However, this share declined sharply for subsequent cohorts—dropping from 90% to 80% for those graduating between 2000 and 2010, and stabilizing at around 80% for the 2010–2020 cohorts. This sharp 10–15 percentage point decrease during the 2000–2010 period may substantially affect the model's reliability.

This trend appears to be primarily driven by second-generation migrants from North Africa and Sub-Saharan Africa, approximately 80% of whom retain names reflecting their ethnicity (Figure A.4). In contrast, the share of second-generation migrants with foreign-sounding names is considerably lower among individuals of Asian and Southern European descent, at approximately 20%.

Further evidence is provided by the threshold analysis presented in Appendix A.5. While the box plots of the predicted probabilities for individuals classified as *non_French* remain relatively stable

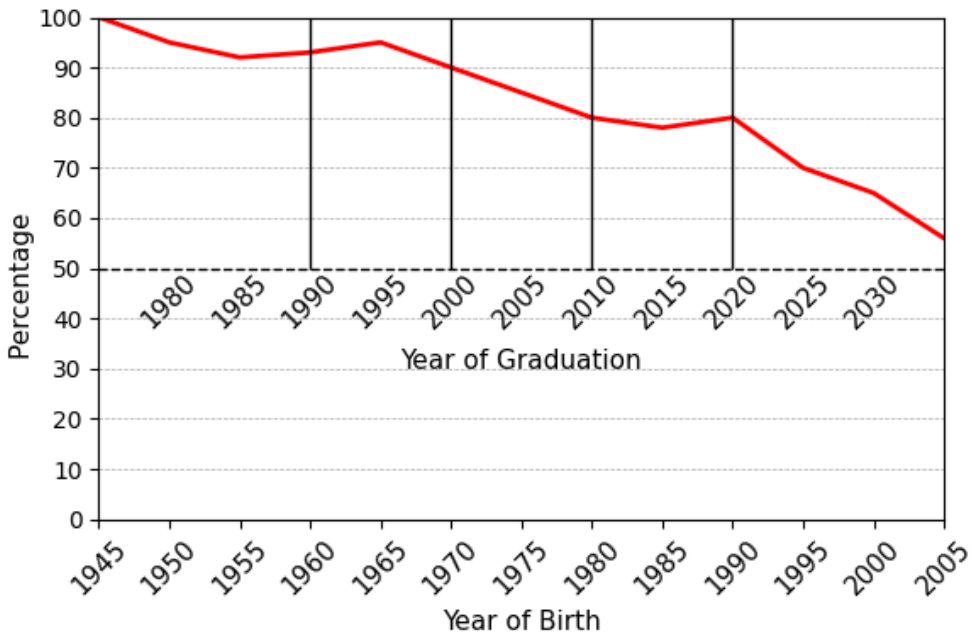


Figure 5: Proportion of newborns with a French Name by Year of Birth (1945–2005), with inferred Graduation Years

Source: Coulmont and Simon (2019). Adjusted by the author to include graduation years information.

over time, the distribution for those classified as *French* becomes noticeably more dispersed after 2000. This indicates that the range of predicted probabilities within the *French* class broadens, potentially reflecting cultural or demographic shifts in naming conventions—names may be less stereotypically "French" while still being confidently assigned to that class by the model.

Nevertheless, the model lacks the ability to distinguish between first- and second-generation migrants born in France if both their given names and surnames have foreign phonetic characteristics. Overall, the model demonstrates stable performance up to the year 2000.

To address the increased ambiguity in classifications post-2000, one potential solution involves adopting a Bayesian adjustment to reweight the classifier's output based on known or estimated class proportions in the population (Saerens et al., 2002). However, this correction would be most effective in cases where names and surnames are mixed—that is, when at least one of them still carries French phonetic features. For the time being, I proceed with a fixed threshold of 0.5, which yields a total of 174,338 individuals classified as *non_French* doctoral graduates.

4.2 Requiem_2 - 12 Regions of Origin

I then apply the **Requiem_2** model to the 174.338 individuals previously classified as *non_French* by **Requiem_1**. As in the previous step, I adopt a classification threshold of 0.5, based on the threshold analysis, which indicates that the distribution of the highest predicted class probabilities consistently lies above this value (see Figure A.6).

Among the classified regions, *Arabic* emerges as the largest region of origin, accounting for approximately 45% of individuals in 1990 and declining to 27% by 2020—a downward trend consistent with official statistics (MESR, 2020). The second-largest region is *Niger_Congolese*, which remains relatively stable at around 13% throughout the period. This is followed by the *East_Asian* region, which shows a steady increase over time, reflecting the rise in Chinese student migration from the 2000s onward. Other regions such as *Spanish*, *Italian*, and *West_North_Germanic* consistently hover around 10%, while the remaining regions of origin each account for roughly 5% over time.

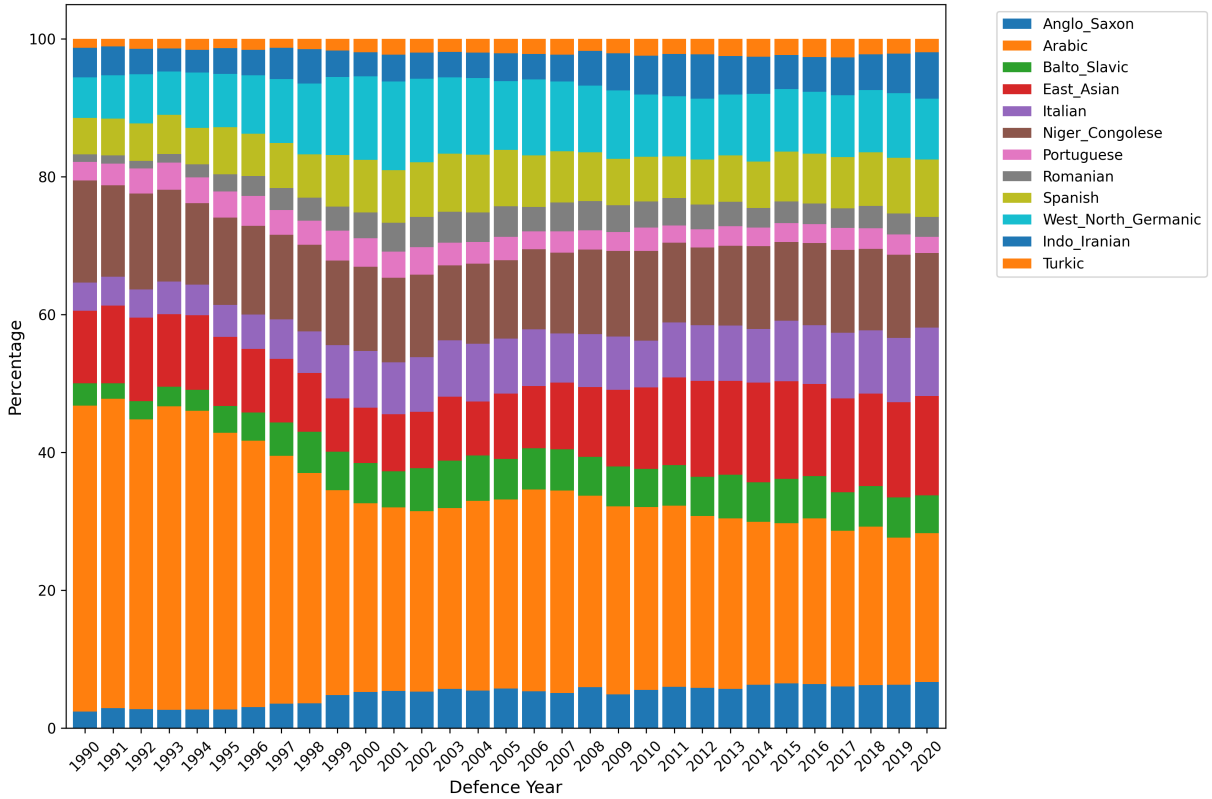


Figure 6: Proportion of regions of origin within foreign Doctoral Graduates in France, 1990-2020

To evaluate the performance of the model, I compare its output with the earliest available official statistics disaggregated by country of origin, as reported in Table 7. Since official statistics refer to all enrolled PhD students, while my dataset captures students at the time of graduation, I compute the average region of origin composition of doctoral graduates between 2011 and 2014 to ensure a consistent comparison. The most notable discrepancy is the underestimation of Arabic doctoral graduates, with the model assigning 24.85% compared to the official figure of 40.02%. This gap may, in part, be explained by the presence of graduates classified as *Indo_Iranian* (5.89%) and *Turkic* (2.37%), two categories not represented in the official statistics at the time. These regions likely include individuals of Arabic origin who were misclassified by the model.

This underestimation could also stem from the way the model is trained—specifically, the penalization of the Arabic class, which is the largest in the dataset (see Table 2). Other discrepancies include the underestimation of Italian and Portuguese graduates by approximately 2.5 percentage points each. Conversely, the *West-North Germanic* region appears significantly overestimated, possibly due to misclassification of French individuals with similar naming conventions. Similar patterns of overestimation are observed for the *Balto-Slavic* and *Anglo-Saxon* regions. The *East_Asian* region is slightly underestimated (by roughly one percentage point), while the estimates for *Niger-Congolese* and *Romanian* graduates closely align with official statistics.

Overall, aside from the substantial underestimation of the Arabic region, the model demonstrates solid performance. These results also highlight the inherent trade-offs involved in developing a classification model over a 30-year period, during which migration patterns and countries of origin have undergone significant transformations—particularly in a historically migration-rich country such as France.

Table 7: Regions of origin and percentage of foreign PhD students in France, 2011–2012

Regions of origin	**% of Foreign PhD Students 2011–2012 (Official)**	**Requiem_2 % 2011–2014**
Anglo_Saxon	2.16	5.96
Arabic	40.02	24.85
Balto_Slavic	2.81	5.93
East_Asian	14.59	13.69
Italian	11.57	7.97
Niger_Congolese	11.64	11.61
Portuguese	5.79	2.67
Romanian	3.13	3.48
Spanish	2.07	6.52
West_North_Germanic	4.97	9.05
Indo_Iranian	0.00	5.89
Turkic	0.00	2.37

Source: Campus France (2017) and results from **Requiem_2**. Official statistics aggregated by ethnicity from original country-level data.

5 Conclusions and Further Developments

This paper developed and applied a sequential classification approach based on a two-stage LSTM model (**Requiem_1** and **Requiem_2**) to identify the regions of origin of doctoral graduates in France from 1990 to 2020. Building upon methodologies from previous innovation and migration studies (Niggli, 2023; Ye et al., 2017), the model classified graduates using name-based linguistic features, tailored specifically to reflect France's unique migration patterns.

The results revealed a gradual shift in the composition of international doctoral graduates over three decades, particularly highlighting a notable decline in the share of Arabic graduates and a simultaneous increase in East Asian graduates. The model performed effectively overall, demonstrating robustness for most regions of origin. However, substantial classification challenges emerged from distinguishing between French natives and second-generation migrants, notably after the year 2000 due to changing naming conventions linked to increased immigration.

This research contributes methodologically by demonstrating how machine learning can enhance migrant classification when administrative data are scarce or incomplete. Practically, it sheds light on long-term migration trends within the French higher education system, offering policymakers valuable insights into demographic shifts in academic mobility.

This study presents several limitations. First, the classification model was trained using country-of-birth data rather than nationality or ethnicity. While country-of-birth information is stable and broadly available, nationality could better reflect actual migration status and align more closely with official statistics. This choice introduces potential inaccuracies, particularly among individuals who moved to France at younger ages or acquired French nationality later in life. Second, the availability and granularity of official statistics for foreign doctoral graduates are limited, with disaggregated data only accessible from 2011 onwards. This limitation restricts validation of the model's long-term accuracy and potentially obscures shifts in migration dynamics before that period. Moreover, using deceased individuals as a training dataset could introduce biases related to historical migration trends, mortality rates, or name-changing behaviors among certain migrant groups, thereby affecting classification accuracy. Finally, the model faced challenges in distinguishing between native-born French individuals and second-generation migrants whose names retain linguistic markers. This limitation, particularly pronounced in recent cohorts, implies that the model may systematically overestimate the share of foreign doctoral graduates post-2000.

Future research could extend this analysis by integrating alternative datasets that more directly capture nationality or educational migration trajectories. One promising avenue involves training the LSTM model on publicly available LinkedIn profiles, leveraging individuals' educational histories—particularly the country where graduates obtained their bachelor's degrees—as a more direct proxy for nationality or migration status. This approach could better reflect actual migration trajectories, capturing nuances in graduates' academic and geographic backgrounds that influence naming patterns.

References

Alba, R. and R. Silberman (2002): "Decolonization immigrations and the social origins of the second generation: the case of North Africans in France," *International migration review*, 36, 1169–1193.

Bade, K. J., P. C. Emmer, L. Lucassen, and J. Oltmer (2005): "Migration, Integration, and Minorities since the 17th Century. A European Encyclopedia," .

Bilsborrow, R. E. (1997): *International migration statistics: Guidelines for improving data collection systems*, International Labour Organization.

Buenstorf, G., D. P. Heinisch, and M. Kapa (2022): "International doctoral graduates as inventors in the German innovation system," *Industry and Innovation*, 29, 488–510.

Campus France (2017): "Les chiffres clés: 2017," Accessed: 2024-10-29.

——— (2024): "La mobilité étudiante dans le monde: Chiffres clés," Agence française pour la promotion de l'enseignement supérieur, l'accueil et la mobilité internationale, https://ressources.campusfrance.org/noindex/chiffres_cles_2024_fr.pdf (last access: July 2024).

Chellaraj, G., K. E. Maskus, and A. Mattoo (2008): "The contribution of international graduate students to US innovation," *Review of International Economics*, 16, 444–462.

Chen, W., K. Yang, Z. Yu, Y. Shi, and C. P. Chen (2024): "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, 57, 137.

CJC (2010): "Les jeunes chercheurs étrangers en France," https://cjc.jeunes-chercheurs.org/expertise/etrangers/2012-09-sondage-JC-etrangers.pdf, confédération des Jeunes Chercheurs.

Cohen, E. (2001): "Un plan d'action pour améliorer l'accueil des étudiants étrangers en France: Diagnostic et propositions," *Rapport au ministre de l'Éducation nationale et au ministre des Affaires étrangères*, 124, 24.

Coulmont, B. and P. Simon (2019): "Quels prénoms les immigrés donnent-ils à leurs enfants en France?" *Population & Sociétés*, 1–4.

de Rassenfosse, G. and R. Hosseini (2020): "Discrimination against foreigners in the US patent system," *Journal of International Business Policy*, 3, 349–366.

Doran, K., A. Gelber, and A. Isen (2014): *The effects of high-skilled immigration policy on firms: Evidence from H-1B visa lotteries*, National Bureau of Economic Research Cambridge, MA.

Gaulé, P. and M. Piacentini (2013): "Chinese graduate students and US scientific productivity," *Review of Economics and Statistics*, 95, 698–701.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep Learning*, MIT Press, http://www.deeplearningbook.org.

Gray, R. D. and Q. D. Atkinson (2003): "Language-tree divergence times support the Anatolian theory of Indo-European origin," *Nature*, 426, 435–439.

Hambrick, D. C., S. C. Davison, S. A. Snell, and C. C. Snow (1998): "When groups consist of multiple nationalities: Towards a new understanding of the implications," *Organization studies*, 19, 181–205.

Hunt, J. et al. (2010): "Skilled immigrants' contribution to innovation and entrepreneurship in the United States," .

INSEE (2023): "Définition: Immigré," Accessed: 2024-01-29.

Kerr, W. (2008): "The ethnic composition of US inventors," .

Kerr, W. R. and W. F. Lincoln (2010): "The supply side of innovation: H-1B visa reforms and US ethnic invention," *Journal of Labor Economics*, 28, 473–508.

López, V., A. Fernández, S. García, V. Palade, and F. Herrera (2013): "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information sciences*, 250, 113–141.

Mateos, P., R. Webber, and P. Longley (2007): "The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names," .

MESR (1994): "Repères et références statistiques sur les enseignements, la formation et la recherche," https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/53925/reperes-et-references-statistiques-sur-les-enseignements-et-la-formation-edition-1994-mini

——— (2005): "La mise en place du L.M.D. (licence-master-doctorat)," https://www.education.gouv.fr/la-mise-en-place-du-lmd-licence-master-doctorat-41234.

——— (2020): "Repères et références statistiques sur les enseignements, la formation et la recherche (Éditions 2005–2020)," https://cpesr.fr/wp-content/uploads/2020/09/, includes data from 2005 to 2020.

Miguelez, E., C. Fink, et al. (2013): *Measuring the international mobility of inventors: A new database*, vol. 8, WIPO.

Mu, G. M. (2015): "A meta-analysis of the correlation between heritage language and ethnic identity," *Journal of Multilingual and Multicultural Development*, 36, 239–254.

Mulders, A. M., B. Hofstra, and J. Tolsma (2024): "A matter of time? Gender and ethnic inequality in the academic publishing careers of Dutch PhDs," *Quantitative Science Studies*, 1–29.

Niggli, M. (2023): "'Moving On'—investigating inventors' ethnic origins using supervised learning," *Journal of Economic Geography*, 23, 921–947.

OECD (2000): *Education at a Glance. OECD Indicators*, Organisation for Economic Co-operation and Development.

——— (2022): *Education at a Glance. OECD Indicators*, Organisation for Economic Co-operation and Development.

Parasurama, P. (2021): "raceBERT–A Transformer-based Model for Predicting Race and Ethnicity from Names," *arXiv preprint arXiv:2112.03807*.

Peri, G., K. Shih, and C. Sparber (2015): "Foreign and Native Skilled Workers: What Can We Learn from H-1B Lotteries?" Tech. rep., National Bureau of Economic Research.

Phinney, J. S. (2013): "Ethnic identity in adolescents and adults: Review of research," *Readings in ethnic psychology*, 73–99.

Pommier, S., M. Talby, M. Auffray-Seguette, M. Dalaut, H. Eijsberg, et al. (2022): "Le doctorat en France: Regards croisés sur la formation doctorale," Rapport de recherche, Réseau National de Collèges Doctoraux.

Ruhlen, M. (1994): *On the origin of languages: studies in linguistic taxonomy*, Stanford University Press.

Saerens, M., P. Latinne, and C. Decaestecker (2002): "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure," *Neural computation*, 14, 21–41.

Stephan, P. E. and S. G. Levin (2001): "Exceptional contributions to US science by the foreign-born and foreign-educated," *Population research and Policy review*, 20, 59–79.

UNESCO (2023): "Migrants, réfugiés ou déplacés?" Accessed: 2024-01-29.

Vaswani, A. (2017): "Attention is all you need," *Advances in Neural Information Processing Systems*.

Vossen, R. and G. J. Dimmendaal (2020): *The Oxford handbook of African languages*, Oxford University Press, USA.

Vourc'h, R. (2010): "Les doctorants: Profils et conditions d'études," Tech. Rep. 24, Observatoire national de la vie étudiante (OVE).

Ye, J., S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena (2017): "Nationality classification using name embeddings," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1897–1906.

Zheng, Y. and O. Ejermo (2015): "How do the foreign-born perform in inventive activity? Evidence from Sweden," *Journal of Population Economics*, 28, 659–695.

Appendices

A Additional Figures

Table 2. Principal immigrant groups in France, 1946–1999

	1946 (%)	1954 (%)	1962 (%)	1968 (%)	1975 (%)	1982 (%)	1990 (%)	1999 (%)
Spaniards	21.6	20.5	24.5	25.5	15.3	9.1	6.2	5.3
Italians	32.2	36.0	34.8	24.0	14.3	9.5	7.3	6.6
Poles	30.2	19.0	9.8	5.5	2.9	1.8	1.4	1.1
Portuguese	1.6	1.4	2.8	12.4	23.4	21.5	18.7	18.2
Algerians	1.6	15.0	19.3	19.9	21.9	22.5	17.6	15.6
Moroccans	1.1	0.8	1.8	3.5	8.0	12.3	16.5	16.6
Tunisians	0.1	0.4	1.5	2.6	4.3	5.3	5.9	5.1
Other Africans	0.9	0.1	1.0	1.4	2.5	4.4	6.9	9.3
Asians	5.0	2.9	2.0	1.9	3.2	8.1	12.2	13.5
Turks	0.6	0.4	*	0.3	1.6	3.4	5.7	6.8
Yugoslavs	1.5	1.2	1.2	2.0	2.2	1.7	1.5	1.6
Russians	3.6	2.4	1.4	0.8	0.4	0.2	0.1	0.4
Total principal nationalities (000s)	1,401	1,413	1,810	2,379	3,243	3,575	3,479	3,047

* Not available.

Source: Calculated from INSEE, censuses published in Blanc-Chaléard, *Les immigrés et la France*, 51.

Figure A.1: Principle immigrant groups in France

Source: Bade et al. (2005) - pag. 59

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		END
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure A.2: Example of the Encoded Name Romain Gary

Taxonomy of ethnic origins and olympic national teams

Ethnic origin	National teams/countries
Anglo-Saxon	Great Britain, Ireland
Chinese	China
Dutch	Netherlands
French	France
German	Germany
Hispanic-Iberian	Spain, Portugal, Mexico
India	India
Italian	Italy
Japanese	Japan
Korean	Korea
Arabic	Egypt, Syria, Saudi Arabia, Jordan,
	UAE, Tunisia, Algeria, Morocco
Persian	Iran
Slavic-Russian	Russia, Ukraine, Belarus
East-Europe	Poland, Czechoslovakia, Hungary
Balkans	Serbia, Croatia, Yugoslavia
Scandinavian	Sweden, Norway, Finland, Denmark, Iceland
Southeast Asia	Vietnam, Thailand, Malaysia, Indonesia,
	Laos, Cambodia
Turkish	Turkey

Figure A.3: Taxonomy used by Niggli (2023)

Source: Niggli (2023)

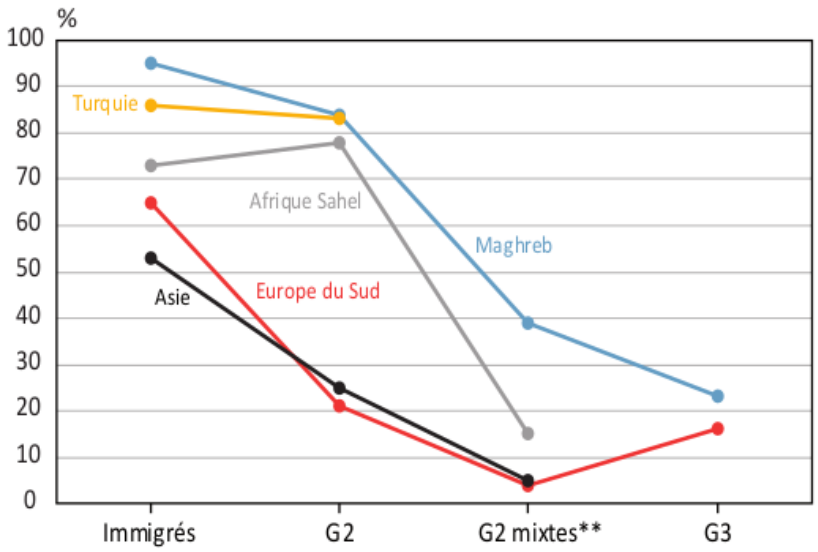


Figure A.4: Proportion of names specific to region of origin by generation

Source: Coulmont and Simon (2019).
Note: Immigrés: First generation migrants.
G2: Second generation migrants.
G2-mixtes: One partent migrant - one parent native.
G3: Newborns identifies as migrant descendants.

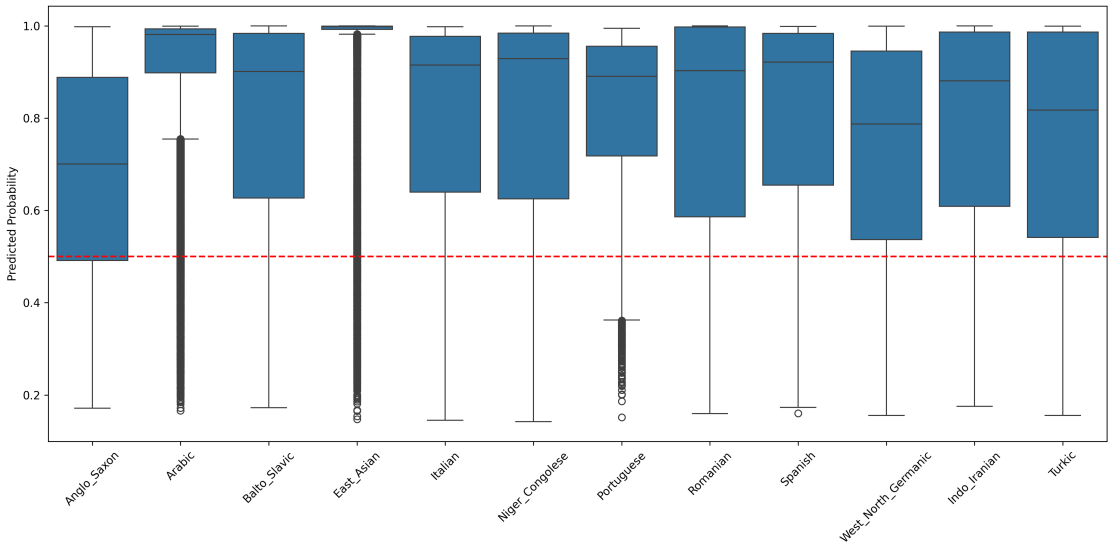


Figure A.6: Distribution of the Highest Predicted Class Probability Across Regions of Origin

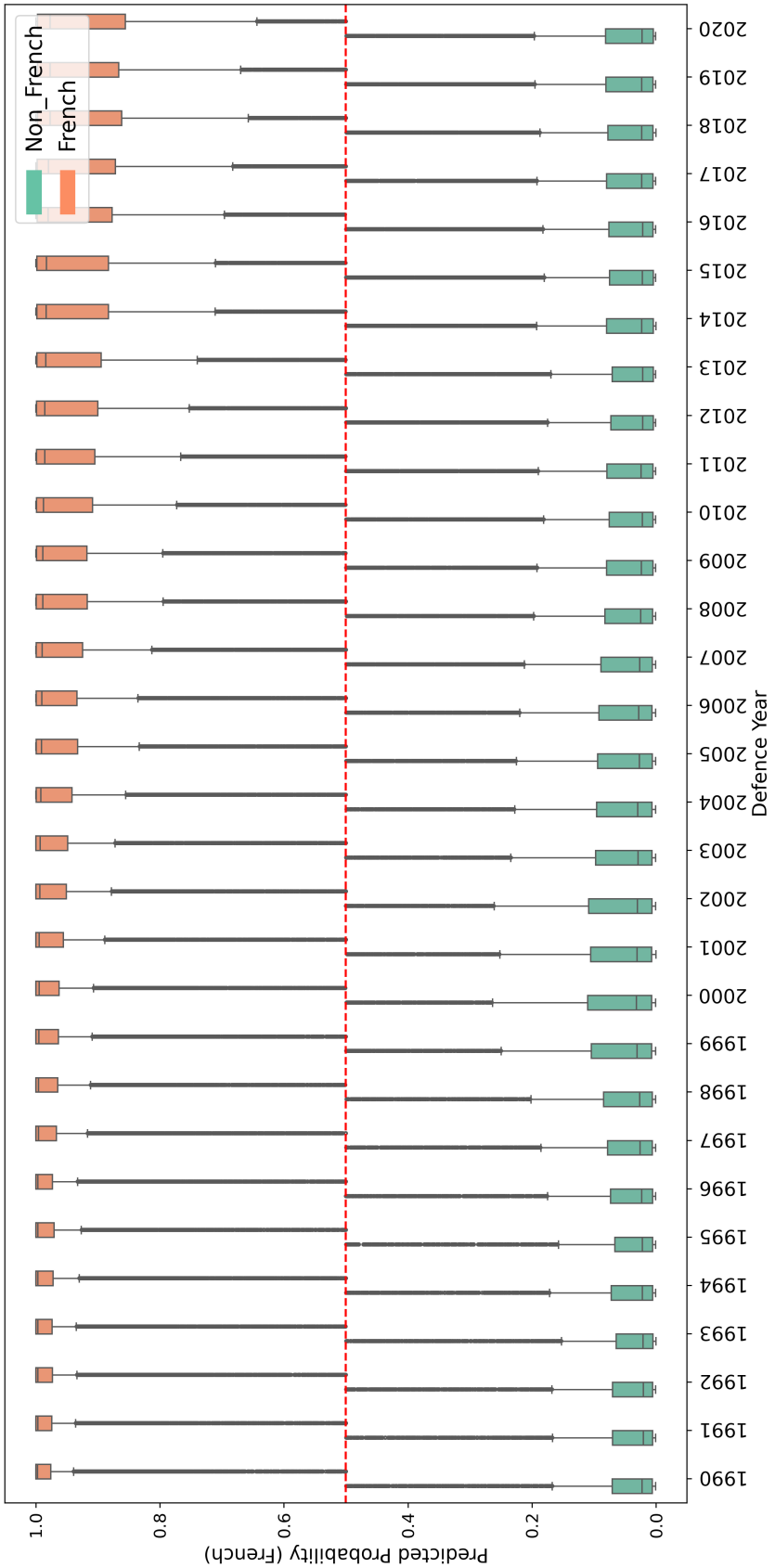


Figure A.5: Threshold Analysis Requiem_1

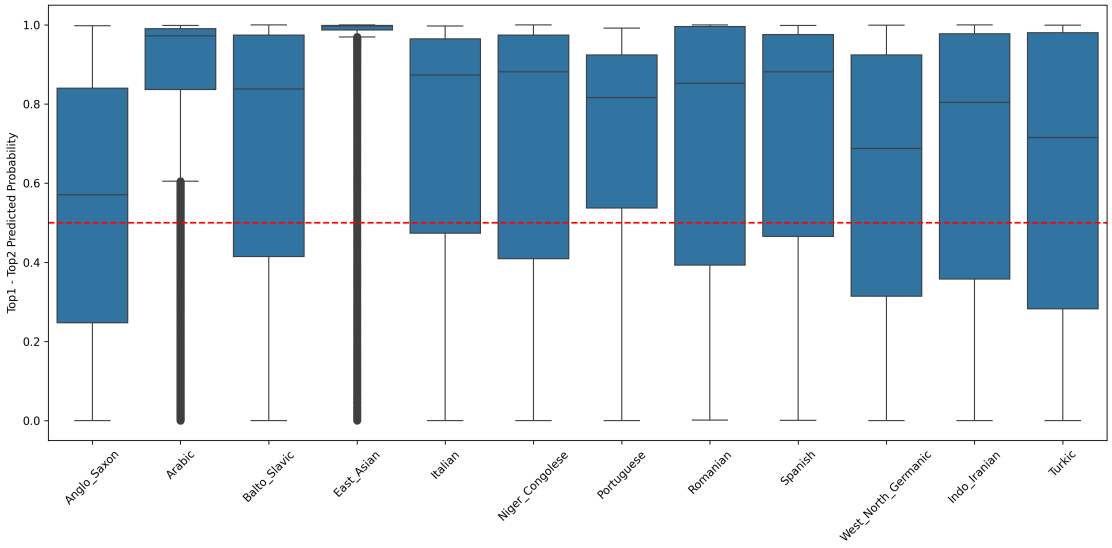


Figure A.7: Distribution of the 2nd-Highest Predicted Class Probability Across Regions of Origin

B Additional Tables

Table B.1: Regions of origin and countries of birth of deceased individuals in France, born between 1950 and 1990

Regions of origin	**Country of Birth of Foreign Deceased Individuals in France**
Anglo_Saxon	United States, Australia, Ireland, Comoros (UK)
Arabic	Algeria, Morocco, Tunisia, Egypt, Lebanon, Syria, Mauritania, Iraq, Libya, Saudi Arabia, Jordan, Yemen, United Arab Emirates, Qatar, Palestinian Territories, Bahrain, Oman, Sudan
Balto_Slavic	Latvia, Lithuania, Estonia, Poland, Czech Republic, Slovakia, Russia, Ukraine, Belarus, Serbia, Bulgaria, Bosnia and Herzegovina, Croatia, Montenegro, North Macedonia, Slovenia
East_Asian	China, Vietnam, Cambodia, Laos, Thailand, Myanmar (Burma), Japan, South Korea, North Korea, Taiwan, Malaysia, Indonesia
Italian	Italy
Niger_Congolese	Senegal, Comoros, Democratic Republic of the Congo, Mali, Madagascar, Côte d"Ivoire, Cameroon, Republic of the Congo, Mauritius, Cape Verde, Guinea, Angola, Togo, Ghana, Benin, Central African Republic, Guinea-Bissau, Nigeria, Gabon, Burkina Faso, South Africa, Rwanda, Gambia, Burundi, Mozambique, Kenya, Liberia, Sierra Leone, Equatorial Guinea, Zimbabwe, Tanzania, Uganda, Niger, Malawi, Zambia, Namibia, Lesotho, Eswatini
Portuguese	Portugal, Brazil
Romanian	Romania
Spanish	Spain, Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela
West_North_Germanic	Germany, Switzerland, Austria, Luxemburg, Denmark, Sweden, Iceland, Aruba (Dutch)
Indo_Iranian	India, Iran, Pakistan, Afghanistan, Bangladesh, Tajikistan
Turkic	Turkey, Azerbaijan, Kazakhstan, Uzbekistan, Turkmenistan, Kyrgyzstan

C Foreign Doctoral Graduates Official Statistics

While in my data I observe PhD students at their graduation, official statistics are available for all the PhD students per cohort. The first cohort-information available on the percentage of foreign PhD students in France is 2004-2005 cohort, as shown in Figure C.8. For that cohort, 33% of PhD students were foreign, a share that increased steadily to 41% by 2010 (MESR, 2020)[15]. To estimate the proportion of foreign PhD students for earlier years, I assume a correlation between the share of foreign students in French universities overall—available since the 1985-1986 cohort—and that of foreign PhD students. This assumption is supported by a survey conducted by the *Confédération*

[15]The various census surveys conducted in France identify foreign students based on self-declared nationality. This includes students who have completed their secondary education within the French school system (MESR, 2020)

des Jeunes Chercheurs[16], which reports that approximately 75% of foreign PhD students in France obtained their master's degree in France, and 24% completed their bachelor's degree there as well (CJC, 2010).

A potential caveat to this inference is the change in the higher education system following the Bologna Process. Prior to its implementation, the structure was more complex[17], though it still comprised a progression through undergraduate (*Licence*) and graduate (*Maîtrise*) degrees. The LMD reform (Licence–Master–Doctorat) merely aligned the French system with international standards (MESR, 2005).

Naturally, there is a lag between the proportion of foreign students in higher education generally and their eventual representation at the PhD level. Assuming a typical five-year trajectory from bachelor's to master's degree, I introduce a temporal shift to align the two trends. The optimal lag, which maximizes the correlation between the two curves (correlation coefficient = 0.84), is four years. This is plausible given that most foreign students entering PhD programs in France had previously enrolled in master's, rather than bachelor's, programs. Based on this lag, I estimate the share of foreign PhD students to range between 30% and 38% over the 1990-2004 period. While these estimates may not be exact, it is reasonable to infer that the proportion did not fall below 30% during this time.

[16]For more information, see: https://cjc.jeunes-chercheurs.org/, last accessed 10 April 2025.

[17]The first two years of undergraduate studies granted the *DEUG* diploma, followed by the *Licence* in the third year. The master's level consisted of the *Maîtrise* in the first year and either the *DEA* or *DESS* in the second year. The doctorate followed thereafter (MESR, 2005).

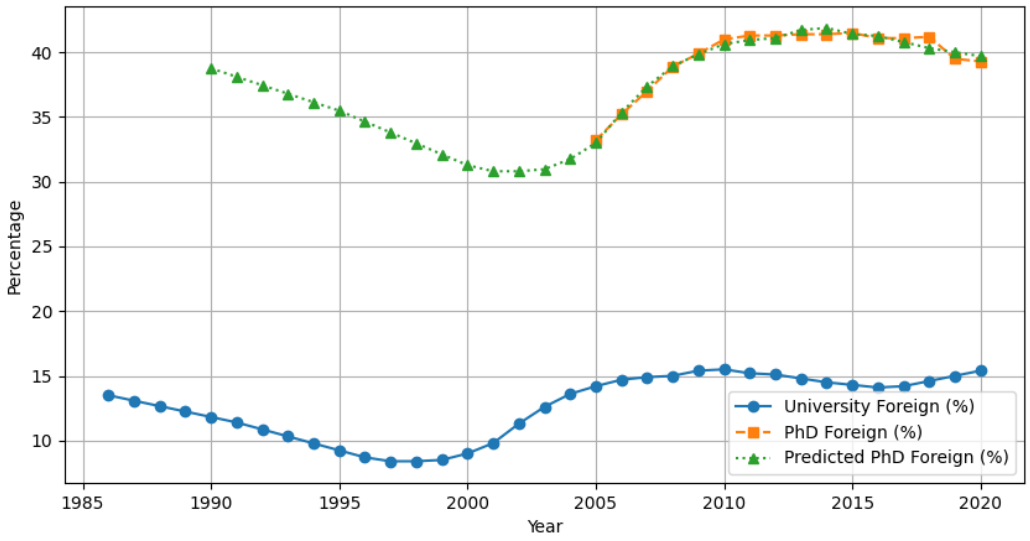


Figure C.8: Proportion foreign PhD students in France 1990-2020

Source: MESR (2020).
Note: Proportion foreign PhD students from 1990 to 2005 is inferred.

As for the composition of foreign PhD students by country of origin, disaggregated data only becomes available starting with the 2011-2012 cohort (Campus France, 2017). I use these statistics to construct a taxonomy of foreign doctoral graduates in France. Table C.2 presents the correspondence between regions of origin and countries of origin. Turkey and Iran are also included, as they were among the leading countries of origin for foreign university students in France during the 1990s (MESR, 1994).

Table C.2: Campus France statistics for Cohort 2011-2012

Rank	Country of Origin	PhD students 2011-2012	% of foreign PhD students	2015-2016 Rank	Regions of origin
1	Tunisia	2275	14.41	2	Arabic
2	Algeria	1934	12.25	4	Arabic
3	Italy	1826	11.57	3	Italian
4	China	1815	11.50	1	East-Asian
5	Lebanon	1054	6.68	5	Arabic
6	Morocco	1054	6.68	6	Arabic
7	Brazil	743	4.71	7	Portuguese
8	Senegal	673	4.26	10	Niger-Congolese
9	Germany	553	3.50	11	West-North Germanic
10	Romania	494	3.13	14	Romanian
11	Vietnam	487	3.09	8	East-Asian
12	Ivory Coast	451	2.86	16	Niger-Congolese
13	Russia	444	2.81	12	Balto-Slavic
14	Cameroon	387	2.45	13	Niger-Congolese
15	United States	341	2.16	18	Anglo-Saxon
16	Spain	327	2.07	9	Spanish
17	Gabon	326	2.07	15	Niger-Congolese
18	Belgium	232	1.47	17	West-North Germanic
19	Madagascar	198	1.25	20	Niger-Congolese
20	Portugal	170	1.08	19	Portuguese

Source: Campus France (2017). Adjusted by the author to reflect 2011-2012 student composition.

D LSMT Model

The process of the LSTM model proposed by Niggli (2023), could be summarized as follows:

- **Data Encoding**: Prior to model training, names are transformed into a machine-readable format using one-hot encoding (For an example see Figure A.2)[18] This ensures the input is structured appropriately for the LSTM model.
- **Training and Test Sets**: The dataset is partitioned into a *Training Set* used for learning and a *Test Set* used for evaluation.
- **LSTM Model Architecture**: The model consists of three LSTM layers, followed by a dense output layer. The first layer captures long-range dependencies, the second layer refines learned patterns, and the third layer distills relevant information for classification. The final hidden state is passed to a dense output layer with a softmax activation function, generating probability distributions over possible name origins.
- **Model Training**: During training, an optimization algorithm adjusts network weights to

[18]Each character is mapped to a unique position in a predefined character dictionary, consisting of lowercase Latin letters (a–z), spaces, and an "END" character for padding. Names are then represented as sequences of one-hot vectors, stored in a three-dimensional tensor.

minimize the loss function, which quantifies the model's predictive accuracy.

- **Model Evaluation**: The trained model is tested on the unseen dataset, with performance assessed through precision, recall, and F1-score metrics. These measures indicate how effectively the model predicts name origins and where it encounters classification challenges.

E Data Collection for Observed Location Information

Interns constructed a hand-curated sample of doctoral graduates from the 1992–2002 cohorts, retrieving information on their geographic background using multiple sources. For each cohort, 80 individuals were selected, resulting in an initial sample of around 880 graduates. Over half of the information was obtained through LinkedIn profiles, while additional sources included personal and university websites, CVs, blogs, and online journals.

The retrieved location indicators include country of birth, and the location of high school, bachelor's degree, and master's degree. Each doctoral graduate in the sample has at least one of these data points available. We exclude the variable *native language*, as all 17 individuals in that category report French, offering no variation for classification. Additionally, we exclude the 2002 cohort due to its unusually low share of non-French graduates (5%), which contrasts sharply with the roughly 20% observed in other years. Table E.3 reports the proportion of non-French individuals in the validated sample by year, prior to the exclusion of the 2002 cohort. The cause of this anomaly is unclear.

After applying these exclusions, we retain a total of 786 observations with at least one usable location indicator for origin classification. Table E.4 summarizes the distribution of these observations by the first available type of location information.

Table E.3: Proportion of Non-French Graduates in the Validation Sample, by Defence Year, 1992–2002

Defence Year	Proportion Non-French
1992	0.1625
1993	0.1875
1994	0.2250
1995	0.2125
1996	0.1625
1997	0.2375
1998	0.1875
1999	0.1875
2000	0.1500
2001	0.2250
2002	0.0500

Table E.4: First Available Location Indicator for Origin Classification (Post-Exclusion)

First Available Location Type	Number of Observations
Bachelor Location	282
Master Location	191
High School Location	164
Country of Birth	146
Total	**786**

Chapter 2

The Chosen: Researcher Visas and Foreign PhD Graduates' Careers in France

Andriy Romanyuk[1,2,*], Alberto Corsini[3] & Francesco Lissoni[1,4,5]

[1]BSE UMR CNRS 6060 – Université de Bordeaux
[2]DiECO – Università dell'Insubria
[3]IPP CSIC – Spanish National Research Council, Madrid
[4]Dept. of Economic Policy – Università Cattolica del Sacro Cuore, Milan
[5]Institut Convergences Migrations, Paris

Abstract

We investigate whether and how the introduction of the first visa scheme dedicated by France to foreign researchers (the *carte de sejour "scientifique-chercheur"* of 1998) affected the post-graduation stay rate of doctoral graduates of French universities from outside the European Union (EU), most of whom coming from former colonies in Arab-speaking countries and Western Africa. We build a novel dataset containing the scientific publications of all doctoral graduates in STEM disciplines (Science, Technology, Engineering and Mathematics) from 1992 to 2002; and identify as stayers those with a post-PhD publishing activity and a French university affiliation (while leavers include both those with no postdoc academic record or active outside France). Alternatively, we identify as stayers those who become themselves doctoral supervisors, by 2025. Through name analysis, we distinguish between non-EU graduates, to whom the *carte de sejour* applied, and French or other EU ones, to whom it did not. Based on a difference-in-differences estimation strategy with repeated cross-sectional data, we find no meaningful impact of the policy under study. The results hold for various versions of our name analysis and our dependent variable. Stay rates eventually converge, but much later than both 1998 and with no apparent connection to visa policy.

JEL classification: F22; K37; O15; O3
Keywords: International migration; Migration policy; Scientific research; Graduate studies

We thank Pauline Menu, Carla Fournier, Valentine Petzold, Baptiste Comby, Aman Araissi, Lucien Boucart, Samuel Desneulin and Clovis Sourisse for their valuable research assistance. We received valuable comments on early drafts of the paper from the participants to the from the participants to the 18th Workshop on the Organization, Economics and Policy of Scientific research (WOEPS), at the University of Strasbourg; the 7th Global Conference on Economic Geography (GCEG), at Clark University; the REGIS Summer School 2025, at the University of Bordeaux; and the 40th meeting of the European Economic Association (EEA), also in Bordeaux. More comments came from students and colleagues attending seminars at the Bordeaux School of Economics and the Department of Political Economy of the Università Cattolica in Milan. We acknowledge funding from the European Patent Office's Academic Research Programme (DOC-TRACK project). However, the results and comments presented in this paper are solely ours and do not engage the Office's views. The Agence bibliographique de l'enseignement supérieur (ABES) kindly provided us with access to the dump version of the data and has always been available for clarification and advice. Lissoni also acknowledges funding from the French State in the framework of the Investments for the Future programme IdEx Université de Bordeaux / GPR HOPE. All errors are ours.

1 Introduction

The international circulation of tertiary students has increased incessantly over the past few decades. In 2021, there were more than 6.4 million foreign students enrolled in higher education programs worldwide, up from less than 2 million in 1998, the first year for which we have reliable statistics on a global scale (UIS, 2000). A number of European Union (EU) countries, including France, appear in the top destination list, well beyond the United States but alongside English-speaking education powerhouses such as the United Kingdom, Australia and Canada.

While only a minority of international students ultimately decide or have the possibility to stay in the host countries after graduation, those who do so are often found to contribute significantly to local science and innovation (Hunt and Gauthier-Loiselle, 2010; Hunt, 2011; Stuen et al., 2012; Ganguli et al., 2020; Romanyuk and Lissoni, 2025). One reason is that, relative to natives, they are more likely to enroll in Science, Technology, Engineering and Mathematics (STEM) programmes (OECD, 2022), especially when it comes to graduate studies and, in particular, doctoral ones. According to UNESCO estimates for the year 2012, 53% of foreign doctoral students worldwide were enrolled in either "Science" or "Engineering, manufacturing and construction" programmes, against only 29% of non doctoral ones (UNESCO, 2016).

Evidence for the United States suggests that one key factor affecting the students' decision to stay in the host countries after completing their studies is the cost and ease of getting a residence permit soon after the expiration of their student visas (Roach and Skrentny, 2019, 2021). This poses a policy problem to governments that both wish to attract STEM foreign students and invest heavily in innovation, but whose general immigration policies tend to be restrictive, as it is often the case in Europe (Hawthorne, 2018). One possible solution, which the EU endorses and many member countries have adopted, consists in the creation of special visas and permits for foreign graduates of local universities, often as part of special programmes for selected categories of highly-skilled immigrants (Klaus, 2022).

In this paper, we evaluate the impact of one such scheme, introduced by France in 1998 as part of a major immigration policy reform (best known as the *Loi Reseda*), by which any holder of either a master or doctoral degree and a work contract with a French university or public laboratory could obtain both a visa and a residence permit (*carte de sejour "scientifique-chercheur"*) through a

simplified procedure, at a time when the release of ordinary work permits was subject to extremely restrictive conditions. Despite some revisions and changes of name, most notably in 2006 and 2016, this visa-cum-permit has become and still is a pillar of French selective immigration policies (d'Albis and Boubtane, 2021). Pointedly, it has survived all the successive policy reforms of the last two decades, even the most restrictive ones.

According to estimates based on the number of scientific permits assigned to previous student permit holders, in each year from 2009 to 2015 (the only years for which we have found official data), this policy measure has allowed around 4,000 foreign graduates to stay in France for a period ranging from a few months to two years, or even longer, after their graduation. Around 28% of them had entered France with a doctoral contract. To date, however, no systematic policy evaluation effort has been undertaken, absent any reliable data on the inflows and stay rates of either STEM students or scientists dating back to before 2007 (OECD, 2017).

To address this gap, we have compiled a comprehensive dataset covering all doctoral students in France since 1992, including their post-doctoral publication activity (authorship of scientific publications), their affiliation at the time of publication (inside or outside France), and the full population of supervisors up to 2025. This allows us to identify those graduates who are likely to have undertaken an academic career in France, either because they keep publishing - after graduating - with a French affiliation or because they appear, a few years later, as supervisors of other doctoral graduates (a role to which French academia admits only its full professors or the equivalent of tenured associate ones, namely the *maîtres de conférences* with a special habilitation). Based on this information, as well as on supplementary information on the students' region of origin, we can exploit a quasi-experimental setting that allows for a dynamic difference-in-differences estimation of the impact of the 1998 policy change on the probability of foreign students remaining in French academia in the years following their graduation. The setting rests on the distinction between students from either France or the European Union (as per its 1995 borders; EU) and all others (non-EU). Of these two groups, only the latter needed - back in 1998 as of today - a residence permit to stay in France after graduation. In particular, EU students were and still are entitled to get one (or, more often, to do without it altogether) in force of the Freedom of Movement principle, first established by the Treaty of Rome (1957) and later reformulated by the Treaty of Maastricht (1992) (EC-EMPL, 2000; Rogers et al., 2012).Policy-wise, this implies that,

for the new policy to be deemed successful, we should observe an increase in the post-doctoral stay rates of non-EU students, relative to French and/or EU students. This is the main hypothesis we aim to test. Additional appreciations of the policy's impact concern the length of the post-doctoral stay (whether limited to just a few years after graduation or more extended in time, or possibly permanent) and the long term effect (if they become doctoral supervisors).

Data-wise, our strategy has consisted of extracting from *thèses.fr*—the official online collection of electronic doctoral dissertations managed by ABES, the French Bibliographic Agency of Higher Education—all names, surnames, and unique IDs of each year's STEM graduates, along with information on their dissertations, supervisors, and universities. Using this data, we tracked each graduate's scientific publications available in *Scopus*—a scientific abstract and citation database—along with their country of affiliation. We then identified as "stayers" those graduates who published articles with a French affiliation after their graduation year, further distinguishing between those active in France immediately afterward and those who remained affiliated in later years. As for identifying the doctoral graduates who become supervisors, we relied on their ID, which ABES maintain unchanged across all instances in which they appear not only in *thèses.fr*, but in all other databases the Agency maintains.

Very importantly, we also assigned each graduate to a region of origin, with the main goal of distinguishing between France and foreign countries or groups of countries and, among the latter, between those inside (EU), and outside the EU (non-EU). Since the information provided by ABES in this respect is unreliable, at least for the period of our interest, and no university or ministry is allowed to release personal information on students, we relied on a probabilistic assignment method based on name analysis.

Specifically, we adapted to our France-specific data a machine learning model conceived by Niggli (2023) for inventors worldwide. In particular, we trained the Niggli's model with information from the "Fichiers des personnes décédées," a database released yearly by the French National Institute of Statistics, which provides the full name, place, and year of birth of all individuals who died in France in that year. The model first assign the graduates to one of two categories, namely French versus non-French, then distributes those in the latter across twelve ethno-linguistic groups, a few of which overlap with regions of origins clearly marked as belonging to the EU or not. (For example, Italians or West-North-Germanic in the EU and Arabic or Niger-Congolese outside it).

For our analysis, we rely on a difference-in-differences estimation strategy with repeated cross-sectional data, with graduates from selected non-EU regions of origin as the treatment group and the French graduates as the control ones (EU graduates fail to satisfy the parallel trends assumption); and with the graduation cohorts from 1998 to 2001 as the treated ones, and the preceding ones as the non-treated.

In line with our expectations, we find that the French–and EU graduates–exhibit substantially higher stay rates than the non-EU ones. But, contrary to them, we also find that the 1998 policy change did not significantly reduce this gap. In particular, we find no effects on short-term stay rates (graduates publishing in France from 2 to 4 years after graduating). For the long term stays rates (graduates publishing in France from 5 to 10 years after graduating) and for the probability to become supervisors, we observe an impact of respectively 6 and 5 percentage points, exclusively for the 2001 graduation cohort. However, these results are sensitive to the fine-tuning of our name-based classification model for the regions of origin and do not hold when introducing supervisor fixed effects.

While requiring further checks, these results are the first of their kind for a European country's scientific visa policy and suggest that the policy's impact has been more limited than it was intended. Several factors may explain them.

First, they may assigning nationality through a name- and surname-based classification model introduces attenuation bias, particularly in the case of individuals with non-French-sounding names who may have been born in France and completed their entire education there. While these individuals are likely to be classified as of foreign origin by the model, they are to be considered French for the purpose of this study.

Second, institutional barriers may have limited the policy's effectiveness. Notably, Algerian nationals—who in 2011 constituted the second-largest group of foreign doctoral graduates in France (Campus France, 2019)—were excluded from eligibility for the visa scheme until 2001 (Slama, 2001; GISTI, 2001). Since the Arabic speaking graduates are also the ones most confounded with French graduates by our classification model, we excluded these graduates from our main analysis. However, this comes with a significant loss of observations, since Arabic speaking graduates are more than half our non-EU graduates.

Finally, more than 70% of foreign doctoral graduates enrolled in 2010 had previously obtained a

Master's degree in France (CJC, 2010). Given that a Master's typically lasts two years and a PhD around four, these students often spent at least six years in the country. Since French citizenship requires five years of residence, the policy may have mainly facilitated the path to citizenship for graduates who were already likely to remain in France.

The paper proceeds as follows. In section 2, we place the policy change of our interest in the context of France's general immigration policies and its historical importance as a destination for foreign doctoral students. In section 3, we present our quasi-experimental strategy. In section 4, we provide essential information on our data acquisition and treatment methods, plus a number of descriptive statistics on our sample. In section 5 we report and discuss our results. Section 6 concludes.

2 Migration policy and foreign PhDs in France

2.1 France's dual immigration policy

Like all EU member countries, France has a dual immigration policy system, open for EU citizens and selective for all others. Concerning the EU, it adheres to the Freedom of Movement principle, which guarantees to all its citizens "the right to move and reside freely within the territory of the Member States" (art.21 of the Treaty on the functioning of the European Union). This right ensures that immigrants from within the EU are not subject to the same immigration laws that concern the non-EU ones. This right is extended to citizens from Iceland and Norway - which adhere to the European Economic Area - as well as to Switzerland, due to a series of treaties, the last of which signed in 1999 (Cristelli and Lissoni, 2020).

Importantly for this study, this exemption dates back at least to 1992, when the Treaty of Maastricht first introduced the concept of EU citizenship and its associated rights; but it builds upon the right to travel and reside for work in any country of the European Economic Community first established in 1957. In France, two of the most important implementation steps of this fundamental right date back to 1968, when it was established that European Community's workers not holding a residence permit were only punishable with a fine (their right to reside being intangible); and to the 1970s, with the introduction of the right for these workers to keep residing in France after

retirement or when losing their job, along with their families (Rodier, 2001).

This liberalisation of movements within the EU took place in a context of progressive restriction of immigration flows from outside it, dating back to the mid-1970s and politically motivated by the unprecedented rise of unemployment in those years (OECD, 2017; DILA, 2024a). In 1974, all labour immigration was suspended. In 1975, the legislator introduced, as a general principle, the requirement for any job offer to immigrant workers to pass a labour market test ("opposabilité de la situation de l'emploi" or OSE), wherein the prospective employers must demonstrate that they have first made a good faith, unsuccessful effort to recruit a French resident. Although its implementation has undergone a number of changes over the years, this procedure is still a cornerstone of French laws on labour immigration. For a while, it affected also the foreign students willing to stay in France after getting their university degree. In 1977, an administrative circular limited their possibility to obtain a residence permit after completing their studies (Kabbanji and Toma, 2020), while legislation in the 1980s introduced increasingly severe controls and penalties for administrative infractions. In 1993, a new wave of laws and administrative regulations also limited, among others, the right of foreign students to be joined by their families before graduating; and further complicated the procedures for moving from a student permit to a permanent one. This was part of a more general strategy to achieve a "zero immigration" goal.[1]

In the public discourse of the successive years, this goal was replaced by a less draconian commitment to a selective immigration policy, with the main goal of banning low-skilled immigration. This paved the way for the introduction of a number of policy schemes aimed at shielding highly skilled individuals from the increasing restrictions on labour immigration. In particular, following the recommendations of a highly publicized official report (Weil, 1997), the *Loi Reseda* (Reseda law) of May 1998 introduced, among others, a special procedure for obtaining both a visa and a residence permit dedicated to scientific researchers (*carte de sejour "scientifique-chercheur"*).[2]

[1] On this goal, see the interviews to then Minister of the Interiors Charles Pasqua, now collected and made available by DILA (2024b). As for the OSE procedure, in 1998, our main year of interest, this involved - also for highly skilled workers - the French National Employment Agency (ANPE), which examined the job offer, published an advertisement, and made sure that no French applicant met the required profile. In case of a favourable decision, this forwarded the file to the French International Migration Office (OMI), which checked for the existence of any administrative or legal impediment to the future workers' immigration, subjected them to medical examination in their countries of origin and organised the travel arrangements. All costs were up to the employers and workers, with an uncertain timing (Herzberg, 1998).

[2] "Reseda" is an acronym standing for law "relative to the entry and permanence of foreigners in France and to the asylum right" (*Relative à l'entrée et au séjour des étrangers en France et au droit d'asile*). In some texts, the same law, whose official number is 98-349, also goes under the name of the Minister of the Interiors of the time,

The new permit was available to both doctoral students and post-doc researchers, and it could be obtained without undergoing the OSE procedure, under the condition of holding at least a master degree and of being recruited by an accredited research organisation. The list of such organizations included all French universities and public laboratories as well as a number of non-profit entities (but no business company, no matter how R&D intensive). Differently from the selective permits introduced by the same law for other categories of highly skilled workers, its beneficiaries were not required to earn a minimal wage or income, as long as their doctoral or post-doc grant respected the legal minimal sum. Four years later, a follow-up decree introduced new rules for facilitating the change of status for students who had a promise of employment of particular technological and commercial interest (Math et al., 2006).

The scientific permit introduced in 1998 is still in place and relies, despite some modifications, on the same key principles (exemption from any labour market test, no minimal income requirement, and no quotas)[3]. From 2018 to 2022, it has been granted to no less than 3000, occasionally 5000 applicants per year, most of whom are either doctoral students or doctoral graduates with post-doc contracts (Campus France, 2024).

The dual structure of the French immigration system we have just described presents a few exceptions. These are due to its colonial past and the ensuing decolonization process, which has produced a number of bilateral treaties between France and a number of African countries. Of these, the most important one is with Algeria, first signed in 1968 and frequently revised. This is particularly restrictive regarding labour immigration and, until 2001, it prevented Algerian students from accessing the *carte de sejour "scientifique-chercheur"* after graduation (Slama, 2001; GISTI, 2001, 2020). On the contrary, immigrants from Togo, Gabon and the Centrafrican Republic were exempted, until 2000, from any labour entry test (OECD, 2017).

Jean-Pierre Chevenement. Besides the scientific permit, it introduced special permits for some cultural and artistic professions, under the condition of a minimal income higher than the French average, as well as for private investors. For an up-to-date list of the accredited employers of scientific permit applicants, see MESR (2024).

[3]The most notable modifications occurred in 2006 and 2016. In 2006, as part of a radical revision of the French immigration law - whose political goal was to move from a supposedly "endured immigration" to a "chosen" one ("immigration subie" versus "choisie") - scientific research was included in a broader list of professions exempted from any labour market test; and the scientific permit was included in a broader set of "competences and talent" permits, with a simplification of the release procedure and a duration extension. More recently, in 2016, along with the other permits for highly skilled workers, the scientific permit was re-branded as one mention of a more general "passport for talents" (*passeport talent*). In the meantime, France had transposed the European directive on the Blue Card for highly skilled immigrants, which - since 2012 - has given masters and doctorate graduates an additional opportunity to obtain a residence permit at the end of their studies (OECD, 2017).

2.2 International PhD students in France

Students' immigration has been a defining feature of the French tertiary education system for a long time. In our focal year of interest, 1998, France attracted 11% of all foreign students enrolled in OECD countries, well beyond the United States, but behind the United Kingdom and Germany. In the same year, foreign students accounted for 7% of tertiary education enrolments. Over the following years, France has become less attractive for foreign students, relative to Anglo-Saxon countries (in 2021, it hosted only 4% of foreign students in OECD countries); but the share of foreign students in its tertiary education institutions has nevertheless increased (to over 9%; OECD, 2000, 2022).

Figures for doctoral students are even more remarkable. In 2021, France was still the fourth most important destination country worldwide, with around 24,000 foreign doctoral students, who represented 5% of all foreign PhD students in OECD countries and 37% of the total PhD population in France (Campus France, 2024). This figure goes up to 50% or more for Information and Communication Technologies (ICT) and Engineering.

The regions of origin of foreign students in France, including doctoral ones, reflect both the country's colonial past and its geographical position within Europe. In the 2011-2012 academic year, the MENA–Arabic speaking–and Sub-Saharan Africa regions accounted for, respectively, 40% and 13% of foreign doctoral students in France. Europe accounted for 23% (with around 12% from Italy only), Asia with Oceania for 17% (of which 11% from China), and the Americas for 7% (Campus France, 2017).

As for individual countries of origin, the top ten - with percentages ranging from 14% to 3% - were, in decreasing order, Tunisia, Algeria, Italy, China, Lebanon, Morocco, Brazil, Senegal, Germany and Romania (see Table A.2 in Appendix A.1).

These figures indicate that the scientific residence permit introduced in 1998 did not affect - besides the French doctoral students - about one-quarter of the foreign students, namely those from the EU, half of whom Italians. However, it concerned the majority of foreign students, especially those from Middle East and North African (except, for a while, the Algerians), as well as Sub-Saharan countries for whom France is a traditional destination not only for studying but also for work.

3 Identification strategy

We evaluate the impact of the *Loi Reseda* of 1998 on the stay rates of non-EU doctoral graduates in France, conditional on undertaking a research career and considering the entire population of graduates in STEM disciplines from 1992 to 2002. In order to do so, we both compare the non-EU graduates' stay rates to those of the French and other EU graduates in the same graduation cohorts, the former being the only ones potentially affected by the policy change (in other words, we use the French and other EU graduates as a control group). In addition, we distinguish between graduation cohorts. Most notably, we make the hypothesis that non-EU graduates who started and completed their doctoral studies before 1998 belong to what we call our pre-treatment graduation cohorts, by which we mean that they were not or were only marginally affected by the policy change, the new *carte de sejour "scientifique-chercheur"* not being yet available when they graduated. We assume instead that the the *Loi Reseda* affected the stay decisions of graduates who also had started their studies before 1998, but completed them after the law had come into force. These form our treated graduates cohorts. As for following cohorts, those composed by graduates who started their doctoral studies after 1998 (post-treatment cohorts), we assume that they were also affected by the policy change but with a possible self-selection bias, as their members took their decision to enrol in a French university - differently than their predecessors - with the new legal framework already in place.[4]

Fine-tuning the definition of pre-treatment, treated, and post-treatment cohorts requires some further assumptions and data handling. While we know the graduation years of all STEM doctoral graduates in France, we do not have information on when they enrolled in their PhD programs (see Section 4 below). To estimate it, we use the national statistics for doctoral programs in France, according to which the average length of a STEM PhD is four years, including the year of graduation (Corsini et al., 2022). Therefore, we assume that all the graduates started their PhD

[4]Self-selection refers here to the possibility that the introduction of the *carte de sejour "scientifique-chercheur"* may have made taking a PhD in France an especially attractive option for those non-EU graduates willing to migrate permanently to France, relative to those keener on returning home after completing their studies, or moving elsewhere. Hence, were we observing a higher post-graduation stay rate of the graduates in the post-treatment cohorts, relative to those in the pre-treatment ones, this could depend on not only on the lower costs of staying (strictly the effect of the new policy, to which we are interested) but also on the higher average propensity to stay of the new graduates' cohorts (that is, a change in the mix of doctoral graduates willing or not willing to stay in France, regardless of the immigration conditions).

three years before their graduation and, accordingly, define their doctoral studies period as running from $c-3$ to c included, where c is the graduation year (to which we also refer as "cohort"). As illustrated in the timeline in Figure 1, we identify the treated cohorts as those of non-EU graduates who presumably started their studies graduated between 1994 and 1997 and graduated from 1998 to 2001. graduates who presumably started their studies before 1994 (and graduated before 1998) form our pre-treatment cohorts. The only post-treatment cohort we retain for our analysis consists of the 2002 graduates, as it may contain some graduates who completed their thesis in five years and therefore enrolled in 1998, possibly in the months before the approval of the *Loi Reseda*.

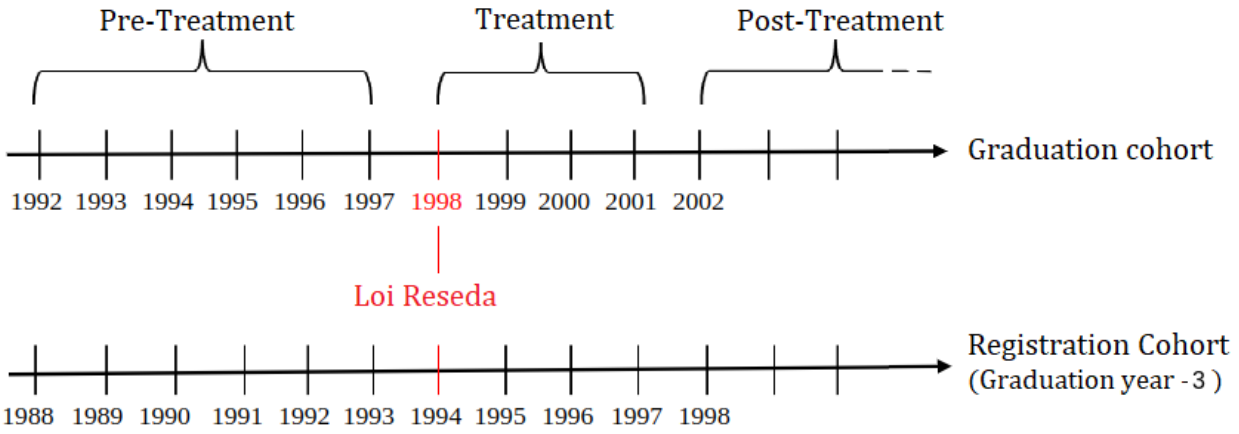


Figure 1: Experiment Timeline

Based on this quasi-experimental setting, we adopt a difference-in-differences estimation strategy with repeated cross-sectional data (or quasi-panel), similar to the one used by Waldinger (2010).[5] The empirical specification is given by Equation 1.

$$\Pr(Y_{i,c+t}=1 \mid X_{i,c}) = \alpha + \gamma_i \cdot Non_EU_i + \sum_{\substack{c \in Cohorts \\ c \neq 1997}} \theta_c \cdot Cohort_c +$$

[5]In the case of Waldinger (2010), the objective of his study consisted in evaluating the impact of the expulsion of Jewish professor from Nazi Germany in 1933 on the doctoral students' careers. Like us, Waldinger distinguishes between students from cohorts not affected by the treatment (who completed their studies before the expulsion) and those affected (who started their studies before the expulsion but completed them afterwards). The control group is given by students from any cohort, who attended universities not affected by expulsion. As explained by the author, while consisting of repeated cross-sections, this model follows the same logic of a dynamic difference-in-differences study, the main difference being the impossibility of introducing fixed effects for the students.

$$+ \sum_{\substack{c \in Cohorts \\ c \neq 1997}} \delta_{i,c} \cdot (Non_EU_i \times Cohort_c) + \phi_i \cdot \mathbf{X}_i + FE_i + \epsilon_{i,c} \quad (1)$$

In the Equation, $Y_{i,c+t}$ represents the outcome of interest for graduate i who graduated in year c, measured t years after c. The variable Non_EU_i is a binary indicator equal to 1 if graduate i comes from outside the EU, and 0 if the graduate comes from France or another EU member country (as per the 1995 membership). $Cohort_c$ is a binary indicator identifying the graduation cohort of calendar year c, for each cohort in our time window. The interaction term $Non_EU_i \times Cohort_c$ is our variable of interest, with its coefficient $\delta_{i,c}$ measuring the impact of the policy reform on non-EU graduates relative to their French or EU peers in each graduation cohort, and with the 1997 cohort (just one year before the *Loi Reseda*, as the reference one.

As for the vector $\mathbf{X}_i$, it includes various control variables for graduates and their supervisors, as follows: $Male_i$, a binary variable taking value 1 for male graduates (and 0 otherwise); $Double_degree_i$, a binary variable taking value 1 for graduates enrolled in double degree programs (and 0 otherwise)[6]; $Published_before_PhD_i$, a binary variable taking value 1 if graduate i has published at least once before the beginning of her PhD program (and 0 otherwise); $Nr_publications_during_PhD_i$, an integer variable which counts the number of articles published by graduate i during her PhD period[7]; and $Foreign_affiliation_during_PhD_i$, a binary variable taking value 1 if the graduate appeared in any such publication with a non-French affiliation (and 0 otherwise). Supervisors are included in the model as fixed effects. In a future version, we will construct productivity variables after linking supervisors to Scopus records.

Moreover, the model includes two sets of individual fixed effects (FE_i), namely: university fixed effects, which account for time-invariant differences (such as variations in institutional quality and available resources, in the decade 1992-2002) across the different institutions delivering the doctoral degrees; and field fixed effects, which capture heterogeneity across academic disciplines, especially for what concerns the graduate's propensity to publish and the career opportunities in France and abroad. Finally, ε is the idiosyncratic error term.

[6]Double degrees are doctoral titles assigned jointly by two universities, under the co-supervision ("cotutelle") of one supervisor from each institution.

[7]To calculate $Nr_publications_during_PhD_i$, we consider the articles published from one year after the beginning of the PhD period (or, symmetrically, two years before the graduation year) to one year after the graduation year, assuming one year of lag in the publication process (Powell, 2016).

To determine whether a doctoral graduate has pursued an academic career in France, we first analyse her post-graduation publication history, which we define as her publication record starting from two years after the graduation year c onwards. In particular, we construct and experiment with two main binary dependent variables (see Figure 2). The first one, *Post_Doc_Stay*, takes value 1 if graduate i's institutional affiliation on her most recent Post Doc–from $c+2$ to $c+4$–publication is a French one, and 0 otherwise. Likewise, *Career_Stay*, is a binary variable that takes value 1 if graduate i affiliation in the last publication is French between five and ten years after graduation, i.e., from $c+5$ to $c+10$, and 0 otherwise. The former is meant to capture a short-term effect, such as the possibility to stay in France for a postdoc. The second refers to a longer-term effects, such as the possibility to start a more permanent career, which - back in 1998 - usually started with the acquisition of a tenured position as assistant professor (*maître de conférences*) in a university or an equally tenured position as junior researchers (*chargé de recherche*) in a public laboratory.

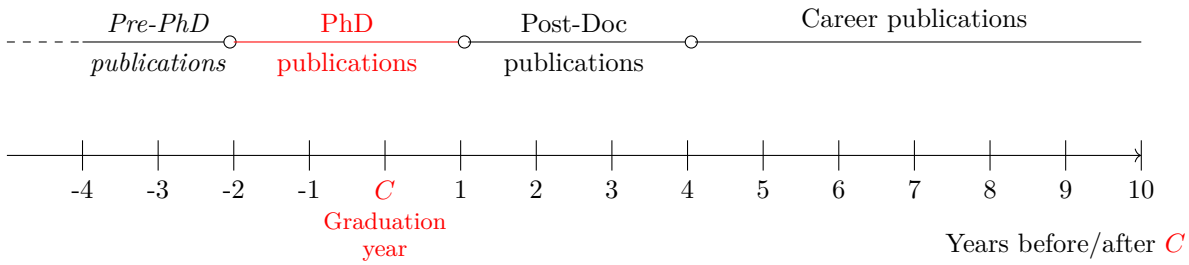


Figure 2: PhD students' publication timeline and classification

We classify publications dated up to three years before graduation year c ($\leq c-3$) as the output of research undertaken by the student before starting the PhD (Pre-PhD, in *italics*) and those from two years before to one year after c ($\geq c-2, \leq c+1$) as produced during the PhD (in red). As for the successive publications, we consider those dated from two to four years after graduation ($\geq c+2, \leq c+4$) as the output of research most likely undertaken as part of a Post-Doc contract, and those undertaken from five to ten years after the graduation year ($\geq c+5, \leq c+10$) as resulting from research undertaken as faculty (we stop observing each individual 10 years after his/her graduation).

Both these dependent variables underestimate the stay rates of the graduates in our sample. First, we do not count as stayers those graduates who ended up working in a French university but never published any article, at least none among those collected in Scopus. To overcome this limitation, in a future version of this paper, we will enlarge our search for publications by matching the graduates of our sample also to authors of publications written in the French language or not published in peer-reviewed journals, as collected by HAL, a dedicated repository managed,

among others, by the French Ministry of Universities and Research (MESRI). We will also explore other bibliographic sources, such as OpenAlex. This should allow us to count as stayers all the graduates who, as post-docs or during their entire academic career, published mostly or only in French or in outlets not covered by Scopus. For what concerns the doctoral graduates with no publication activity at all, but working for a university, these can be tracked only by collecting data on the university and public laboratory staff in the 2000s and 2010s. We will try to obtain them, but it should be stressed that the ultimate objective of the introduction of the *carte de sejour "scientifique-chercheur"* was not to favour the retention of foreign graduates working in university exclusively as teachers, but of those actively engaged in research.

Second, even if not part of our research question, we cannot track graduates who remain in France but not in academia (thus, not publishing any article). However, it is important to recall that the introduction of the *carte de sejour "scientifique-chercheur"* did not aim to increase the stay rate of all foreign doctoral graduates in general, but only of those interested in a research career (in fact, the list of potential employers for the permit applicants did not include any business company, but only universities and public research institutes). Therefore, even if we could measure stay rates outside of academia, they would not directly relate to the policy under study. In particular, we have no reason to expect any change in the *Post_Doc_Stay*, since companies could not use the new *carte de sejour* to recruit any foreign graduates. As for the *Career_Stay*, there is the possibility that - after staying in France with a post-doc contract - some foreign graduates would get a job in a business company (and stop publishing). In a future version of this paper, we will explore this possibility by looking into patent data.

In addition to the two publication-based dependent variables just described, we also experiment with a non-publication one, namely *Supervisor*. This is a binary variable that takes value 1 if, by 2025, graduate i re-appears on *thèses.fr* as a doctoral supervisor of at least one dissertation and 0 otherwise. This variable also underestimates the stay rate of graduates, to the extent that not every researcher in a French university or public laboratory ends up supervising a dissertation, and that those who do not are likely to publish less than those who do. In this respect, *Supervisor* has the same limitations of the publication-based measures and possibly a few more (some graduates may publish early in their career, but give up later and never become supervisors). But it also has the advantage of being entirely based on the disambiguation carried out by ABES, and the unique

ID assigned to all graduates and scholars active in France (while the publication-based measures rely on our name-based matching of graduates and Scopus authors; see section 4 below).[8]

4 Data

Our database results from linking two bibliographic data sources, *theses.fr* (for doctoral graduates and their supervisors) and Elsevier's *Scopus* (for their scientific publications), further integrated with synthetic biographical information on both graduates and supervisors obtained through name analysis.

theses.fr is the official French repository of Electronic Doctoral Theses (EDT), managed by the French Bibliographic Agency for Higher Education (ABES), under the supervision of the French Ministry for Higher Education and Research. It consists of a catalogue of all French doctoral dissertations archived in printed or electronic format since 1985 and of an archive of the electronic ones. These were first obtained by merging a number of pre-existing catalogues maintained by different organizations and later on by making the submission of information on both ongoing and completed (succesfully defended) electronic dissertations compulsory for all universities. As of today, *theses.fr* contains entries for over 450,000 completed dissertations (most of them downloadable) and around 82,00 ongoing ones. The catalogue reports the name and surname of both the doctoral graduates and their supervisors, along with details such as the dissertation title, abstract, field of study, and graduation year.

Concerning the field of study, all dissertations are catalogued by the library of the degree-granting universities according to the Dewey Decimal Classification. We use the classification's 2-digit codes both to distinguish between STEM and non-STEM dissertations and to distribute the former across five broad domains, namely: Engineering, Life Sciences, Mathematics, Medicine and Physics (none of which corresponds, however, to the one-digit domains of the original Dewey classification). Notice that Physics include not only the physical sciences *stricto sensu*, but also all

[8]The French academic ladder consists of only two tenured positions, namely those of *professeur des universités* (full professor) and of *maître de conférences*. These are mirrored, in public laboratories, by those of *directeur des recherches* and *chargé des recherches*. While the former of each pair are entitled by default to supervise doctoral dissertations (and qualify as such *via* a strong publication record), the latter need to pass an examination (*habilitation à diriger des recherches*), also based on their publication record and mostly seen as the first step towards getting a promotion. This *habilitation* is also necessary for professors nominated as such by private universities, whatever their rank.

natural sciences not included in other domains.[9]

For what concerns the degree-granting universities, they come both with the name they had in the graduation year, plus an identifier that allows keeping track of the frequent changes of name due break-ups and mergers that have occurred over the years. In a very limited number of cases, the degree-granting universities are two, one French the other a foreign one, as part of a *Double_degree*.[10]

theses.fr is just one element of an articulated database system that tracks the publishing activity of all researchers active in the French universities and public laboratories (albeit with an uneven coverage, especially in the years of our interest). As part of it, each researcher receives a unique identifier, *IdRef*, including when appearing on a doctoral dissertation as supervisor. From around 2002 all doctoral graduates are associated to an *IdRef* upon graduation; before then, their identifier was created retrospectively, whenever they produced a publication included the ABES system. In both cases, this allows us to track the graduates who eventually become supervisors, by examining all records up to 2025, as per the *Supervisor* variable described in the previous section.[11]

Scopus is a comprehensive bibliographic database published by Elsevier since 2004 (with retrospective coverage), which contains over 97.3 million records on articles published on scientific journals, conference proceedings and other academic publications, with information on their titles, abstracts, authors and their affiliations (as declared by the authors in the title byline). Additional information created by database curators includes a field classification of journals, a citation metrics for both each publication and the journals, and - most importantly for us - a unique ID for each author (a feature still missing in other bibliographic resources).[12]

[9]For the use of Dewey Decimal Classification in *theses.fr*, see https://documentation.abes.fr/aidethesesfr/index.html#ListeFacettesTheses:schId160; last visit: February 2025. In addition, the various doctoral schools provide a classification by discipline (corresponding to the degree they issue), but this varies from one institution to another and does not serve our purposes.

[10]For more information on these university identifiers, see https://documentation.abes.fr/sudoc/regles/CodesUnivEtab.htm. Last visit: February 2025.

[11]IdRef is a French authority data platform that enables querying, creating, and editing standardized authority records for individuals and institutions. It is maintained by the academic documentation networks in France (e.g., SUDOC, Calames, STAR) and is primarily used by authorized professionals in higher education and research institutions. For more information on IdRef see: https://www.idref.fr. For more information on the consistency and history of *theses.fr*, see https://theses.fr/fr/apropos. For both websites, last visit was on February 2025.

[12]On *Scopus* field classification of journals, see https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/ (last visit: February 2025) and, for an evaluation of its accuracy, Thelwall and Pinfield (2024). On *Scopus* citation metrics and author IDs, see https://www.elsevier.com/products/scopus/metrics and https://www.elsevier.com/products/scopus/author-profiles (last visit: February 2025). The presence of a unique ID for authors in *Scopus* is important for us because it greatly simplifies the matching between doctoral graduates and publication authors (see below and the Appendix).

We exploit these two bibliographic sources as follows. First, we extract from *theses.fr* all the dissertations in STEM fields from 1992 to 2002, for a total of 80,903. Since no dissertation is co-authored by multiple students, this is also the baseline number of graduates in our sample. Second, absent any information on the country of birth or nationality of these graduates, we assign each of them a region of origin, based a probabilistic treatment of their names and surnames. We also assign a gender to each of them, again based on their names. We proceed similarly for their supervisors. Third, we name-match the graduates in our sample to the authors of publications in *Scopus*, then validate the true matches by means of a machine learning model based on their graduation year, field of study, degree-granting university and supervisor (see Appendix B). We assume that the graduates for which we cannot find any match did not produce any scientific publications, while we use the publications of the matched ones to calculate the first two dependent variables of our model ($Post_Doc_Stay_{i,c+t}$ and $Career_Stay_{i,c+t}$) and a number of explanatory ones ($Published_before_PhD_i$, $Nr_publications_during_PhD_i$ and $Foreign_affiliation_during_PhD_i$). In a future version of the paper, we will proceed similarly for the supervisors and the variables that concern them.

In what follows, we provide some essential details on the second and third of these operations, including some indications on how we plan to refine them in the immediate future (further details in the Appendix).

4.1 Region of origin

Our identification strategy requires distinguishing between French and/or other EU graduates and non-EU ones. However, due to France's extremely strict implementation of European directives on the protection of sensitive data, it is not possible to obtain information on either nationality or place of birth from social security data or other administrative sources. As for other sources, some information on nationality is contained in graduates' *IdRef* files, but its historical coverage, for the 1990–1999 period, is very limited, with missing information for approximately 30–40% of graduates. Nor it is accurate, as it suffers of selection bias.[13]

[13]This is because the *IdRef* identifiers are primarily assigned to individuals who have published and/or held academic positions in France, so that the available data reflect only those graduates who remain active in academia or research after graduation. As such, we are effectively observing individuals post-outcome. When interviewed, the ABES staff explained that, when the nationality is unknown during data entry, the default classification of nationality is set to "French". As a result, *IdRef* substantially underrepresents the share of foreign doctoral graduates. Based on our estimates, for the roughly 60% of graduates from this period with an *IdRef* record containing nationality data, only 5–7% are identified as foreign. This contrasts sharply with the only official statistic available for that decade,

Due to these difficulties, we resorted to name analysis, as in a number of studies on foreign inventors (survey by Lissoni and Miguelez, 2024) and, more recently, students (Beine et al., 2024). In particular, we adapted to our specific needs an artificial recurrent neural network with long short-term memory (LSTM) first devised by Niggli (2023).[14]

A key step in developing this classification model consists in pre-determining the regions of origin one aims to identify. To guide this choice, we relied on the earliest available official statistics reporting the sending countries of foreign doctoral graduates in France, specifically for the 2011–2012 cohort (Campus France, 2019) and selected the top ones (see column 2 in Table 1). We expanded this initial set of countries by including also Iran and Turkey, which represented a substantial share of foreign students across all levels of higher education in the 1994–1995 cohorts (MESR, 1994).

Table 1: Taxonomy of Regions of origin and countries of origin of foreign PhD (2011-2012) and University (1994-1995) Students

Regions of origin	**Countries**	**Population-Cohort**
Anglo_Saxon	United States	PhD Stud.s 2011-2012
Arabic	Tunisia, Algeria, Lebanon, Morocco	PhD Stud.s 2011-2012
Balto_Slavic	Russia	PhD Stud.s 2011-2012
East_Asian	China, Vietnam	PhD Stud.s 2011-2012
Italian	Italy	PhD Stud.s 2011-2012
Niger_Congolese	Senegal, Ivory Coast, Cameroon, Gabon, Madagascar	PhD Stud.s 2011-2012
Portuguese	Brazil, Portugal	PhD Stud.s 2011-2012
Romanian	Romania	PhD Stud.s 2011-2012
Spanish	Spain	PhD Stud.s 2011-2012
West_North_Germanic	Germany, Belgium	PhD Stud.s 2011-2012
Indo_Iranian	Iran	University Stud.s 1994-1995
Turkic	Turkey	University Stud.s 1994-1995

Our classification model uses the sequential structure of characters in names and surnames to infer the likely region of origin. For instance, character patterns such as the suffix "-yuk" or "-iy" are highly frequent in Ukrainian names. The model thus heavily relies on linguistic features embedded in names to produce a probabilistic assignments of origins. To formalize this approach, we group the selected countries into broader linguistic regions, which have a corresponding geographic denomination. For example, the "Arabic" linguistic group refers to countries in the Middle East and

which reports a 27% share of foreign doctoral graduates in the 1999 cohort (Cohen, 2001).

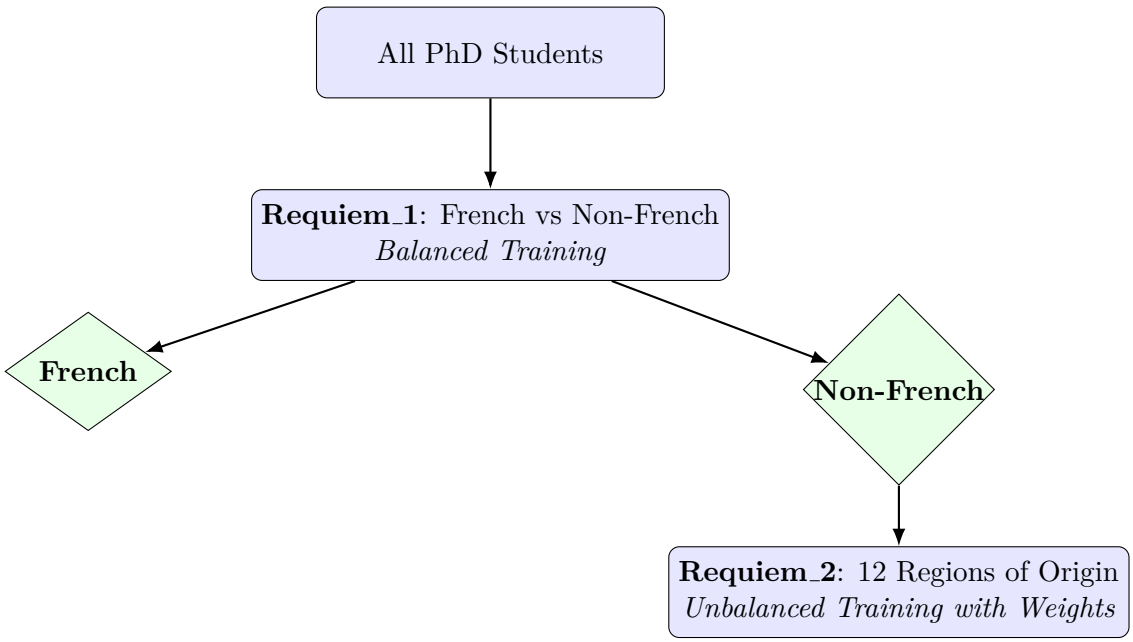
[14]For a summary introduction to LSTM models, see dedicated chapters in Subasi (2020) and Theodoridis (2020). For a more detailed one, see Hochreiter (1997).

North African region; "Niger-Congo" encompasses Sub-Saharan African countries; "Portuguese" covers Portugal and Brazil; and "Spanish" includes Spain and Spanish-speaking countries in Latin America. Column 1 in Table 1 report the regions associated to each of the twenty most common countries of origin for foreign doctoral graduates.

Based on this, we followed a two-stage classification strategy (see Figure 3). Given that French nationals comprise 60–70% of the dataset, a single-step multi-class classifier would suffer from severe class imbalance. Therefore, we opted for first distinguishing between *French* and *non-French* doctoral graduates (**Requiem_1**), using a balanced binary classification; and then, in a second stage (**Requiem_2**), for assigning the *non-French* graduates to one of 12 regions of origin in Table 1. To mitigate the imbalance in this multi-class step, we applied class weighting, with the weights inversely proportional to each class's frequency in the training dataset. This approach is meant to improve the performance on underrepresented groups and yield a classification pipeline aligned with the population's distributional characteristics.

To train the classification models, we set up a novel dataset based on the **Register of Deceased Persons in France** maintained by INSEE (*Fichier des personnes décédées*), which includes names, surnames, and countries of birth. We extract records of individuals born between 1950 and 1990 who died in France between 1990 and 2023, resulting in over 2.5 million observations. After cleaning and applying a name-based nationality classifier to exclude likely French individuals born abroad, the final dataset includes 228,328 foreign-born individuals assigned to one of 12 regions of origin relevant to the population of foreign PhD graduates in France. (For the correspondence between the countries of birth of deceased individuals in France and regions of origin see Table A.1 in Appendix A.) To train the binary classifier (**Requiem_1**), and the multi-class classifier (**Requiem_2**) we followed the workflow presented in Figure 3. This assigns first a probability distribution across classes such that the probabilities sum to one. In our baseline analysis, we apply a threshold of 0.5 for both **Requiem_1** (French vs. non-French) and **Requiem_2** (among non-French classes), as the predicted class probabilities are generally well above this threshold (see Figures A.2 and A.3 in Appendix A.2). Robustness checks using more conservative thresholds are presented in Appendix G.1.

Figure 4 displays the estimated proportion of non-French doctoral graduates produced by **Requiem_1**, in comparison with official data. Unfortunately, the latter are available only for year



Class weights used in Requiem_2:

$$\text{class_weight}_j = \frac{n_{\text{total_samples}}}{n_{\text{samples}_j} \times n_{\text{classes}}}$$

Figure 3: LSTM-Based Two-Stage Name Classification Model

1999, where the share of non-French graduates stood at 27% (Cohen, 2001), and from 2005 onward. Therefore, for the years of our interest, we estimated the non-French share of graduates by using data on the overall share of foreign students enrolled in French universities in all degree levels (Bachelor, Master, and PhD), as detailed in Appendix A.1. When compared to our model's estimate—approximately 33% on average—**Requiem_1** appears to perform reasonably well. However, for 1999 specifically, the model produces a foreign doctoral graduate share of 33.6%, thus overestimating the official figure by roughly seven percentage points.

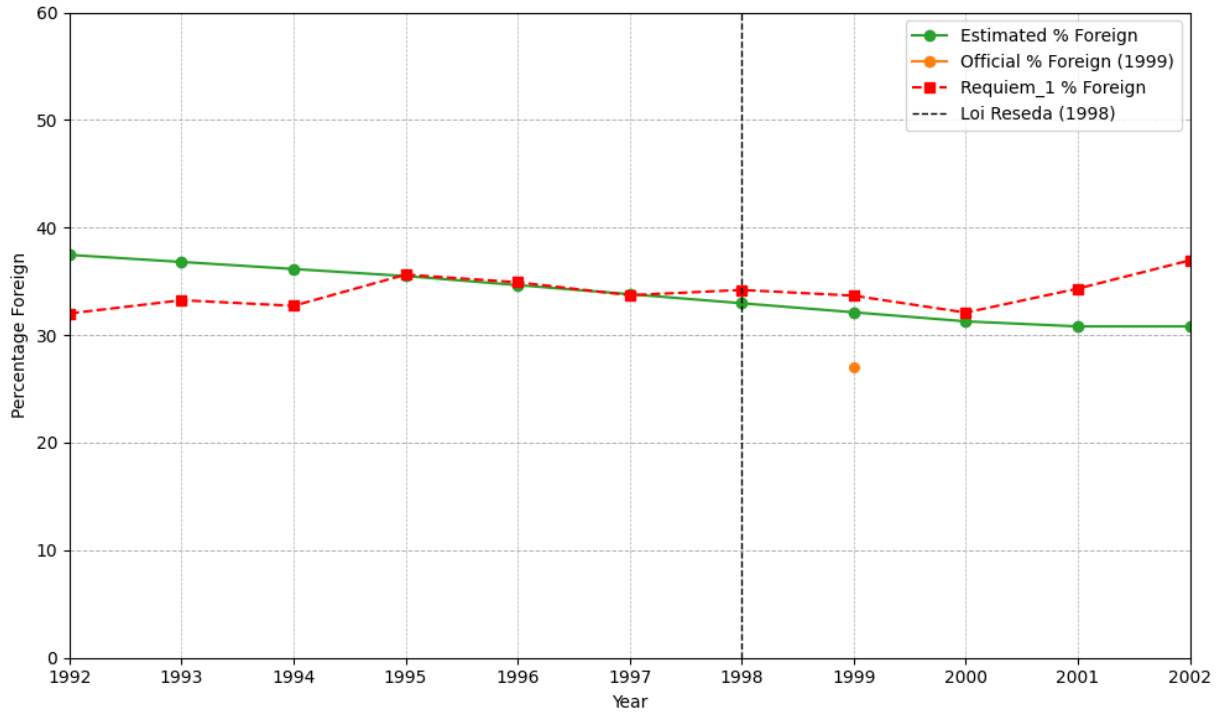


Figure 4: Percentage of foreign doctoral graduates in France from 1992 to 2002

Note: The percentage of foreign doctoral graduates is an estimate based on official statistics (details in Appendix A.1)

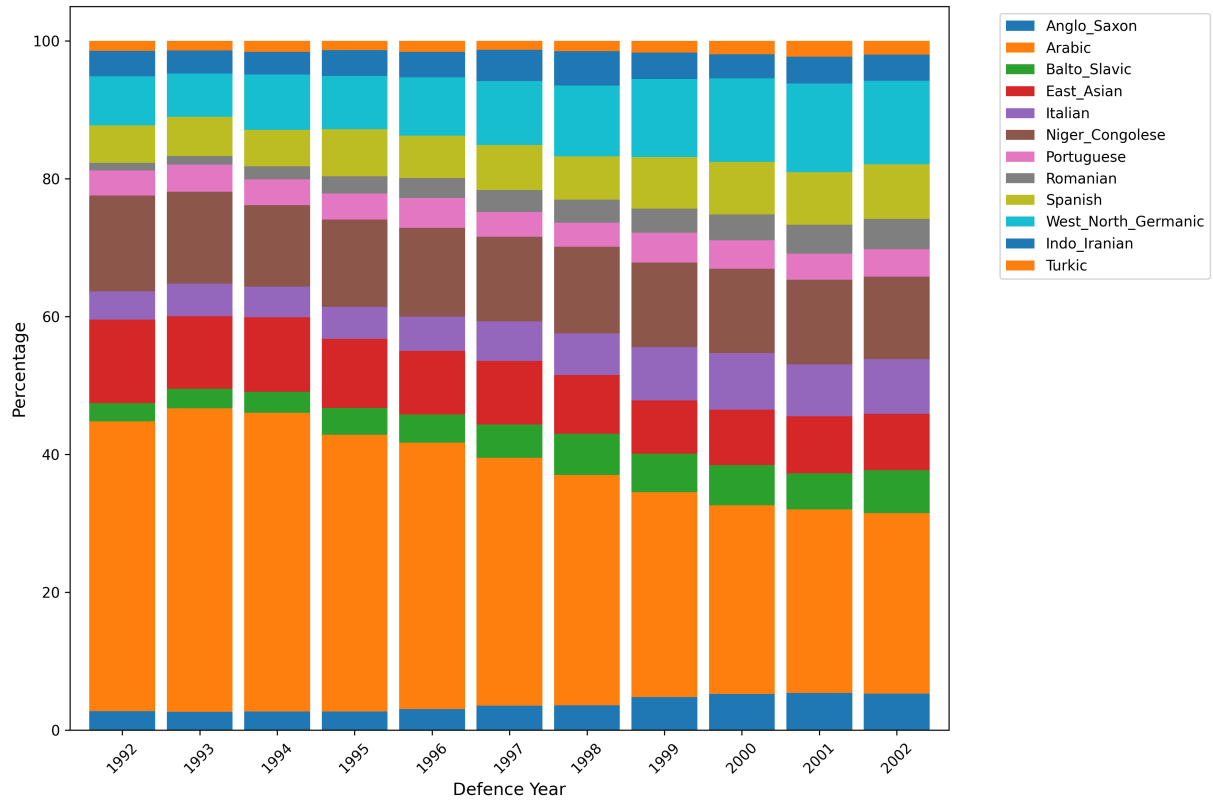


Figure 5: Proportion of regions of origin within foreign Doctoral Graduates in France, 1992-2002

Table 2: Regions of origin and percentage of foreign PhD students in France, 2011–2012

Regions of origin	**% of Foreign PhD Students 2011–2012 (Official)**	**Requiem_2 % 2011–2014**
Anglo_Saxon	2.16	5.96
Arabic	40.02	24.85
Balto_Slavic	2.81	5.93
East_Asian	14.59	13.69
Italian	11.57	7.97
Niger_Congolese	11.64	11.61
Portuguese	5.79	2.67
Romanian	3.13	3.48
Spanish	2.07	6.52
West_North_Germanic	4.97	9.05
Indo_Iranian	0.00	5.89
Turkic	0.00	2.37

Source: Campus France (2017) and results from **Requiem_2**.
Official statistics aggregated by ethnicity from original country-level data.

As for the classification results of the **Requiem_2** step, these are presented in Figure 5. As expected, we find a predominance of graduates from *Arabic*-speaking regions. However, their share declines notably over the decade, from 42.1% in 1992 to 26.2% in 2002. The *Niger_Congolese* region remains relatively stable, fluctuating between 13% and 12%. In contrast, the share of graduates from *West_North_Germanic* regions shows a steady increase, rising from 7.2% to 12.1%. The *East_Asian*, *Italian*, and *Spanish* groups each hover around 7-8% throughout the period, while all remaining regions of origin individually account for 4% or less.

In Table 2 we compare the **Requiem_2**'s predictions to the earliest available official statistics for the 2011-2012 cohort used for our taxonomy (see Table A.2 in Appendix A.1). As the official figures refer to enrolled PhD students while our dataset captures graduates, we compute the average regional composition of doctoral graduates from 2011 to 2014 for comparability. The most notable discrepancy is the underestimation of Arabic-origin graduates, with our model assigning 24.85% versus 40.02% in official statistics. This may partly reflect misclassifications into adjacent categories not present in the official taxonomy, such as *Indo_Iranian* (5.89%) and *Turkic* (2.37%). It may also stem from algorithmic penalization due to class imbalance, as the Arabic group is the largest in the training data. Other deviations include a slight underestimation of *Italian* and *Portuguese* graduates (by approximately 2.5 percentage points each), and overestimation of the

West-North Germanic, *Balto-Slavic*, and *Anglo-Saxon* categories—likely due to naming similarities with French individuals. The *East Asian* group is underestimated by about one percentage point, while estimates for *Niger-Congolese* and *Romanian* graduates closely match the official data. [15]

Lastly, we evaluate the model's performance using location information available on LinkedIn profiles, personal websites and other online sources for approximately 880 doctoral graduates (details in Appendix A.3.2). In this sample, the share of foreign graduates appears to be underestimated at 19%, compared to the official estimate of roughly 27% in 1999. The model achieves very high performance for the *French* class, with precision around 0.99 and recall between 0.90 and 0.80. For *non-French* graduates, precision starts at 0.77 when country of birth is used as the reference, but declines to 0.60 when using high school or bachelor's degree location. Due to the small number of non-French individuals in this validation sample, it is not feasible to compute reliable performance metrics within specific foreign-origin groups, as per the **Requiem_2** step.[16]

4.2 Gender

Compared to the attribution of a country or region of origin, the classification of the doctoral graduates by gender is a rather trivial task, due to the widespread availability of gender-name dictionaries covering multiple countries, as well as of well-tested methodologies. We follow Corsini et al. (2022) and proceed in two steps. First, we match our graduates' first names in the gender dictionary of French names extracted from the "Behind the Name" website and published on the official public data platform of the French government.[17]. Second, we assign a gender to the remaining, non-matched graduates, based on the name-gender dictionary produced by Lax-Martinez et al. (2016) for the classification of inventors worldwide.

[15] Appendix A.3.1 reports a confusion matrix (Provost and Kohavi, 1998), detailing misclassifications between the *French* class and the various regions of origin (Table A.5), as well as among the regions themselves (Figures A.4 and A.5). The majority–42.89%–of misclassified French individuals are assigned to the *Arabic* region of origin, likely reflecting the presence of second-generation immigrants—individuals born in France to Arabic-origin parents. The *Italian*, *Niger_Congolese*, *Spanish*, and *West_North_Germanic* regions each account for approximately 9% of misclassified French individuals, while the remaining regions exhibit very low misclassification rates. Regarding misclassifications among the regions of origin of foreign doctoral graduates, there is intuitive confusion between *Italian* and *Spanish*, *Arabic* and *Niger_Congolese*, as well as between *Anglo_Saxon* and *West_North_Germanic*.

[16] Precision measures the accuracy of positive predictions: Precision $= \frac{TP}{TP+FP}$, where TP (True Positives) are correctly predicted positive cases, and FP (False Positives) are incorrectly predicted positives. Recall assesses the model's ability to identify all actual positives: Recall $= \frac{TP}{TP+FN}$, where FN (False Negatives) are actual positives missed by the model.

[17] See, respectively, https://www.behindthename.com/ and https://www.data.gouv.fr/fr/datasets/liste-de-prenoms/, last visited in February 2025

4.3 Publications

We construct the graduate–publication link through a multi-stage matching procedure between our baseline sample of 80,903 doctoral recipients and the *Scopus* bibliographic database. We proceed in three steps as displayed in Figure 6, combining automated search with algorithmic filtering and supervised validation. The full technical description is provided in Appendix B.

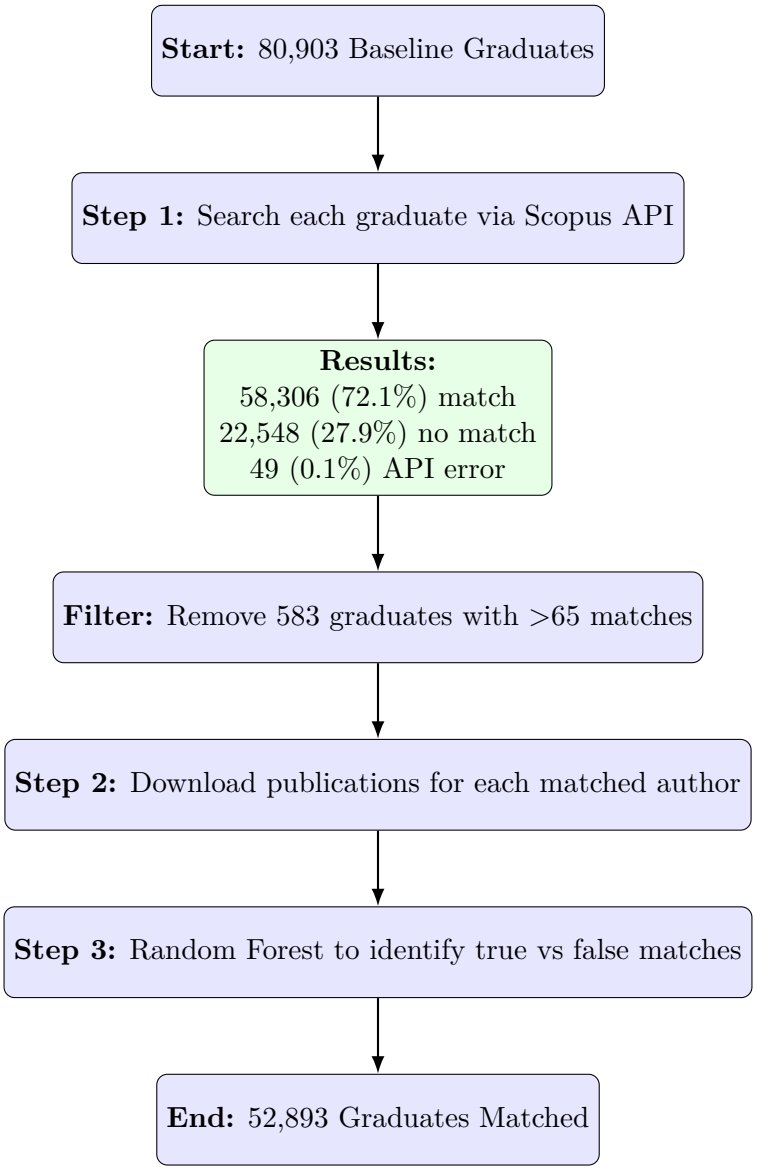


Figure 6: Workflow Matching Graduates to Scopus Authors

In the first step, we submit each graduate's full name (surname and given name) to the Scopus Author Search API. This query returns a set of candidate Scopus Author IDs, which can vary

in number depending on the distinctiveness of the name and the coverage of the database. For most individuals, the API returns either a small number of plausible matches or no match at all. Overall, 72.1% of graduates have at least one potential match, 27.9% have no match, and 0.1% (49 cases) trigger an API error due to extremely ambiguous names generating more than 5,000 possible matches.

In the second step, we apply a computational filter to reduce the burden of processing implausible cases. Specifically, we exclude 583 graduates for whom the API returned more than 65 Author IDs. These cases typically correspond to graduates with very common names (mostly Asian), where reliable identification is infeasible without additional biographical information. After this filtering, we are left with 57,723 graduates and a total of approximately 240,000 candidate Author IDs.

In the third step, we retrieve the full publication records of all candidate Author IDs linked to the remaining graduates. Each record contains detailed metadata, including publication year, journal, co-authors, institutional affiliations, and subject area. This rich information allows us to assess the likelihood that a given Scopus Author ID corresponds to the doctoral graduate in question. We then implement a Random Forest classifier to distinguish between true and false positive matches. The model is trained on manually validated cases and leverages features such as the presence of the graduate's doctoral supervisor among co-authors, consistency between doctoral field and Scopus subject categories, and overlap between the graduate's university and the author's listed affiliations.

The Random Forest classification yields a final set of validated author–graduate links. Out of the 57,723 candidates, 52,893 graduates are assigned at least one validated Scopus Author ID, while 4,703 graduates remain unmatched because all candidate profiles are deemed false positives. For individuals with multiple validated Author IDs, we retain only the highest-scoring profile to avoid double-counting. In Appendix G.2, we adopt two alternative strategies: retaining only the top three scoring matches and excluding all graduates with multiple matches.

The resulting linkage rate is 59% of all doctoral graduates in our baseline sample. While this is somewhat lower than the 68% reported by Corsini et al. (2022) in a comparable study of Italian PhD holders, the difference is largely explained by a strong cohort trend. In our data, the probability of a successful match increases sharply over time: only 40% of graduates in 1992 can be linked to Scopus, compared to 78% in 2002. This evolution reflects both a gradual rise in publication requirements

during doctoral training and the growing adoption of English as the primary language of scientific dissemination.

4.4 Descriptive statistics

Our baseline sample comprises 80,271 individuals who graduated between 1992 and 2002, distributed across regions of origin as shown in Table 3. This figure corresponds to the initial 80,903 graduates, minus the 632 individuals excluded during the publication-matching procedure described in Section 4.3.

For our baseline regression analysis, we define the treated group (non-EU graduates) as consisting of *Niger_Congolese*, *East_Asian*, *Balto_Slavic*, *Indo_Iranian* and *Turkic* graduates, totaling 7,622 individuals. We exclude *Arabic* graduates for two reasons. First, as explained in section 2, Algerians - who likely constitute the majority of the graduates from this region - remained ineligible for the *carte de sejour "scientifique-chercheur"* until 2001. Second, as discussed above, this is the region of origin with the highest degree of confusion with French graduates, which could produce an attenuation bias in our results.

Regions of Origin	**Count**
French	55,007
Arabic	10,531
Niger_Congolese	2,466
West_North_Germanic	2,441
East_Asian	1,836
Spanish	1,564
Italian	1,384
Balto_Slavic	1,125
Indo_Iranian	967
Anglo_Saxon	861
Portuguese	861
Romanian	843
Turkic	385
Total	**80,271**

Table 3: Results Classification Regions of Origin, 1992-2002

As for the control groups, we consider two alternative one. The first, to which we refer as the EU one, consists of 2,441 *West_North_Germanic* and 1,384 *Italian* individuals. We do not consider

Spanish and *Portuguese* graduates because we cannot reliably distinguish those from Europe from those from South America. The second, alternative control group consists of the 55,007 *French* graduates.

Figure 7 reports the distribution of all three groups by region of origin and field of study. Doctoral graduates from non-EU regions are more concentrated in Engineering and, to a lesser extent, Mathematics, relative to their peers from both France and the selected EU regions. Conversely, the latter are relatively more concentrated in Life Sciences. As for the other domains, French graduates are significantly more concentrated in Medicine, while both non-EU and EU graduates in Physics.

Table 4 provides summary statistics for both the dependent and explanatory variables in our baseline regressions. On average, the French and EU doctoral graduates exhibit significantly higher stay rates than the non-EU ones, in both postdoctoral positions (*Post_Doc_Stay*) and longer-term career positions (*Career_Stay*), with statistically significant differences of 10 percentage points, respectively, when compared to non-EU graduates. In turn, for both outcomes, the values for EU graduates are lower than those for the French graduates. The same applies to the share of those who become supervisors (*Supervisor*), albeit with smaller (but still significant differences). Overall, these results reassure us with respect to the our name-based classification model's capacity to distinguish meaningfully between treatment and control groups.

Table 4: Summary Statistics, for selected areas of origin

Variable	**Average**	**Std. Dev.**	**non_EU (1)**	**EU (2)**	**French (3)**	**(1 vs 2)**	**(1 vs 3)**
Post Doc Stay	0.34	0.47	0.24	0.34	0.38	***	***
Career Stay	0.31	0.46	0.20	0.30	0.36	***	***
Supervisor	0.17	0.37	0.12	0.16	0.19	***	***
Female	0.37	0.48	0.38	0.44	0.39	***	
Double Degree	0.01	0.08	0.02	0.01	0.00	*	***
Published before PhD	0.11	0.32	0.10	0.16	0.12	***	***
Nr of Publications during PhD	2.01	3.66	1.66	2.66	2.19	***	***
Foreign Affiliation dur. PhD	0.09	0.28	0.12	0.19	0.08	***	***

Columns (1) to (3) report mean values for the indicated area of origin. Columns (1 vs 2) and (1 vs 3) report significance levels for mean differences between non_EU and EU, and non_EU and French, respectively. (****: 0.01, **: 0.05, *: 0.1*).

As for the explanatory variables, the EU graduates exhibit a higher proportion of women compared to both non-EU and French graduates. Non-EU graduates have the highest enrollment in double degree programs (*Double_Degree*), albeit very small in absolute value. More interestingly,

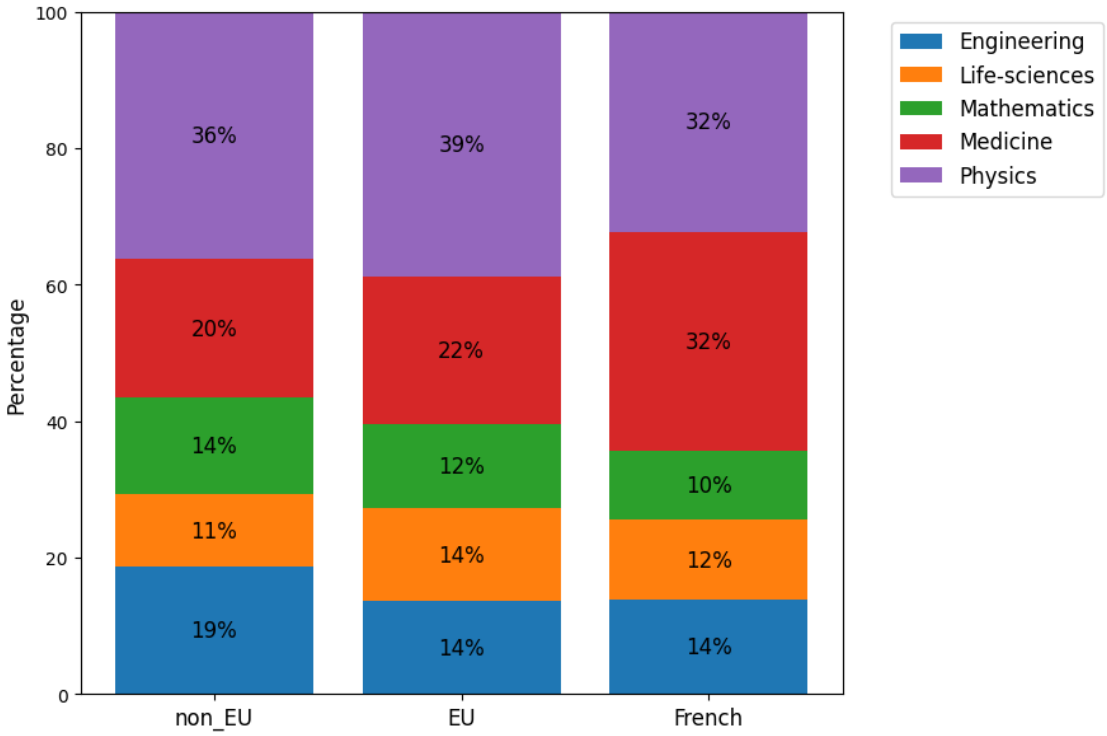


Figure 7: Doctoral graduates from Treated and Control groups by field, 1992 - 2002

both the EU and non-EU graduates are more likely than the French ones to have a foreign affiliation during their PhD (*Foreign_Affiliation_during_PhD*), which again suggests that our name-based classification results are meaningful.

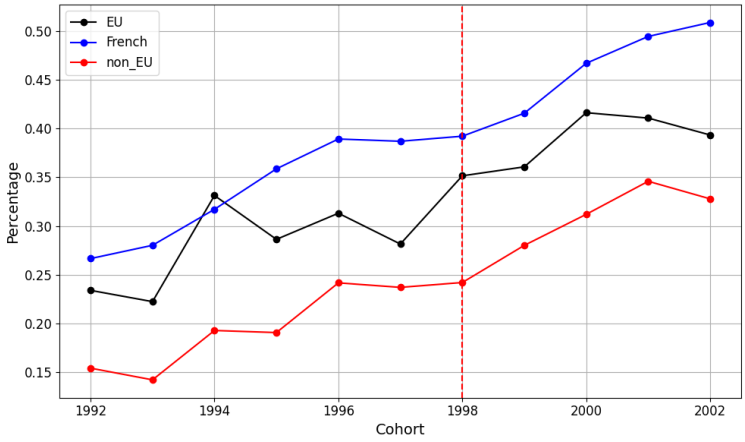
The EU and French graduates are more likely than both the non-EU and French ones to have published prior to the likely start of their PhD (*Published_before_PhD*) and are more productive during their doctoral studies (*Nr_Publications_during_PhD*). However, the gap between EU and French graduates is smaller than that between the latter and the non-EU ones.

Figure 8 illustrates the proportion of post-doctoral and career stayers as well as graduates who become supervisors (share of students with, respectively, *Post_Doc_Stay*, *Career_Stay* and *Supervisor* equal to one) over 1992 to 2002 cohorts, comparing doctoral graduates from non-EU to those from EU and France.

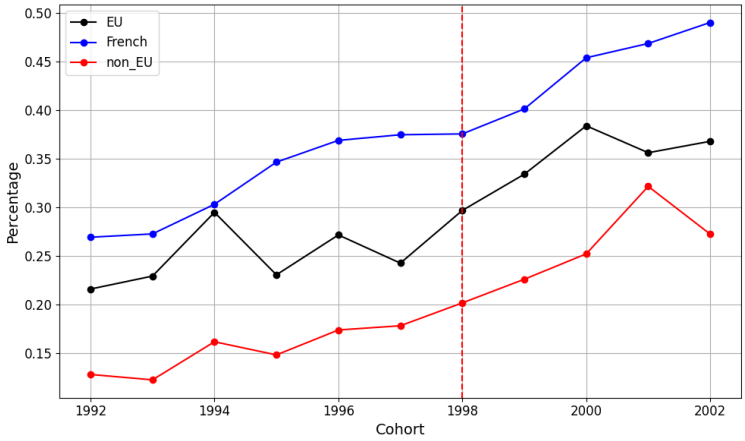
As already shown in Table 2, both EU and French doctoral graduates exhibit a higher stay rate than non-EU for all outcome variables. For the post-doc stayers, no clear convergence is observed (Figures 8a). For the career stayers, instead, the rates of non-EU graduates converge with those of both French and EU graduates, especially in 2001, but diverge again in 2002 (Figure 8b). Lastly,

the proportion of non-EU graduates that will become supervisors converges in 2000 to both French and EU graduates and overtakes the latter for the 2001 cohort (Figure 8c).

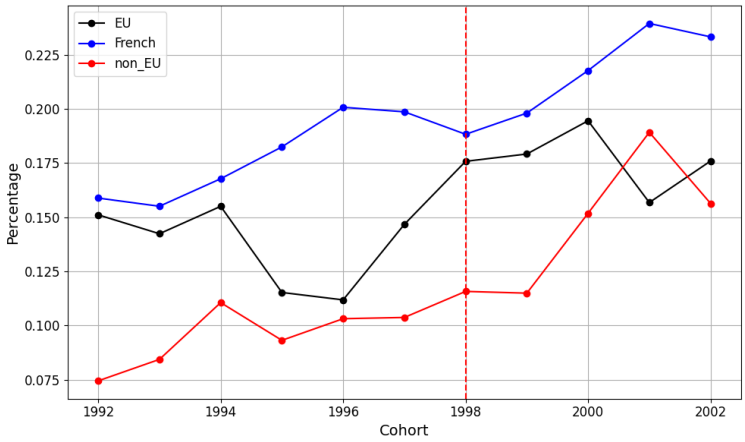
For all indicators, pre-treament trends for non-EU and French graduates are rather parallel. Instead, the lines for EU ones are more irregular, especially for *Supervisor*. This is possibly due to the relatively small sample for this group.



(a) Post Doc Stay



(b) Career Stay



(c) Supervisor

Figure 8: Proportion of Stayers, by EU, non-EU and French status, and graduation cohort (1992–2002)

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$

5 Results

We estimate Equation 1 by means of a linear probability model, with either *Post_Doc_Stay*, *Career_Stay* or *Supervisor* as the binary outcome variables. Our primary focus is on the value of the various δ coefficients in Equation 1, one for each interaction term between cohorts c and the treatment groups ($Non_EU_i \times Cohort_c$). Figure 10 plots the estimated values of such δ coefficients in three different regressions, which differ for the choice of the dependent variable (*Post_Doc_Stay*, *Career_Stay* and *Supervisor*). Table 5 reports all the other estimated coefficients.[18]

For our main analysis, we employ French graduates as the control group, as their outcomes conform to the parallel trends assumption when inspected graphically (see Figure 8). In contrast, the EU control group exhibits erratic pre-trends. As documented in Appendix C, several pre-trend coefficients are significantly different from zero, and the non-EU coefficient, though statistically insignificant, carries a sign opposite to theoretical expectations. Likewise, the coefficient on the male indicator is negative, whereas the existing literature consistently reports a positive association with academic outcomes. These anomalies are most plausibly attributable to the small sample size of the EU control group. However, if we try to add *Spanish* and *Portuguese* to increase size, the shares for the outcome variables for the EU group go significantly down suggesting that graduates from South America outweigh those from Spain and Portugal (see Appendix E). By contrast, the French control group produces smooth pre-trend dynamics, with the principal coefficients statistically significant and aligned with theoretical priors.

Overall, our provisional results suggest that the introduction of the *carte de séjour "scientifique-chercheur"* did not have a significant effect on the stay rates of non-EU doctoral graduates. For the *Post_Doc_Stay* outcome, coefficients are all not significantly different from zero (see Figure 10a).

A positive effect is observed for the long term stay for the 2001 cohort (*Career_Stay*), with an estimated 6 percentage points higher share of non-EU graduates remaining in France. For graduates who become *Supervisors*, the only observable effects also occur for the 2001 cohort: an estimated 5 percentage points increase following the policy. Results remain consistent across alternative specifications of the *Post_Doc_Stay* and *Career_Stay* variables, whether retaining up to the top three matches or excluding graduates with multiple matches entirely (see Appendix G.2).

[18] For the exact value of the interaction terms' coefficients, see Appendix D.

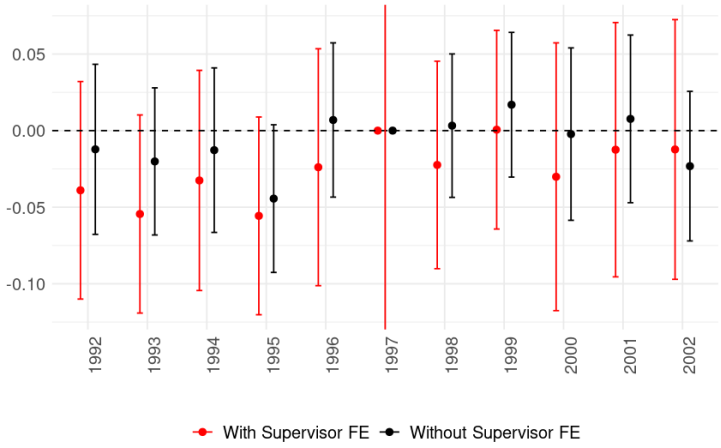
However, these effects are limited to the 2001 cohort and are sensitive to the application of a more stringent threshold for assigning graduates to a region of origin (see Appendix G.1). The *Career_Stay* effect for the 2001 cohort becomes statistically insignificant, while the *Supervisors* effect loses significance at the 5% level but remains significant at the 10% level (see Figure G.3). We do not apply more stringent thresholds in order to avoid further reducing the number of non-EU observations, which would undermine the statistical power of the model. While the 2001 coefficients are not significant the overall trend is similar if we use 1998 as reference year (see Appendix G.3). When conditioning on having at least one publication during the relevant years of the dependent variable—that is, $\geq c+2$ to $\leq c+4$ for *Post_Doc_Stay* and $\geq c+5$ to $\leq c+10$ for *Career_Stay*—thus restricting the sample to individuals who pursue an academic career either abroad or in France, the results for *Post_Doc_Stay* remain unchanged. In contrast, for *Career_Stay* the effect becomes more pronounced and robust, albeit with a less favorable pre-trend (see Appendix G.4).

Results for *Arabic* are displayed in Appendix F. The null results could be attributed both to the confusion between Arabic speaking–treated–and–French–graduates, as well as to the exclusion of Algerian graduates from the visa up to 2001.

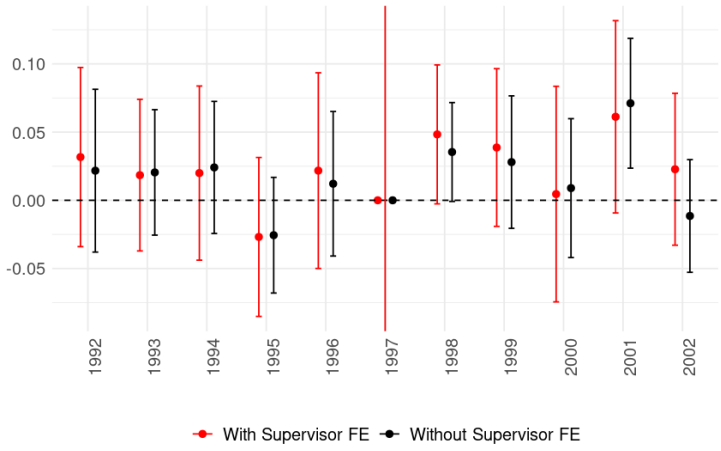
Table 5: Dynamic Difference-in-Differences: non-EU vs French, LPM, With and Without Supervisor FE, 1992-2002

	Post Doc Stay		Career Stay		Supervisor	
non_EU	−0.109***	−0.078**	−0.165***	−0.157***	−0.076***	−0.068***
	(0.019)	(0.026)	(0.018)	(0.023)	(0.016)	(0.019)
Male	0.028***	0.039***	0.066***	0.068***	0.060***	0.055***
	(0.004)	(0.005)	(0.005)	(0.006)	(0.004)	(0.005)
Double_Degree	−0.099**	−0.147**	−0.063	−0.066	−0.036*	−0.046
	(0.033)	(0.053)	(0.050)	(0.079)	(0.018)	(0.043)
Published_before_PhD_1	0.123***	0.107***	0.118***	0.103***	0.034***	0.025**
	(0.011)	(0.013)	(0.011)	(0.014)	(0.006)	(0.009)
Nr_Publications_during_PhD_1	0.042***	0.043***	0.040***	0.040***	0.026***	0.026***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Foreign_Affiliation_during_PhD_1	−0.131***	−0.130***	−0.047***	−0.053***	0.014	0.004
	(0.008)	(0.012)	(0.009)	(0.015)	(0.008)	(0.012)
Cohort FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Field FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Institute FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Supervisor FE	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Observations	62,512	62,512	62,512	62,512	62,512	62,512
R^2	0.176	0.485	0.178	0.485	0.106	0.437

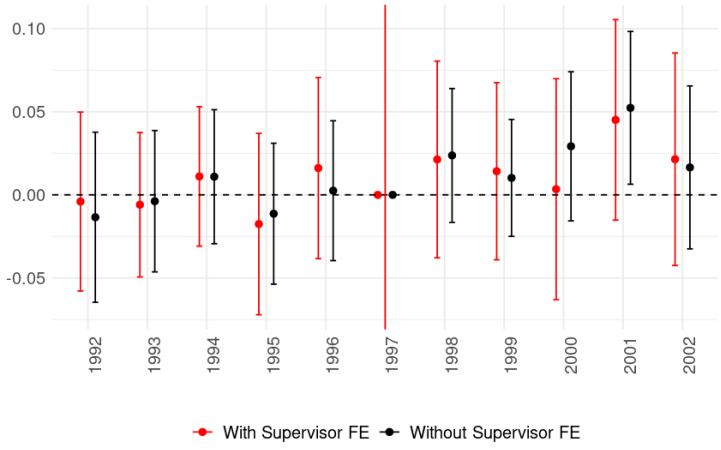
*p<0.1; **p<0.05; ***p<0.01. Standard errors in parentheses.



(a) Post Doc Stay - non-EU vs French



(b) Career Stay - non-EU vs French



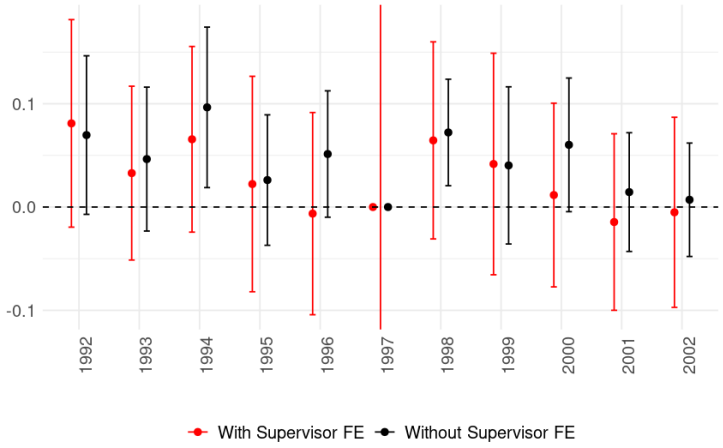
(c) Supervisor - non-EU vs French

Figure 10: Estimated δ Coefficients (95% C.I.) by Cohorts: All outcome variables and treated vs control groups, 1992 – 2002, non-EU vs French

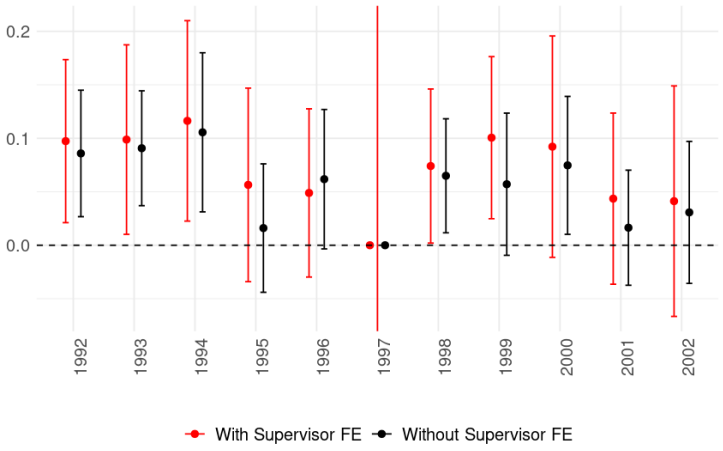
Note - Grey lines: regressions without supervisor FE; Red lines: regressions with supervisor FE

Table 5 reports all the estimated coefficients besides δ, from the same regressions underlying Figure 10. The coefficient for non-EU in Table D.1 indicates that, on average, PhD students from non-EU are less likely to remain in postdoctoral or career positions compared to their French peers, with a smaller likelihood between 11 and 8 percentage points for *Post_Doc_Stay* and around 16 percentage points for *Career_Stay*. Similarly, they are circa 7 percentage points less likely to become supervisors. Gender also plays a significant role, as male PhD students have a higher probability of staying in both post-docs (3 to 4 percentage points more) and career positions (7 percentage points more) as well as supervisor ones (6 percentage points more) than female PhDs, highlighting historical gender disparities in academic careers in France. Holding a double degree is associated with a lower probability of remaining in the French academic sector, probably due to the higher mobility opportunities of these students. As measured by publications before and during the PhD, academic productivity is positively associated with the stay rates. Students who published before starting their PhD are significantly more likely to remain in post-docs and start an academic career in France (around 11 percentage points), suggesting that early academic success is a key predictor of long-term academic engagement (possibly due to greater ability and/or motivation). The effect is smaller (circa 3 percentage points) but still significant for becoming a supervisor. Moreover, as expected, the number of publications during the PhD is also positively associated with the stay rates in French academia, with an increase in the probability of staying by 4 percentage points for every one-unit increase in articles published for post-doc and career stayers and less than 3 for future supervisors. Finally, holding a foreign affiliation during the PhD is negatively associated with a higher probability of staying in both post-doc and career positions, while it has no impact on becoming or not a supervisor.

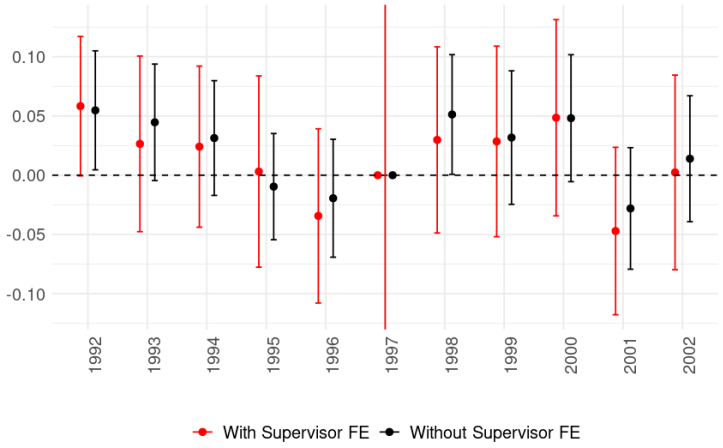
In Figure 12 we report the results of a simple placebo test, performed on sample where we include only the French and EU graduates, with the latter taking the place of the non-EU. We run regressions with *Post_Doc_Stay*, *Career_Stay* and *Supervisor* as dependent variables. We obtain some positive coefficients for all the variables, however all of the variables appear to have a pre-trend, invalidating the results.



(a) Post Doc Stay - EU vs French



(b) Career Stay - EU vs French



(c) Supervisor - EU vs French

Figure 12: Placebo Test; Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated (EU) vs Control (French), 1992 - 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$

6 Conclusions

This paper assesses the impact of France's 1998 Loi Reseda law, which introduced a scientific permit (*carte de sejour "scientifique-chercheur"*), with the specific intent of increasing the retention of foreign doctoral graduates interested in an academic career in France. Despite the importance of the law, which is still a pillar of French selective immigration policy, no systematic policy evaluation effort has been undertaken to date, due to the absence of any reliable data on the inflows and stay rates of either STEM students or scientists (OECD, 2017). Leveraging a novel dataset we have constructed, we assess in this paper the effect of the law on short- and long-term academic stays of foreign PhD students, with a focus on non-EU graduates as the ones affected by the policy.

Our analysis employs an identification strategy based on a dynamic difference-in-differences design. It is underpinned by a rich and unique dataset that integrates multiple data sources, which allows us to consider the entire population of PhD graduates in STEM disciplines in France between 1992 and 2002, proxy their short- and long-term permanence in the French academic system, and identify their region of origin by using a tailored machine-learning model that we developed to overcome data limitations stemming from strict privacy laws.

Our provisional findings suggest that the introduction of the *carte de séjour "scientifique-chercheur"* had no clear overall effect on the stay rates of non-EU doctoral graduates. For *Post_Doc_Stay*, we find no effect.

Some evidence of positive effects emerges for the 2001 cohort, particularly in *Career_Stay* and *Supervisors*, but these results are sensitive to the stringency of the criteria for assigning graduates to origin groups and lose significance under stricter specifications. We refrain from applying more restrictive rules to preserve statistical power. Lastly, the interaction coefficients evaluating the impact of the law are not robust to supervisor fixed effects.

For Arabic-speaking graduates, the absence of significant effects may reflect classification challenges and the exclusion of Algerian graduates from the visa scheme until 2001. Moreover, many foreign doctoral graduates had already accumulated enough years of study in France to meet the five-year residency requirement for citizenship, suggesting that the policy primarily benefited those who were already likely to stay.

The robustness of our findings is, however, subject to certain limitations. Misclassification bias

remains a potential concern, as our tailored machine learning model, is not immune to errors. Misclassification may have weakened the observed effects of the policy by unintentionally including treated individuals in the control group or vice versa.

From an econometric perspective, we aim to better deal with the parallel trend assumption by adopting strategies to relax it. We will also try other model specifications, such as non-linear models, to estimate the effect of the policy change.

References

Beine, M., G. Peri, and M. Raux (2024): "The Contribution of Foreign Master's Students to US Start-Ups," Tech. rep., National Bureau of Economic Research.

Biau, G. and E. Scornet (2016): "A random forest guided tour," *Test*, 25, 197–227.

Campus France (2017): "Les chiffres clés: 2017," Accessed: 2024-10-29.

——— (2019): "Les doctorants à l'international - Tendances de la mobilité doctorale en France et dans le monde," Agence française pour la promotion de l'enseignement supérieur, l'accueil et la mobilité internationale, https://ressources.campusfrance.org/publications/notes/fr/note_60_fr.pdf (last access: July 2024).

——— (2024): "La mobilité étudiante dans le monde: Chiffres clés," Agence française pour la promotion de l'enseignement supérieur, l'accueil et la mobilité internationale, https://ressources.campusfrance.org/noindex/chiffres_cles_2024_fr.pdf (last access: July 2024).

CJC (2010): "Les jeunes chercheurs étrangers en France," https://cjc.jeunes-chercheurs.org/expertise/etrangers/2012-09-sondage-JC-etrangers.pdf, confédération des Jeunes Chercheurs.

Cohen, E. (2001): "Un plan d'action pour améliorer l'accueil des étudiants étrangers en France: Diagnostic et propositions," *Rapport au ministre de l'Éducation nationale et au ministre des Affaires étrangères*, 124, 24.

Corsini, A., J. Koenig, B. Özgün, A. Romanyuk, G. Buenstorf, F. Lissoni, E. Miguelez, M. Pezzoni, and C. Martinez (2025): "Tracking Knowledge Production Among European PhD Graduates: Publication and Exit Patterns Over Time," Manuscript in preparation.

Corsini, A., M. Pezzoni, and F. Visentin (2022): "What makes a productive Ph. D. student?" *Research Policy*, 51, 104561.

Cristelli, G. and F. Lissoni (2020): "Free movement of inventors: open-border policy and innovation in Switzerland," *Available at SSRN 3728867*.

DILA (2024a): "Chronologie : les lois sur l'immigration depuis 1974," Direction de l'information légale et administrative, https://www.vie-publique.fr/eclairage/20162-chronologie-les-lois-sur-limmigration-depuis-1974 (last access: July 2024).

——— (2024b): "Interviews de M. Charles Pasqua, ministre de l'intérieur et de l'aménagement du territoire, dans 'Le Monde' le 2 juin, à RTL le 15 et dans 'Le Parisien' le 18 juin 1993, sur la politique d'immigration et la police," Direction de l'information légale et administrative, https://www.vie-publique.fr/discours/241463-charles-pasqua-02061993-la-politique-dimmigration-et-la-police (last access: July 2024).

d'Albis, H. and E. Boubtane (2021): "L'immigration professionnelle en France depuis 2000," *Annales des Mines-Realités industrielles*, 2 (Mai), 40–43.

EC-EMPL (2000): *Thirty years of free movement of workers in Europe*, European Commission - Directorate for Employmeny, Social Affairs and Inclusion.

Ganguli, I., S. Kahn, and M. J. MacGarvie, eds. (2020): *The Role of Immigrants and Foreign Students in Science, Innovation, and Entrepreneurship*, National Bureau of Economic Research - University of Chicago Press.

GISTI (2001): "Accord franco-algérien : ce qui va changer," Accessed: 2025-08-25.

——— (2020): "Statut des Algériennes et des Algériens en France," Groupe d'information et de soutien des immigrés, https://www.gisti.org/spip.php?article6450 (last access: July 2024).

Hawthorne, L. (2018): *Attracting and retaining international students as skilled migrants*, Oxford University Press.

Herzberg, N. (1998): "Le gouvernement allège les procédures d'attribution de visas pour les scientifiques étrangers," *Le Monde*, June 6.

Hochreiter, S. (1997): "Long Short-term Memory," *Neural Computation MIT-Press*.

Hunt, J. (2011): "Which Immigrants Are the Most Innovative and Entrepreneurial? Distinctions by Entry Visa," *Journal of Labor Economics*, 29, 417–457.

Hunt, J. and M. Gauthier-Loiselle (2010): "How Much Does Immigration Boost Innovation," *American Economic Journal: Macroeconomics*, 2, 31–56.

Kabbanji, L. and S. Toma (2020): "Politiques migratoires et sélectivité des migrations étudiantes en France: une approche sociodémographique," *Migrations Société*, 32, 37–64.

Klaus, S. (2022): "20 Years of EU Directives for Attracting Highly Qualified and Skilled Workers," in *Encyclopedia of Contemporary Constitutionalism*, ed. by J. Cremades and C. Hermida del Llano, Springer.

Lax-Martinez, G., J. D. Raffo, and K. Saito (2016): "Identifying the gender of PCT inventors," *World Intellectual Property Organization (WIPO) Economic Research Working Paper Series*.

Lissoni, F. and E. Miguelez (2024): "Migration and innovation: Learning from patent and inventor data," *Journal of Economic Perspectives*, 38, 27–54.

Math, A., S. Slama, A. Spire, and M. Viprey (2006): "La fabrique d'une immigration choisie. De la carte d'étudiant au statut de travailleur étranger (Lille et Bobigny, 2001-2004)," *La Revue de l'IRES*, 50, 27–62.

MESR (1994): "Repères et références statistiques sur les enseignements, la formation et la recherche," https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/53925/reperes-et-references-statistiques-sur-les-enseignements-et-la-formation-edition-1994-mini

——— (2005): "La mise en place du L.M.D. (licence-master-doctorat)," https://www.education.gouv.fr/la-mise-en-place-du-lmd-licence-master-doctorat-41234.

——— (2020): "Repères et références statistiques sur les enseignements, la formation et la recherche (Éditions 2005–2020)," https://cpesr.fr/wp-content/uploads/2020/09/, includes data from 2005 to 2020.

MESR (2024): "Accueil en France des scientifiques étrangers," Ministère de l'enseignement supèrieur et de la recherche, https://www.enseignementsup-recherche.gouv.fr/fr/accueil-en-france-des-scientifiques-etrangers-46403 (last access: July 2024).

Niggli, M. (2023): "'Moving On'—investigating inventors' ethnic origins using supervised learning," *Journal of Economic Geography*, 23, 921–947.

OECD (2000): *Education at a Glance. OECD Indicators*, Organisation for Economic Co-operation and Development.

——— (2017): *Le recrutement des travailleurs immigrés : France*, Organisation for Economic Co-operation and Development.

——— (2022): *Education at a Glance. OECD Indicators*, Organisation for Economic Co-operation and Development.

Provost, F. and R. Kohavi (1998): "On applied research in machine learning," *MACHINE LEARNING-BOSTON-*, 30, 127–132.

Roach, M. and J. Skrentny (2019): "Why foreign STEM PhDs are unlikely to work for US technology startups," *Proceedings of the National Academy of Sciences*, 116, 16805–16810.

——— (2021): "Rethinking immigration policies for STEM doctorates," *Science*, 371, 350–352.

Rodier, C. (2001): "Les grandes étapes de la construction de l'≪espace européen≫ de Rome à Amsterdam en passant par Schengen," *Plein droit*, 49, 36–41.

Rogers, N., R. Scannell, and J. Walsh (2012): *Free movement of persons in the enlarged European Union*, Sweet & Maxwell.

Romanyuk, A. and F. Lissoni (2025): "Stay rates and careers of foreign doctorate recipients from Dutch universities: A bibliometric analysis," Mimeo, Bordeaux School of Economics.

Rose, M. E. and J. R. Kitchin (2019): "pybliometrics: Scriptable bibliometrics using a Python interface to Scopus," *SoftwareX*, 10, 100263.

Slama, S. (2001): "Tapis rouge pour les élites," *Plein droit*, 47-48, 36–39.

Stuen, E. T., A. M. Mobarak, and K. E. Maskus (2012): "Skilled immigration and innovation: evidence from enrolment fluctuations in US doctoral programmes," *The Economic Journal*, 122, 1143–1176.

Subasi, A. (2020): *Practical machine learning for data analysis using python*, Academic Press.

Thelwall, M. and S. Pinfield (2024): "The accuracy of field classifications for journals in Scopus," *Scientometrics*, 129, 1097–1117.

Theodoridis, S. (2020): *Machine learning: a Bayesian and optimization perspective*, Academic press.

UIS (2000): "Unesco Institute for Statistics," United Nations Educational, Scientific and Cultural Organization,(https://data.uis.unesco.org/; last access: July 25, 2024).

UNESCO (2016): *UNESCO Science Report: Towards 2030*, United Nations Educational, Scientific and Cultural Organization.

Waldinger, F. (2010): "Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany," *Journal of political economy*, 118, 787–831.

Weil, P. (1997): "Mission d'étude des législations de la nationalité et de l'immigration : rapports au Premier ministre," Direction de l'information légale et administrative, https://www.vie-publique.fr/rapport/25205-mission-detude-des-legislations-de-la-nationalite-et-de-limmigration (last access: July 2024).

Appendices

A Region of origin assignment

Table A.1: Regions of origin and countries of birth of deceased individuals in France, born between 1950 and 1990

Regions of origin	**Country of Birth of Foreign Deceased Individuals in France**
Anglo_Saxon	United States, Australia, Ireland, Comoros (UK)
Arabic	Algeria, Morocco, Tunisia, Egypt, Lebanon, Syria, Mauritania, Iraq, Libya, Saudi Arabia, Jordan, Yemen, United Arab Emirates, Qatar, Palestinian Territories, Bahrain, Oman, Sudan
Balto_Slavic	Latvia, Lithuania, Estonia, Poland, Czech Republic, Slovakia, Russia, Ukraine, Belarus, Serbia, Bulgaria, Bosnia and Herzegovina, Croatia, Montenegro, North Macedonia, Slovenia
East_Asian	China, Vietnam, Cambodia, Laos, Thailand, Myanmar (Burma), Japan, South Korea, North Korea, Taiwan, Malaysia, Indonesia
Italian	Italy
Niger_Congolese	Senegal, Comoros, Democratic Republic of the Congo, Mali, Madagascar, Côte d"Ivoire, Cameroon, Republic of the Congo, Mauritius, Cape Verde, Guinea, Angola, Togo, Ghana, Benin, Central African Republic, Guinea-Bissau, Nigeria, Gabon, Burkina Faso, South Africa, Rwanda, Gambia, Burundi, Mozambique, Kenya, Liberia, Sierra Leone, Equatorial Guinea, Zimbabwe, Tanzania, Uganda, Niger, Malawi, Zambia, Namibia, Lesotho, Eswatini
Portuguese	Portugal, Brazil
Romanian	Romania
Spanish	Spain, Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela
West_North_Germanic	Germany, Switzerland, Austria, Luxemburg, Denmark, Sweden, Iceland, Aruba (Dutch)
Indo_Iranian	India, Iran, Pakistan, Afghanistan, Bangladesh, Tajikistan
Turkic	Turkey, Azerbaijan, Kazakhstan, Uzbekistan, Turkmenistan, Kyrgyzstan

A.1 Official Statistics on Foreign PhDs in France

While in our data we observe PhD students at their graduation, official statistics are available for all the PhD students per cohort. The first cohort-information available on the percentage of foreign PhD students in France is 2004-2005 cohort, as shown in Figure A.1. For that cohort, 33% of PhD

students were foreign, a share that increased steadily to 41% by 2010 (MESR, 2020)[19]. To estimate the proportion of foreign PhD students for earlier years, we assume a correlation between the share of foreign students in French universities overall—available since the 1985-1986 cohort—and that of foreign PhD students. This assumption is supported by a survey conducted by the *Confédération des Jeunes Chercheurs*[20], which reports that approximately 75% of foreign PhD students in France obtained their master's degree in France, and 24% completed their bachelor's degree there as well (CJC, 2010).

A potential caveat to this inference is the change in the higher education system following the Bologna Process. Prior to its implementation, the structure was more complex[21], though it still comprised a progression through undergraduate (*Licence*) and graduate (*Maîtrise*) degrees. The LMD reform (Licence–Master–Doctorat) merely aligned the French system with international standards (MESR, 2005).

Naturally, there is a lag between the proportion of foreign students in higher education generally and their eventual representation at the PhD level. Assuming a typical five-year trajectory from bachelor's to master's degree, we introduce a temporal shift to align the two trends. The optimal lag, which maximizes the correlation between the two curves (correlation coefficient = 0.84), is four years. This is plausible given that most foreign students entering PhD programs in France had previously enrolled in master's, rather than bachelor's, programs. Based on this lag, we estimate the share of foreign PhD students to range between 30% and 38% over the 1990-2004 period. While these estimates may not be exact, it is reasonable to infer that the proportion did not fall below 30% during this time.

[19]The various census surveys conducted in France identify foreign students based on self-declared nationality. This includes students who have completed their secondary education within the French school system (MESR, 2020)

[20]For more information, see: https://cjc.jeunes-chercheurs.org/, last accessed 10 April 2025.

[21]The first two years of undergraduate studies granted the *DEUG* diploma, followed by the *Licence* in the third year. The master's level consisted of the *Maîtrise* in the first year and either the *DEA* or *DESS* in the second year. The doctorate followed thereafter (MESR, 2005).

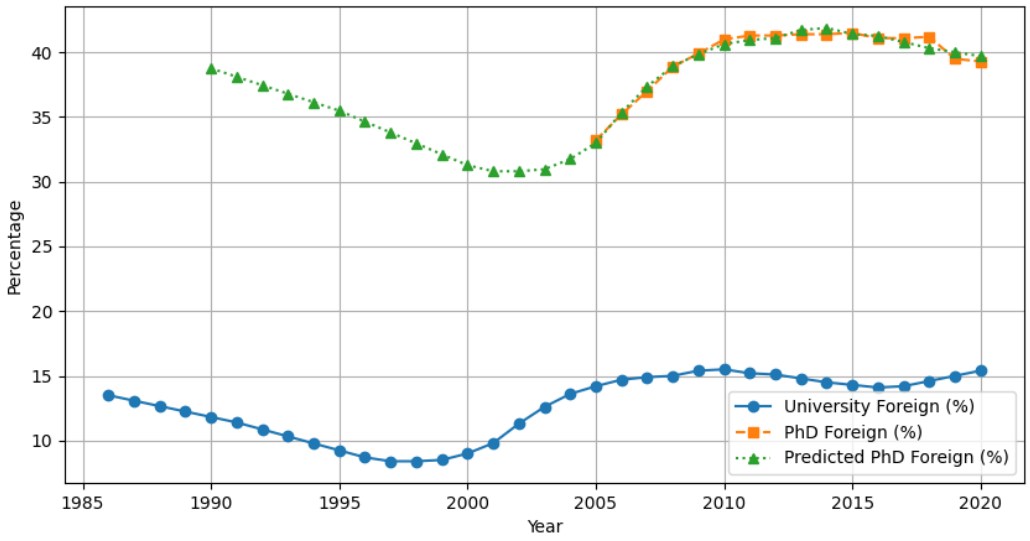


Figure A.1: Proportion foreign PhD students in France 1990-2020

Source: MESR (2020).
Note: Proportion foreign PhD students from 1990 to 2005 is inferred.

As for the information on the sending countries of foreign doctoral graduates in France the first official statistics available are for the 2011-2012 cohort, as displayed in Table A.2.

Table A.2: Campus France statistics for Cohort 2011-2012

Rank	Country of Origin	PhD students 2011-2012	% of foreign PhD students	2015-2016 Rank	Regions of origin
1	Tunisia	2275	14.41	2	Arabic
2	Algeria	1934	12.25	4	Arabic
3	Italy	1826	11.57	3	Italian
4	China	1815	11.50	1	East-Asian
5	Lebanon	1054	6.68	5	Arabic
6	Morocco	1054	6.68	6	Arabic
7	Brazil	743	4.71	7	Portuguese
8	Senegal	673	4.26	10	Niger-Congolese
9	Germany	553	3.50	11	West-North Germanic
10	Romania	494	3.13	14	Romanian
11	Vietnam	487	3.09	8	East-Asian
12	Ivory Coast	451	2.86	16	Niger-Congolese
13	Russia	444	2.81	12	Balto-Slavic
14	Cameroon	387	2.45	13	Niger-Congolese
15	United States	341	2.16	18	Anglo-Saxon
16	Spain	327	2.07	9	Spanish
17	Gabon	326	2.07	15	Niger-Congolese
18	Belgium	232	1.47	17	West-North Germanic
19	Madagascar	198	1.25	20	Niger-Congolese
20	Portugal	170	1.08	19	Portuguese

Source: Campus France (2017). Adjusted by the author to reflect 2011-2012 student composition.

A.2 Threshold Analysis

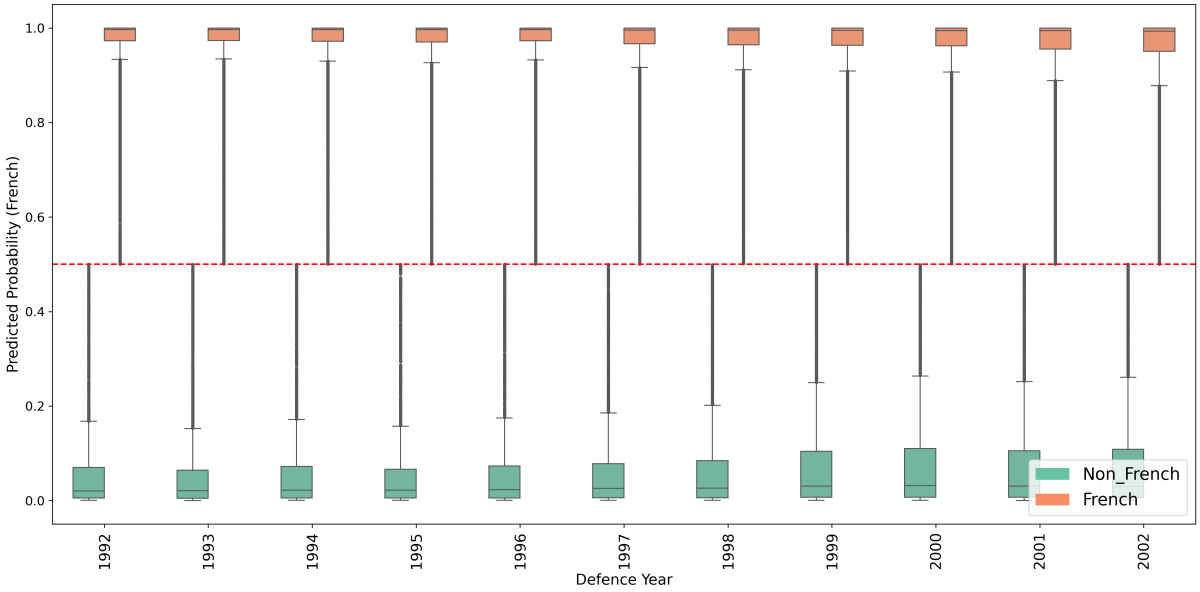


Figure A.2: Distribution of the Highest Predicted Class Probability Across French and non-French, 1992-2002

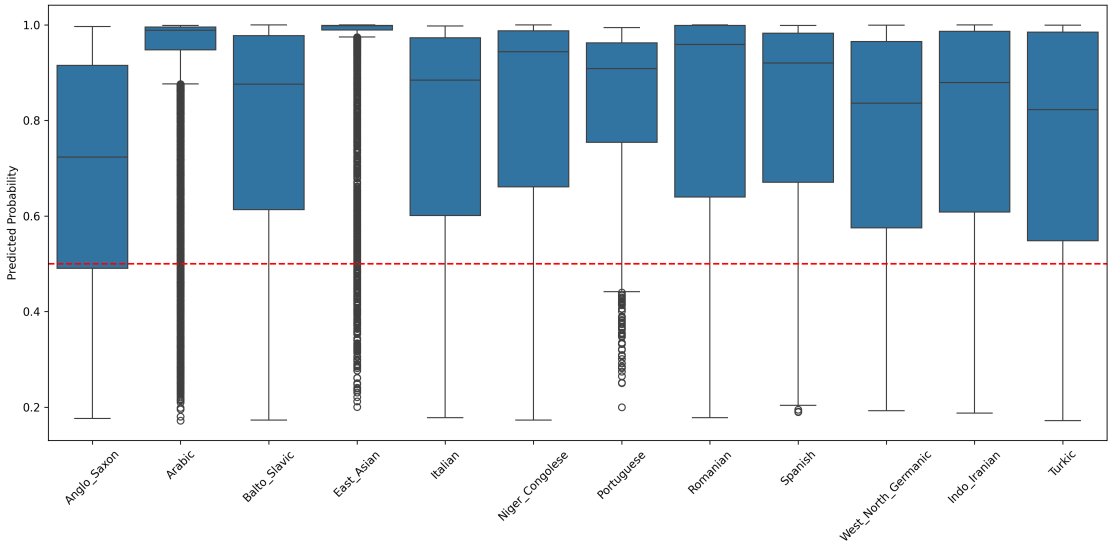


Figure A.3: Distribution of the Highest Predicted Class Probability Across Regions of Origin, 1992-2002

A.3 Model's performance

A.3.1 In-Sample Performance

The classification report for **Requiem_1** in Table A.3 demonstrates excellent model performance in distinguishing between *French* and *non_French* classes. Both classes exhibit high precision (all at or above 0.95), indicating that the model is both accurate and consistent across the two categories. Notably, the precision for the *French* class (0.968) is slightly higher than for the *non_French* class (0.958), whereas the *non_French* class has a marginally higher recall (0.957 vs. 0.968). This balance results in an almost identical F1-score of 0.963 and 0.962 for the two classes. The overall accuracy, macro average, and weighted average scores all stand at 0.963, underscoring the robustness and fairness of the model across a balanced dataset.[22]

Table A.3: Classification Report - **Requiem_1**

Class	Precision	Recall	F1-Score	Support
non_French	0.958	0.968	0.963	22733
French	0.968	0.957	0.962	22733
Accuracy			0.963	45466
Macro Avg	0.963	0.963	0.963	45466
Weighted Avg	0.963	0.963	0.963	45466

Table A.4 presents the classification performance of **Requiem_2** across the twelve regions of origin. The overall accuracy of the model is high at 0.898, with a strong weighted F1-score of 0.900, indicating solid general performance across a diverse set of classes. The macro average F1-score of 0.866 suggests that the model maintains relatively balanced performance even across less represented classes.

Arabic, *Portuguese*, and *East_Asian* achieve the highest F1-scores (above 0.92), indicating highly reliable classification for these classes. In contrast, performance is weakest for the *Indo_Iranian*, with a notably lower F1-score of 0.692, suggesting difficulty in distinguishing this class—likely due to a combination of lower support and higher class overlap. Similarly, *Anglo-Saxon* and

[22] Model performance is evaluated using standard classification metrics. Precision measures the accuracy of positive predictions: Precision $= \frac{\text{TP}}{\text{TP+FP}}$, where TP (True Positives) are correctly predicted positive cases, and FP (False Positives) are incorrectly predicted positives. Recall assesses the model's ability to identify all actual positives: Recall $= \frac{\text{TP}}{\text{TP+FN}}$, where FN (False Negatives) are actual positives missed by the model. F1-score is the harmonic mean of Precision and Recall: F1 $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$, balancing both concerns. Accuracy captures overall correctness: Accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, with TN (True Negatives) being correctly predicted negatives.

West_North_Germanic classes have comparatively modest scores, with F1-scores of 0.844 and 0.802 respectively.

The model performs particularly well on classes with larger support, such as *Arabic* (8316) and *Portuguese* (3072), suggesting that the amount of training data positively influences classification performance. The *Turkic* and *Italian* classes, despite having fewer examples, still achieve F1-scores close to or above 0.85, highlighting good model generalization.

Table A.4: Classification Report - **Requiem_2**

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Anglo_Saxon	0.792	0.903	0.844	824
Arabic	0.969	0.907	0.937	8316
Balto_Slavic	0.943	0.903	0.923	1627
East_Asian	0.932	0.919	0.925	1045
Italian	0.777	0.960	0.859	1005
Niger_Congolese	0.867	0.819	0.842	2988
Portuguese	0.936	0.955	0.945	3072
Romanian	0.853	0.874	0.863	318
Spanish	0.857	0.859	0.858	1309
West_North_Germanic	0.765	0.843	0.802	1064
Indo_Iranian	0.596	0.824	0.692	369
Turkic	0.844	0.960	0.898	796
Accuracy			0.898	22733
Macro Avg	0.844	0.894	0.866	22733
Weighted Avg	0.905	0.898	0.900	22733

Table A.5 presents a breakdown of *False Positives*—*non_French* individuals classified as *French*—and *False Negatives*—*French* individuals classified as *non_French*—based on predictions from **Requiem_1**. The table displays the misclassification patterns with respect to the regions of origin defined in **Requiem_2**.

For the *False Positives*, the true country of birth is known, allowing us to identify, for each region of origin, the proportion of individuals incorrectly classified as French. The highest proportion is found among individuals classified as *West_North_Germanic* (9.82%), likely driven by individuals from Belgium. This is followed by *Anglo_Saxon* individuals and, notably, by regions of origin associated with countries with a strong history of migration to France—first from Africa and the Middle East (e.g., *Niger_Congolese*, *Arabic*, *Indo_Iranian*), and then from Europe (e.g., *Spanish*, *Italian*).

In the case of *False Negatives*, since French individuals do not have an assigned region of origin by default, we infer their region of origin by reclassifying them using **Requiem_2**.[23] I then compute, for each region of origin, its proportion over the total number of *False Negatives*. The majority (42.89%) are classified as *Arabic*, which likely reflects the presence of second-generation immigrants—individuals born in France but with Arabic parents. Similar reasoning may apply to individuals classified as *Italian* (11.34%), *Niger_Congolese* (9.48%), and *Spanish* (9.48%). Finally, the *West_North_Germanic* class (8.45%) again likely includes individuals from Belgium.

Table A.5: Breakdown of False Positives and False Negatives by Regions of Origin for **Requiem_1**

Regions of origin	**Region X Misclassified as French**	**French Misclassified as Region X**
Anglo_Saxon	5.68	4.23
Arabic	3.98	42.89
Balto_Slavic	0.84	4.12
East_Asian	2.17	2.16
Italian	1.74	11.34
Niger_Congolese	4.39	9.69
Portuguese	0.47	3.40
Romanian	0.34	1.65
Spanish	2.20	9.48
West_North_Germanic	9.82	8.45
Indo_Iranian	2.98	1.96
Turkic	0.37	0.62
Formula	$\frac{\text{False Positives for Region X}}{\text{Total Region X}} \times \%$	$\frac{\text{False Negatives classified as Region X}}{\text{Total False Negatives}} \times \%$

Note: False Positives are non-French individuals classified as French. False Negatives are French individuals classified as non-French. To know where they would be confounded in the second stage, I classify them with **Requiem_2** (selecting the class with the highest value for each individual).

Another way to interpret the Classification Report (Table A.4) is through the visualization of the corresponding confusion matrix, a table that summarizes a classification model's performance by showing the number of correct and incorrect predictions for each class. This representation is particularly useful for understanding how misclassified individuals are distributed across classes. Figure A.4 presents a row-normalized confusion matrix, effectively illustrating the *recall* for each class as reported in the classification report. Diagonal values correspond to the recall scores for each class, while off-diagonal entries indicate the proportion of instances from a given class that were misclassified into others. Overall, recall values are high across all classes, never falling below

[23] Each individual is assigned to the region of origin with the highest probability output by **Requiem_2**.

0.82. Notable patterns include modest confusion between *Anglo_Saxon* and *West_North_Germanic* (around 0.03–0.04), as well as misclassification of *Niger_Congolese* individuals primarily as *Portuguese* and *Arabic* (both around 0.04). Similarly, *Spanish* individuals are often confused with *Portuguese* and *Italian* (approximately 0.04), while *Indo_Iranian* individuals are frequently misclassified as *Arabic* (0.06) or *Niger_Congolese* (0.04).

Figure A.5 provides a complementary view by normalizing the confusion matrix column-wise, thereby representing the *precision* for each class. The lowest precision is observed for the *Indo_Iranian* class, with 10% of its true instances misclassified as *Niger_Congolese* and 19% as *Arabic*. The *Arabic* class, in particular, appears to attract the highest number of misclassifications from other classes. This may be attributed to its disproportionately large size during training (see Table **??**), potentially biasing the classifier toward this region of origin.

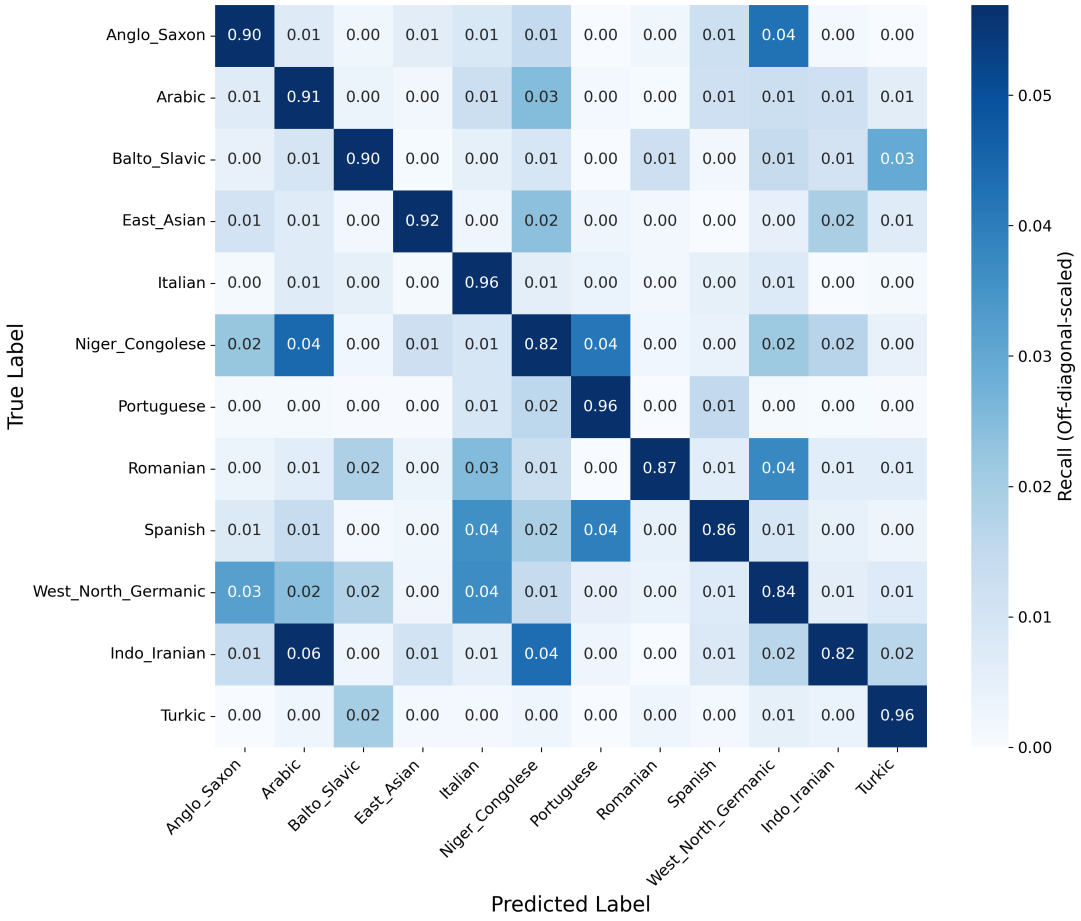


Figure A.4: Requiem_2 - Confusion Matrix - Recall

Note: The Confusion Matrix is normalized row-wise in order to represent Recall. The heatmap is scaled on the off-diagonal values.

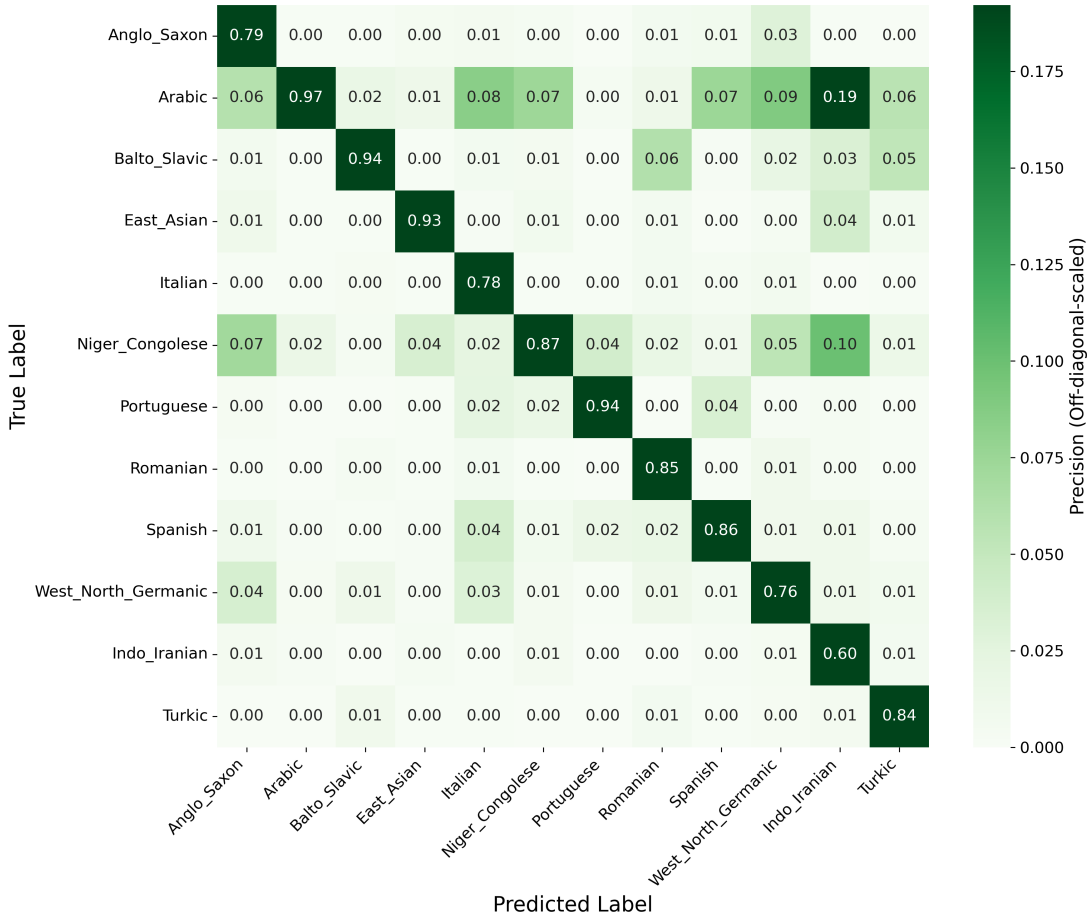


Figure A.5: Requiem_2 - Confusion Matrix - Precision

Note: The Confusion Matrix is normalized column-wise in order to represent Precision. The heatmap is scaled on the off-diagonal values.

A.3.2 Out-of-Sample Performance

With the help of research assistants, we constructed a hand-curated sample of doctoral graduates from the 1992–2002 cohorts, retrieving information on their geographic background using multiple sources. For each cohort, 80 individuals were selected, resulting in an initial sample of around 880 graduates. Over half of the information was obtained through LinkedIn profiles, while additional sources included personal and university websites, CVs, blogs, and online journals.

The retrieved location indicators include country of birth, and the location of high school, bachelor's degree, and master's degree. Each doctoral graduate in the sample has at least one of these data points available. We exclude the variable *native language*, as all 17 individuals in that

category report French, offering no variation for classification. Additionally, we exclude the 2002 cohort due to its unusually low share of non-French graduates (5%), which contrasts sharply with the roughly 20% observed in other years. According to the only official statistic available for the period, the proportion of foreign doctoral graduates in the 1999 cohort was around 27% (Cohen, 2001). This comparison suggests that the hand-curated sample underestimates the true share of foreign graduates. Table A.6 reports the proportion of non-French individuals in the validated sample by year, prior to the exclusion of the 2002 cohort. The cause of this anomaly is unclear.

After applying these exclusions, we retain a total of 786 observations with at least one usable location indicator for origin classification. Table A.7 summarizes the distribution of these observations by the first available type of location information.

Table A.6: Proportion of Non-French Graduates in the Validation Sample, by Defence Year, 1992–2002

Defence Year	Proportion Non-French
1992	0.1625
1993	0.1875
1994	0.2250
1995	0.2125
1996	0.1625
1997	0.2375
1998	0.1875
1999	0.1875
2000	0.1500
2001	0.2250
2002	0.0500

Table A.7: First Available Location Indicator for Origin Classification (Post-Exclusion)

First Available Location Type	Number of Observations
Bachelor Location	282
Master Location	191
High School Location	164
Country of Birth	146
Total	**786**

For these 786 doctoral graduates from the 1992–2001 cohorts we can compute external performance statistics. Table A.8 presents the classification performance by type of available information.

The model achieves consistently high precision for French doctoral graduates, exceeding 0.95 across all information sources. Recall is similarly strong, reaching 0.88 when country of birth is known, 0.90 for high school, and declining slightly to 0.78 and 0.74 for bachelor and master data, respectively.

For non-French graduates, the model performs well only when country of birth is available, with a precision of 0.77. Precision drops to 0.61 when based on high school location, 0.59 for bachelor, and as low as 0.13 for master's degree location. However, recall remains relatively stable around 0.88–0.93 for all sources except the master level, where it decreases to 0.78. This decline likely reflects the known fact that many international doctoral graduates complete their master's studies in France, making this variable a poor indicator of origin (CJC, 2010).

From the model's perspective, the most relevant and consistent signal is the country of birth, which aligns closely with the region of origin classification based on name and surname analysis. However, if the focus is on post-graduation careers, high school location may be more informative, as individuals who attended high school in France—regardless of their country of birth—are likely to have undergone similar educational socialization within the French system.

Unfortunately, given that non-French individuals represent only around 20% of the sample, it is not possible to replicate the full classification report across all 12 non-French classes. The limited sample size lacks the statistical power and class-specific numerosity required for such disaggregated analysis.

Table A.8: Classification Report - **Requiem_1**: Using Observed Location Information, 1992-2001

Info Type	**Precision**		**Recall**		**Accuracy**	**Support**
	Non-French	French	Non-French	French		
Country of Birth	0.77	0.95	0.89	0.88	0.88	146
High School	0.61	0.98	0.88	0.90	0.90	164
Bachelor	0.59	0.97	0.93	0.78	0.82	282
Master	0.13	0.99	0.78	0.74	0.74	191

B Graduates Scopus Matching

B.1 Scopus API Matching and Publication Retrieval

We retrieve publication records for the 80,903 doctoral graduates in our study using the `pybliometrics` Python package, which interfaces with the Scopus database (Rose and Kitchin, 2019). To identify the authors, we first use the `AuthorSearch` function, which performs fuzzy name matching on the Scopus database by looking for Scopus authors with names similar to each of our doctoral graduates. As displayed in Table B.1, at least one match is successfully retrieved for 58,306 graduates (72.1%), while 22,548 individuals (27.9%) remain unmatched. In 49 cases (0.1%), the Scopus API returned an error due to more than 5,000 potential matches for a single name-surname combination—a limitation of the system. Upon manual inspection, these cases were all identified as Chinese graduates.

Table B.1: Summary of Graduate Scopus API Matching Results

Result	**Number of Graduates**	**Percentage**
Matched	58,306	72.1%
Not Matched	22,548	27.9%
API Error (5000 entries)	49	0.1%
Total	**80,903**	**100%**

The 58,306 matched graduates correspond to a total of 506,268 Scopus Author IDs. Therefore, each graduate is matched on average with circa nine Scopus authors. To manage computational complexity, we exclude graduates in the top 1st percentile–see Table B.2 for the percentiles distribution–in terms of the number of matched Author IDs (i.e., more than 65 matches), removing 583 individuals (0.72%). This results in a working dataset of 57,723 graduates linked to 130,327 unique Scopus Author IDs for publication retrieval.

We download publication metadata for the 130,327 Author IDs. Due to Scopus-side technical issues, 81 Author IDs could not be retrieved, and an additional 514 Author IDs were automatically merged by Scopus, resulting in a final set of 129,813 Author IDs. This yields a total of 3,352,157 publication records. Post-processing reveals that 166,276 publications (5.0%) do not include the associated Author ID in the `authors_id` field, due to Scopus's limit of listing only the first 100

Table B.2: Distribution of Number of Author ID Matches Across Percentiles

Percentile	Number of Matches
25%	1
50%	1
75%	2
95%	10
99%	65

authors per article. As a result, affiliation data cannot be retrieved for these cases. These publications are linked to 4,445 Author IDs. Among them, 3,955 Author IDs are also associated with other publications that include affiliation data, while 490 have no such coverage. Furthermore, among the 3,185,881 publications that do include the Author ID, 78,743 lack valid affiliation data due to missing entries in the `author_afids` field on Scopus. These records are linked to 23,818 Author IDs, all of whom nonetheless have at least one other publication with retrievable affiliation information.

In the final step, we identify 57,596 graduates who have at least one Scopus Author ID with associated publication and affiliation data. This implies that only 127 graduates out of the filtered set could not be matched with usable publication records.

B.2 Match Verification with Random Forest

From the initial matching with the Scopus API, 57,596 graduates are each associated, on average, with approximately nine Scopus author profiles. To identify the correct matches, we apply a Random Forest classification algorithm developed by Corsini et al. (2025) within the framework of the research project *DOC-TRACK: STEM Doctoral Graduates and Inventive Activities in European Countries*, funded by the European Patent Office Academic Research Program (EPO-ARP 2021).

A Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their outputs to improve prediction accuracy and mitigate overfitting. Each tree is trained on a bootstrap sample of the data, and at each node, a random subset of features is considered for splitting. For classification, the final prediction is determined by majority voting across trees; for regression, by averaging the predictions. This combination of randomness and aggregation makes Random Forests robust and effective across a wide range of tasks (Biau and

Scornet, 2016).

In the *DOC-TRACK* project, we train a Random Forest on 1,500 theses.fr STEM graduates (2000–2020) whose matches with Scopus were hand-validated, splitting the data into training (80%) and validation (20%) sets. The model predicts match validity using features from publication records, such as supervisor co-authorship, institutional and national affiliations, and similarity in titles, disciplines, and author names. On the 2000–2020 cohorts, the model achieves 96.1% precision and 94.2% recall. We assume these performance levels extend to our study period, despite covering graduates up to eight years earlier.

The Random Forest algorithm identifies valid matches for 52,893 out of the 57,596 initially matched graduates, implying that 4,703 graduates were associated only with false positives[24]. As shown in Figure B.1, nearly 92% of the 52,893 graduates are linked to a single Scopus author, around 7% are matched with two authors, and fewer than 1% are matched with three or more.

To avoid potential bias from the small number of graduates with multiple matches, we retain only the match with the highest Random Forest score for each graduate in the main analysis. Robustness checks, however, are conducted by alternatively retaining up to the top four matches or excluding all graduates with multiple matches (see Appendix G.2).

[24] A false positive in matching occurs when the algorithm incorrectly identifies two records as a match, even though they refer to different individuals.

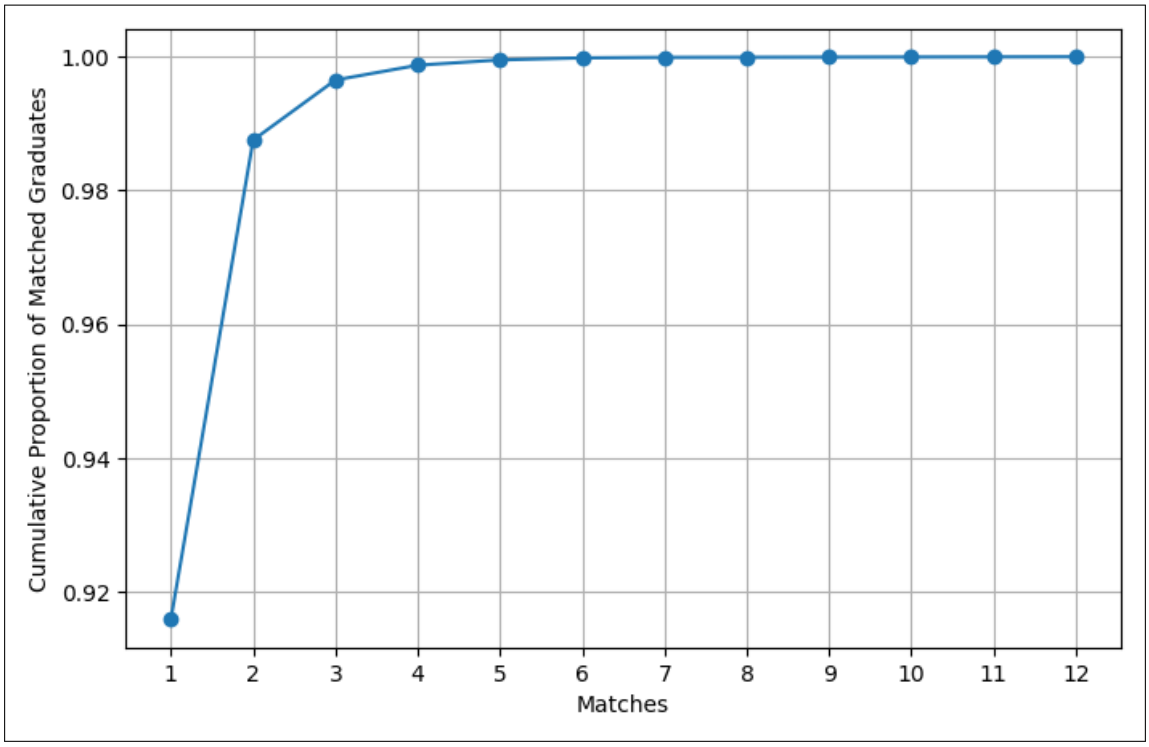


Figure B.1: Cumulative Distribution Function of the Number of Matches of the Matched Graduates

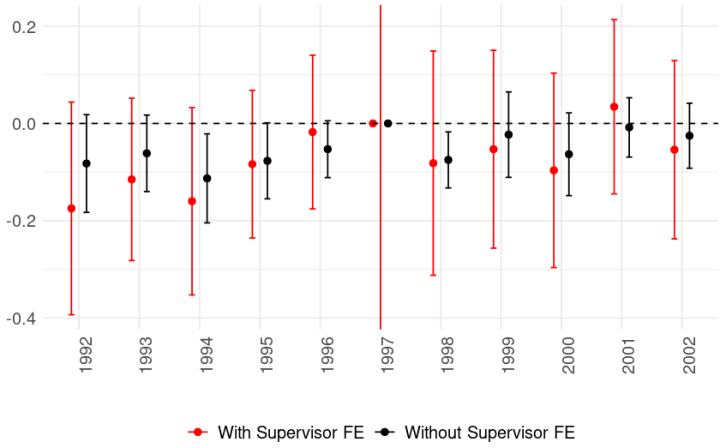
C EU as Control Group

As displayed in Table C.1, coefficients for the treated group non-EU are non significant and have mixed signs when we use the EU graduates as control group. Moreover, the *Male* control variable while significant only once is almost always negative. This reduces the robustness of the results for the EU control group, likely because of the low number of graduates in it. Figure C.1 presents the estimated interaction coefficients for the non-EU cohort variables. With the exception of the *Supervisor* outcome in 2001, none of the coefficients are statistically significant. In that year, the probability of non-EU graduates becoming supervisors increased by approximately one percentage point following the introduction of the law.

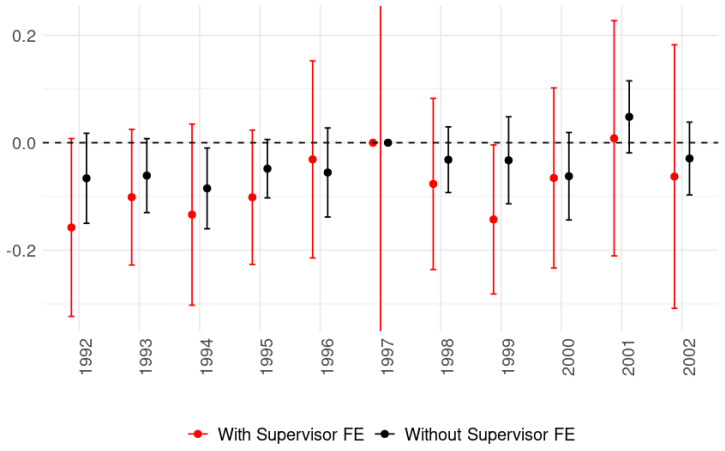
Table C.1: Dynamic Difference-in-Differences: Non-EU vs EU, LPM, With and Without Supervisor FE, 1992-2002

	Post Doc Stay		Career Stay		Supervisor	
non_EU	0.002	0.043	−0.022	0.037	−0.012	0.025
	(0.028)	(0.063)	(0.024)	(0.055)	(0.022)	(0.051)
Male	−0.017*	−0.005	−0.004	−0.014	0.006	−0.004
	(0.009)	(0.030)	(0.008)	(0.024)	(0.007)	(0.019)
Double_Degree	−0.144**	−0.184*	−0.146***	−0.129	−0.073**	−0.088
	(0.044)	(0.101)	(0.040)	(0.101)	(0.027)	(0.101)
Published_before_PhD	0.098***	0.118*	0.042*	0.045	0.027*	0.028
	(0.016)	(0.049)	(0.019)	(0.065)	(0.013)	(0.061)
Nr_Publications_during_PhD	0.039***	0.038***	0.037***	0.036***	0.025***	0.025***
	(0.003)	(0.010)	(0.003)	(0.010)	(0.002)	(0.006)
Foreign_Affiliation_during_PhD	−0.168***	−0.171**	−0.127***	−0.127**	−0.042***	−0.057*
	(0.016)	(0.053)	(0.011)	(0.038)	(0.011)	(0.030)
non_EU × Cohort1992	−0.082	−0.175	−0.066	−0.158	−0.062*	−0.087
	(0.051)	(0.112)	(0.043)	(0.084)	(0.030)	(0.088)
non_EU × Cohort1993	−0.061	−0.115	−0.061*	−0.101	−0.041	−0.048
	(0.040)	(0.085)	(0.035)	(0.064)	(0.031)	(0.067)
non_EU × Cohort1994	−0.113*	−0.160	−0.085*	−0.134	−0.016	−0.050
	(0.047)	(0.098)	(0.038)	(0.086)	(0.026)	(0.057)
non_EU × Cohort1995	−0.077*	−0.084	−0.048*	−0.102	0.003	−0.045
	(0.040)	(0.078)	(0.028)	(0.064)	(0.023)	(0.056)
non_EU × Cohort1996	−0.053	−0.018	−0.055	−0.031	0.024	0.024
	(0.030)	(0.081)	(0.042)	(0.094)	(0.034)	(0.070)
non_EU × Cohort1998	−0.075*	−0.082	−0.032	−0.077	−0.027	−0.028
	(0.029)	(0.118)	(0.031)	(0.081)	(0.035)	(0.075)
non_EU × Cohort1999	−0.023	−0.053	−0.033	−0.143*	−0.018	−0.028
	(0.045)	(0.104)	(0.041)	(0.071)	(0.026)	(0.055)
non_EU × Cohort2000	−0.063	−0.096	−0.062	−0.066	−0.011	−0.036
	(0.043)	(0.102)	(0.042)	(0.086)	(0.036)	(0.074)
non_EU × Cohort2001	−0.008	0.034	0.048	0.008	0.083*	0.098
	(0.031)	(0.091)	(0.034)	(0.112)	(0.034)	(0.074)
non_EU × Cohort2002	−0.025	−0.054	−0.029	−0.063	0.018	−0.064
	(0.034)	(0.094)	(0.035)	(0.125)	(0.033)	(0.087)
Cohort FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Field FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Institute FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Supervisor FE	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Observations	11,353	11,353	11,353	11,353	11,353	11,353
R^2	0.144	0.725	0.137	0.728	0.092	0.717

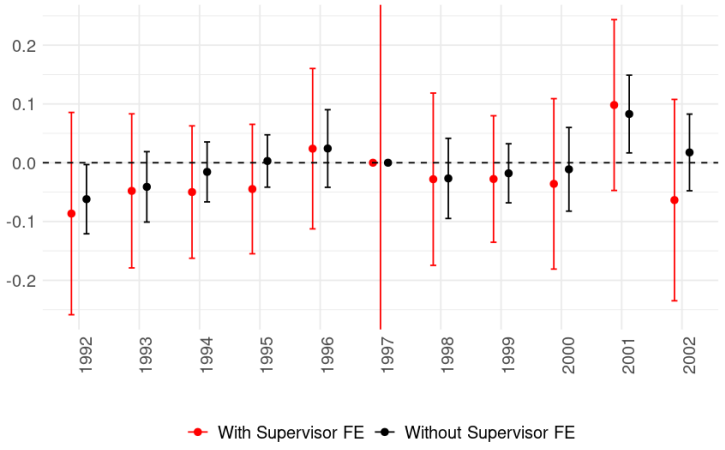
*p<0.1; **p<0.05; ***p<0.01. Standard errors in parentheses.



(a) Post Doc Stay - non-EU vs EU



(b) Career Stay - non-EU vs EU



(c) Supervisor - non-EU vs EU

Figure C.1: Estimated δ Coefficients (95% C.I.) by Cohorts: All outcome variables and treated vs control groups, 1992 – 2002, non-EU vs EU

Note - Grey lines: regressions without supervisor FE; Red lines: regressions with supervisor FE

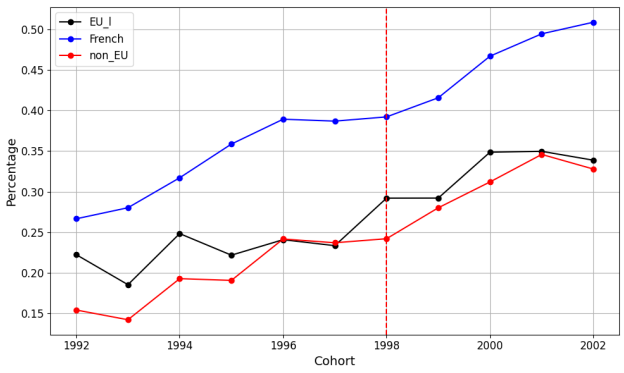
D Full Regressions Tables

Table D.1: Dynamic Difference-in-Differences: non-EU vs France, LPM, With and Without Supervisor FE, 1992-2002

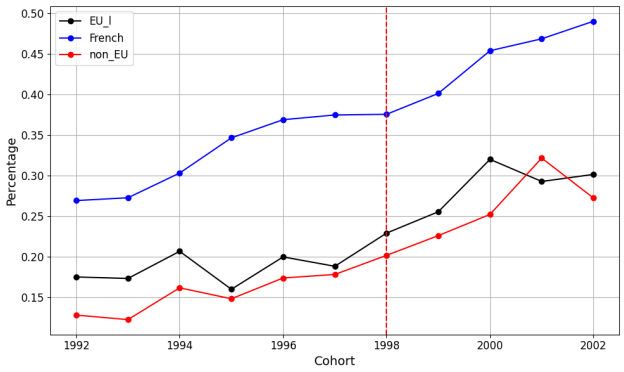
	Post Doc Stay		Career Stay		Supervisor	
non_EU	−0.109***	−0.078**	−0.165***	−0.157***	−0.076***	−0.068***
	(0.019)	(0.026)	(0.018)	(0.023)	(0.016)	(0.019)
Male	0.028***	0.039***	0.066***	0.068***	0.060***	0.055***
	(0.004)	(0.005)	(0.005)	(0.006)	(0.004)	(0.005)
Double_Degree	−0.099**	−0.147**	−0.063	−0.066	−0.036*	−0.046
	(0.033)	(0.053)	(0.050)	(0.079)	(0.018)	(0.043)
Published_before_PhD_1	0.123***	0.107***	0.118***	0.103***	0.034***	0.025**
	(0.011)	(0.013)	(0.011)	(0.014)	(0.006)	(0.009)
Nr_Publications_during_PhD_1	0.042***	0.043***	0.040***	0.040***	0.026***	0.026***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
Foreign_Affiliation_during_PhD_1	−0.131***	−0.130***	−0.047***	−0.053***	0.014	0.004
	(0.008)	(0.012)	(0.009)	(0.015)	(0.008)	(0.012)
non_EU × Cohort1992	−0.012	−0.039	0.022	0.032	−0.014	−0.004
	(0.028)	(0.036)	(0.031)	(0.034)	(0.026)	(0.027)
non_EU × Cohort1993	−0.020	−0.054	0.020	0.018	−0.004	−0.006
	(0.025)	(0.033)	(0.024)	(0.028)	(0.022)	(0.022)
non_EU × Cohort1994	−0.013	−0.033	0.024	0.020	0.011	0.011
	(0.027)	(0.037)	(0.025)	(0.033)	(0.021)	(0.021)
non_EU × Cohort1995	−0.044*	−0.056*	−0.026	−0.027	−0.011	−0.018
	(0.025)	(0.033)	(0.022)	(0.030)	(0.022)	(0.028)
non_EU × Cohort1996	0.007	−0.024	0.012	0.022	0.003	0.016
	(0.026)	(0.040)	(0.027)	(0.037)	(0.022)	(0.028)
non_EU × Cohort1998	0.003	−0.022	0.035*	0.048*	0.024	0.021
	(0.024)	(0.035)	(0.019)	(0.026)	(0.021)	(0.030)
non_EU × Cohort1999	0.017	0.001	0.028	0.039	0.010	0.014
	(0.024)	(0.033)	(0.025)	(0.030)	(0.018)	(0.027)
non_EU × Cohort2000	−0.002	−0.030	0.009	0.005	0.029	0.003
	(0.029)	(0.045)	(0.026)	(0.040)	(0.023)	(0.034)
non_EU × Cohort2001	0.008	−0.013	0.071**	0.061*	0.052*	0.045
	(0.028)	(0.042)	(0.024)	(0.036)	(0.023)	(0.031)
non_EU × Cohort2002	−0.023	−0.012	−0.012	0.023	0.017	0.021
	(0.025)	(0.043)	(0.021)	(0.028)	(0.025)	(0.033)
Cohort FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Field FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Institute FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Supervisor FE	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Observations	62,512	62,512	62,512	62,512	62,512	62,512
R^2	0.176	0.485	0.178	0.485	0.106	0.437

*p<0.1; **p<0.05; ***p<0.01. Standard errors in parentheses.

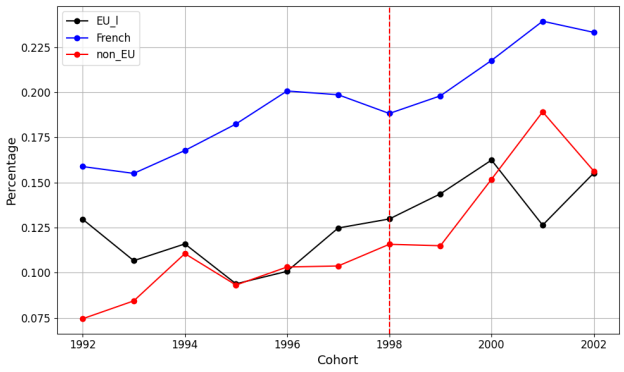
E EU "Large" Graphs with Spanish and Portuguese



(a) Post Doc Stay



(b) Career Stay

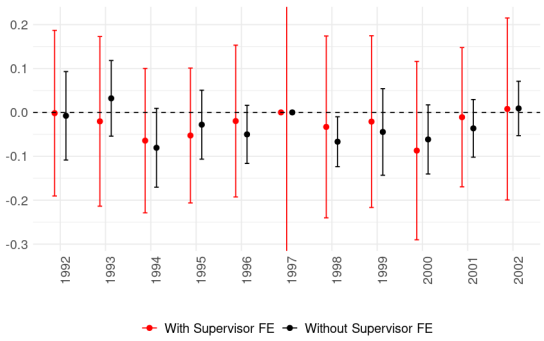


(c) Supervisor

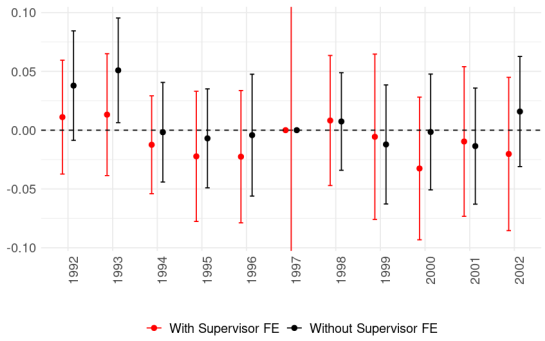
Figure E.1: Proportion of Stayers, by EU, non-EU and French status, and graduation cohort (1992–2002)

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$. In EU *Spanish* and *Portuguese* graduates are added.

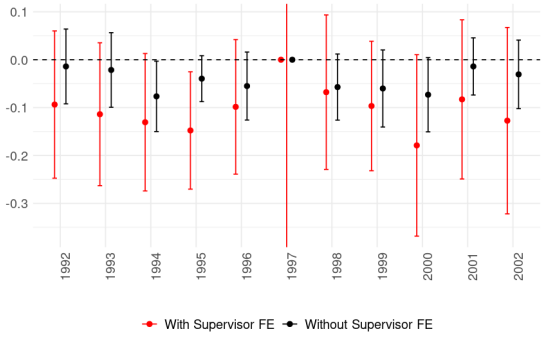
F Arabic Results



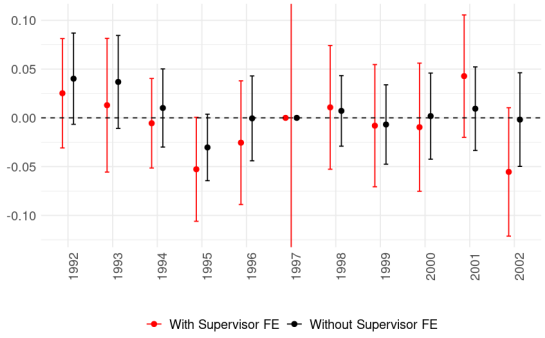
(a) Post Doc Stay - Arabic vs EU



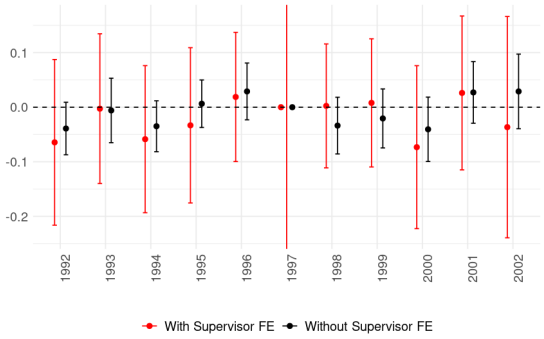
(b) Post Doc Stay - Arabic vs French



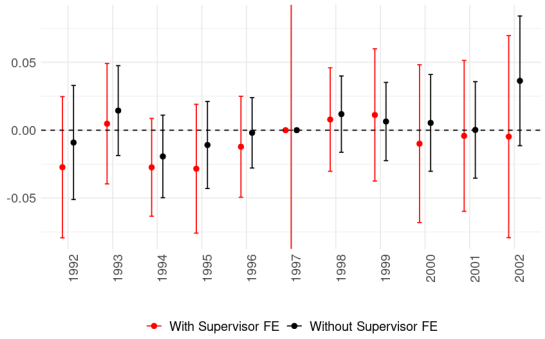
(c) Career Stay - Arabic vs EU



(d) Career Stay - Arabic vs French



(e) Supervisor - Arabic vs EU



(f) Supervisor - Arabic vs French

Figure F.1: Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated vs Control, 1992 – 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$

G Robustness Check

G.1 More Stringent Thresholds

In our main analysis, given the distribution of predicted probabilities shown in the box plots in Figures A.2 and A.3 for the two-stage classification models (**Requiem_1** and **Requiem_2**)—used to predict the region of origin of doctoral graduates—we adopt a threshold of 0.5 to determine whether an individual belongs to a given class (region of origin) or not.

A more conservative approach would apply stricter thresholds. Figure G.1 shows the distribution of probabilities for **Requiem_1** divided into twenty bins. The first and last bins, containing values from 0 to 0.05 (very low probability of being French) and from 0.95 to 1 (very high probability of being French), together account for roughly 75% of graduates. As shown in Table G.1, when we evaluate model performance on a validation sample with known location information (see Appendix A), applying a 0.5 threshold yields a precision for non-French graduates of 0.61 and a recall of 0.89. Tightening the thresholds to 0.05/0.95 for **Requiem_1** increases precision to 0.72 with only a marginal decrease in recall to 0.88.

Similarly, Figure G.2 plots the distribution of the highest predicted probability across regions of origin for **Requiem_2**. The last bin (0.95–1) contains almost 55% of non-French graduates. Applying stringent thresholds to both stages—0.05/0.95 for **Requiem_1** and 0.95 for **Requiem_2**—further increases precision for non-French graduates from 0.72 to 0.75, though recall declines from 0.88 to 0.83. Figure G.3 presents the regression results from the main analysis under these stricter thresholds. Overall, compared to the baseline results in Figure **??**, the effect on *Career_Stay* vanishes, while the effect on *Supervisors* loses significance at the 5% level but remains significant at the 10% level.

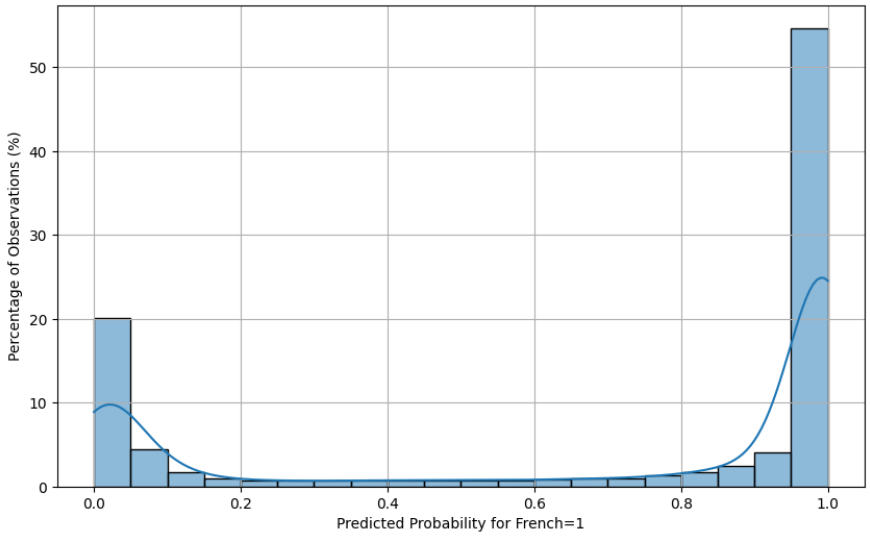


Figure G.1: Threshold Distribution **Requiem_1** Predictions of French equal 1, 1992-2002

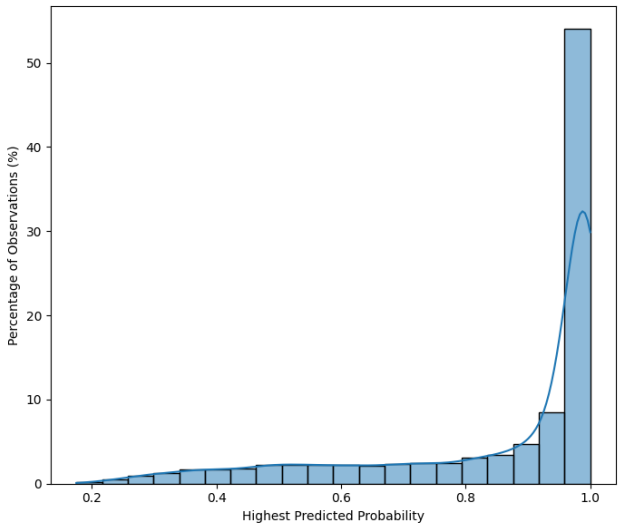
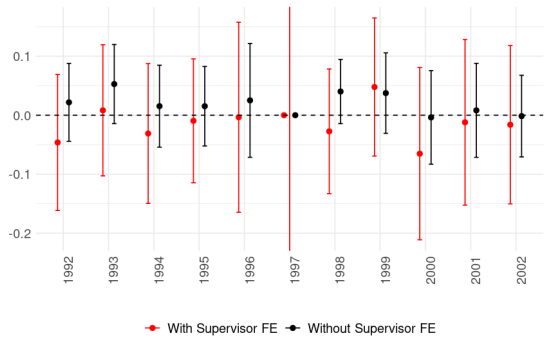


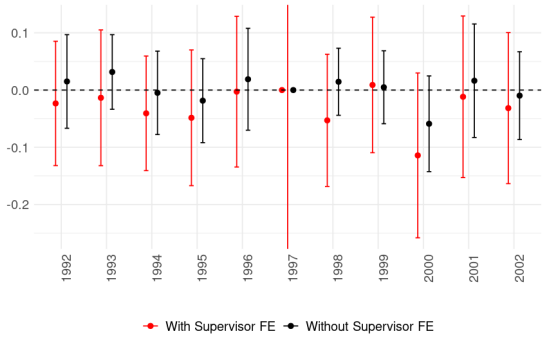
Figure G.2: Threshold Distribution **Requiem_2** Predictions of the Highest Region of Origin Value, 1992-2002

Table G.1: Classification performance for different **Requiem_1** and **Requiem_2** threshold settings

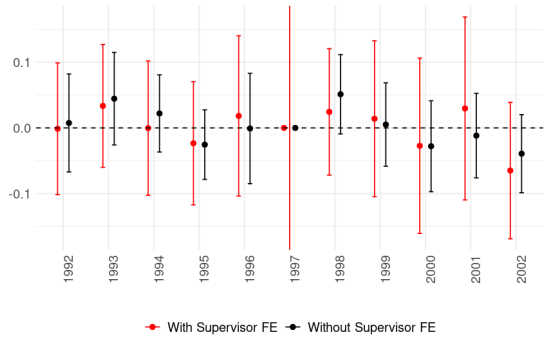
Requiem_1	Requiem_2	Class	Precision	Recall	F1-score	Support	Accuracy
0.50	0.50	non_French	0.61	0.89	0.73	149	0.85
		French	0.96	0.84	0.90	513	
0.05/0.95	0.50	non_French	0.72	0.88	0.79	111	0.89
		French	0.96	0.90	0.93	384	
0.05/0.95	0.95	non_French	0.75	0.83	0.79	76	0.92
		French	0.96	0.94	0.95	366	



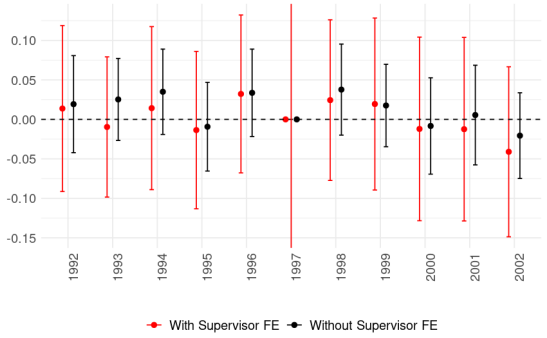
(a) Post Doc Stay - **Requiem_1** 05-95 Threshold



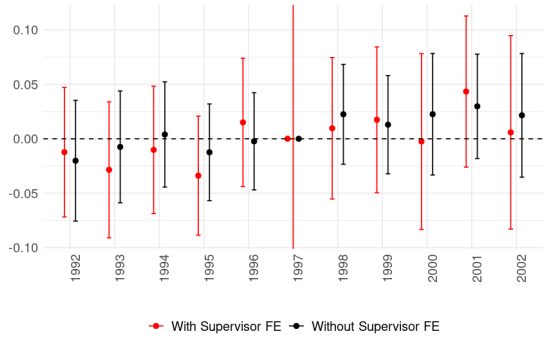
(b) Post Doc Stay - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold



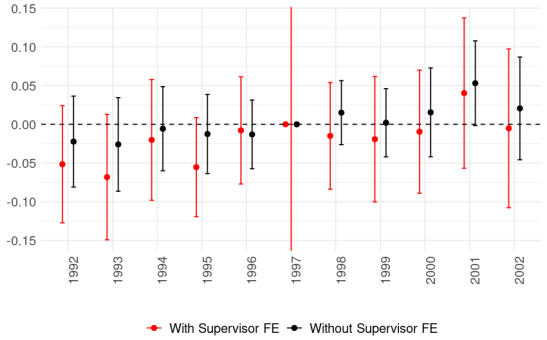
(c) Career Stay - **Requiem_1** 05-95 Threshold



(d) Career Stay - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold



(e) Supervisor - **Requiem_1** 05-95 Threshold



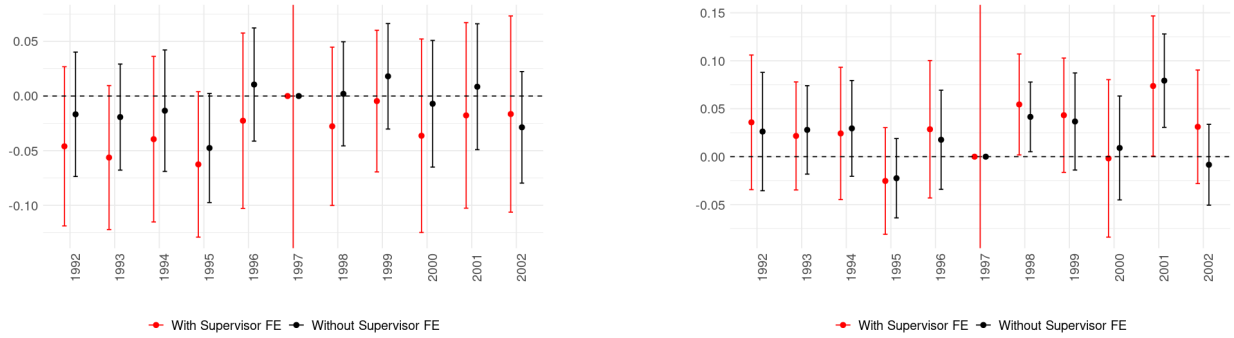
(f) Supervisor - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold

Figure G.3: Estimated Coefficients (95% Confidence Interval) by Cohorts: non-EU vs French, 1992 – 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$

G.2 Multiple Matches Alternative Approaches

G.2.1 Retention of the Top Three Matches per Graduate



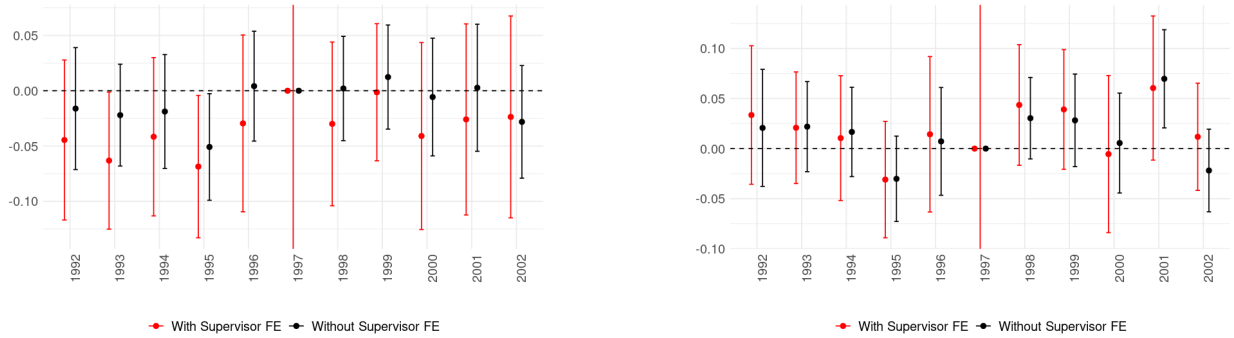
(a) Post Doc Stay 3 - non-EU vs French

(b) Career Stay 3 - non-EU vs French

Figure G.5: Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated vs Control, 1992 – 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$. If a graduate has more than one match the top three matches are retained.

G.2.2 Excluding Graduates with Multiple Scopus Author Matches (Post Random Forest)



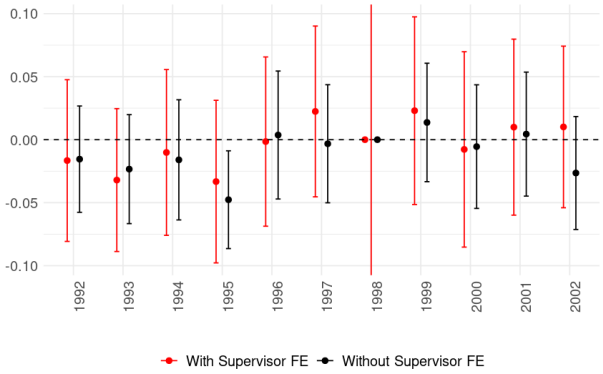
(a) Post Doc Stay - non-EU vs French

(b) Career Stay - non-EU vs French

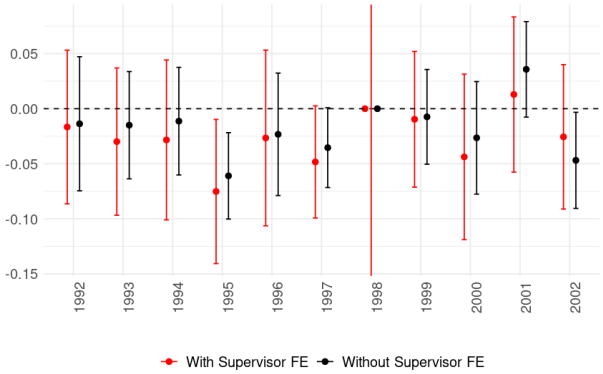
Figure G.7: Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated vs Control, 1992 – 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort c with $Supervisor_{i,c+t} = 1$. Graduates with more than one match are excluded.

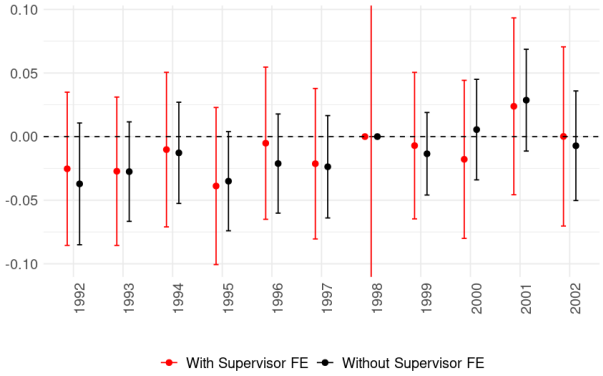
G.3 1998 as Reference Year



(a) Post Doc Stay - non-EU vs French



(b) Career Stay - non-EU vs French



(c) Supervisor - non-EU vs French

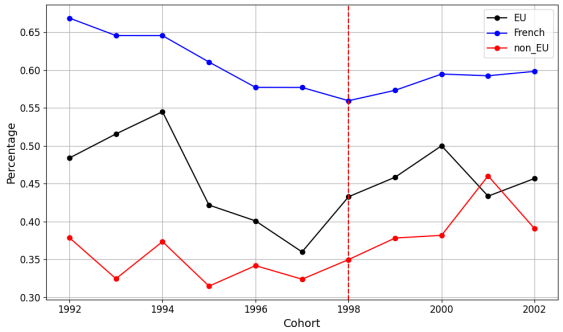
Figure G.9: Estimated δ Coefficients (95% C.I.) by Cohorts: All outcome variables and treated vs control groups, 1992 – 2002, non-EU vs French, 1998 as Reference Year

Note - Grey lines: regressions without supervisor FE; Red lines: regressions with supervisor FE

G.4 Conditioning on Publishing in Post Doc & Career Periods



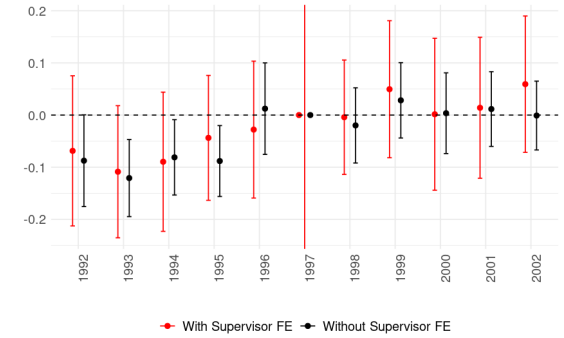
(a) Post Doc Stay; Conditional on Publishing in $\geq c+2, \leq c+4$



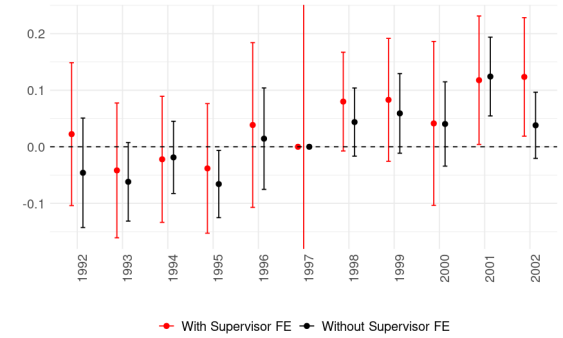
(b) Career Stay; Conditional on Publishing in $\geq c+5, \leq c+10$

Figure G.11: Proportion of Stayers, by EU, non-EU and French status, and graduation cohort (1992–2002); Conditional on Publishing in Post Doc & Career Periods

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$.



(a) Post Doc Stay - non-EU vs French; Conditional on Publishing in $\geq c+2, \leq c+4$



(b) Career Stay - non-EU vs French; Conditional on Publishing in $\geq c+5, \leq c+10$

Figure G.13: Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated vs Control, 1992 – 2002

Post Doc Stay: share of students in cohort c with $Post_Doc_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort c with $Career_Stay_{i,c+t} = 1$.

Chapter 3

Stay rates and careers of foreign doctorate recipients from Dutch universities: A bibliometric analysis

Andriy Romanyuk [1,2,*] and Francesco Lissoni[2,3,4]

[1]Dipartimento di Economia, Università dell'Insubria (Varese)
[2]Université de Bordeaux, CNRS, BSE, UMR 6060
[3]ICRIOS – Bocconi University
[4]French Collaborative Institute on Migration

Abstract

Based on a unique dataset combining doctoral theses and scientific publications, we examine the publishing activity and cross-country mobility of native and foreign-born STEM doctoral graduates of Dutch universities from 2000 to 2020. After controlling for composition effects, native and foreign-born graduates show similar probabilities of publishing post-graduation, with a minor exception (-3%) for European ones. The average number of publications is the same. However, conditional on publishing, foreign-born graduates are 11% to 23% less likely than natives to publish from the Netherlands, depending on field, university, supervisor, and origin. Europeans are generally less likely to leave upon graduation than non-Europeans, though not always significantly; and the difference is largely driven by the influx of Chinese students in the 2010s, many with funding schemes binding them to return home once graduated. The introduction of the "knowledge migrant" visa in 2004 and its 2008 reform appear to have had no impact on stay-or-leave decisions. Finally, we find no evidence of self-selection among those who leave versus those who stay, except for native Dutch graduates. For them, scientific productivity during the PhD is mildly but positively associated with publishing from abroad, especially from the United States and other R&D-intensive Anglo-Saxon countries. Compared to similar studies on the United States, the main and most studied destination of foreign PhD students, the Netherlands exhibits comparable stay rates with (so far) less variation due to migration policy changes.

JEL classification: F22, I23, J61, O15

Keywords: Academic careers, International graduate students, Highly skilled migration, Netherlands

We benefited of useful comments and suggestions from participants to the WICK Workshop (2022, Turin), the 1st EoS Workshop (2023, Bordeaux), the BETA workshop (2023, Strasbourg), the REGIS Summer School (2023, Pisa), the KID Summer School (2023, Nice), the STI Summer School (2023, Barcelona), the EmELHE workshop (Kassel, 2025) and the PhD seminar programme at BSE (Bordeaux). Marije Sluiskes, Frank Corvers, Frank van Rijnsoever, Jarno Hoekman and Koen Frenken provided key references on Dutch doctoral funding systems and/or migration policies. Pauline Menu provided key research assistantship, along with Aman Araissi, Lucien Boucart, Baptiste Comby, Samuel Desneulin, and Valentine Petzold. We acknowledge funding from the European Patent Office's Academic Research Programme (DOCTRACK project). Lissoni also acknowledges funding from the French State in the framework of the Investments for the Future programme IdEx Université de Bordeaux / GPR HOPE. The results and comments presented in this paper are solely ours and do not engage the sponsors' views. All errors remain ours, too.

1 Introduction

Higher education in Science, Technology, Engineering, and Mathematics (STEM) plays a significant role both in producing the human capital needed for innovation-driven economic growth (Stephan et al., 2004) and in attracting it from abroad. Many advanced countries increasingly fill their STEM university programs with foreign students, especially at the graduate level. This phenomenon results from long-standing trends. First, the share of foreign university students in developed countries has grown steadily since the late 1970s. Second, this share increases when transitioning from undergraduate to graduate programs and is highest for PhDs (OECD, 2023). Third, in the transition, the foreign students' propensity to choose a STEM field also increases (OECD, 2020).

Evidence for the United States, so far the main attraction pole for international students worldwide, indicate that foreign STEM students play a key role in the country's science and innovation system, as measured by the publications and patents they produce both before and after their graduation (Chellaraj et al., 2008; Hunt and Gauthier-Loiselle, 2010; Hunt, 2011; Gaule and Piacentini, 2013). The extent of their post-graduation contribution depends crucially on their permanence in the host country, which in turn is affected – among others – by the ease with which they can move from a student visa to a labour one (Roach and Skrentny, 2019, 2021; Kahn and MacGarvie, 2020). Evidence for European countries and more generally for other countries than the United States is sparse, with no more than one study per country, such as Zheng and Ejermo (2015) on Sweden and Buenstorf et al. (2022) on Germany, if none at all.

In this paper we examine the scientific productivity and stay rates of foreign STEM doctoral graduates from Dutch universities, relative to their native peers. In particular, we investigate whether – conditional on undertaking a career in science – foreign graduates who stay in Netherlands are positively or negatively self-selected, based on their scientific production during the PhD. Furthermore, we analyze trends in stay rates from 2000 to 2020, also considering the impact of changes in migration policies.

Netherlands represents a compelling case study for research on STEM migration. It is a small yet R&D-intensive country, with both a high demand for STEM workers and a number of world class uni-

versities, which attract an increasingly high number of foreign students in graduate STEM programmes (OECD, 2023). Based on the Free Movement of Persons principle, all students from within the European Economic Area (EEA) can settle freely in the country, both before and after completing their studies. But the same does not apply, in principle, to non-EEA ones, who nevertheless account for an increasing share of total students. For this reason, since the early 2000s, all Dutch governments have multiplied the immigration policy measures aimed at easing their transition into the Dutch labour market (OECD, 2016; Buers et al., 2018). Despite these efforts, however, some past research indicates that the Netherlands' retention rate of foreign doctoral recipients is still lower than that of the United States (Rud et al., 2015); and not higher than in comparable European countries (SVR, 2012). This adds policy salience to our study.

Our analysis relies on an original dataset combining information on the doctoral dissertations in Science, Technology, Engineering or Mathematics (STEM) produced in the Netherlands between 2000 and 2020 with information on their authors' country of birth and gender. By linking this information to publications data, we follow the scientific production of both native and foreign doctoral recipients, both before, during and after their graduation. We do the same for their supervisors, too. Based on this, we address four research questions.

First, we investigate whether foreign-born STEM doctoral graduates have a higher or lower propensity to undertake a scientific career, relative to natives, as measured by their probability to publish on scientific journals in the years following their graduation.

Second, we investigate whether the foreign students who leave the Netherlands after completing their studies are positively or negatively self-selected relative to their research skills, as measured by their scientific productivity during their PhD. More plainly, we test whether the Netherlands loses or retains the best and brightest of their foreign doctoral recipients or, on the contrary, it loses them out to other countries.

Third, and relatedly, we examine the countries to which the leavers relocate, considering whether these are their countries of origin, to which they return after completing their studies abroad, or third countries competing with the Netherlands in the global race for talent.

Last, we investigate the effects of changes in the Netherlands' immigration policies, most notably the introduction of a new Knowledge Migrant Scheme in 2004 and, in 2008, of a specific measure for its application to scientific researchers, based on the implementation of the European Union directive 2005/71.

In all our exercises, we control for a number of confounding factors such the graduates' gender and age; their co-affiliations, if any, to a foreign institution (before or during their doctoral studies); and the characteristics of their field, PhD-granting institutions and supervisors (via fixed effects).

Our findings suggest that foreign-born individuals tend to be concentrated in fields, such as engineering, that provide more career options outside academia compared to other disciplines. Field and supervisor sorting mechanisms explain their lower probability to publish after graduation. Moreover, once these sorting mechanisms are accounted for, the research productivity of foreign and native graduates is equivalent.

Regarding stay rates, foreign-born students are significantly more likely than native ones to leave the Netherlands after graduation, as expected. We also find some evidence that students who – regardless of origin – begin publishing before completing their degrees are more likely to emigrate upon graduation, suggesting a slight positive self-selection among those who leave. This difference however, is largely explained by Chinese students, whose publishing activity we possibly overestimate due to the high number of homonymous scientific authors. Examining the destinations of emigrants, we observe that more productive native graduates are more likely to relocate to English-speaking countries, while no clear pattern emerges between productivity and destination choice among foreign-born graduates, with the only exception of Chinese ones. This exception may partly result -once again- from a measurement error, but also from a substantial factor, namely the type of funding received by many Chinese students over the years 2010s, which bound them to return home upon graduation. Prior international experience is the variable most strongly correlated with a higher likelihood of emigration overall, which suggest that a large number of leavers either moved to the Netherlands for their PhD while already holding positions abroad or took part to joint degrees between Dutch and foreign institutions. Concerning highly skilled immigration policies, we do not find any significant effect, which is coherent with the same policies' lack

of a specific focus on STEM students.

The paper proceeds as follows. In section 2 we review the relevant literature and discuss our research questions in relation to it. In section 3 we provide contextual information on the Netherlands' higher education and immigration policies. In section 4 we describe the dataset and our econometric strategy. Section 5 contains our results and their discussion. Section 6 concludes. An online Appendix provides more details on both our data, methods, and results.

2 International STEM students and Innovation

In the past 15 years or so, innovation scholars have devoted increasing efforts to explore the many ties linking migration and innovation, including the direct contribution of highly skilled migrants to the advancement of science and technology in their destination countries (Kerr et al., 2016; Breschi et al., 2020a; Capoani et al., 2023). This contribution can take many forms. One of these is knowledge transfer, with immigrants bringing along scientific or technical notions otherwise absent or hard-to-access in the host countries (Moser et al., 2014; Ganguli, 2015; Docquier and Rapoport, 2012; Miguelez and Morrison, 2023). Another one is self-selection of highly skilled migrants into science and technology jobs, which increases the local supply of the necessary skills for R&D. General evidence in this sense is the over-representation of immigrants among STEM students worldwide (OECD, 2020), STEM jobs (Hanson and Slaughter, 2019; Hunt, 2015; Hunt and Gauthier-Loiselle, 2010; Freeman, 2015), and highly productive scientists and inventors (Stephan and Levin, 2001; Bernstein et al., 2022; Cristelli and Lissoni, 2025).

Immigrants' self-selection into R&D jobs and high productivity may have two complementary reasons. First, as per the classic Roy-Borjas model, when migration occurs between countries with, respectively, less and more unequal skill-based income distribution, it is the most skilled who move (Borjas, 1991; Parey et al., 2017). This may be the case of home-educated STEM professionals trying to enter a foreign labour market of an R&D-intensive country, but also of good students who decide to get or complete their STEM education abroad and enter the global labor market from a vantage position. Second, a dynamic division of labour mechanism may be in place, by which immigrants tend to specialize in tasks

that are less demanding in terms of mastery of local language, social capital or cultural integration, with respect of which natives hold a relative advantage. Getting a STEM education and applying for a R&D job is therefore a good strategy to follow for a foreign student whose ultimate purpose is to migrate permanently (Peri and Sparber, 2011). Based on aggregate data from the US Census Bureau, Chellaraj et al. (2008) observe a positive impact of international foreign graduate students on US patenting activities and attribute part of it to these students' higher concentration in STEM fields. Similarly, by using a combination of individual level data from a national survey of college graduates and aggregated patent data, Hunt and Gauthier-Loiselle (2010) impute the higher patenting rate of foreigners entirely to them holding disproportionately STEM degrees. The same, however, does not apply to their higher publication rate, which resist to all controls for composition effects (Hunt, 2011).

Self-selection may be reinforced by the immigration policies of destination countries, which often include provisions favoring highly skilled individuals (Skeldon, 2018). In most cases, these policies consist in introducing exceptions to otherwise restrictive and demand-driven labor immigration systems. Such systems typically include a labor market test of some kind, by which employers are required to apply for work permits on behalf of foreign workers and to demonstrate that no domestic candidates are available to fill the positions. Additionally, they may limit the maximum number of foreign workers admitted annually (De Haas et al., 2018). Within this context, selective policies offer special visas to professionals possessing highly demanded skills, with no need of labor market tests nor caps. The "knowledge migrant" scheme (*kennismigranten regeling*) introduced in the Netherlands in 2004 is a typical example (see next section), albeit less well known than the H-1B visa in the United States (Kerr and Lincoln, 2010; Shih, 2016; Glennon, 2023; Mayda et al., 2023) or the European Union (EU) Blue Card, to which it is often compared (de Lange, 2018; De Lange and Vankova, 2022).[1]

Attracting STEM students from abroad is not enough for a country to benefit from them in terms of innovation. It is also necessary both to retain them after graduation and to ease their access to R&D occupations. This depends, in turn, on both contingent economic conditions and immigration policies. Concerning the former, Grogger and Hanson (2015) find that foreign PhD students in the US tend to

[1] In countries with supply-driven immigration policies—such as Canada, Australia, and New Zealand—employers play a lesser or no role in the process. Instead, prospective immigrants are screened using point-based systems, which assign higher scores—among others—to applicants with advanced education, particularly in STEM fields (Koslowski, 2018).

stay during times of strong economic growth or when their home country's economy is struggling. As for the latter, a number of studies find that the transition from student to job-related visas in many countries is far from smooth, due to the rigidity and instability of labour immigration schemes, even in the presence of favourable schemes for the highly skilled (Roach and Skrentny, 2021; Han et al., 2015; Hawthorne, 2018).

Stay rates of foreign STEM graduates depend also on the labor market and more general on economic and political conditions of their home countries. For example, according to Finn (2010) stay rates of Indian and Chinese STEM doctoral students in the United States in the 1990s and early 2000s were much higher than those their European peers, which is less the case in recent years. It may also be the case that foreign students in/from specific countries move abroad thanks to funding schemes that bind them to return home upon graduation, as with the Fulbright scholarships for studying in the United States (Kahn and MacGarvie, 2016) or the Chinese Scholarship Council grants (see next section). This implies that the geographic composition of foreign student inflows has a decisive influence on the total stay rates.

Overall, the literature reviewed so far suggests that foreign STEM students, and especially those in graduates programs, can contribute to innovation in their host countries, and provides some contingencies under which this is more or less likely. However, this evidence comes with a number of limitations.

First, there is a disproportion in the amount of studies on the United States relative to those on other countries, including some with a strong presence of foreign STEM students, including many European ones. In addition, the few available studies on such countries cast doubts on the possibility to generalise from the US-related findings. Based on a study on the entire population of Sweden, Zheng and Ejermo (2015) find that the immigrants are generally less likely to patent than natives and this holds true also after controlling for their education level and whether they were educated in Sweden or abroad. Buenstorf et al. (2022) study the patenting activities of foreign STEM doctoral graduates at German universities and find them to patent mostly for international companies.

Several factors suggest that differences exist between the United States and European countries, even though both are R&D-intensive, attract many international students, and maintain selective immigration

systems. First, the international students of European countries originate from both inside and outside the EEA. The former have no or minimal limitations to entry and stay, while the latter may face severe restrictions. This may affect considerably their career and stay decisions, albeit in a non-straightforward manner. On the one hand, EEA graduates may be more likely to stay, given that they face less costs (in terms of distance from their home countries) and no restrictions; but, on the other hand, they are also less constrained to take a definitive decision, being free to move to any other EEA country at any time. Instead, for non-EEA students, staying in the country where they graduated may be the only alternative to returning home. Second, when it comes to competing for talents, Europe is historically lagging behind the United States, which do not only exert a stronger attraction in Asia and South America, but also attract many European students and possibly the best ones (Veugelers, 2017; Parey et al., 2017). Hence, when studying the contribution of foreign STEM students to innovation in Europe, one should take into account that most of them, when it comes to their career decisions, do not only consider whether to stay in their host country or return home, but also whether they want and can move to the United States or other attractive countries in the EEA as well as the United Kingdom, Canada, Australia, and Switzerland (Stephan et al., 2004; Van Der Wende, 2015).

A second limitation of the literature is the scarcity of studies, whether in the United States or elsewhere, based on micro-data enabling the researchers to follow the STEM graduates' transition from education into R&D-related jobs at the individual level. This makes it difficult to explore the microeconomic determinants of their stay and career decisions, especially for what concerns information on the various individuals' level of competences, as measured for example by their scientific production before or during the PhD or the quality of their supervisors.

In this respect, the doctoral dissertations are an important data source, as they contain or allow to retrieve most of such information. However, as showed in Gaule and Piacentini (2013), who analyze the doctoral dissertations of Chinese chemistry PhD students in the United States, the information-retrieving process is rather complex and it often comes at the price of either losing observations or making drastic methodological choices.

A third limitation of the literature concerns the impact analysis of immigration policies. Most of

the available evidence concerns the H-1B visa, which allows firms and universities in the United States to employ foreign workers in specialty occupations and is vastly used by foreign doctoral students to prolong their stay once they graduate, waiting for a more permanent solution (Roach and Skrentny, 2021). Exploiting the fact that H-1B visas come with national quotas, Shih (2016) shows that the number of international students coming to from each country is positively related to the number of visas made available for that country. Other authors, in turn, find a positive association between the availability of H-1B visas, which varies over time, and the innovative activities of U.S. cities and firms (Kerr and Lincoln, 2010; Peri et al., 2015; but see Doran et al., 2022 for contrary evidence). As for the available evidence on stay rates, this is mostly based on student surveys, which collect information of the respondents' intentions to stay or leave upon graduation (Roh, 2015; Khosla, 2018; Kahn and MacGarvie, 2020).

Our paper aims at filling some of these gaps in the literature. First, we focus on a European country, the Netherlands, for which – to our knowledge – no in-depth study has ever been produced, despite the high internationalization rate of its studentship and its strong R&D orientation. Second, we build and use an original and extensive dataset of individual doctoral recipients in STEM fields, with detailed information on their place of birth as well as thesis, university, supervisor, and scientific publications, both before, during and after their studies. This allows us to consider, in our analysis, a number of theoretical and methodological inputs from the economics of science, especially for what concerns the influence of supervisors on STEM graduates' careers (Corsini et al., 2022; Horta and Santos, 2016; Gaule and Piacentini, 2018). Third, we study the effects of two policy measures introduced in the Netherlands (including the "knowledge migrant" we mentioned above), that -in principle- may have had an impact on the stay rates of foreign doctoral graduates from outside Europe.

3 The Netherlands as a case study

For researchers investigating the links between migration and innovation, the Netherlands offer at least two compelling reasons for interest. First, the Dutch R&D labour market is disproportionately

large, relative to the country's size, and increasingly dependent on foreign talent. As of 2022, it counted almost 6000 full-time R&D workers per-million inhabitants, the 8th highest score worldwide (Dutta et al., 2022). This research personnel has almost doubled between 2011 and 2021 (Eurostat, 2018), also due to the presence of a number of world class research universities and large multinational corporations, which the literature suggests to be strong employers of international STEM graduates (Kerr et al., 2016).[2] According to Lissoni and Miguelez (2021), in 2006-2010 foreign nationals accounted for around 15% of the inventors with a Dutch address, up from 5% in 1996-2000 (and with an estimate of 20% for 2011-15). In 2023, 48% of the scientific personnel in Dutch universities (51% of assistant professors) were foreign nationals, up from 27% (18%) in 2007, one third of which from outside Europe. In Engineering and the Natural Sciences, the same figures were, respectively, 57% and 51% (Rathenau Instituut, 2024a,b).

Second, since at least the early 2000s, successive Dutch governments have acknowledged this growing dependence, which the Free Movement of People within the EEA has proven insufficient to address. In response, they have implemented a series of policies targeting non-EEA nationals (Obradović, 2013; OECD, 2016; Kulu-Glasgow et al., 2018). The first and best known of these, introduced in 2004, has been the Knowledge Migrant Scheme (*Kennismigranten regeling*). This is best understood as a set of exceptions to the general admission policy for workers from outside the EEA. Under such policy, labour migrants need both a work and a residence permit, which are granted only under stringent conditions (such as passing a labour market test, by which the employer proves that the vacancy cannot be filled by a Dutch or EEA national) and after navigating a very lengthy process. In contrast, knowledge migrants – as defined by the scheme – are not subject to any labour market test and have their applications processed swiftly, provided their employer meets certain criteria. These include offering a gross salary significantly above the Dutch median (in 2024: €5,331 per month for workers aged 30 or more and €3,909 for younger ones) and registering as a legitimate sponsor with the Dutch Immigration and Naturalisation Service (IND), based on reliability, continuity, and financial health (at present, the register comprises more than 10,000 organizations, and includes all the Dutch universities and research institutes). Contrary to what

[2]These include domestic companies (such as Philips, Akzo Nobel, ASML, DSM, Shell plc, and Unilever) as well as foreign ones, which in 2019 comprised 30% of the overall private R&D expenses in the Netherlands (Gilsing and Erken, 2002; Soete et al., 2020; see also: https://www.government.nl/latest/news/2020/02/19/record-number-of-foreign-companies-choose-the-netherlands, last visited: October 2, 2023).

the reference to 'knowledge' may suggest, however, no migrant's attributes such as the level of education, or specific degree or expertise are necessary to qualify for the scheme, as an assumption is made that high salaries correspond to high skills.[3]

On the one hand, this simplicity has made the scheme highly popular. Today, it is the largest program for permanent labor immigration in the Netherlands. Moreover, its introduction has rendered the EU Blue Card, launched in 2011, largely redundant due to the similarity of the two schemes (Berkhout et al., 2015). On the other hand, the emphasis on salary requirements tends to favor high-wage sectors, such as Business and Financial Services, rather than the most knowledge-intensive ones, such as R&D and Higher Education.

Over the years, however, the Knowledge Migrant Scheme has been both modified and complemented with additional measures explicitly targeting researchers. In particular, since 2008 and in application of a European Union directive (EU 2005/71), applicants for a Knowledge Migrant permit are exempted from meeting the minimum salary requirement, as long as their sponsor is a university or other research establishment included in the IND list. By 2014, this "Researcher EU 2005/71" scheme was providing the second largest entry channel for permanent labor immigration in the Netherlands, being second only to the income-based Knowledge Migrant scheme and having successfully replaced all pre-existing programs for non-EEA foreign academic researchers (Obradović, 2013; OECD, 2016). Approximately one half of the permits were granted to individuals aged 30 or younger, with the average recipient being 32 years old and the modal one between 25 and 28. It is worth noticing that, under the Researcher EU 2005/71 scheme, the sponsor may not be the researcher's employer, but simply a host for the duration of the research period. This makes the scheme accessible to visiting researchers from foreign institutions and foreign students enrolling in a Dutch PhD programme. Notice also that, reflecting a more general trend, about one third of the permit beneficiaries were Chinese nationals.[4]

[3]Other important advantages for knowledge migrant permit holders -compared to other labor permit holders- are the long duration of the permit (up to five years, conditional on not changing employer) and the possibility for partners to enter and work in the Netherlands with no minimal salary requirement, and to bring children.

[4]It is also worth adding that, one year after the implementation of the EU directive 2005/71, the Dutch government extended to doctoral graduates the possibility to apply, withing three years from the graduation, for a Orientation Year residence permit until then reserved to bachelor and master degree holders. Applicants for this permit do not need to have any work contract, but only to look actively for one and to have graduated from either a Dutch institution or (since 2009, for master and doctoral graduates) a top foreign one (one included in the top 200 lists of the 'Times Higher Education World Ranking', the 'QS World University Ranking' or the 'Shanghai Jiao Tong'). One last notation concerns our choice to refer to this particular policy as the Researcher EU 2005/71 scheme, while other sources -not specific to the Netherlands-

These and other policies were subsequently restructured and unified under the Modern Migration Policy Act (MoMi), a comprehensive legal framework introduced in 2013 to streamline and simplify the processing of immigration requests for work and study purposes.

Within this context, Dutch universities play a crucial role both as employers of foreign graduates and as attractors of international students. In the former role, they are potentially affected by the policies described above. In the latter role, they have actively engaged in a global competition since at least the early 2000s. Evidence of this includes the widespread use of English as a *lingua franca*; the restructuring of education programs into a two-tier cycle (a more globally recognized framework; Boer et al., 2010; Kolster, 2014); and the strategic support provided by government organizations such as Nuffic, the Netherlands Organisation for International Cooperation in Higher Education.[5]

Overall, this has brought, between 2000 and 2020, to a rapid increase in the share of foreign doctoral students. In 2005, they accounted for 31% of the total enrollments. By 2013, this share had already increased to 44% and it is currently at 56% (Rud et al., 2015; Rathenau Instituut, 2024a). This places the Netherlands well above the OECD average and also above traditional attractors such as the UK and Australia (OECD, 2023).

However, there has been some debate regarding the repayment of public investments made to sustain these trends. A study by Rud et al. (2015) on doctorate recipients employed by Dutch universities shows that their 10-year stay rate after graduation is significantly lower than that of their peers in the United States. In 2010, a survey of knowledge migrants employed as either doctoral students or researchers indicated that – while they evaluated their experience in the Netherlands positively – no more than 30% had decided to renew their residence permit. Furthermore, 27% reported considering a move to the United States, with an equal proportion considering Germany, the United Kingdom, or Canada. Notably, none of the respondents attributed their decisions to the immigration policy regimes of the different countries (Berkhout et al., 2010).

refer to it by using mor general terms such "Scientific Visa package" and "Researcher Directive" (Colussi, 2016; ICMPD, 2012). This is because, in 2008, the Netherlands already had in place - as mentioned - other specific programmes for scientific researchers, albeit little used. In our choice, we follow a Dutch source such as Obradović (2013); OECD (2016).

[5]Nuffic plays a particularly significant role in attracting students from developing nations such as India and China (De Haan, 2014; Luijten-Lub et al., 2005). Regarding the use of English in Dutch universities, it has become so pervasive that it has recently faced political criticism. In response, the Dutch government has committed to partially reverting to greater use of Dutch, at least in undergraduate courses. For more information, see: https://open.overheid.nl/documenten/ronl-683f03212afac8d426a56f343e42395cad6bbe75/pdf, last visited: 23 October, 2023.

One final observation concerns the dual structure of the Dutch higher education system, which comprises two main groups of institutions: the Research Universities (*Wetenschappelijk Onderwijs*) and the Universities of Applied Sciences (*Hoger Beroepsonderwijs*). The former, totaling 14 institutions, are the only ones authorized to grant doctoral certificates, and it is to them that we refer - in remainder of this paper - when speaking generically as "universities", due to our exclusive focus on doctoral graduates.

Collectively, these institutions are represented by the Dutch universities' association (*Universiteiten van Nederland*, or UNL), established in 1985 and reorganized in 2021. Eight of these universities host a faculty of medicine. Among those that do not, three – Delft, Eindhoven, and Twente (formerly Enschede) – specialize in hard sciences and technology, while one, Wageningen, focuses on agriculture (Cohen and van der Steege, 1982). These heterogeneity implies strong composition effects in the different career orientations of foreign versus native graduates, of which our analysis needs to take account.

4 Data and methods

We investigate the scientific careers of doctoral recipients of Dutch universities by combining information extracted from their doctoral dissertations (or "theses", for brevity), with information on their publications and the publications of their supervisors, before and after their graduation year. The dataset we produce follows the publishing activity of doctoral graduates from Dutch universities in years 2000-2020, from the graduation year (or before, where applicable) until 2022. Biographical information includes, with various degrees of accuracy, the graduates' name, gender, and country of origin, as well as a unique ID for the university granting the degree and the graduation year. The dataset contains similar information for the graduates' supervisors, as well as information on the scientific or technical fields and contents of both dissertations and publications, as well as their forward citations.

To the best of our knowledge, the only dataset comparable to ours has been produced (at the same time, but independently from us) by Mulders et al. (2024). Similar, but less detailed datasets for France and Germany have been produced, respectively, by Corsini et al. (2022) and Buenstorf et al. (2022). Otherwise, most data sources on doctoral graduates' career consists of national or international surveys,

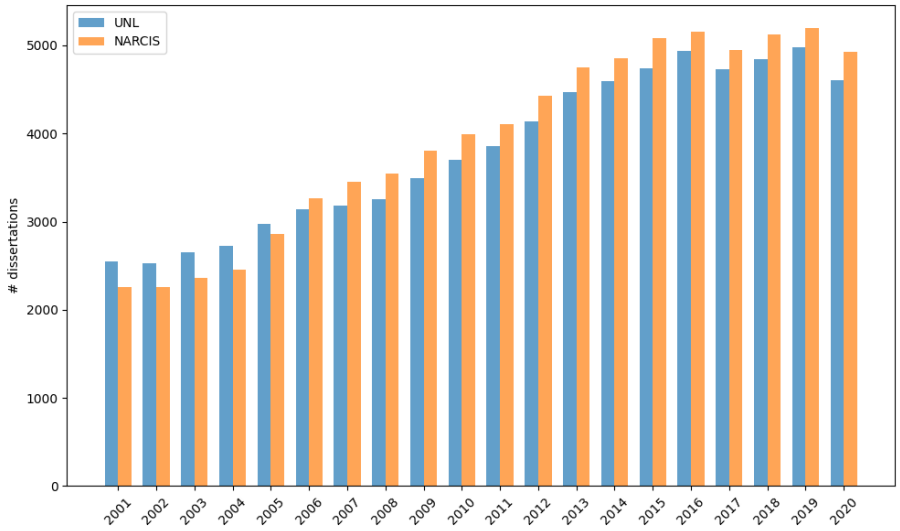


Figure 1: Number of dissertations by year, 2000-2020: NARCIS vs UNL count

Sources: NARCIS archive; Universiteiten van Nederland (UNL).

which do not come with bibliometric information nor – in most cases – with any longitudinal dimension and/or access to individual records.[6]

4.1 Dissertation data

Concerning the doctoral theses, our source is NARCIS (National Academic Research and Collaboration Information System), an online portal for searching Dutch scientific research publications and data created and maintained, until recently, by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NWO), through their joint Data Archiving and Network Services (DANS). NARCIS – now dismissed – operated by harvesting and harmonizing the information made available by the Dutch university libraries on their own web portals, including their Electronic Dissertations and Theses (EDT) repositories.[7]

The total number of dissertations in NARCIS for 2000-2020 is 81571 (before deduplication). Figures

[6]For an example of international survey, see the OECD's database on the Careers of Doctoral Holders (CDH; https://www.oecd.org/innovation/inno/careers-of-doctorate-holders.htm, last visited: October 2, 2023. At the national level, for the Netherlands, see the Survey of Doctorate Recipients (SDR; https://www.nsf.gov/statistics/srvydoctoratework/, last visited: October 2, 2023), conducted since 1973 and focused on the STEM disciplines, whose information is a useful complement to ours.

[7]Started in 2004, but including information from the 1990s, NARCIS has been decommissioned and taken offline in July 2023. A new portal is now maintained by the UKB - a partnership of Dutch University Libraries and The Royal Library of the Netherlands - and SURF, an ICT cooperative of Dutch education and research institutions.

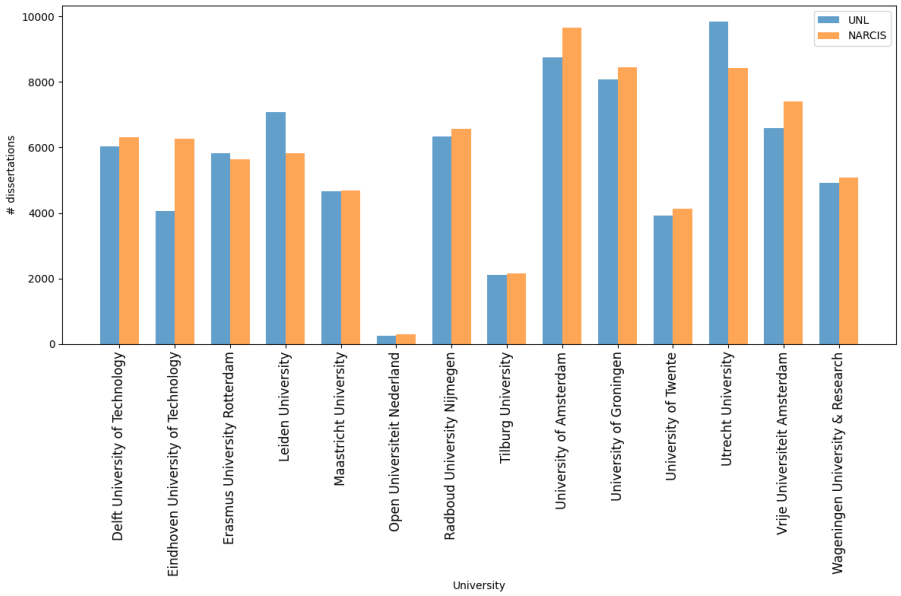


Figure 2: Number of dissertations by university, 2000-2020: NARCIS vs UNL count

Sources: NARCIS archive; Universiteiten van Nederland (UNL).

1 and 2 report their distribution over time and per university. By comparing these descriptives to those released by UNL, the Dutch universities' association, we notice that, for the period of our interest, some universities are under-represented and others slightly over-represented (see Figure A.1 in Appendix A.1). The under-representation cases are mostly concentrated in the early 2000s, possibly due to some technical issues with NARCIS's access to universities' portals. Instead, the over-representation cases are both less notable and entirely due to some duplication of the observations in NARCIS (multiple records for the same theses), which we detected and corrected.

The raw data we extract from NARCIS come with a number of limitations. The most important one is the absence of a disciplinary classification for the theses as well as, for around 68% cases, of the first names of students and/or supervisors (in these cases, we have only the initials). While we can and do remedy to these drawbacks, this comes at a cost in terms of loss of observations. In addition, as it is common with EDT repositories, no information is provided on the graduates' and supervisors' gender and country of origin, for which we had to find complementary data sources. In what follows we describe synthetically our main data treatment operations, with all details in Appendix A.

Concerning the disciplinary classification of the theses, we first distinguish between STEM and non-

STEM ones, to retain only the former. To this end, we rely on the classification system of Theses.fr, a French EDT repository that provides circa 222.000 dissertations in English, classified by field (Corsini et al., 2022). In particular, we consolidate the Theses.fr classification into a STEM/non-STEM binary variable that we associate to the dissertations' titles. We then use the resulting dataset to train a machine learning algorithm and apply it to NARCIS titles, most of which in English (for details: Appendix A.2). In this way we identify around 54.000 STEM versus 27.000 non-STEM theses (see Figure 3), with a very high precision rate (see Table A.2 in Appendix A.2). Since the former are the only ones we retain for our analysis, from now on, when speaking generically of theses, we will exclusively mean the STEM ones.[8]

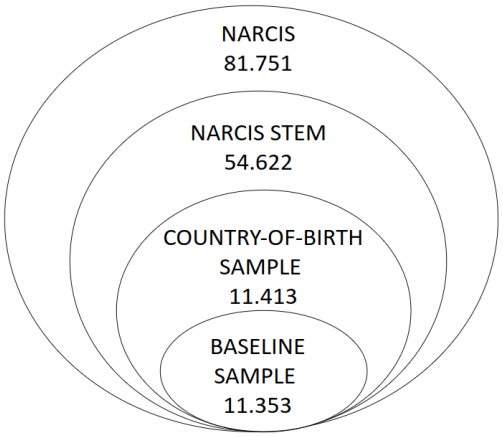


Figure 3: Venn diagram of our sampling process, from the NARCIS archive (2000-2020) to our baseline sampling, based on retrieval of information on graduates' country of birth and manual corrections for matches to Scopus authors

When it comes to a more fine-grained classification, unfortunately, the algorithm based on Theses.fr suffers from lack of precision. As a remedy, we make the assumption that PhD students conduct their research in the same field of their supervisors. Hence, we first classify the supervisors according to the disciplinary fields of their publications (as provided by Elsevier Scopus, see below), and then extend the classification to their doctoral students' theses. Our preferred classification is a rather broad one, with only three macro fields (Health Sciences, Life Sciences and Physical sciences, the last including also Engineering). These can be broken down into 26 subfields, but at the cost of having the same

[8]Theses.fr and its classification system are maintained by ABES, the Agence Bibliographique de l'Enseignement Superieur. For more info: https://abes.fr/, last visited: October 17, 2023

supervisors publishing in many of them, which make the classification less precise (see next section and Appendix A.6).

Regarding the country of origin of the PhD graduates, we leverage the online availability of approximately 36,000 STEM theses in their original .pdf format. For administrative reasons, all Dutch theses contain a short biography of their authors. This allows us to extract information on the doctoral graduate's city and/or country of birth for around 11,413 cases, which constitutes our "country-of-birth" sample (see again Figure 3). Notice that no thesis reports the country of birth of supervisors and we preferred not to assign one on the sole basis of name analysis, as in other papers on migration and innovation (for a discussion, see: Coda-Zabetta et al. (2022, 2023)).

A preliminary analysis of the year distribution of such "country-of-birth" sample indicates no significant sampling bias (see Figure A.6 in Appendix A.3). In contrast, the distribution by university is quite uneven, with coverage varying from as high as 50% in some cases to none in others (see Figure A.7 in Appendix A.3). Informal discussions with Dutch academics - however - suggest that the treatment of foreign-born doctoral students does not vary across universities.

When compared to official statistics, the "country-of-birth" sample appears to be representative in terms of the proportion of foreign versus Dutch doctoral recipients per year and across universities (see Figures A.8 and A.9 in Appendix A.3). However, it over-represents students from outside the EEA relative to EEA ones, which we define as those countries that were part of the European Economic Area prior to 2000, plus Switzerland and the countries accessing the EU in 2004 or later. Based on comments we received from university administrators and students, this over-representation may be due to a certain number of students born outside the EEA who acquired a EEA nationality before moving to the Netherlands, based on legislation that grants this possibility to children or even grandchildren of old European emigrants (as with, for example, the Italian nationality law of 1992; see: Blanchard, 2025).[9]

From the same .pdf files used to built the "country-of-origin" sample, we also extract the first

[9]Concerning for the countries that joined the EU in the 2004 enlargement and the following ones, we treat them as part of the EEA only from the year in which the Netherlands lifted any entry restrictions for their citizen, in compliance with Free Movement principle, namely 2007 for the 2004 enlargement and 2014 for the 2007 enlargement (Kahanec et al., 2016). Since our sample does not extend beyond 2022, Croatia - whose citizens fully enjoy Free Movement since that year - is never considered in the EEA. See Table C.4 for the list of countries by EEA enlargement and the corresponding years when the Netherlands lifted labor market restrictions. For a robustness check using the EU enlargement year—rather than the restriction-lifting year—as the transition point for defining the EEA boundaries, see Appendix **??**.

names of graduates and supervisors, when missing in NARCIS. When missing also in the .pdf files, we recover them through web scraping or via the graduates' or supervisors' publications (additional details in Appendix A.7). Based on the first names so retrieved, we determine the gender of doctoral recipients. To do so, we rely on the World Gender-Name Dictionary (WGND 2.0), developed by the World Intellectual Property Organization for the gender analysis of inventors worldwide (Martínez et al., 2021). Last, accessing the .pdf files also allows us to identify the source of funding, when acknowledged by the graduates and corresponding to a well-known public programme.

4.2 Publication data

After cleaning, filtering and enriching the NARCIS archive, we match all the 11.413 graduates and supervisors in our "country-of-birth" sample to the authors of STEM publications in Elsevier Scopus. This is a commercial bibliographic repository that comprises more than 26.000 active peer-reviewed journals and around 11.7 million conference papers dating back to well before our period of interest, with information on their authors[10] The matching is based on a combination of surnames and first and middle names (or, absent them, their initials). One feature of Scopus of particular importance for this matching is the presence of a unique ID for each author in the database. This both allows us to track the mobility of authors across institutions and countries (provided that they have published) and simplifies the matching task. For each STEM doctoral graduate, we first match and extract from Scopus the authors of articles and conference proceedings published within a narrow time window around the dissertation year (three years before/after it). Based on this, we refine the match by using more detailed information, including (from NARCIS) the names of each graduate's supervisor, those of fellow graduates from the same year and with the same supervisor, and the university of graduation; and (from Scopus) on the authors' institutional affiliations and the names of co-authors of each publication. Using a hand-curated subset of matched data, we develop and validate a random forest algorithm, achieving a precision rate of 87% and a recall rate of 86% (see A.4 for further details and alternative methodologies). Applying this algorithm to the full database allows us to use the Scopus author ID to retrieve all publications of validated matches beyond the original matching interval (all detail in Appendix A.4). It is important

[10]For more info see: https://www.elsevier.com/?a=69451, last visited: October 23, 2023.

to stress that the precision rate of our matching varies across countries of birth, with that for China being lower than average. This is a well-known issue for all researchers engaged in matching exercises like ours, due to the high number of homonyms in and the size of the Chinese STEM community (and population in general; see Lissoni and Miguelez, 2024).

We rely on Elsevier Scopus also for retrieving the supervisors' publications. This is an easier task, as supervisors typically exhibit a higher publication output than their doctoral graduates, and that we know for sure (thanks to NARCIS) at least one of the universities to which they have been affiliated throughout their careers. We therefore validate the original name-cum-surname matches using a triple-criteria approach: (1) co-authorship with students (via their NARCIS records of supervised graduates); (2) co-authorship with colleagues (other academics associated with the thesis in question, for example those seating in the PhD defense committee, on which NARCIS also provides information); and (3) affiliation at the time of the dissertation. Overall, this process identifies 2,622 supervisors with publication, corresponding to around 80% of doctoral recipients (see Appendix A.5).

Once completed these operations, we exclude 60 observations concerning graduates matched to four or more Scopus authors each (51 of which Chinese), while we manually disambiguate those matched to two or three Scopus authors. This leaves us with 11.353 graduates, either matched to a Scopus author or not, which amount to 20% of the whole STEM graduates between 2000 and 2020. They constitute our "baseline sample" , to be used for regression analysis (and subject to further reductions in size due to missing values in some variables; see again Figure 3)).

For what concerns its composition, our baseline sample includes 6.531 native Dutch graduates and 4.822 foreign-born, 1.745 born inside the EEA and 3.077 outside it, as defined in footnote 10 above. Gender-wise, female graduates are 42% overall, with very similar percentages for natives and foreigners. The average age at graduation is also similar across the two groups (32.7 years for foreign-born, 32.46 for natives–see Figure in Appendix 6).

Figure 4.2 presents the trend for the top five countries of birth by count of foreign doctoral graduates in our baseline sample (see Table C.3 in Appendix C for the top twenty countries of origin). Consistent with previous research (Rud et al., 2015), we find that the growth of foreign doctoral recipients in

the years 2010s is predominantly driven by Chinese graduates. In particular, this increase is largely attributable to the students funded by the Chinese Scholarship Council (CSC), which is acknowledged by one third of the 676 Chinese graduates in our baseline sample report CSC funding. The program is widely understood as a mechanism to encourage, and sometimes require, recipients to return to China upon completion of their studies (Fedasiuk, 2020)[11].

In our data, CSC-funded graduates exhibit much higher return migration rates compared to non-CSC ones. Within four years of graduation, 52% of CSC-funded Chinese PhD recipients pursuing academic careers return to China, compared to 31% of those without CSC support. Extending the observation window to seven years increases return rates to 70% for CSC recipients and 45% for non-CSC graduates.

The main difference between foreign-born and native graduates concerns their distribution across academic fields (Figure 5). Consistent with the literature on other countries, foreign graduates are over-represented in the physical sciences (65.5%, versus 35.1% for natives). Conversely, 47.7% of natives hold degrees in health sciences (16.8% for foreign-born) (Kolman, 2024). When splitting the foreign-born between EAA and non-EEA the asimmetry increases, with the share of physical sciences graduates rising to 69.7% for the non-EEA (58% for the EEA); and the share in health science going down to just 15.8% for the non-EEA (21% for the EEA). Gender differences are also notable, with the female graduates over-represented in health sciences (44% of degrees versus 27.7% for males); and under-represented in the physical sciences (34.7% versus 57.5%). Summary statistics for all variables – including missing values – and group-specific averages are presented in Table C.1 in Appendix C (see Table C.2 for CSC vs non-CSC Chinese graduates statistics).

[11]Established in 1996, the CSC primarily supports doctoral and postdoctoral studies. By 2018, approximately 7% of Chinese PhD graduates abroad had received CSC funding. Within the CSC, programs such as the National Construction High-Level University Postgraduate Program (NCHUPP) impose a mandatory return period of at least two years, while others -such as the National Outstanding Self-Funded International Student (NOSIS) - simply encourage it. Some sources indicate that, when funded otherwise, Chinese students do not exhibit an especially high home return rate; see for example (Zwetsloot et al., 2020).

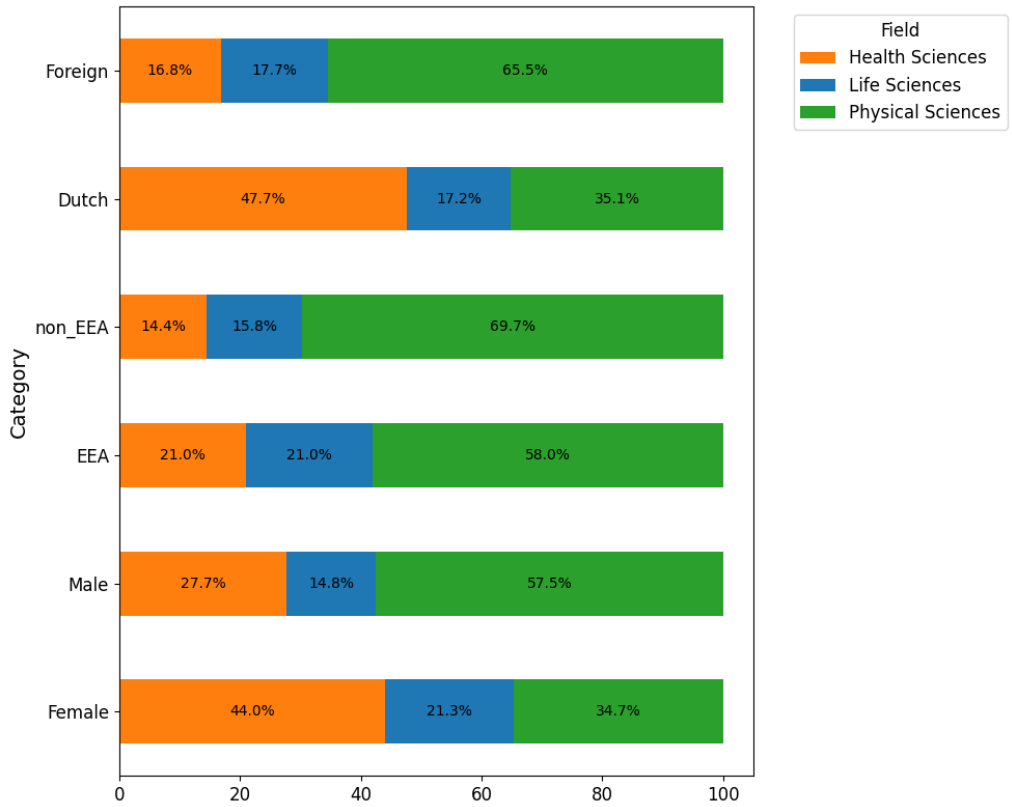


Figure 5: Percentage distribution of doctoral recipients by field and category, 2000-2020

Categories:
✓ Foreign-born vs Dutch ✓ EEA vs non_EEA (foreign-born only) ✓ Gender

Inspecting the publishing activity of the graduates in our sample requires first to distinguish between the publications produced before, during, and after the doctoral studies, also in view of the econometric analysis.

Based on estimates by Nuffic, we assume an average duration of PhD studies of four years, including the graduation year.[12] This allows us to draw a publication timeline centered on the graduation year, denoted by g in Figure 6. Based on this, we consider all publications dated four or three years before g as the output of research the graduate has undertaken before starting the PhD (*pre_PhD*); and those from two years before to one year after g as produced during the PhD (in red). As for the successive publications, we limit our analysis to those dated from two to four years after graduation, which we label *Post_Doc*, being – most likely – the output of research undertaken as part of a postdoc contract.

[12]Source: Nuffic, the Dutch organization for internationalization in education: https://www.studyinnl.org/dutch-education/phd, accessed January 7, 2024.

The main reason for this limitation consists in mitigating the right truncation bias due to our Scopus data ending in 2022. For robustness check purposes, however, we replicate our results using an extended seven-year window, at the cost of losing observations, but we no noticeable differences in results (see Appendix 4.2).

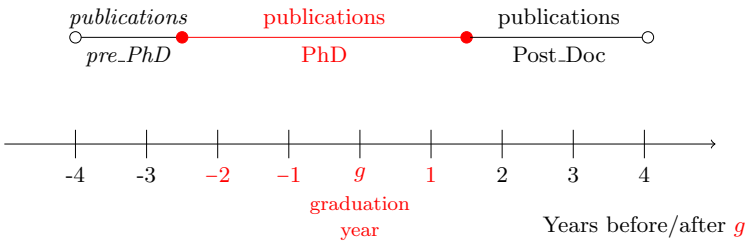
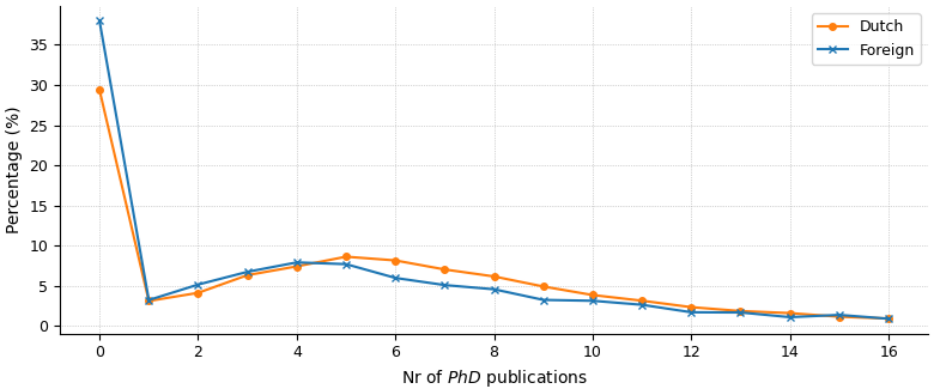


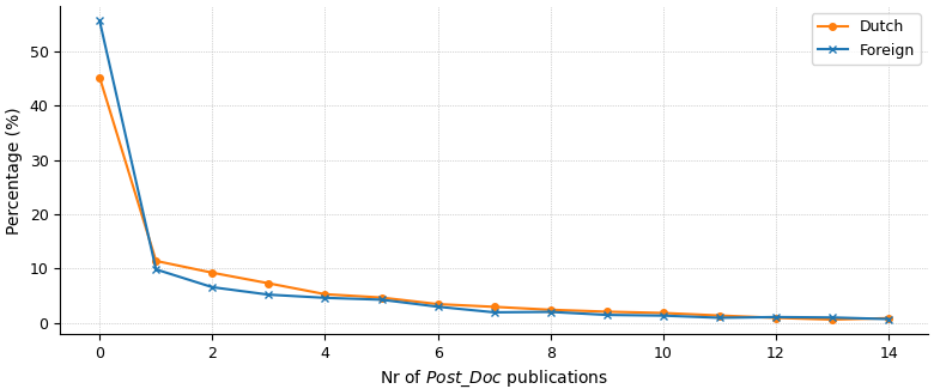
Figure 6: Doctoral graduates' publication timeline and classification

Note: We assume the doctoral studies to last four years and distinguish between *pre_PhD*, *PhD* and *Post_Doc* publications, based on the publications year (respectively: $\in [g-4, g-3]$; $\in [g-2, g+1]$; and $\in [g+2, g+4]$, where g is the graduation year)

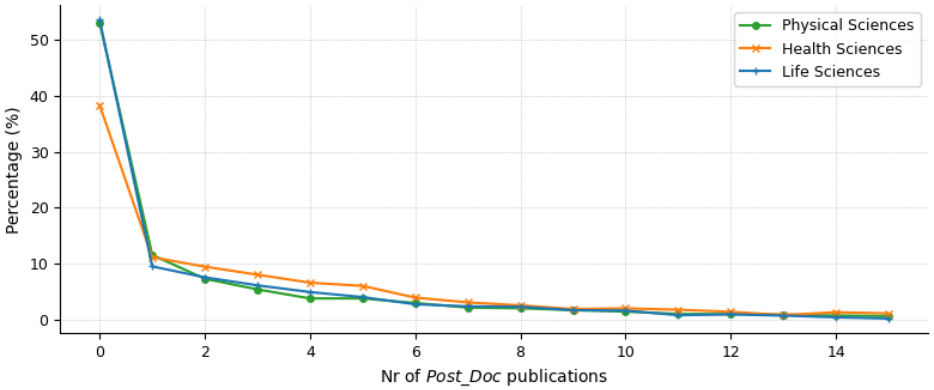
Fifty-five percent of graduates have no *pre_PhD_publications*, with a higher incidence among foreign-born graduates (61.12%) compared to natives (50.42%). On average, native graduates produce 0.12 more *pre_PhD_publications* than their foreign-born counterparts (1.26 versus 1.14). While natives exhibit a slightly greater average number of pre-PhD publications, the distribution tails converge rapidly, indicating similar publication counts among the most productive individuals in both groups (see Figure 6 in Appendix C).



(a) Foreign-born vs. natives (PhD period)



(b) Foreign-born vs. natives (Post-doc)



(c) By disciplinary field (Post-doc)

Figure 7: *PhD* (2000-2020) and *Post_Doc* publications frequency (2000–2018)

Note: We assume the doctoral studies to last four years and distinguish between *pre_PhD*, *PhD*, and *Post_Doc* publications, based on the publication year (respectively: $\in [g-4, g-3]$; $\in [g-2, g+1]$; and $\in [g+2, g+4]$, where g is the graduation year).

Figure 7a, trimmed at the 95th percentile for clarity, indicates that the share of foreign-born graduates without any *PhD_publications* is around 7% higher than that of native counterparts. Among those with five to nine publications, the curve for native graduates is slightly more pronounced, representing roughly 38% of all observations. Beyond this range, the distributions for both groups converge rapidly, with a minor reversal in their relative positions at the end of the right tail.

Figure 7b, similarly truncated, reveals that the proportion of foreign-born graduates with no *Post_Doc* publications exceeds that of natives by about 10%. Native graduates are also more represented among those with one to three publications (approximately 30%). As with PhD publications, the two curves align more closely at higher publication counts, with a modest inversion at the end of the right tail.

4.3 Estimation strategy

We estimate five cross-sectional equations, each one addressing a specific research question, with individual graduates (i) as our observations. In all cases but one, the dependent variable is a binary one and–if not otherwise specified–we apply a linear probability model for estimation (see Appendix 4.9.1 for robustness checks based on logit regressions)[13]

The first equation examines whether foreign graduates differ from native (Dutch) ones in their decision to pursue an academic path via a postdoctoral position. Specifically, the dependent variable *Post_Doc* takes value 1 if the individual has a least one *Post_Doc* publication (0 otherwise), as defined in Figure 6 above:

$$\Pr(Post_Doc_i = 1) = \alpha + \beta\, foreign_i + \gamma X_i + \phi_{(u,g,f,s)} + \epsilon_i \tag{1}$$

where *foreign* is binary variable taking value one when i is foreign-born (zero if born in the Netherlands). In an alternative specification, we replace this variable with two different ones, respectively *EEA* (equal one if i was born in a different European Economic Area country than the Netherlands) and *non_EEA* (equal one if born outside the EEA). The vector of controls X_i includes i's gender (*female*), age at graduation (*age*), and publications before and during the doctoral studies (*pre_PhD* and *PhD*), while ϕ is a vector of fixed effects for, respectively, i's PhD-granting university u, graduation year (cohort) g ,

[13]Regressions from equation (1) to (4) are estimated with *fixest* R package to include the supervisor fixed effects (Bergé, 2018).

field of study f, and supervisor s. ϵ is the error term.

To assess academic productivity more directly, we replace the binary dependent variable with the count of i's *Post_Doc* publications and the linear dependent model with a Poisson one:

$$\mathrm{E}(Post_Doc_Publications_i) = \exp(\alpha + \beta\, foreign_i + \gamma X_i + \phi_{(u,g,f,s)}). \tag{2}$$

Next, we explore the factors affecting the doctoral graduates' stay or leave decisions, after completing their studies. To do so, we first introduce a new binary variable, *Leave*, which equals 1 if i's institutional affiliation on her most recent *Post_Doc* publication is a foreign one, conditional on having at least one such publication. If multiple publications exist in the same year, we consider as most recent the one with the latest publication month.

Based on this definition, our baseline sample includes 1.151 graduates who both published and left the Netherlands. Of these, 387 are natives and 764 are foreign-born, of which 271 are from the EEA and 493 non-EEA. Figure 4.3 reports the share of migrants in each group (EEA, non-EEA, and natives). As expected, the migration rate is significantly higher among the foreign-born graduates (40% on average over the entire period versus 15% for the natives), but with considerable fluctuations across years (while the values for the natives are relatively stable).

We investigate these patterns by estimating the following equation:

$$\Pr(Leave_i = 1 \mid Post_Doc_i = 1) = \alpha + \beta\, foreign_i + \gamma X_i + \phi_{(u,g,f,s)} + \epsilon_i \tag{3}$$

where the vector controls X_i now includes an additional binary variable, *prior_foreign_affiliation*, which takes value one if at least one of her *pre_PhD* and *PhD publications* (with the exception of those published in $g+1$), i results affiliated to a foreign institution.[14] As in the previous equations, we estimate an alternative specification, with the two dummies *EEA* and *non_EEA* replacing *foreign*.

Finally, we analyze the destination countries of those who, either native or foreign-born, leave the Netherlands after graduating (that is, with *Leave* = 1). In particular, we distinguish between the coun-

[14] We exclude $g+1$ from this range to limit the risk of misclassifying an early postdoctoral transition as a pre-graduation characteristic.

tries whose academic labor markets compete with the Dutch one for the attraction of foreign talent (in particular, the United States and other Anglophone countries hosting top universities), the EEA countries (some of which also compete for talent with the Netherlands) and – in the case of the foreign-born – the graduates' countries of origin. More precisely, for the native graduates, we consider two possible foreign destinations, namely: (i) *Anglosaxon* countries (United States, Canada and Australia); and (ii) the *EEA* (excluding the Netherlands). For foreign-born graduates, we introduce the additional category of *Origin* countries, which takes value one in case of return migration (including to any Anglosaxon or EEA country), and zero otherwise.

Figure 4.3 presents the proportions of leavers by such destination categories. Among Dutch doctoral recipients, the majority relocate to other EEA countries (55.8%) or to Anglo-Saxon countries (38.2%; the remaining 6% go to other destinations). For EEA-born graduates, 37.5% return to their country of origin, 40.7% migrate to other EEA countries, and 18.6% relocate to Anglo-Saxon countries. In contrast, return migration is notably higher among non-EEA graduates, with 49.7% returning to their home countries. Only 26.6% of non-EEA graduates move to other EEA countries, and 15.9% migrate to Anglo-Saxon countries.

We investigate these patterns by estimating the following equation:

$$\Pr(Country_set_i = 1 \mid Post_Doc_i = 1) = \alpha + \gamma X_i + \phi_{(u,g,f,s)} + \epsilon_i, \tag{4}$$

where *Country_set* is a binary variable indicating migration to a specific country or group of countries. We run separate regressions for Dutch versus foreign-born graduates and for different specifications of the dependent variable.

One set of additional regressions analyzes the effect of immigration policy changes on the probability of doctoral recipients leaving the Netherlands. We modify equation (3) by incorporating it in a dynamic difference-in-differences (DID) framework based on repeated cross-sectional data (as in Waldinger's 2010 study on the impact of the expulsion of Jewish professors on PhD student outcomes in Nazi Germany). In particular, we focus on the introduction of the Knowledge Migrants Scheme in 2004 and of the "Researcher EU 2005/71" in 2018, as described in Section 3.

In order to do so, we reduce our sample to a limited number of student cohorts (graduation years g from 2000 to 2007 for the study of the Knowledge Migrants Scheme; and from 2000 to 2011 for the Researcher EU 2005/71) and to the foreign-born graduates. Of these, we distinguish between those from outside the EEA, whose permanence in the Netherlands depend on obtaining a residence permit and was therefore affected by the policy changes, and those from within the EEA, whose permanence in the Netherlands is regulated by the Free Movement of Persons principle and therefore unaffected. In the DID, the former represent our treated group (with the treatments being administered, respectively, in 2004 and 2008), and the latter the control group. As for the choice of the cohorts to include in the sample, this descends from our assumption of the doctoral studies lasting - on average - four years. This lead us to assume that all cohorts before the introduction of the policy change (from 2000 to 2003 for the Knowledge Migrants Scheme; and to 2007 for Researcher EU 2005/71 provision) were unaffected by it, the affected ones being instead those up to three years after the policy change (respectively, 2007 and 2011). These cohorts, in fact, comprise the students who enrolled before the policy change was introduced (which allows us to assume they did not choose to study in the Netherlands in view of benefiting from it), but graduated afterwards, thus finding themselves facing lower costs for staying in the Netherlands, especially if recruited as postdoc by a university or research institution. As for the successive cohorts, they also faced these lower costs, but were aware of it before moving to the Netherlands, which may change their composition in terms of leave-or-stay preferences.

We therefore estimate the following equation:

$$\Pr(Leave_i = 1 \mid Post_Doc_i = 1) = \alpha + \beta non_EEA_i + \sum_{\substack{g=2000 \\ g \neq 2004(08)}}^{2007(11)} \theta \cdot non_EEA_i + \gamma X_i + \phi_{(u,g,f,s)} + \epsilon_i \qquad (5)$$

which differ from equation (3) in two respects. First, the *foreign* dummy is replaced from the start by the *non_EEA* one (=1 for the non-EEA graduates and =0 for the EEA ones). Second, it includes the interaction terms between *non_EEA* and the student cohorts g comprised in the sample (from 2000 to: 2007 when estimating the effect of the Knowledge Migrants Scheme; 2011 for the Researcher EU 2005/71 reform), with the cohort coinciding with the policy change as the reference yar (respectively: 2000 and

2004).

Our estimation strategy comes with a number of limitations. First, and most important, when moving from equations (1) and (2) to equations (3), (4) and (5) we encounter a sample selection problem (Heckman, 1979). Namely, to the extent that the foreign-born versus native status (or the foreign-born's specific country of origin, EEA versus non_EEA) affects both the probability to undertake an academic career (i.e. to have at least one *Post_Doc* publication) and the subsequent productivity, the results in the second group of equations may be biased. This may occur, for example, if obtaining a PostDoc requires overcoming a different skill threshold for the foreign-born versus the natives (or the EEA versus the non_EEA), or individuals with the same skills but belonging to different groups face different costs or incentive thresholds for deciding to get one. Intervening on this problem with an Heckman correction requires at least one regressor in the selection equation not affecting the outcome one, which so far it proved impossible to find (Puhani, 2000). For this reason, we opted for estimating equations (3) to (5) with the simple warning that they are conditional on publishing after the doctoral graduation.

Second, the best estimation strategy for the research question addressed by equation (4) would have possibly consisted in using a multinomial logit or probit regression, with the graduates leaving the Netherlands choosing between different destinations. Two obstacles, however, stand in the way. First, the choice of destinations may be hierarchical (with, for example, the return decision coming first), which would violate the independence of irrelevant alternatives hypothesis (Hausman and McFadden, 1984). Second, we have a high number of fixed effects, which create problems of convergence when running maximum-likelihood estimations. In any case, we run a number of robustness checks using a multinomial logistic regression (see Appendix 4.9.2), with no appreciable differences in our results.

Last, the high number of supervisor fixed effects expose our estimations to a risk of overfitting (see Figure A.14 in Appendix A.5 for details). For this reason, in some estimates, we are forced to drop them and consider, in their place, a time-varying indicator of the *supervisor_productivity*, defined as the number of publications the student's supervisor has in $[g-4, g+1]$.

5 Results and discussion

Table 1 presents our estimation results for equation 1. In columns from (1) to (3), the explanatory variable of interest is *foreign*, which in columns (4) and (5) we replace with the two dummies *EEA* and *non_EEA* (with Dutch natives always the reference case). Absent the controls for the disciplinary field (column 1), the foreign-born graduates' probability to produce at least one *PostDoc* publication is around 5.75 percentage points less than that of natives. Once introduced these controls (column 2), the estimated coefficient for *foreign* goes down to 4.57. When further controlling for the supervisor fixed effect (column 3) the coefficient halves and becomes not significant. A foreign-born vs natives difference reappears in column (4), but only for the graduates from EEA countries and with a smaller size (less than 4 percentage points). The difference between the EEA and non-EEA coefficients is statistically significant only when Chinese graduates are excluded (in column 5) or when a different EEA specification is applied (as in Table 4.32 in Appendix 4.5). Overall, these results suggest that, once considering all composition effects (field and supervisor), the foreign-born graduates from Dutch universities have the same or almost the same probability to work as PostDoc than their native homologues. This indicates no or a very limited self-selection into academic careers (for EEA graduates only). This is both interesting *per se* (as it signals no major differences in access to academic careers) and reassuring from the methodological viewpoint (it implies no or limited truncation bias in the estimates for equations from 3 to 5; see the discussion at the end of the previous section).

When examining the control variables at the individual level, we notice that female graduates' probability to produce at least one *PostDoc* publication is lower than that of males, with a difference of around 2 percentage points, with the effect turning not significant once the Chinese graduates are excluded from the sample (see Table 4.1 in Appendix 4.1). This suggests that the effect is likely caused by the lower precision of the publication matching for Chinese graduates.[15]

[15]When Chinese are split into those who were financed by the CSC and those who were not, and only the latter are excluded, the difference between the two coefficients is again not significant, see Table 4.5 in Appendix 4.1.3

Table 1: Probability to Take a Postdoctoral Position by Foreign Status, 2000-2018

Dependent Variable:	Post_Doc				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	-0.0575***	-0.0457**	-0.0256		
	(0.0183)	(0.0174)	(0.0218)		
non_EEA				-0.0166	0.0070
				(0.0291)	(0.0234)
EEA				-0.0396**	-0.0379*
				(0.0157)	(0.0182)
female	-0.0172*	-0.0196*	-0.0257	-0.0248	-0.0276
	(0.0094)	(0.0102)	(0.0160)	(0.0155)	(0.0184)
age	-0.0017**	-0.0019*	-0.0034***	-0.0034***	-0.0029**
	(0.0008)	(0.0009)	(0.0008)	(0.0008)	(0.0009)
pre_PhD_publications	-0.0066	-0.0061	-0.0061	-0.0060	-0.0148
	(0.0109)	(0.0108)	(0.0139)	(0.0138)	(0.0123)
PhD_publications	0.0291***	0.0262***	0.0232***	0.0232***	0.0334***
	(0.0027)	(0.0025)	(0.0039)	(0.0039)	(0.0051)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	6,639	6,639	6,258
R^2	0.22026	0.21504	0.48533	0.48545	0.52666
Within R^2	0.20015	0.18111	0.15112	0.15132	0.20081

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. Column (5) excludes Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant, while significant at 5% level in column (5).

One additional year of age also diminishes the probability to publish as a PostDoc, of around 0.3 percentage points and the results do not change in any meaningful way when introducing a quadratic term (see Table 4.35 in Appendix 4.6). *pre_PhD_publications* appear not to be play a role, while the opposite holds for *PhD_publications*. In this case, each additional publication increases the probability to

publish as a PostDoc of around 2-3 percentage points. This last results suggest that, after controlling for composition effects, the probability that a doctoral graduate gets a PostDoc contract is mostly correlated to his/her productivity as a student.

Results are robust to extending the time period for the definition of *Post_Doc* publications (i.e. $[g+2, g+7]$ instead of $[g+2, g+4]$; see Table 4.13 in Appendix 4.2), and to weighing the *pre_PhD_publications* and *PhD_publications* for the number of forward citations (see Table 4.18 in Appendix 4.3.1). The same holds when replacing the 3-field classification of theses with a more granular one (see Table 4.25 in Appendix 4.4.1) or adopting a logit model specification instead of OLS (see Table 4.43 in Appendix 4.9.1).

Table 2 reports our estimates for equation 2, based on a Poisson regression. The coefficients are exponentiated, so to obtain – for each variable – its Incidence Rate Ratio (IRR). This indicates the variable's multiplicative effect on the expected publication counts, *ceteris paribus*. Based on column (1), the foreign-born graduates are expected to publish approximately 0.01 less articles than natives, and the estimated coefficient is not statistically significant. When the field fixed effects are included in column (2), the IRR for the foreign status rises to 1.04; and when supervisor fixed effects are added it goes to 1.1 and becomes statistically significant at the 5% level. However, it is again non significant when Chinese graduates are excluded from the sample (see Table 4.2 in Appendix 4.1).[16]

Column (4) further distinguishes between EEA and non-EEA foreign graduates. The non-EEA graduates appear to be more productive, but the difference between the two coefficients is not significant. This result is consistent across all the robustness checks in Appendix 4 and the effect disappears once Chinese graduates are excluded in column (5).

Overall, these results confirm that no major difference exists, in terms of productivity as PostDocs, between native and foreign graduates nor between EEA and non-EEA ones. Once again, this downplays the risk of a major truncation bias for the estimation of equations from 3 to 5.

Across all specifications, female graduates consistently exhibit lower productivity than males. The IRR for *female* around 0.8, which implies that women produce only about four fifths of the men's *PostDoc* publications. These estimates are highly statistically significant. Neither the coefficient for age effects

[16]The significance appears to be equally driven by Chinese graduates that are financed by the CSC and those who are not (see Table 4.6 and Table 4.9 in Appendix 4.1)

nor that for *pre_PhD_publications* are not statistically significant. Instead, the number *PhD_publications* once again strongly predict the future *PostDoc* publications, with one of the former increasing of more than 0.07 the expected number of the latter, once considered all fixed effects.

Table 2: Postdoctoral Productivity by Foreign Status (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications (IRRs)				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	0.9912	1.0420	1.1002**		
	(1.0867)	(1.0844)	(1.0472)		
non_EEA				1.1407*	1.0703
				(1.0746)	(1.0510)
EEA				1.0466	1.0733
				(1.0549)	(1.0639)
female	0.7861***	0.7817***	0.8301***	0.8323***	0.8369***
	(1.0416)	(1.0306)	(1.0621)	(1.0622)	(1.0644)
age	1.0060*	1.0032	1.0029	1.0028	1.0029
	(1.0034)	(1.0032)	(1.0040)	(1.0040)	(1.0046)
pre_PhD_publications	1.0082	1.0061	0.9951	0.9958	0.9933
	(1.0096)	(1.0083)	(1.0201)	(1.0199)	(1.0199)
PhD_publications	1.0230***	1.0232***	1.0742***	1.0742***	1.0789***
	(1.0027)	(1.0023)	(1.0056)	(1.0057)	(1.0097)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	5,902	5,902	5,510
Squared Correlation	0.16951	0.19521	0.62006	0.62042	0.67116
Pseudo R^2	0.18042	0.19922	0.48212	0.48225	0.48228
BIC	67,061.9	54,591.2	46,687.1	46,687.2	42,318.7

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_fields* are broad disciplinary categories. Fewer observations in columns with supervisor fixed effects are due to Poisson dropping groups with no within-group variation in the outcome. Column (5) excludes Chinese graduates. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (4) and (5) is not statistically significant.

These results hold both when extending to $g+7$ the time period for the *Post_Doc_Productivity* variable (see Table 4.14 in Appendix 4.2) and when classifying theses with more granular fields (see

Table 4.26 in Appendix 4.4.1). When weighing the pre-PhD and PhD publications with citations, the estimated coefficient for age increases to approximately 1.12 and becomes statistically significant across all specifications for PhD citations (see Table 4.19 in Appendix 4.3.1). This suggests that age has a distinct and stronger positive association with citation impact compared to its effect on publication quantity.

Table 3, presents the results for equation 3, based on which we estimate the probability to leave the Netherlands (in short, the migration probability), for both native and foreign-born graduates. Our main variable of interest is *foreign*, which we expect to have a positive coefficient, but we are also interested in the graduates' publication activity before and during the PhD, whose coefficient can be interpreted as indicative of migrants' self-selection, either positive or negative.

As expected, the foreign-born graduates have a much higher probability to leave the Netherlands (conditional on undertaking a PostDoc and publishing) than the native ones. In column (1), the difference is estimated at around 23 percentage points. This, however, is not entirely explained by their migrant status, but also by some composition effects. First, as for previous results, when controlling for the field of study, the difference reduces of two percentage points (column 2), which become seven when controlling also for supervisor fixed effects (column 3). To the extent that supervisors may differ with respect to the extent of their international contacts, these could explain at the same time the attraction of foreign doctoral graduates and their job placement outside the Netherlands.

In Column (4), we distinguish again between foreign-born from *EEA* and *non_EEA* countries. The former's migration probability is around 12 percentage points higher than that of natives, while for latter it is almost 19 points higher, with the difference between the estimated coefficients significant at 10%. Once we exclude the Chinese graduates, in column (5), the difference between the two coefficients is no more significant, but it becomes significant again, at 5%, when we exclude only the Chinese graduates not funded by the CSC, namely those without a return requirement (column 6). Overall, this suggest that differences in migration probability between European and non-European graduates are entirely driven by Chinese graduates, partly due to a measurement error bias, partly to a substantive reason,

namely the implications of CSC funding [17]

Notice that, when moving from the two first equations to the third one, we introduce *prior_foreign_affiliation* as an additional explanatory variable. This turns out to be the strongest predictor in the model, as having a foreign affiliation prior to graduation increases the likelihood of leaving the Netherlands afterwards by approximately 27 percentage points. Nonetheless, the effect does not appear to be primarily driven by foreign or non-EEA graduates. Specifically, in Table 4.40 in Appendix 4.7, we examine interactions between the *foreign*/*non_EEA* and *prior_foreign_affiliation* indicators. The interaction term between *foreign* status and prior foreign affiliation becomes statistically insignificant once supervisor fixed effects are included (see column 3). In contrast, the corresponding interaction involving *non_EEA* status remains statistically insignificant across all specifications. Meanwhile, the main effect of *prior_foreign_affiliation* remains robustly significant and substantively large across all model specifications. This variable might be capturing different features of the graduate programmes attended by each individual, such as his/her insertion in international teams, an exchange or a double degree with foreign universities or, for a foreign one, the persistence of ties with the origin country, based on pre-PhD academic experiences.

The estimated coefficient for *PhD_publications* is always positive and statistically significant effect, which at first sight might signal positive self-selection into migration. Each additional article published during the PhD is associated with an increase in the migration probability of about 0.16 to 0.26 percentage points, depending on the specification. This effect – however – is driven by a small number of individuals in the top 5th percentile of those publishing during the PhD and is not statistically significant once Chinese graduates are excluded from the sample (column 5). [18]

Concerning the control variables, both age and gender appear to be negatively correlated to the migration probability, but the estimated coefficients become statistically insignificant when controlling for the disciplinary field (also when introducing a quadratic term for age; see Table 4.37 in Appendix 4.6).

[17] All results are robust to the changing the definition of EEA (see Table 4.33 in Appendix 4.5).

[18] Up to the 95th percentile *PhD_publications*), the two distributions of those who leave and do not leave the Netherlands are overlapping, see Figure 6 in Appendix C. In Appendix 4.8 we also interact *PhD_publications* with either *foreign* or EEA/non-EEA, and the interaction coefficient turns out to be insignificant once controlling for the supervisor fixed effects (see Table 4.41). This suggests that there is no self-selection for both native and foreign graduates.

Table 3: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
foreign	0.2314***	0.2110***	0.1584***			
	(0.0160)	(0.0146)	(0.0273)			
non_EEA				0.1878***	0.1934***	0.1930***
				(0.0290)	(0.0296)	(0.0286)
EEA				0.1173***	0.1175**	0.1208***
				(0.0354)	(0.0432)	(0.0360)
female	-0.0242***	-0.0075	-0.0011	0.0012	0.0056	0.0072
	(0.0075)	(0.0105)	(0.0114)	(0.0114)	(0.0090)	(0.0100)
age	-0.0019**	-0.0008	-0.0003	-0.0002	-0.0007	-0.0006
	(0.0009)	(0.0010)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0013	-0.0015	0.0017	0.0019	0.0043	0.0036
	(0.0017)	(0.0018)	(0.0036)	(0.0036)	(0.0042)	(0.0041)
PhD_publications	0.0012**	0.0016***	0.0026*	0.0026*	0.0016	0.0017
	(0.0005)	(0.0004)	(0.0012)	(0.0012)	(0.0025)	(0.0024)
prior_foreign_affiliation	0.2871***	0.2929***	0.2706***	0.2700***	0.2510***	0.2619***
	(0.0238)	(0.0296)	(0.0245)	(0.0248)	(0.0239)	(0.0261)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21012	0.20627	0.59089	0.59196	0.59484	0.59528
Within R^2	0.17223	0.15959	0.12736	0.12964	0.11556	0.12132

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 10% level in column (4) and at the 5% level in column (6), but not significant in column (5).

Results are robust to extending the time period for the *Leave* variable to $g+7$ (see Table 4.15 in Appendix 4.2), to considering more granular fields (see Table 4.27 in Appendix 4.4.1), and to using a logit model specification (see Table 4.44 in Appendix 4.9.1). When weighing the pre-PhD and PhD publications with their forward citations (see Table 4.20 in Appendix 4.3.1), the coefficient for PhD

publications becomes significant at 10%, but only in the specification of columns (3) and (4), and goes back to insignificance when we exclude Chinese graduates.

We investigate the destination countries of migrants by splitting our baseline sample into two groups, natives and foreign-born, and by estimating equation 4 with separate regressions.

Table 4 reports the results for the sample including only the Dutch graduates, for three different binary outcome variables, namely: *Leave* (column 1), which is the same as in the previous table and simply takes value 1 if the graduate leaves the Netherlands, conditional on having at least one *Post_Doc* publication; *EEA*, which takes value 1 if the graduate leaves the Netherlands for a EEA country (and zero in case she stays in the Netherlands, return home or move outside the EEA; column 2); and Anglosax, which takes value 1 if the graduate leaves the Netherlands for either the United States, Canada or Australia (column 3). Notice that we do not include supervisor fixed effect to avoid overfitting (see Appendix A.14), and that we replace them with a measure of *supervisor_productivity* (see Appendix 4.4.2 for results with supervisor fixed effects).

Being the sample entirely composed of natives, our main variables of interest are those indicative of possible self-selection into migration, based on their quality as researchers. In particular, the coefficient for *PhD publications* has a positive sign, which becomes significant in column 3 and may be indicative of positive self-selection into migration to the United States or other English-speaking countries, in accordance with Borjas (1991) and Parey et al. (2017). The size of the effect is nonetheless small, with one additional paper authored during the PhD adding slightly more than 0.1 percentage points to the probability of moving. Besides, the publication activity preceding the PhD (*Pre_PhD publications*) is negatively associated with both the probability to migrate in both column 1 and 3). Rather than of negative self-selection, we interpret this last result as indicative of pre-existing research activity within the Netherlands and – therefore – of a higher personal orientation not to leave it, *ceteris paribus*. When we weigh pre-PhD and PhD publications with citations, however all coefficients lose significance across all columns (see Table 4.21 in Appendix 4.3.1).

Table 4: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
female	0.0213*	0.0238**	0.0016
	(0.0106)	(0.0092)	(0.0080)
age	-0.0030***	-0.0013***	-0.0016***
	(0.0006)	(0.0004)	(0.0004)
pre_PhD_publications	-0.0045***	0.0010	-0.0051***
	(0.0013)	(0.0016)	(0.0012)
PhD_publications	0.0010	-0.0004	0.0013***
	(0.0009)	(0.0007)	(0.0003)
prior_foreign_affiliation	0.2157***	0.1052*	0.0949***
	(0.0570)	(0.0483)	(0.0157)
supervisor_productivity	-0.0002	-0.0001**	0.0000
	(0.0001)	(0.0001)	(0.0001)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.07112	0.04015	0.04154
Within R^2	0.04691	0.02001	0.02461

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

As for gender, it appears to bear a positive effect on leaving (column 1) and it is mostly due to female graduates leaving for the *EEA* (column 2). Age equally negatively affects the probability to leave in general and both towards the EEA and Angosaxon countries (once again, the effect is small, with one additional year of age decreasing the migration probability of 0.3 percentage points at most). When considering the effect of having a prior foreign affiliation we got similar results to those of the previous

regressions. The effect of *supervisor_productivity* is negative, but extremely small and statistically significant pnly for migration to the EEA (the same holds when weighing the supervisor's publication with citations; see Table 4.23 in Appendix 4.3.2).

Results mostly hold when extending the time period for the *Leave* variable to $g+7$ (see Table 4.16 in Appendix 4.2), with the main exception being the loss of significance for coefficient of *prior_foreign_affiliation*. This might suggest that having foreign connections matters more for the migration decision early in one's career rather then in the long run. The introduction of supervisor fixed effects cancels the significance of almost all coefficients, likely due to overfitting (see Table 4.30 in Appendix 4.4.2).

Table 5 reports another set of estimation results for equation 4, for a sample including only the foreign-born graduates. In this case we are interested in both the distinction between EEA and non-EEA ones and, as before, in the possibility of self-selection into migration, also relative to the destination (which now includes also the graduate's country of origin, in columns 4 to 6).

The estimated coefficient for *non_EEA* in column 1 is in line with the results in column 5 of Table 5, as it indicates a difference in migration probability between European and non European graduates, of around 6 percentage points. More interestingly, the estimated coefficients in the other columns (non significant in columns 2 and 3, and significant and also around 0.065 in column 4) indicate that the difference is entirely driven by non-EEA graduates' higher propensity to return migration. However, when Chinese graduates are excluded, in column (5), the coefficient becomes not significant. When we re-include in the sample the Chinese graduates funded by CSC (column 6) the coefficient regains some significance (at 10%). Overall, this confirms that differences between European and non-European graduates are entirely due to Chinese graduates, which in turn may result from a combination of measurement error bias (due to the low precision of publication matching for Chinese graduates) and a substantial effect (the presence among Chinese graduates of home-bound CSC funds recipients). Unfortunately, it is hard to tell the two determinants apart.[19]

[19]When we exclude from the sample only the Chinese graduates with CSC funding, the coefficient is smaller (around 0.05) but still significant at 10% (see Table 4.8 in Appendix 4.1). Estimations with multinomial logistic regression produce similar results (see Table 4.49 in Appendix 4.9.2), and the same holds when extending the time period to $g+7$ (see Table 4.17 in Appendix 4.2).

Table 5: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax	Origin		
Sample:		Full		Full	Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.0580**	-0.0299	0.0044	0.0652**	0.0430	0.0602*
	(0.0238)	(0.0306)	(0.0152)	(0.0275)	(0.0336)	(0.0312)
female	-0.0498***	0.0104	-0.0046	-0.0426**	-0.0447**	-0.0471**
	(0.0150)	(0.0094)	(0.0173)	(0.0168)	(0.0199)	(0.0181)
age	0.0039	-0.0003	-0.0036*	0.0066**	0.0072**	0.0067**
	(0.0031)	(0.0015)	(0.0020)	(0.0027)	(0.0027)	(0.0027)
pre_PhD_publications	-0.0023	-0.0068**	0.0018	0.0027	0.0006	-0.0007
	(0.0026)	(0.0030)	(0.0020)	(0.0041)	(0.0056)	(0.0054)
PhD_publications	0.0026**	0.0023	0.0003	0.0002	0.0009	0.0011
	(0.0010)	(0.0014)	(0.0003)	(0.0014)	(0.0030)	(0.0029)
prior_foreign_affiliation	0.3336***	0.0250	0.0618***	0.2393***	0.2031***	0.2231***
	(0.0392)	(0.0150)	(0.0188)	(0.0179)	(0.0253)	(0.0227)
supervisor_productivity	-0.0004	0.0000	0.0000	-0.0004**	-0.0004**	-0.0004*
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.15865	0.02907	0.03720	0.14552	0.12839	0.13404
Within R^2	0.12179	0.00796	0.02048	0.11552	0.09642	0.10228

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

Return migration, however, is mostly explained by the affiliations the doctoral graduates may have had with home universities before or during their studies, as indicated by the large difference in size (and, in one case, significance) of the estimated coefficient for *prior_foreign_affiliation* in columns (1) and (4-6)

versus (2) and (3). Also gender and age play a role, with female and older graduates being more likely to go back to their country of origin. Patterns for *pre_PhD publications* and *PhD publications* are more confused and do not lend themselves to a straightforward interpretation. When we weight them with citations, however, we find hints of a possible, albeit weak negative self-selection for return migrants, as the probability to return home increases of 0.14 percentage points per citation (see Table 4.22 in Appendix 4.3.1). When excluding Chinese graduates, with or without CSC funding, the coefficient does not change, but becomes weakly significant. The *supervisor_productivity* has a very small negative association with return migration, too.

Results are robust to classifying theses with more granular fields (see Table 4.29 in Appendix 4.4.1); while the introduction of supervisor fixed effects cancels – once again – the significance of almost all coefficients besides *prior_foreign_affiliation* (see Table 4.31 in Appendix 4.4.2).

We conclude this section by reporting the estimation results for equation 5, based on a regression sample including only the foreign-born graduates, with a distinction between those born inside and outside the EEA. The exercise is meant to estimate the effects of the two main immigration reforms introduced in the years 2000s and potentially affecting foreign-born, non-EEA doctoral graduates' stay or leave decisions. These are respectively the introduction of the Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008. We test the hypothesis that one or another reform has decreased the non-EEA academic graduates' probability to leave, conditional on publishing after the PhD and relative to the EEA graduates, not affected by the reform.

Figure 11 report the estimated values and 95% confidence intervals for the θ coefficients, which in equation 5 capture the interaction of the *non_EEA* dummy with those for years 2000-2007 in panel (a) (with year 2004 as reference case) and years 2000-2011 in panel (b) (with year 2008 as reference case). (Full regression results are reported in Appendix B.)

In neither cases we find a significant impact of the policy. In particular, no θ coefficient appears to be significant, with the exception of that for year 2003, which is negative and indicative of a temporary reduction in the *non_EEA-EEA* difference of migration probability. Going back to figure 4.3 in the previous section, we notice that this correspond to an anomaly in our data, with an out-of-trend peak

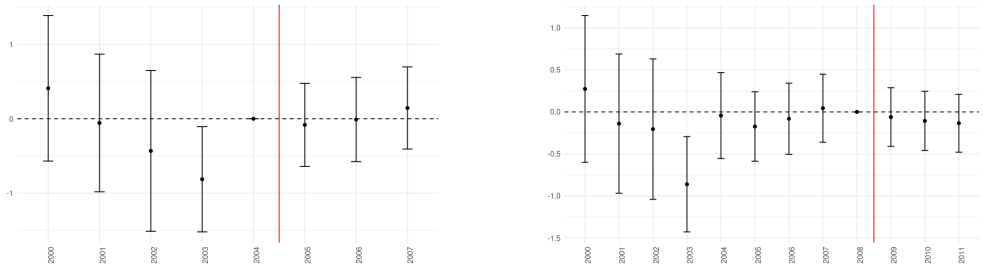
in the emigration rate of EEA graduates and a plunge in that of non-EEA ones, for which we have no ready explanation.[20]

Results are consistent across all the alternative samples, that is with no Chinese graduates with and without CSC funding, as well as to alternative EEA specifications and extended time periods (see Appendix 4.10).

The results for the Knowledge Migrants Scheme is not unexpected, since - as discussed in section 3 - this policy did not target academic researchers, but more generally migrants with high salaries (possibly too high for being at reach of PostDoc researchers). That is, we should intend it more as policy introducing an exception to an otherwise restrictive immigration policy (one directed to well-paid, less politically controversial migrants), than a measure designed for attracting or retaining scientific talent.

The results for the Researcher EU 2005/71 permit, on the contrary, should be taken less for granted. In this case, the doctoral graduates formed – with master ones – the specific policy target. It suggests than the policy either did not reduce substantially the costs associated to staying in the Netherlands, for both the doctoral graduates and the Dutch universities willing to employ them, or that such costs depend less on the bureaucracy associated to obtaining a residence permit, and more to the career prospects, language barriers, or family issues (such as the children's education or the partner's access to the labour market).

[20]Notice also that, due to fewer observations for the early graduation cohorts, the confidence intervals for the corresponding years are larger than for the following ones.



(a) Knowledge Migrants Scheme (2004) (b) Researcher EU 2005/71 (2008)

Figure 11: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates. Full regression results are in Appendix B.

6 Conclusions

The Netherlands is an R&D-intensive country with a strong focus on attracting foreign students to its higher education system., especially in STEM fields. However, prior research indicates that foreign doctoral recipients in the Netherlands have lower stay rates compared to those in the United States (Rud et al., 2015). While the two countries differ significantly in institutional and policy contexts, a key question remains: who are the doctoral graduates that the Netherlands is losing, and what are the channels and factors associated with their migration and destination choices?

To address these questions, we have constructed a novel dataset containing detailed information on doctoral recipients, including country of birth, gender, age at graduation, and supervisor's identity, as well as – through a machine learning match – their publications before, during, and after the PhD studies. Our findings indicate that foreign-born doctoral graduates in Dutch universities are disproportionately represented in less publication-oriented fields. This sorting pattern explains entirely or almost entirely their lower likelihood of publishing after graduation, depending on whether they come from inside or outside the EEA. Besides, once we account for this sorting, foreign-born graduates exhibit similar productivity levels than their native counterparts.

Not surprisingly, the foreign-born doctoral recipients are much more likely to leave the Netherlands after graduation than their native peers. More interestingly, we do not observe major differences between non-EEA and EEA native graduates, despite the former facing less administrative costs and hurdles when deciding to stay, thanks to their reliance on the Free Movement principle. Nor we find any change in the relative propensity to stay in the Netherlands of the two groups of foreign graduates following the introduction of the Knowlege Migrant Scheme in 2004 and the Researcher EU 2005/71 permit four years later. The only exception is represented by Chinese graduates, who appear to be more likely to leave and, in particular, to return to the home country. This finding has an explanation in the form of funding received by many such students, which comes with a return requirement; but we find it difficult to assess the size of this requirement's effect, due to some measurement error bias affecting our match of Chinese graduates to publications.

We also find little or no trace of an either positive or negative self-selection, by which the most or least productive foreign graduates would leave the Netherlands upon graduation, not even for those returning home. Interestingly, these results apply also to the native (Dutch) graduates, with the possible exception of those moving to the United States and other English-speaking countries outside the EEA; for which we find – in some regressions – a small-sized, positive self-selection effect.

Overall, the strongest predictor of a decision to leave the Netherlands upon graduation, and especially of return migration, is the existence of a foreign affiliation prior to the completion of the PhD. This may be indicative of the graduate's engagement in a double degree with a foreign institution or of an established collaboration with other researchers outside the Netherlands, most likely in his/ger home country. Alongside the null or weak results on the importance of the immigration regime for non-EEA versus EEA-born graduates, this suggests that other costs besides those related to obtaining a residence permit in the Netherlands are those that matter the most, including the sunk costs linked to having a position or some collaborations in one's home country.

Our results do not come without limitations, which we plan to address in our future research.

First and foremost, we observe the doctoral graduates – after their graduation year – only through their publications. This means that we do not know the rate of stay-or-leave of those who do not

undertake an academic career (do not publish), nor the determinants of their decisions, including the costs associated to getting a residence permit. This implies that we do not know whether the Knowledge Migrant Scheme, largely irrelevant for academic careers, may have affected instead the decisions of the graduates with jobs in the business sector. We suspect not, because if this was the case some foreign-born graduates from outside the EEA could have turned to a career in industry, instead of an academic one, in order to secure a residence permit via a higher salary. But this does not appear to be the case, as our descriptive statistics do not suggest a decline in their probability to produce one or more post-doc publications. As for the Researcher EU 2005/71 permit, this addressed explicitly the graduates pursuing an academic career, which makes our lack of data less problematic.

Second, conditioning our results on stay or leave decisions on publishing after the PhD may come with sample selection bias. We discuss this point extensively at the end of section 4. However, in section 5, we also find that the foreign-born's probability to publish after the PhD does not differ substantially from that of natives, which should limit the extent of the bias, at least for our main variable of interest.

Third, our definition of EEA versus non-EEA graduates is entirely based on their country of birth, as we have no data on their nationality. As some doctoral students from Latin America observed, when discussing our evidence at seminars, a certain number of them can obtain a EU nationality due to "jus sanguinis" legislation, such as that until recently in vigour in Italy for descendants of Italian grand or event great-grandparents; and do so precisely when moving abroad. This could introduce a measurement error in our *non_EEA* dummy and an attenuation bias in our estimates differences between EEA and non-EEA graduates.

Our plans for future research address such limitations. In particular, we plan to obtain information of non-academic careers, based on their profiles published on social media, especially those linked to job search and placement. This, however, will limit our research to the more recent cohorts of graduates, whose members are more likely to have and maintain such profiles (Breschi et al., 2020b).

References

Bergé, L. (2018): "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm," *CREA Discussion Papers*.

Berkhout, E., A. Heyma, M. Volkerink, and S. van der Werff (2015): "Attracting and retaining highly skilled migrants in the Netherlands," Tech. Rep. SEO-rapport No. 2015-88, SEO Economisch Onderzoek.

Berkhout, E., T. Smid, and M. Volkerink (2010): "Wat beweegt kennismigranten? Een analyse van de concurrentiekracht van NL bij het aantrekken van kennismigranten," Tech. Rep. SEO-rapport No. 2010-03, SEO Economisch Onderzoek.

Bernstein, S., R. Diamond, A. Jiranaphawiboon, T. McQuade, and B. Pousada (2022): "The contribution of high-skilled immigrants to innovation in the United States," Tech. rep., National Bureau of Economic Research.

Blanchard, M. (2025): "Bloodline gateways: Italian ancestral citizenship as a resource to bypass European migration policies," *Journal of Ethnic and Migration Studies*, 51, 2734–2750.

Boer, H. d., R. Kolster, and H. Vossensteyn (2010): "Motives underlying bachelors–masters transitions: The case of Dutch degree stackers," *Higher Education Policy*, 23, 381–396.

Borjas, G. J. (1991): "Immigration and self-selection," in *Immigration, trade, and the labor market*, ed. by J. M. Abowd and R. B. Freeman, University of Chicago Press, 29–76.

Breschi, S., C. Lawson, F. Lissoni, A. Morrison, and A. Salter (2020a): "STEM migration, research, and innovation," *Research Policy*, 49, 104070.

Breschi, S., F. Lissoni, and E. Miguelez (2020b): "Return Migrants' Self-Selection," *The Roles of Immigrants and Foreign Students in US Science, Innovation, and Entrepreneurship*, 17–48.

Buenstorf, G., D. P. Heinisch, and M. Kapa (2022): "International doctoral graduates as inventors in the German innovation system," *Industry and Innovation*, 29, 488–510.

Buers, C., J. Klaver, B. Witkamp, S. Duysak, E. Verbeek, M. Bouterse, and A. Mack (2018): "Aantrekkelijkheid van Nederland voor kennismigranten," Tech. rep., Wetenschappelijk Onderzoek en Datacentrum (WODC), Dutch Ministry of Justice.

Capoani, L., V. Chabert, A. B. Valle, and F. Rinaldi (2023): "Immigration and Innovation: Review on the Effects on Innovation Outputs and Education," *Migration and Diversity*, 2, 225–278.

Chellaraj, G., K. E. Maskus, and A. Mattoo (2008): "The contribution of international graduate students to US innovation," *Review of International Economics*, 16, 444–462.

Coda-Zabetta, M., C. Chacua, F. Lissoni, E. Miguelez, J. Raffo, and D. Yin (2022): *The missing link: international migration in global clusters of innovation*, Oxford University Press.

Coda-Zabetta, M., E. Miguelez, F. Lissoni, and S. Hegarty (2023): "Foreign-origin Managers and the FDI Location Choice," Mimeo.

Cohen, A. and M. van der Steege (1982): "An historical overview of the state and higher education in the Netherlands," *European journal of Education*, 271–281.

Colussi, T. (2016): "The Impact of the Implementation of Council Directives on Labour Migration Flows from Third Countries to EU Countries," *OECD Social, Employment, and Migration Working Papers*, 0_1.

Corsini, A., M. Pezzoni, and F. Visentin (2022): "What makes a productive Ph. D. student?" *Research Policy*, 51, 104561.

Cristelli, G. and F. Lissoni (2025): "Free movement of inventors: open-border policy and innovation in Switzerland," *Journal of the European Economic Association (forthcoming)*.

De Haan, H. (2014): "Can internationalisation really lead to institutional competitive advantage?–a study of 16 Dutch public higher education institutions," *European Journal of Higher Education*, 4, 135–152.

De Haas, H., K. Natter, and S. Vezzoli (2018): "Growing restrictiveness or changing selection? The nature and evolution of migration Policies1," *International Migration Review*, 52, 324–367.

de Lange, T. (2018): "Intersecting Policies of Innovation and Entrepreneurship Migration in the EU and the Netherlands," in *EU External Migration Policies in an Era of Global Mobilities: Intersecting Policy Universes*, ed. by S. Carrera, A. P. L. den Hertog, M. Panizzon, and D. Kostakopoulou, Brill Nijhoff, 224–243.

De Lange, T. and Z. Vankova (2022): "The Recast EU Blue Card Directive: Towards a Level Playing Field to Attract Highly Qualified Migrant Talent to Work in the EU?" *European Journal of Migration and Law*, 24, 489–515.

Docquier, F. and H. Rapoport (2012): "Globalization, brain drain, and development," *Journal of economic literature*, 50, 681–730.

Doran, K., A. Gelber, and A. Isen (2022): "The effects of high-skilled immigration policy on firms: Evidence from visa lotteries," *Journal of Political Economy*, 130, 2501–2533.

Dutta, S., B. Lanvin, L. Rivera León, and S. Wunsch-Vincent, eds. (2022): *Global innovation index 2021*, Geneva: World Intellectual Property Organization.

EUROSTAT (2018): "R&D personnel in Europe," https://ec.europa.eu/eurostat/statistics-explained/index.php?title=R%26D_personnel&oldid=551400, accessed on July 27, 2023.

FEDASIUK, R. (2020): "The China scholarship council: An overview," *Center for Security and Emerging Technology*, 614–621.

FINN, M. G. (2010): "Stay rates of foreign doctorate recipients from US universities, 2007," Tech. rep., Oak Ridge Inst. for Science and Education (ORISE), Oak Ridge, TN (United States).

FREEMAN, R. B. (2015): "Immigration, international collaboration, and innovation: Science and technology policy in the global economy," *Innovation Policy and the Economy*, 15, 153–175.

GANGULI, I. (2015): "Immigration and ideas: what did Russian scientists "bring" to the United States?" *Journal of Labor Economics*, 33, S257–S288.

GAULE, P. AND M. PIACENTINI (2013): "Chinese graduate students and US scientific productivity," *Review of Economics and Statistics*, 95, 698–701.

——— (2018): "An advisor like me? Advisor gender and post-graduate careers in science," *Research Policy*, 47, 805–813.

GILSING, V. AND H. ERKEN (2002): "Trends in corporate R&D," Tech. rep., Ministry of Economic Affairs, The Hague.

GLENNON, B. (2023): "How do restrictions on high-skilled immigration affect offshoring? Evidence from the H-1B program," *Management Science*.

GROGGER, J. AND G. H. HANSON (2015): "Attracting talent: Location choices of foreign-born PhDs in the United States," *Journal of Labor Economics*, 33, S5–S38.

HAN, X., G. STOCKING, M. A. GEBBIE, AND R. P. APPELBAUM (2015): "Will they stay or will they go? International graduate students and their decisions to stay or leave the US upon graduation," *PloS one*, 10, e0118183.

HANSON, G. H. AND M. J. SLAUGHTER (2019): "High-skilled immigration and the rise of STEM occupations in US employment," in *Education, skills, and technical change: Implications for future US GDP Growth*, ed. by C. R. Hulten and V. A. Ramey, University of Chicago Press, 465–494.

HAUSMAN, J. AND D. MCFADDEN (1984): "Specification tests for the multinomial logit model," *Econometrica: Journal of the econometric society*, 1219–1240.

HAWTHORNE, L. (2018): "Attracting and retaining international students as skilled migrants," in *High-skilled migration: Drivers and policies*, ed. by M. Czaika, Oxford University Press.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, 153–161.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd (2020): "spaCy: Industrial-strength Natural Language Processing in Python," .

Horta, H. and J. M. Santos (2016): "The impact of publishing during PhD studies on career research publication, visibility, and collaborations," *Research in Higher Education*, 57, 28–50.

Hunt, J. (2011): "Which immigrants are most innovative and entrepreneurial? Distinctions by entry visa," *Journal of Labor Economics*, 29, 417–457.

——— (2015): "Are immigrants the most skilled US computer and engineering workers?" *Journal of Labor Economics*, 33, S39–S77.

Hunt, J. and M. Gauthier-Loiselle (2010): "How much does immigration boost innovation?" *American Economic Journal: Macroeconomics*, 2, 31–56.

ICMPD (2012): "Implementation and Impact of the Researchers Directive (Directive 2005/71/EC)," Tech. rep., International Centre for Migration Policy Development.

Kahanec, M., M. Pytlikova, and K. F. Zimmermann (2016): *The free movement of workers in an enlarged European Union: Institutional underpinnings of economic adjustment*, Springer.

Kahn, S. and M. MacGarvie (2016): "Do return requirements increase international knowledge diffusion? Evidence from the Fulbright program," *Research Policy*, 45, 1304–1322.

——— (2020): "The impact of permanent residency delays for stem PhDs: Who leaves and why," *Research Policy*, 49, 103879.

Kerr, S. P., W. Kerr, Ç. Özden, and C. Parsons (2016): "Global talent flows," *Journal of Economic Perspectives*, 30, 83–106.

Kerr, W. R. and W. F. Lincoln (2010): "The supply side of innovation: H-1B visa reforms and US ethnic invention," *Journal of Labor Economics*, 28, 473–508.

Khosla, P. (2018): "Wait time for permanent residency and the retention of immigrant doctoral recipients in the US," *Economic Analysis and Policy*, 57, 33–43.

Kolman, M. (2024): "The impact of internationalization on research performance," https://www.elsevier.com/connect/the-impact-of-internationalization-on-research-performance, accessed on April 17, 2024.

Kolster, R. (2014): "Academic attractiveness of countries; a possible benchmark strategy applied to the Netherlands," *European journal of higher education*, 4, 118–134.

Koslowski, R. (2018): "Shifts in selective migration policy models," in *High-Skilled Migration: Drivers and Policies*, ed. by M. Czaika, Oxford University Press.

Kulu-Glasgow, I., D. Schans, and M. Smit (2018): "The Dutch battle for highly skilled migrants: policy, implementation and the role of social networks," *Migration Letters*, 15, 517–532.

Lissoni, F. and E. Miguelez (2021): "International migration and innovation: France in comparative perspective," Tech. Rep. CAE Focus No. 071-2021, Report for the Conseil d'Analyse Économique.

——— (2024): "Migration and innovation: Learning from patent and inventor data," *Journal of Economic Perspectives*, 38, 27–54.

Luijten-Lub, A., M. Van der Wende, and J. Huisman (2005): "On cooperation and competition: A comparative analysis of national policies for internationalisation of higher education in seven Western European countries," *Journal of studies in international education*, 9, 147–163.

Martínez, G. L., H. S. De Juano-i Ribes, D. Yin, B. Le Feuvre, I. Hamdan-Livramento, K. Saito, and J. Raffo (2021): "Expanding the World Gender-Name Dictionary: WGND 2.0," Tech. rep., World Intellectual Property Organization-Economics and Statistics Division.

Mayda, A. M., F. Ortega, G. Peri, K. Shih, and C. Sparber (2023): "Coping with H-1B shortages: Firm performance and mitigation strategies," *ILR Review*, 76, 919–943.

Miguelez, E. and A. Morrison (2023): "Migrant inventors as agents of technological change," *The Journal of Technology Transfer*, 48, 669–692.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013): "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*.

Moser, P., A. Voena, and F. Waldinger (2014): "German Jewish émigrés and US invention," *American Economic Review*, 104, 3222–3255.

Mulders, A. M., B. Hofstra, and J. Tolsma (2024): "A matter of time? Gender and ethnic inequality in the academic publishing careers of Dutch PhDs," *Quantitative Science Studies*, 1–29.

Obradović, E. (2013): "Monitor Kennismigranten Kwantitatieve analyse," Tech. rep., Ministerie van Veiligheid en Justitie, Immigratie- en Naturalisatiedienst (IND).

OECD (2016): *Recruiting Immigrant Workers: The Netherlands 2016*, Organisation for Economic Co-operation and Development.

——— (2020): *Education at a Glance 2020*, Organisation for Economic Co-operation and Development.

OECD (2023): *Education at a Glance 2022*, Organisation for Economic Co-operation and Development.

Parey, M., J. Ruhose, F. Waldinger, and N. Netz (2017): "The selection of high-skilled emigrants," *Review of Economics and Statistics*, 99, 776–792.

Peri, G., K. Shih, and C. Sparber (2015): "STEM Workers, H-1B Visas, and Productivity in US Cities," *Journal of Labor Economics*, 33, S225–S255.

Peri, G. and C. Sparber (2011): "Highly educated immigrants and native occupational choice," *Industrial Relations: a journal of economy and society*, 50, 385–411.

Puhani, P. (2000): "The Heckman correction for sample selection and its critique," *Journal of economic surveys*, 14, 53–68.

Rathenau Instituut (2024a): "The share of foreign scientific personnel," https://www.rathenau.nl/en/science-figures/personnel/international-researchers/share-foreign-scientific-personnel, accessed on 26 November, 2024.

——— (2024b): "The share of foreign scientific personnel at Netherlands universities, overall and by area of origin," https://www.rathenau.nl/en/science-figures/personnel/international-researchers/share-foreign-scientific-personnel-netherlands, accessed on 26 November, 2024.

Řehřek, R. and P. Sojka (2010): "Software framework for topic modelling with large corpora," .

Roach, M. and J. Skrentny (2019): "Why foreign STEM PhDs are unlikely to work for US technology startups," *Proceedings of the National Academy of Sciences*, 116, 16805–16810.

——— (2021): "Rethinking immigration policies for STEM doctorates," *Science*, 371, 350–352.

Roh, J.-Y. (2015): "What predicts whether foreign doctorate recipients from US institutions stay in the United States: Foreign doctorate recipients in science and engineering fields from 2000 to 2010," *Higher Education*, 70, 105–126.

Rud, I., B. Wouterse, and R. Van Elk (2015): "Stay rates of foreign PhD graduates in the Netherlands," *CPB Background Document, September, CPB Netherlands Bureau for Economic Policy Analysis, The Hague*.

Shih, K. (2016): "Labor market openness, H-1B visa policy, and the scale of international student enrollment in the United States," *Economic Inquiry*, 54, 121–138.

Skeldon, R. (2018): "High-skilled migration and the limits of migration policies," in *High-skilled migration: Drivers and policies*, ed. by M. Czaika, Oxford University Press.

Soete, L. L., B. Verspagen, and T. H. Ziesemer (2020): "The productivity effect of public R&D in the Netherlands," *Economics of Innovation and New Technology*, 29, 31–47.

Stephan, P. E. and S. G. Levin (2001): "Exceptional contributions to US science by the foreign-born and foreign-educated," *Population research and Policy review*, 20, 59–79.

Stephan, P. E., A. J. Sumell, G. C. Black, and J. D. Adams (2004): "Doctoral education and economic development: The flow of new Ph. Ds to industry," *Economic Development Quarterly*, 18, 151–167.

SVR (2012): "Mobile talents? A comparison of the intentions of international students to stay in five countries of the European Union," Tech. rep., Sachverständigenrat deutscher Stiftungen für Integration und Migration GmbH.

Van Der Wende, M. (2015): "International academic mobility: Towards a concentration of the minds in Europe," *European review*, 23, S70–S88.

Veugelers, R. (2017): "Countering European brain drain," *Science*, 356, 695–696.

Waldinger, F. (2010): "Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany," *Journal of political economy*, 118, 787–831.

Zheng, Y. and O. Ejermo (2015): "How do the foreign-born perform in inventive activity? Evidence from Sweden," *Journal of Population Economics*, 28, 659–695.

Zwetsloot, R., J. Feldgoise, and J. Dunham (2020): "Trends in US intention-to-stay rates of international Ph. D. graduates across nationality and STEM fields," *Center for Security and Emerging Technology*.

Appendices

A Data collection and augmentation

A.1 NARCIS representativeness

NARCIS served as a centralized portal for aggregating scientific research from various institutions in the Netherlands. Among the collected outputs were doctoral dissertations, which often overlapped with publications from other sources. This overlap arises because the fourteen universities in the Netherlands are the sole institutions authorized to confer doctoral degrees, leading to duplication of some dissertations across multiple entries. In theory, if a dissertation is listed by another institution[21], it should also be present in the repository of the degree-granting university. However, this is not consistently the case.

To the best of our ability, we have attributed each dissertation to one of the fourteen universities. This dataset was subsequently compared with official statistics, as presented in Figures 1 and 2. The analysis reveals under-representation of dissertations for earlier years and over-representation for more recent years. The under-representation is primarily due to universities uploading fewer theses online than the number officially reported, as NARCIS merely harvested existing online records. The over-representation, on the other hand, can be attributed to cases of duplication. For instance, as shown in Figure A.1, Eindhoven University exhibits an inflated number of dissertations that closely mirrors the pattern of official statistics, while Utrecht University shows a notable gap in dissertations between 2002 and 2006.

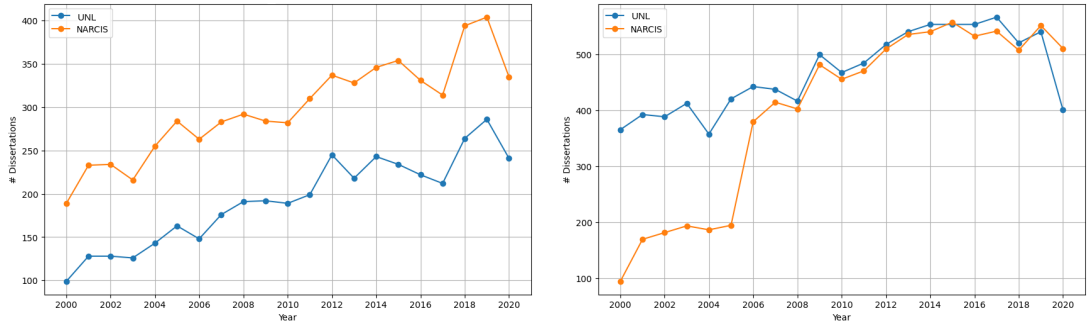


Figure A.1: NARCIS under and over representation by university

[21]The full list of contributing institutions from which dissertations were collected via NARCIS includes: *University of Groningen, University of Amsterdam, Radboud University Nijmegen, Vrije Universiteit Amsterdam, Delft University of Technology, Eindhoven University of Technology, Wageningen University Research, Utrecht University, Erasmus University Rotterdam, Leiden University, Maastricht University, University of Twente, University Medical Center Utrecht, Tilburg University, Leiden University Medical Center (LUMC), TNO, CWI, Universities of Applied Sciences, Open Universiteit Nederland, IHE Delft Institute for Water Education, Nyenrode Business University, DIFFER, University for Humanistics, Breda University of Applied Sciences, Theological University Kampen, Protestant Theological University, Theological University Apeldoorn, and Naturalis Biodiversity Center.*

A.2 Binary Classifier for the identification of STEM dissertations

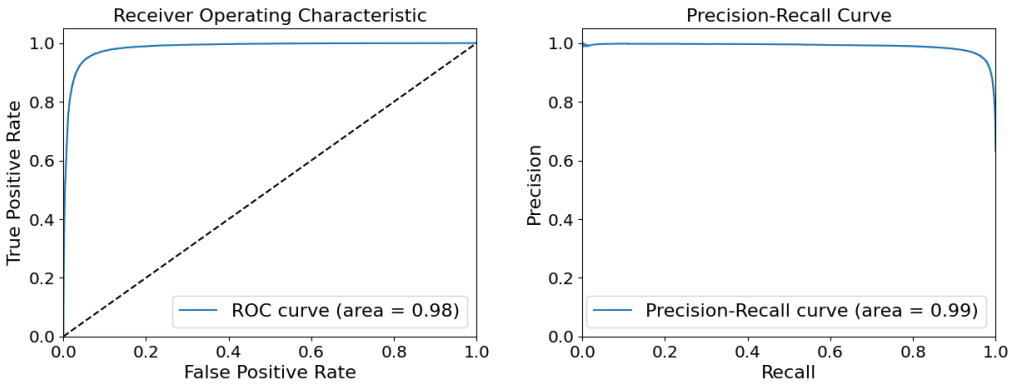


Figure A.2: Model performance on the Theses.fr

Table A.1: Classification Report - Theses.fr

	Metrics			
Class	Precision	Recall	F1-Score	Support
Other	0.94	0.92	0.93	16288
STEM	0.96	0.96	0.96	28127
Accuracy			0.95	44415
Macro avg	0.95	0.94	0.95	44415
Weighted avg	0.95	0.95	0.95	44415

Table A.2: Classification Report - NARCIS

	Metrics			
Class	Precision	Recall	F1-Score	Support
Other	0.88	0.92	0.90	300
STEM	0.96	0.94	0.95	700
Accuracy			0.94	1000
Macro avg	0.92	0.93	0.93	1000
Weighted avg	0.94	0.94	0.94	1000

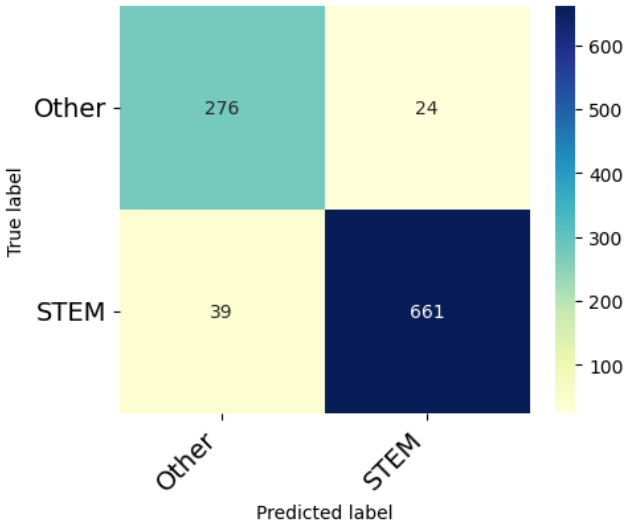


Figure A.3: Confusion Matrix - NARCIS

To assign a field to the dissertations in NARCIS, we employ a binary classification model. This supervised machine learning approach is trained on a labeled dataset, where the objective is to predict class membership for unseen instances. Specifically, the model assigns each dissertation to one of two classes, typically denoted as 0 and 1, based on a set of input features.

We utilize the Word2Vec algorithm introduced by Mikolov et al. (2013) to generate word embeddings, which are then fed into a logistic regression classifier. The implementation leverages the Gensim library (Řehřek and Sojka, 2010).

Our training dataset consists of dissertation titles from Theses.fr, selected due to the database's comprehensive coverage and reliable labeling. Titles, rather than abstracts, are used as input due to the limited availability of abstract data in NARCIS (fewer than 54% of NARCIS dissertations include abstracts). We assume semantic similarity between the corpus of STEM dissertations in English from France (2000–2020) and those from the Netherlands during the same period. The training dataset includes 222,071 French dissertations, approximately 64% of which are labeled as STEM.

Figure A.2 demonstrates the strong performance of our binary classifier on the Theses.fr dataset. An ideal classifier would achieve a *Receiver Operating Characteristic* (ROC) curve with a 0 *False Positive Rate* (FPR) and a 1 *True Positive Rate* (TPR). Our model achieves commendable performance, with an approximate FPR of 0.05 and TPR of 0.95. Similar trends are observed in the *Precision-Recall Curve*, which mirrors the ROC results.

To maximize recall, we set the classification threshold at 0.5. This decision was further validated through tests on additional models based on transistor methodologies, which indicated that misclassification primarily occurred at the intersection of Psychology and Medicine. Increasing the threshold did not significantly improve model performance. The selected threshold provides satisfactory results across

standard evaluation metrics[22], as summarized in Table A.1. All metrics range between 0.92 and 0.96, with the STEM class (our primary interest) achieving the highest performance.

Finally, we applied the trained model to the titles of NARCIS dissertations. To evaluate performance, we manually labeled 1,000 titles as either *STEM* or *Other* and generated a *Classification Report* for NARCIS. Table A.2 highlights a slight decrease in precision for identifying *Other* (Non-STEM) titles, though performance remains high for the *STEM* class. The weighted average precision, recall, and F1-score are 0.94, considering class imbalance. These results are corroborated by a low rate of misclassification, as visualized in the *Confusion Matrix* in Figure A.3.

We also tested deep learning-based classification methods but excluded results from comparison tables due to differences in sample size, as transistor-based methodologies require access to high-performance GPUs. Given the comparable results, we opted to proceed with the Word2Vec-based model for its reproducibility and computational efficiency.

[22]The performance of the model is quantified using four key metrics: Accuracy, Precision, Recall, and the F1-score. Each metric provides distinct insights into model performance:

Accuracy: Measures the proportion of correct predictions to the total predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision: Evaluates the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: Assesses the model's ability to identify positive instances.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score: Combines Precision and Recall into a single metric, offering a harmonic mean of the two:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A.3 Country of Origin retrieval

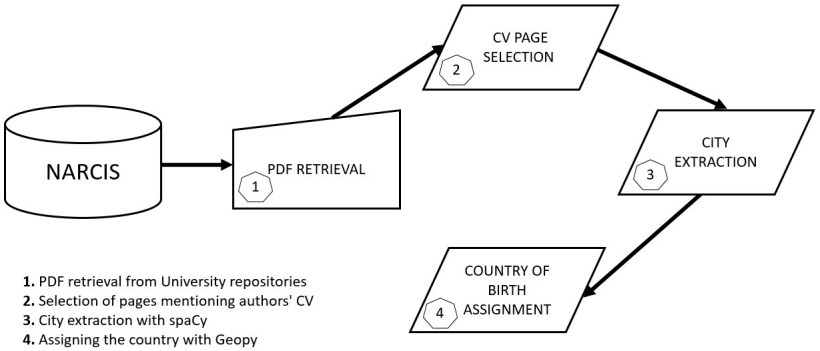


Figure A.4: Country of Birth retrieval - Flowchart



Figure A.5: Example of a Name Entity Recognition task

To retrieve the country of origin of doctoral graduates, we scraped the dissertations available in PDF format. As illustrated in Figure A.4, dissertations with PDF files were identified through the NARCIS repository and subsequently accessed via the respective university repositories. Once all PDFs were collected, we identified pages likely to contain the place of birth by searching for keywords such as *curriculum*, *cv*, and *biography* in both English and Dutch. After locating these pages, a Named Entity Recognition (NER) algorithm was applied to strings surrounding keywords such as *birth*, *born*, *geboorte*, and *geboren*. The NER algorithm, a Natural Language Processing (NLP) tool, is designed to identify and categorize entities, specifically the *location entity* in our context (for a graphical representation, see Figure A.5) (Honnibal et al., 2020).

In most cases, doctoral graduates provided their city of birth rather than their country. To map each city to its corresponding country, we utilized the Geopy package[23].

Overall, we successfully retrieved the country of birth for approximately 11,000 doctoral graduates. However, as shown in Figure A.6, the proportion of retrieved data improves over time, reflecting enhanced coverage and quality of the PDFs. A more significant challenge arises from Figure A.7, which demonstrates substantial variability in data availability across universities. This variability is primarily due to the inconsistent availability of PDFs, as well as differences in bureaucratic compliance. For instance,

[23] https://github.com/geopy/geopy

Utrecht University demonstrates the highest level of compliance, with 50% of all STEM dissertations between 2000 and 2020 being accessible in PDF format.

Finally, we compare our findings with official statistics retrieved from *Universiteiten van Nederland*[24]. In Figure A.8, we observe a clear rise in foreign graduates starting in 2007. The proportion of Dutch-born (orange line) and foreign-born graduates (red line) appears to be representative of the underlying trends. However, when categorizing regions of origin into EER/EU (blue line) and the rest of the world (green line), we note that individuals from the EER/EU are under-represented, while those from the rest of the world are over-represented. This discrepancy may result from foreign-born individuals acquiring EU citizenship before moving to the Netherlands, or it may reflect sample bias.

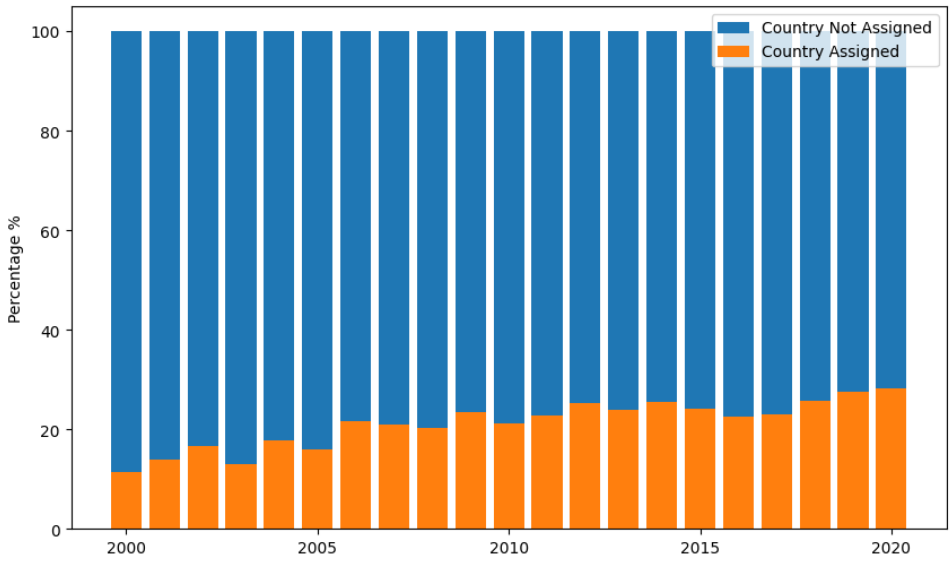


Figure A.6: Percentage of graduates with country of origin information by year, 2000-2020

[24] https://www.universiteitenvannederland.nl

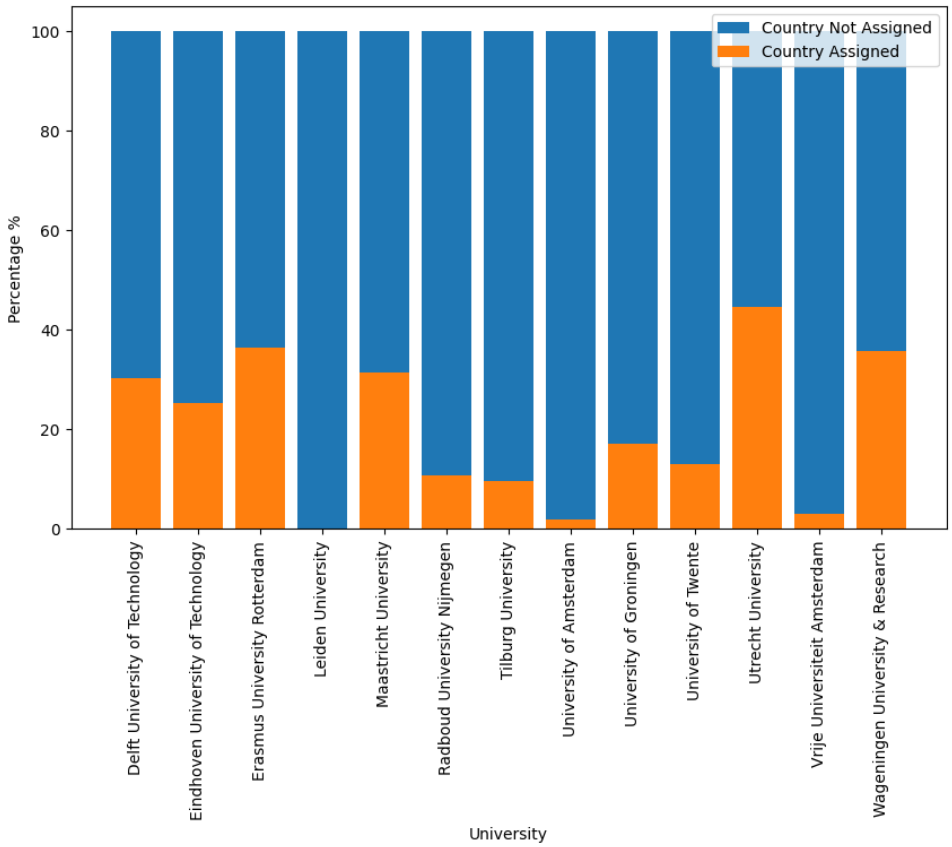


Figure A.7: Percentage of graduates with country of origin information by university, 2000-2020

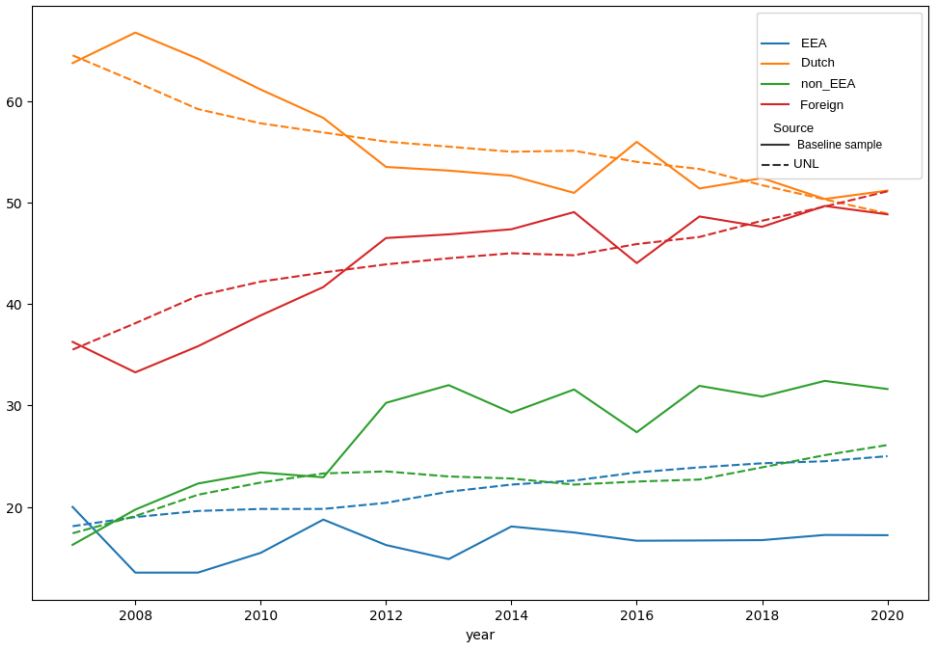


Figure A.8: Share of foreigners by year

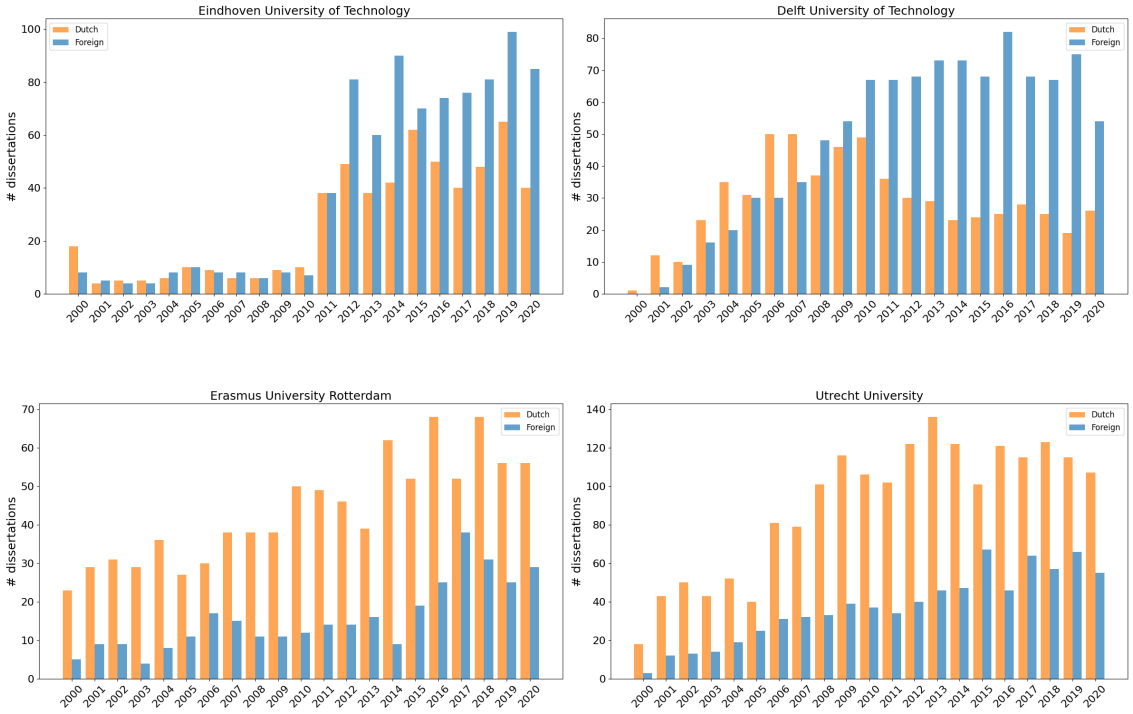


Figure A.9: Comparison number of dissertations by foreign status and university, 2000-2020

A.4 Graduates Scopus Matching

To obtain a dataset with the academic careers of the doctoral graduates we match them with the authors present in Elsevier Scopus. Scopus has the advantage of disambiguated author ids that aggregate all the work done by an author under an unique id. This is especially useful when authors appear in slightly different ways in papers (e.g. Marcus Aurelius Antoninus Augustus; Marcus Aurelius; A.A.A., Marcus etc.). Nonetheless, the presence of different combinations of names and surnames increases the complexity of the matching exercise, as the proportion of 1:1 (One-to-one) matches drastically decreases. Given the enormous difference in size between NARCIS and Scopus, the M:1 (Many-to-one) matches are not troublesome in our case. However, the same cannot be said for the 1:M (One-to-many). We match on several re-combinations of name (initials when absent) and surname information in NARCIS to Scopus authors the 11.326 doctoral graduates for whom we have country of birth information and we end up with 10.151 doctoral graduates having at least one matched scopus id. However, we obtain on average 13 matches (NARCIS author 1:13 Scopus author).

To identify the correct matches we adopt the random forest machine learning technique. We start by obtaining a sample dataset of 1500 NARCIS authors and for computational reasons restricting the matches in a time span around the obtainment of the PhD (3 years prior and 3 posterior). All the matches are then hand validated providing as a set of doctoral graduates for whom we know the correct matches. We create a set of variables for each doctoral recipient (see Figure A.10) that can predict the validity of a match. We then split the resulting dataset in a 80-20 ratio and use it to train and validate a random forest model. The model introduces randomness in both variable selection (feature randomness) and data sampling (bootstrapping), leading to the creation of multiple decision trees. A decision tree has a flowchart-like structure and its internal nodes stand for features, branches depict decision rules, and leaf nodes signify outcomes (see Figure A.11). In short, the model learns which are the features that allow to more accurately make a distinction between the two classes (in our case true or false matches), by decreasing the 'impurity' (how mixed are the classes) at a node through Mean Decrease in Impurity (MDI) (expressed as Mean Decrease Accuracy of Mean Decrease Gini). The impurity is weighted by the proportion of samples that are in that specific node. Intuitively, the greater the amount of data split by the criteria the greater is the influence on the MDI.

Overall, our random forest performs quite well with an Accuracy of 0.97 a Precision of 0.87 and a Recall of 0.86. Moreover, as the confusion matrix in Figure A.12 shows, the number of false positives (30) and false negatives (34) are very close, suggesting that the model does not exhibit a strong bias toward predicting one class incorrectly over the other. In other words, it is roughly equally likely to misclassify a negative instance as positive (type I error) as it is to misclassify a positive instance as negative (type II error). Lastly, we compare the performance of the model with two other common techniques applied for this sort of task, a Logit and a Support Vector Machine models. The random forest model clearly

over-performs both in terms of precision and recall (see Figure A.13).

We then proceed to apply the random forest model to our dataset of 10.151 doctoral graduates matched with Scopus using the recombinations of names and surname information. This results in an 7.981 PhDs matched with at least one Scopus author and 2.170 with no matches. Basically, the 2.170 PhDs were matched to all false positives and no true positives. Of the 7.981 matches we fast 7.503 are matched to one Scopus author, 373 with two and 105 with three of more. The PhDs matched with three or more Scopus authors are mainly Chinese and for the moment we exclude this 105 doctoral graduates. Lastly, we manually check for 57 cases were have the same Scopus author assigned to multiple doctoral graduates. This operation results in an additional drop of 31 doctoral graduates resulting in a total of 7.845 PhDs matched with Scopus.

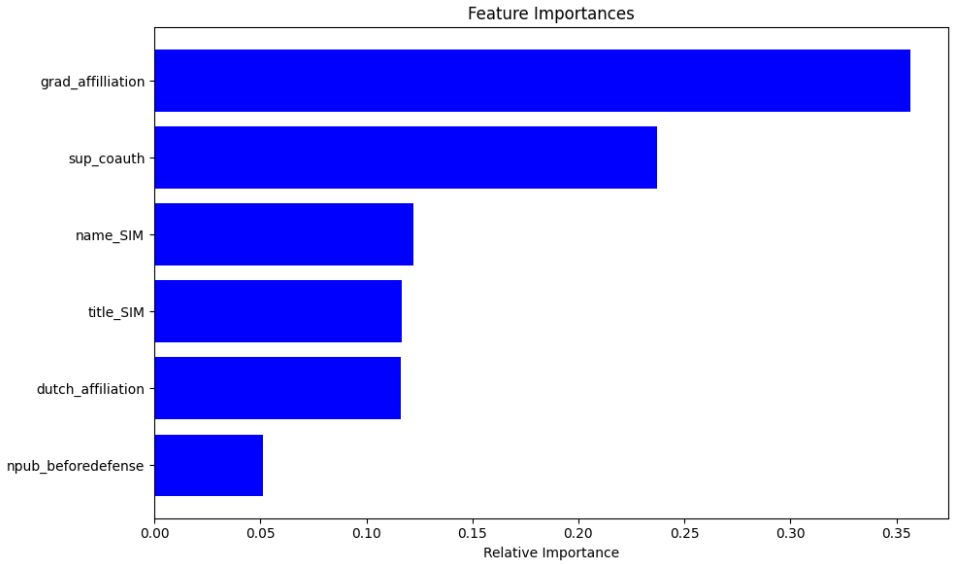


Figure A.10: RF Variables

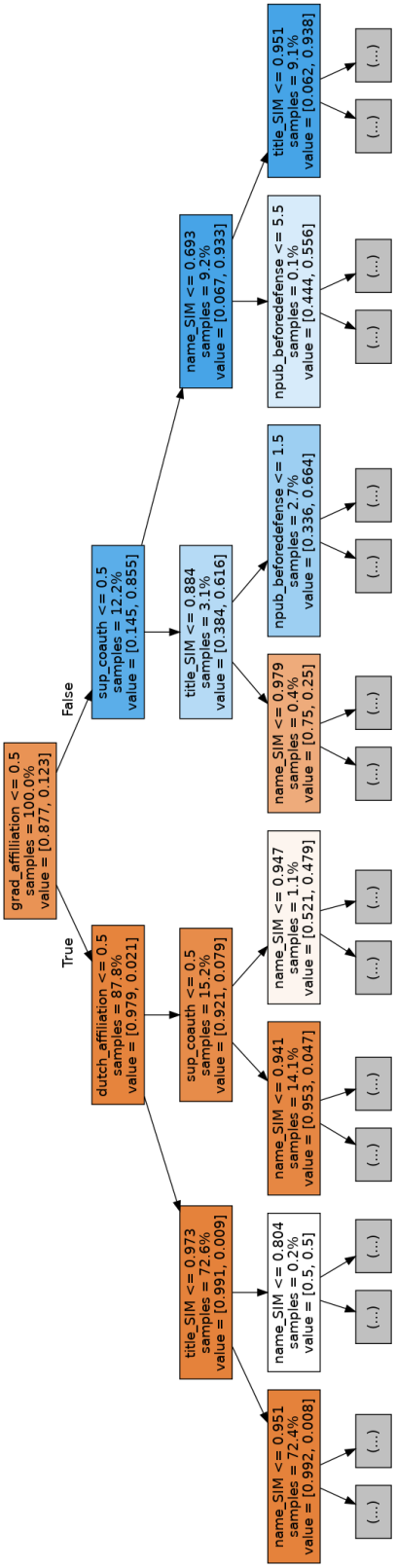


Figure A.11: RF Tree Visualization

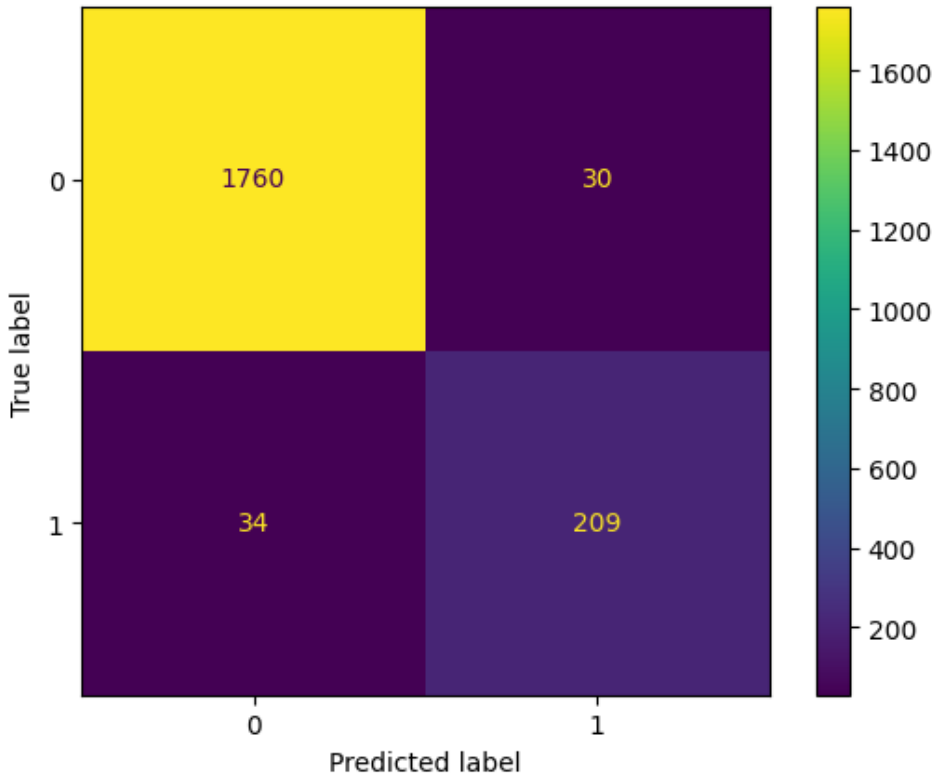


Figure A.12: RF Confusion Matrix

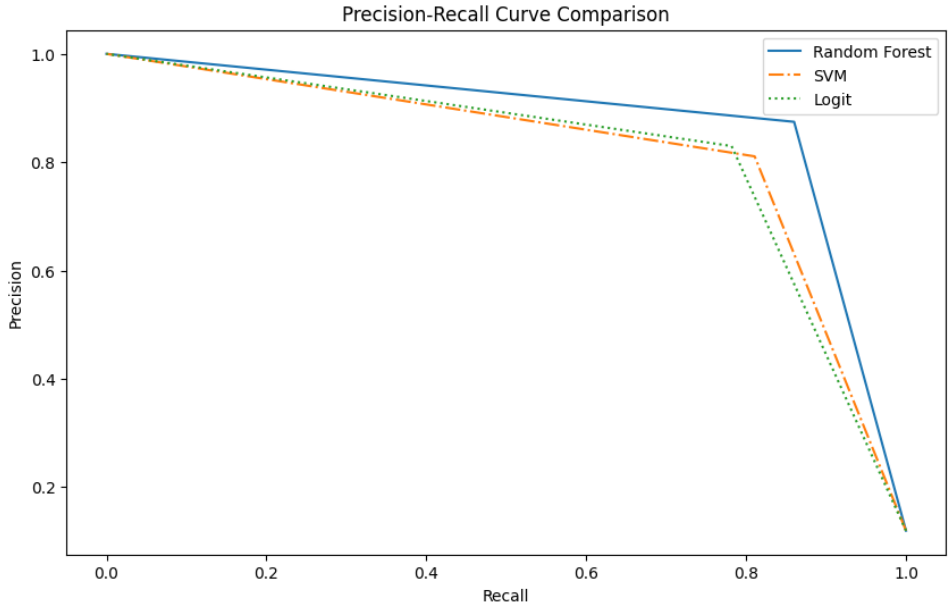


Figure A.13: Alternative Methods Comparison

A.5 Supervisors Scopus Matching

Supervisors have a significantly higher number of publications making the retrieval easier. We assume that the first person indicated in the list of people involved with the thesis in NARCIS is the thesis supervisor. We perform a broad matching with name and surname combinations as in **??** to maximize recall. We associate to each supervisor all the students that indicate her/him in the list of people involved with the thesis, from this list we consider all the other people involved in each thesis as the pool of colleagues, lastly we create an affiliation time span for each supervisor at the university that starts with the first thesis associated and ends with the last. We validate the multiple matches for each supervisor with 3 criteria: (1) they coauthored at least one publication with one of the students, they coauthored at least one publication with one of the colleagues, they have at least one publication with the university affiliation in the time span obtained from the associated doctoral dissertations. Finally we had 444 supervisors assigned with the same Scopus author IDs creating 990 supervisor - Scopus author associations mainly in health sciences. We manually check all the cases and correctly assign the Scopus author ids. We obtain 2.440 supervisors covering the 79.88% of doctoral graduates in our sample. In Figure A.14 we can see how in our sample almost half of the supervisors have only one doctoral recipient, the other bulk has between 2 and 10 students few outliers have more.

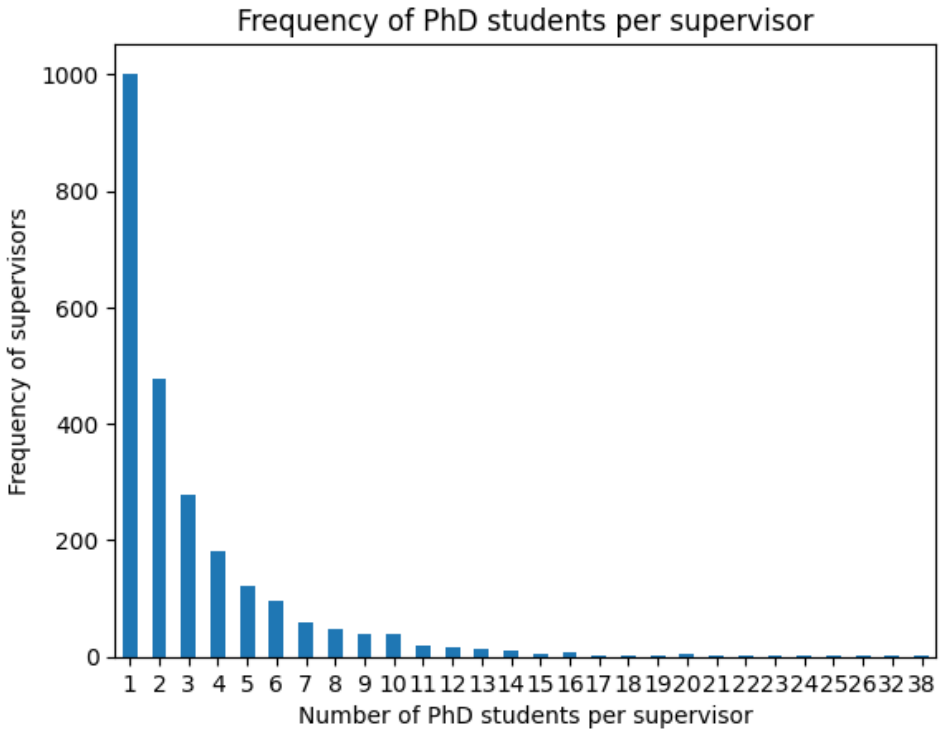


Figure A.14: Supervisors PhD students distribution

A.6 Theses Classification by Macro Fields

We retrieve information on the subject areas of the supervisors' publications from Scopus. Table A.3 presents the correspondence between the Subject Area, which we refer to as *macro_fields*, and the Subject Area Classification, which we refer to as *fields*. To construct the fixed effects, we adopt two approaches. First, for the more granular classification (*fields*), we generate a set of indicator variables, where each dummy equals one if the supervisor has at least one publication in the corresponding field. Second, for a more aggregated approach (*macro_fields*), we associate each Subject Area Classification with its corresponding Subject Area, including an additional category, *Multidisciplinary*, which is indistinguishably present across all four Subject Area Classifications. This classification results in five macro-fields: *Physical Sciences*, *Life Sciences*, *Health Sciences*, *Multidisciplinary*, and *Social Sciences & Humanities*. We then assign to each supervisor the macro-field with the highest cumulative number of publications. However, one observation is classified as *Multidisciplinary*, and 96 as *Social Sciences & Humanities*. Within these 96 cases, the latter field accounts for at least 50% of publications in only 9 observations; in the majority (approximately 60 observations), it accounts for just 25%. Therefore, we reassign these supervisors to their second most frequent field. As a result, only one supervisor remains exclusively assigned to *Social Sciences & Humanities*, which we drop from the sample. The final sample consists of 2,622 supervisors with a uniquely assigned macro-field.

Table A.3: Subject Areas and their Scopus Classifications

Subject Area	**Subject Area Classifications**
Physical Sciences	Chemical Engineering, Chemistry, Computer Science, Earth and Planetary Sciences, Energy, Engineering, Environmental Science, Material Science, Mathematics, Physics and Astronomy, Multidisciplinary
Health Sciences	Medicine, Nursing, Veterinary, Dentistry, Health Professions, Multidisciplinary
Social Sciences	Arts and Humanities, Business, Management and Accounting, Decision Sciences, Economics, Econometrics and Finance, Psychology, Social Sciences, Multidisciplinary
Life Sciences	Agricultural and Biological Sciences, Biochemistry, Genetics and Molecular Biology, Immunology and Microbiology, Neuroscience, Pharmacology, Toxicology and Pharmaceutics, Multidisciplinary

Source: https://service.elsevier.com/app/answers/detail/a_id/12007/supporthub/scopus/

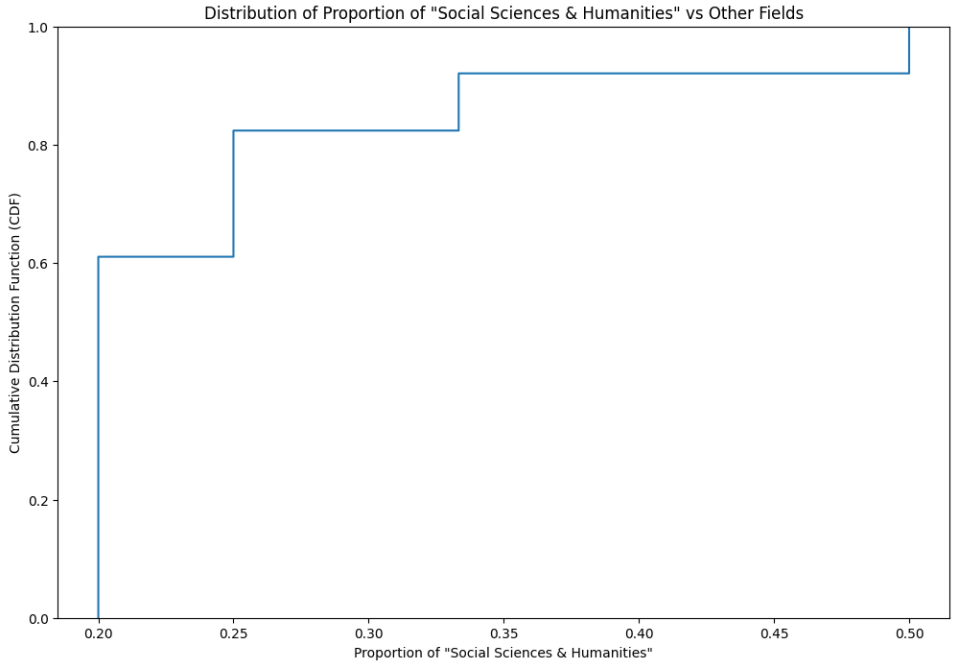


Figure A.15: Supervisors Social Sciences & Humanities distribution

A.7 Name Augmentation

One of the biggest issues for our study is the absence of full names for more than two thirds (69% circa) of the NARCIS dataset. However, we always have the surname as well as the initials. We exploit this information to implement two data augmentation exercises: retrieval of full names through pdf scraping and web scraping. As for the pdf scraping the methodology is quite straightforward. We take the string of the surname and search for it inside the pdf while looking for the closest string starting with the same character as one of or the initial/s, that is the first name. While for web scraping we create a snippet for each doctoral graduates containing: name (surname with initials), university of graduation, graduation year and thesis title. This snipped is fed to an algorithm that through Google APIs[25] returns a dataframe containing the Google results for each doctoral student. This dataframe is processed with a NER (Name Entitiy Recognition) tool provided by spaCy to identify *person entities* (for a graphical representation see Figure A.5) (Honnibal et al., 2020). In this way we isolate the *person entity* associated to each Google result and we obtain a list of possible names of the doctoral recipient. On this list we apply a fuzzy matching algorithm[26] to order the *person entities* from the one resembling the most the original name string from NARCIS in a descending manner. Then, we create a new variable containing the character length of each name entity. Lastly, we take as scraped information the *person entity* having

[25]The APIs have been acquired with ScrapingBee. For more information see: https://www.scrapingbee.com/, last visited: October 20, 2023

[26]We use the *fuzzywuzzy* package in Python. For more info see: https://github.com/seatgeek/fuzzywuzzy, last visited: October 20, 2023

at the same time the highest matching score and greatest length compared to the NARCIS name. When the two lengths match it means that no information is retrieved. Overall, we retrieve the first name for circa 94% of our sample (10.706 out of the 11.413).

B Full Results Estimated Impact of Knowledge Migrant Scheme (2004) and Researcher EU 2005/71 permit (2008)

Table B.1: Dynamic Difference-in-Differences: Migration by Foreign Status (Linear Probability Model)

Dependent Variable:	Leave			
Reference Year:	2004	2004	2008	2008
Sample:	Full	Chinese Excluded	Full	Chinese Excluded
Variables				
non_EEA	0.0322	0.0214	0.1326	0.0975
	(0.2328)	(0.2445)	(0.1395)	(0.1441)
female	-0.0202	-0.0585	-0.0531	-0.0517
	(0.0897)	(0.0995)	(0.0488)	(0.0524)
age	-0.0011	-0.0059	0.0016	0.0010
	(0.0077)	(0.0084)	(0.0036)	(0.0038)
pre_PhD_publications	0.0213	0.0222	0.0021	0.0029
	(0.0135)	(0.0140)	(0.0055)	(0.0083)
PhD_publications	-0.0103	-0.0097	0.0018	0.0018
	(0.0078)	(0.0081)	(0.0019)	(0.0039)
prior_foreign_affiliation	0.3568***	0.3317**	0.3429***	0.3070***
	(0.0973)	(0.1037)	(0.0541)	(0.0576)
supervisor_productivity	-0.0002	0.0001	-0.0004	-0.0004
	(0.0009)	(0.0009)	(0.0004)	(0.0004)
non_EEA × 2000	0.4090	0.4294	0.2737	0.3119
	(0.4987)	(0.5117)	(0.4454)	(0.4525)
non_EEA × 2001	-0.0561	-0.1100	-0.1390	-0.1147
	(0.4716)	(0.5536)	(0.4225)	(0.4911)
non_EEA × 2002	-0.4317	-0.4568	-0.2046	-0.1928
	(0.5515)	(0.5675)	(0.4262)	(0.4417)
non_EEA × 2003	-0.8129*	-0.6582	-0.8603**	-0.6638.
	(0.3606)	(0.4122)	(0.2893)	(0.3402)
non_EEA × 2004	–	–	-0.0433	-0.0241
			(0.2607)	(0.2719)
non_EEA × 2005	-0.0819	-0.0674	-0.1735	-0.1222
	(0.2847)	(0.2975)	(0.2107)	(0.2186)
non_EEA × 2006	-0.0104	-0.0393	-0.0809	-0.0651
	(0.2885)	(0.3026)	(0.2161)	(0.2249)
non_EEA × 2007	0.1450	0.2505	0.0443	0.1774
	(0.2816)	(0.2969)	(0.2063)	(0.2176)
non_EEA × 2008	–	–	–	–
non_EEA × 2009	–	–	-0.0610	-0.1021
			(0.1780)	(0.1863)
non_EEA × 2010	–	–	-0.1063	-0.0132
			(0.1798)	(0.1890)
non_EEA × 2011	–	–	-0.1344	-0.1209
			(0.1758)	(0.1829)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes
supervisor	No	No	No	No
Fit statistics				
Observations	161	148	437	396
R^2	0.2855	0.2791	0.2109	0.1982
Adjusted R^2	0.1069	0.0785	0.1290	0.1053

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1, ".": 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$. *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\epsilon\ [g-4, g]$. *macro_fields* are broad disciplinary categories.

C Additional Tables & Figures

Table C.1: Summary Statistics by Group, 2000-2020

Variable	Average	StdDev	Missing	Average by Group				
				Dutch	Foreign	EEA	non-EEA	Chinese
Controls								
female	0.42	0.49	707	0.43	0.42	0.48	0.38	0.38
male	0.58	0.49	707	0.57	0.58	0.52	0.62	0.62
age	32.56	6.22	893	32.46	32.71	32.27	32.96	31.28
pre_PhD_publications	1.21	2.60	0	1.26	1.14	1.23	1.09	1.37
PhD_publications	5.43	7.52	0	5.67	5.12	5.46	4.92	6.46
prior_foreign_affiliation	0.12	0.33	0	0.07	0.19	0.18	0.20	0.24
supervisor_productivity	92.90	72.12	2278	96.95	87.50	85.65	88.57	90.31
Other Controls								
cit_pre_PhD_publications	8.05	19.20	0	9.34	6.30	7.92	5.38	5.51
cit_PhD_publications	12.13	17.94	0	13.25	10.61	11.75	9.96	10.97
cit_supervisor_productivity	1892.16	2824.09	2278	2069.51	1655.88	1696.97	1632.34	1504.80
Dependent Variables								
Post_Doc	0.53	0.50	1714	0.57	0.47	0.48	0.46	0.49
Post_Doc_Productivity	3.41	7.34	1714	3.43	3.37	3.44	3.32	5.33
Leave	0.25	0.43	6738	0.13	0.45	0.40	0.48	0.54
EEA	0.10	0.30	6738	0.07	0.14	0.16	0.13	0.07
Anglosax	0.06	0.24	6738	0.05	0.07	0.07	0.08	0.07
Origin	0.20	0.40	9644		0.20	0.15	0.24	0.37
Extended Dependent Variables								
Post_Doc_Ext	0.57	0.50	3954	0.61	0.51	0.52	0.50	0.48
Post_Doc_Productivity_Ext	7.37	15.84	3954	7.35	7.42	7.74	7.23	11.40
Leave_Ext	0.33	0.47	8535	0.17	0.59	0.52	0.64	0.76
EEA_Ext	0.12	0.33	8535	0.10	0.16	0.19	0.14	0.08
Anglosax_Ext	0.07	0.26	8535	0.06	0.09	0.09	0.10	0.14
Origin_Ext	0.29	0.45	10309		0.29	0.22	0.34	0.51
Fields								
physical sciences	0.48	0.50	2278	0.35	0.65	0.58	0.70	0.76
health sciences	0.34	0.48	2278	0.48	0.17	0.21	0.14	0.08
life sciences	0.17	0.38	2278	0.17	0.18	0.21	0.16	0.16

Table C.3: Count of Top 20 Countries of birth of foreign graduates, 2000-2020

Country	Count	Country	Count
China	676	United States	148
Germany	430	Indonesia	131
Italy	352	Turkey	115
India	280	Spain	109
Iran	248	Russia	106
Belgium	171	Greece	95
France	149	Brazil	88
United Kingdom	74	Poland	82
Portugal	73	Pakistan	70
Romania	62	Colombia	58

Table C.2: Summary Statistics of Chinese Graduates by Group (CSC vs non-CSC), 2000-2020

Variable	**Average**	**StdDev**	**Missing**	**Average by Group**	
				CSC	**non-CSC**
Controls					
female	0.38	0.48	58	0.40	0.36
male	0.62	0.48	58	0.60	0.64
age	31.28	4.01	51	30.46	31.65
pre_PhD_publications	1.37	4.03	0	1.05	1.52
PhD_publications	6.46	14.80	0	5.29	7.02
prior_foreign_affiliation	0.36	0.48	0	0.39	0.35
supervisor_productivity	90.31	67.20	105	90.45	90.24
Other Controls					
cit_pre_PhD_publications	5.51	15.19	0	5.09	5.71
cit_PhD_publications	10.97	17.00	0	13.02	10.00
cit_supervisor_productivity	1504.80	1778.73	105	1601.69	1456.49
Dependent Variables					
Post_Doc	0.49	0.50	157	0.56	0.46
Post_Doc_Productivity	5.33	15.92	157	4.89	5.47
Leave	0.54	0.50	445	0.63	0.50
EEA	0.07	0.25	445	0.04	0.08
Anglosax	0.07	0.26	445	0.06	0.08
Origin	0.37	0.48	445	0.52	0.31
Dependent Variables Extended					
Post_Doc_Ext	0.48	0.50	331	0.62	0.46
Post_Doc_Productivity_Ext	11.40	29.57	331	12.32	11.23
Leave_Ext	0.76	0.43	558	0.85	0.74
EEA_Ext	0.08	0.27	558	0.07	0.08
Anglosax_Ext	0.14	0.34	558	0.07	0.15
Origin_Ext	0.51	0.50	558	0.70	0.45
Fields					
physical sciences	0.76	0.43	105	0.70	0.79
health sciences	0.08	0.27	105	0.07	0.08
life sciences	0.16	0.37	105	0.23	0.13

Table C.4: Country Group Classification by Enlargement and Labor Market Access

Group	**Countries**	**EU Enlargement**	**Restrictions Lifted (NL)**
EEA (pre-2000)	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, Norway, Iceland, Liechtenstein, United Kingdom, Switzerland	–	–
2004 Enlargement	Cyprus, Czechia, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, Slovenia	2004	2007
2007 Enlargement	Romania, Bulgaria	2007	2014
2013 Enlargement	Croatia	2013	2023*

*Not included in the sample as it falls outside the observation period.

Source: Kahanec et al. (2016)

4 Robustness Checks

4.1 Exclusion Chinese Students

4.1.1 All Chinese Graduates Excluded

Table 4.1: Probability to Take a Postdoctoral Position by Foreign Status - Chinese Graduates Excluded, 2000-2018

Dependent Variable:	Post_Doc			
Model:	(1)	(2)	(3)	(4)
foreign	-0.0490***	-0.0379**	-0.0132	
	(0.0148)	(0.0129)	(0.0199)	
non_EEA				0.0070
				(0.0234)
EEA				-0.0379*
				(0.0182)
female	-0.0168	-0.0168	-0.0295	-0.0276
	(0.0111)	(0.0114)	(0.0185)	(0.0184)
age	-0.0013	-0.0012	-0.0028**	-0.0029**
	(0.0008)	(0.0009)	(0.0010)	(0.0009)
pre_PhD_publications	-0.0118	-0.0121	-0.0148	-0.0148
	(0.0092)	(0.0095)	(0.0123)	(0.0123)
PhD_publications	0.0371***	0.0345***	0.0333***	0.0334***
	(0.0045)	(0.0038)	(0.0050)	(0.0051)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	7,902	6,258	6,258	6,258
R^2	0.26105	0.25602	0.52622	0.52666
Within R^2	0.24213	0.22404	0.20007	0.20081

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the EEA and non-EEA coefficients in column (4) is statistically significant at the 5 percent level.

Table 4.2: Postdoctoral Productivity by Foreign Status - Chinese Graduates Excluded (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications			
Model:	(1)	(2)	(3)	(4)
foreign	0.9901	1.0247	1.0717	
	(0.0546)	(0.0570)	(0.0459)	
non_EEA				1.0703
				(0.0517)
EEA				1.0732
				(0.0639)
female	0.8507***	0.8460***	0.8369***	0.8368***
	(0.0404)	(0.0296)	(0.0509)	(0.0522)
age	1.0127***	1.0116***	1.0028	1.0029
	(0.0033)	(0.0030)	(0.0046)	(0.0046)
pre_PhD_publications	0.9326***	0.9381***	0.9933	0.9933
	(0.0144)	(0.0141)	(0.0193)	(0.0195)
PhD_publications	1.0659***	1.0625***	1.0790***	1.0790***
	(0.0058)	(0.0051)	(0.0104)	(0.0104)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	7,902	6,258	5,510	5,510
Squared Correlation	0.28524	0.31523	0.67115	0.67116
Pseudo R^2	0.25007	0.26541	0.48228	0.48228
BIC	54,944.3	44,511.9	42,310.1	42,318.7

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.212 implies a 21.2% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.3: Probability to Migrate by Foreign Status, Chinese Graduates Excluded, 2000-2018

Dependent Variable:	Leave			
Model:	(1)	(2)	(3)	(4)
foreign	0.2298***	0.2113***	0.1584***	
	(0.0155)	(0.0127)	(0.0285)	
non_EEA				0.1934***
				(0.0296)
EEA				0.1175**
				(0.0432)
female	-0.0208**	-0.0061	0.0030	0.0056
	(0.0080)	(0.0109)	(0.0091)	(0.0090)
age	-0.0020**	-0.0010	-0.0007	-0.0007
	(0.0008)	(0.0010)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0006	-0.0011	0.0043	0.0043
	(0.0017)	(0.0020)	(0.0041)	(0.0042)
PhD_publications	0.0006	0.0012*	0.0015	0.0016
	(0.0008)	(0.0006)	(0.0023)	(0.0025)
prior_foreign_affiliation	0.2689***	0.2741***	0.2509***	0.2510***
	(0.0219)	(0.0273)	(0.0240)	(0.0239)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	3,887	3,141	3,141	3,141
R^2	0.19124	0.18816	0.59366	0.59484
Within R^2	0.15673	0.14541	0.11297	0.11556

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.4: Migration Patterns of Foreign Doctoral Recipients by Destinations, Chinese Graduates Excluded, 2000-2018

Dependent Variables:	Leave	EEA	Anglosax	Origin
Model:	(1)	(2)	(3)	(4)
non_EEA	0.0618*	-0.0119	0.0082	0.0430
	(0.0310)	(0.0310)	(0.0132)	(0.0336)
female	-0.0474**	0.0085	-0.0012	-0.0447**
	(0.0202)	(0.0106)	(0.0193)	(0.0199)
age	0.0034	-0.0009	-0.0040**	0.0072**
	(0.0033)	(0.0016)	(0.0018)	(0.0027)
pre_PhD_publications	-0.0022	-0.0038	0.0010	0.0006
	(0.0052)	(0.0029)	(0.0029)	(0.0056)
PhD_publications	0.0033	0.0011	0.0014	0.0009
	(0.0026)	(0.0017)	(0.0012)	(0.0030)
prior_foreign_affiliation	0.3081***	0.0329	0.0629**	0.2031***
	(0.0516)	(0.0209)	(0.0205)	(0.0253)
supervisor_productivity	-0.0004	-2.75×10^{-6}	-0.0001	-0.0004**
	(0.0003)	(8.68×10^{-5})	(0.0001)	(0.0002)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes
supervisor	No	No	No	No
Fit statistics				
Observations	1,037	1,037	1,037	1,037
R^2	0.14596	0.02838	0.04611	0.12839
Within R^2	0.10383	0.00314	0.02458	0.09642

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, Canada or Australia and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country, and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* refers to broad disciplinary groupings.

4.1.2 Only Chinese with CSC Excluded

Table 4.5: Probability to Take a Postdoctoral Position by Foreign Status - Chinese Graduates with CSC Excluded, 2000-2018

Dependent Variable:	Post_Doc			
Model:	(1)	(2)	(3)	(4)
foreign	-0.0596***	-0.0478**	-0.0250	
	(0.0185)	(0.0175)	(0.0215)	
non_EEA				-0.0174
				(0.0294)
EEA				-0.0362**
				(0.0156)
female	-0.0172	-0.0188	-0.0265	-0.0257
	(0.0098)	(0.0106)	(0.0154)	(0.0149)
age	-0.0017**	-0.0019**	-0.0034***	-0.0034***
	(0.0007)	(0.0008)	(0.0008)	(0.0008)
pre_PhD_publications	-0.0065	-0.0060	-0.0060	-0.0059
	(0.0108)	(0.0107)	(0.0136)	(0.0135)
PhD_publications	0.0289***	0.0260***	0.0230***	0.0230***
	(0.0027)	(0.0024)	(0.0038)	(0.0039)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	8,241	6,546	6,546	6,546
R^2	0.21935	0.21397	0.48650	0.48658
Within R^2	0.19904	0.17977	0.14940	0.14954

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.6: Postdoctoral Publications and Foreign Status - Chinese Graduates with CSC Excluded (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications			
Model:	(1)	(2)	(3)	(4)
Variables				
foreign	0.9758	1.0261	1.0868*	
	(0.0807)	(0.0781)	(0.0477)	
non_EEA				1.1107
				(0.0755)
EEA				1.0558
				(0.0511)
female	0.7912***	0.7881***	0.8415***	0.8427***
	(0.0419)	(0.0316)	(0.0549)	(0.0552)
age	1.0064*	1.0035	1.0038	1.0037
	(0.0033)	(0.0032)	(0.0037)	(0.0037)
pre_PhD_publications	1.0088	1.0066	0.9949	0.9953
	(0.0095)	(0.0082)	(0.0196)	(0.0195)
PhD_publications	1.0230***	1.0232***	1.0731***	1.0731***
	(0.0027)	(0.0024)	(0.0056)	(0.0056)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	8,241	6,546	5,813	5,813
Squared Correlation	0.17359	0.20072	0.62459	0.62472
Pseudo R^2	0.18255	0.20219	0.48372	0.48377
BIC	65,906.1	53,488.1	45,935.7	45,941.5

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.212 implies a 21.2% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.7: Probability to Migrate by Foreign Status, Chinese Graduates with CSC Excluded, 2000-2018

Dependent Variable:	Leave			
Model:	(1)	(2)	(3)	(4)
foreign	0.2300***	0.2103***	0.1561***	
	(0.0146)	(0.0131)	(0.0266)	
non_EEA				0.1884***
				(0.0259)
EEA				0.1127**
				(0.0413)
female	-0.0254**	-0.0082	-0.0021	-3.43×10^{-5}
	(0.0088)	(0.0114)	(0.0136)	(0.0143)
age	-0.0019**	-0.0008	-0.0004	-0.0003
	(0.0008)	(0.0010)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0011	-0.0013	0.0021	0.0023
	(0.0017)	(0.0018)	(0.0038)	(0.0037)
PhD_publications	0.0012**	0.0016***	0.0026*	0.0026*
	(0.0005)	(0.0004)	(0.0012)	(0.0012)
prior_foreign_affiliation	0.2811***	0.2855***	0.2625***	0.2625***
	(0.0215)	(0.0272)	(0.0217)	(0.0212)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	4,029	3,257	3,257	3,257
R^2	0.20302	0.19852	0.59084	0.59206
Within R^2	0.16690	0.15407	0.12314	0.12576

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.8: Migration Patterns of Foreign Doctoral Recipients by Destinations, Chinese Graduates with CSC Excluded, 2000-2018

Dependent Variables:	Leave	EEA	Anglosax	Origin
Model:	(1)	(2)	(3)	(4)
non_EEA	0.0555**	-0.0231	0.0062	0.0519*
	(0.0244)	(0.0309)	(0.0144)	(0.0288)
female	-0.0532**	0.0057	-0.0055	-0.0398*
	(0.0175)	(0.0101)	(0.0164)	(0.0182)
age	0.0038	-0.0006	-0.0038*	0.0070**
	(0.0030)	(0.0015)	(0.0020)	(0.0027)
pre_PhD_publications	-0.0018	-0.0071**	0.0016	0.0037
	(0.0027)	(0.0029)	(0.0020)	(0.0041)
PhD_publications	0.0025**	0.0023	0.0003	0.0002
	(0.0010)	(0.0013)	(0.0003)	(0.0014)
prior_foreign_affiliation	0.3248***	0.0320*	0.0603**	0.2226***
	(0.0401)	(0.0151)	(0.0209)	(0.0175)
supervisor_productivity	-0.0004	8.07×10^{-6}	-6.32×10^{-5}	-0.0004***
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes
supervisor	No	No	No	No
Fit statistics				
Observations	1,153	1,153	1,153	1,153
R^2	0.15255	0.03134	0.03711	0.14174
Within R^2	0.11667	0.00809	0.01964	0.11178

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, Canada or Australia and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country, and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* refers to broad disciplinary groupings.

4.1.3 Only Chinese without CSC Excluded

Table 4.9: Probability to Take a Postdoctoral Position by Foreign Status - Chinese Graduates without CSC Excluded, 2000-2018

Dependent Variable:	Post_Doc			
Model:	(1)	(2)	(3)	(4)
foreign	-0.0467***	-0.0357**	-0.0145	
	(0.0146)	(0.0128)	(0.0204)	
non_EEA				0.0060
				(0.0238)
EEA				-0.0409**
				(0.0184)
female	-0.0168	-0.0175	-0.0288	-0.0268
	(0.0107)	(0.0111)	(0.0194)	(0.0192)
age	-0.0012	-0.0012	-0.0028**	-0.0028**
	(0.0008)	(0.0010)	(0.0010)	(0.0010)
pre_PhD_publications	-0.0119	-0.0122	-0.0149	-0.0148
	(0.0093)	(0.0096)	(0.0126)	(0.0126)
PhD_publications	0.0373***	0.0347***	0.0336***	0.0336***
	(0.0045)	(0.0038)	(0.0049)	(0.0050)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	8,000	6,351	6,351	6,351
R^2	0.26209	0.25729	0.52485	0.52534
Within R^2	0.24341	0.22564	0.20220	0.20302

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the EEA and non-EEA coefficients in column (4) is statistically significant at the 5 percent level.

Table 4.10: Postdoctoral Publications and Foreign Status - Chinese Graduates without CSC Excluded (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications			
Model:	(1)	(2)	(3)	(4)
foreign	1.0104	1.0448	1.0869*	
	(0.0557)	(0.0576)	(0.0464)	
non_EEA				1.1083**
				(0.0517)
EEA				1.0624
				(0.0661)
female	0.8453***	0.8397***	0.8230***	0.8243***
	(0.0396)	(0.0286)	(0.0563)	(0.0572)
age	1.0123***	1.0113***	1.0018	1.0018
	(0.0035)	(0.0031)	(0.0049)	(0.0049)
pre_PhD_publications	0.9319***	0.9374***	0.9935	0.9938
	(0.0144)	(0.0142)	(0.0199)	(0.0197)
PhD_publications	1.0661***	1.0626***	1.0802***	1.0802***
	(0.0058)	(0.0050)	(0.0105)	(0.0106)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	8,000	6,351	5,599	5,599
Squared Correlation	0.28026	0.30936	0.66405	0.66400
Pseudo R^2	0.24832	0.26305	0.48075	0.48078
BIC	55,980.1	45,490.2	43,059.2	43,066.0

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.212 implies a 21.2% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant.

Table 4.11: Probability to Migrate by Foreign Status, Chinese Graduates without CSC Excluded, 2000-2018

Dependent Variable:	Leave			
Model:	(1)	(2)	(3)	(4)
foreign	0.2317***	0.2124***	0.1605***	
	(0.0167)	(0.0140)	(0.0288)	
non_EEA				0.1930***
				(0.0286)
EEA				0.1208***
				(0.0360)
female	-0.0195**	-0.0053	0.0044	0.0072
	(0.0071)	(0.0102)	(0.0112)	(0.0100)
age	-0.0020**	-0.0010	-0.0007	-0.0006
	(0.0009)	(0.0010)	(0.0016)	(0.0017)
pre_PhD_publications	-0.0009	-0.0013	0.0035	0.0036
	(0.0016)	(0.0020)	(0.0040)	(0.0041)
PhD_publications	0.0007	0.0012*	0.0016	0.0017
	(0.0008)	(0.0006)	(0.0023)	(0.0024)
prior_foreign_affiliation	0.2759***	0.2825***	0.2626***	0.2619***
	(0.0241)	(0.0294)	(0.0254)	(0.0261)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes
supervisor	No	No	Yes	Yes
Fit statistics				
Observations	3,938	3,187	3,187	3,187
R^2	0.19933	0.19694	0.59419	0.59528
Within R^2	0.16304	0.15196	0.11895	0.12132

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_field* refers to broad disciplinary groupings. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is statistically significant at the 10 percent level.

Table 4.12: Migration Patterns of Foreign Doctoral Recipients by Destinations, Chinese Graduates without CSC Excluded, 2000-2018

Dependent Variables:	Leave	EEA	Anglosax	Origin
Model:	(1)	(2)	(3)	(4)
non_EEA	0.0657**	-0.0207	0.0065	0.0602*
	(0.0284)	(0.0307)	(0.0142)	(0.0312)
female	-0.0431**	0.0135	-3.97×10^{-5}	-0.0471**
	(0.0180)	(0.0110)	(0.0194)	(0.0181)
age	0.0036	-0.0006	-0.0038*	0.0067**
	(0.0033)	(0.0016)	(0.0018)	(0.0027)
pre_PhD_publications	-0.0028	-0.0034	0.0012	-0.0007
	(0.0050)	(0.0028)	(0.0028)	(0.0054)
PhD_publications	0.0035	0.0010	0.0015	0.0011
	(0.0026)	(0.0017)	(0.0011)	(0.0029)
prior_foreign_affiliation	0.3196***	0.0250	0.0646***	0.2231***
	(0.0482)	(0.0200)	(0.0181)	(0.0227)
supervisor_productivity	-0.0004	1.82×10^{-5}	-0.0001	-0.0004*
	(0.0003)	(0.0001)	(0.0001)	(0.0002)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes
supervisor	No	No	No	No
Fit statistics				
Observations	1,083	1,083	1,083	1,083
R^2	0.15310	0.02604	0.04683	0.13404
Within R^2	0.11051	0.00297	0.02598	0.10228

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, Canada or Australia and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country, and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* refers to broad disciplinary groupings.

4.2 Extended Dependent Variables

Table 4.13: Probability to Take a Postdoctoral Position by Foreign Status - Extended, 2000-2015

Dependent Variable:	Post_Doc_Ext				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	-0.0586***	-0.0537**	-0.0194		
	(0.0175)	(0.0210)	(0.0309)		
non_EEA				-0.0027	0.0257
				(0.0421)	(0.0356)
EEA				-0.0441**	-0.0403
				(0.0197)	(0.0240)
female	-0.0165	-0.0220*	-0.0397	-0.0380	-0.0376
	(0.0117)	(0.0108)	(0.0236)	(0.0238)	(0.0251)
age	-0.0009	-0.0012	-0.0037**	-0.0038**	-0.0028*
	(0.0007)	(0.0008)	(0.0015)	(0.0015)	(0.0013)
pre_PhD_publications	-0.0040	-0.0030	-0.0040	-0.0039	-0.0230
	(0.0146)	(0.0148)	(0.0224)	(0.0223)	(0.0155)
PhD_publications	0.0296***	0.0258***	0.0234***	0.0234***	0.0340***
	(0.0039)	(0.0042)	(0.0073)	(0.0073)	(0.0057)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	6,412	4,842	4,842	4,842	4,605
R^2	0.23025	0.22017	0.52508	0.52546	0.55808
Within R^2	0.21039	0.18578	0.16060	0.16127	0.20137

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. *Post_Doc_Ext* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+7]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. Column (5) excludes Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant, while significant at 5% level in column (5).

Table 4.14: Postdoctoral Productivity by Foreign Status (Exponentiated Coefficients) - Extended, 2000-2015

Dependent Variable:	Post_Doc_Publications_Ext (IRRs)				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	1.0165	1.0320	1.1263**		
	(1.0551)	(1.0582)	(1.0589)		
non_EEA				1.1702	1.1513**
				(1.1022)	(1.0709)
EEA				1.0714	1.1195
				(1.0571)	(1.0740)
female	0.7560***	0.7458***	0.7572***	0.7599***	0.7658***
	(1.0666)	(1.0449)	(1.0306)	(1.0282)	(1.0274)
age	1.0076*	1.0044	1.0017	1.0016	1.0017
	(1.0045)	(1.0047)	(1.0034)	(1.0034)	(1.0039)
pre_PhD_publications	1.0159	1.0135	0.9994	0.9999	0.9901
	(1.0109)	(1.0095)	(1.0313)	(1.0314)	(1.0321)
PhD_publications	1.0230***	1.0233***	1.0773***	1.0775***	1.0832***
	(1.0032)	(1.0029)	(1.0145)	(1.0148)	(1.0190)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	6,412	4,842	4,310	4,310	4,077
Squared Correlation	0.15730	0.19478	0.71247	0.71322	0.72108
Pseudo R^2	0.19991	0.22566	0.55961	0.55975	0.56275
BIC	101,536.9	77,677.7	51,828.2	51,824.0	47,425.7

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications_Ext* is the number of publications in years $[g+2, g+7]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_fields* are broad disciplinary categories. Fewer observations in columns with supervisor fixed effects are due to Poisson dropping groups with no within-group variation in the outcome. Column (5) excludes Chinese graduates. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (4) and (5) is not statistically significant.

Table 4.15: Probability to Migrate by Foreign Status - Extended, 2000-2015

Dependent Variable:	Leave_Ext					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
foreign	0.3246***	0.3042***	0.2665***			
	(0.0267)	(0.0314)	(0.0794)			
non_EEA				0.3109**	0.2847**	0.2995**
				(0.1191)	(0.1185)	(0.1154)
EEA				0.2064***	0.1931***	0.1993***
				(0.0466)	(0.0512)	(0.0525)
female	-0.0014	0.0171	0.0085	0.0130	0.0207	0.0211
	(0.0199)	(0.0183)	(0.0289)	(0.0279)	(0.0323)	(0.0349)
age	-0.0049***	-0.0036**	-0.0034	-0.0033	-0.0028	-0.0029
	(0.0012)	(0.0014)	(0.0037)	(0.0038)	(0.0039)	(0.0038)
pre_PhD_publications	0.0031	0.0020	0.0021	0.0019	0.0065	0.0066
	(0.0028)	(0.0026)	(0.0066)	(0.0065)	(0.0092)	(0.0090)
PhD_publications	-0.0010	-0.0006	0.0004	0.0004	-0.0019	-0.0020
	(0.0008)	(0.0009)	(0.0017)	(0.0016)	(0.0023)	(0.0021)
prior_foreign_affiliation	0.2429***	0.2454***	0.2125**	0.2146**	0.2313**	0.2347**
	(0.0378)	(0.0400)	(0.0937)	(0.0924)	(0.0996)	(0.1003)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	2,493	1,918	1,918	1,918	1,842	1,857
R^2	0.25556	0.26220	0.68995	0.69160	0.69103	0.69050
Within R^2	0.19791	0.18454	0.14159	0.14614	0.13845	0.14708

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+7]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is not statistically significant in columns (4), (5) and (6).

Table 4.16: Migration of Dutch doctoral recipients by destination type - Extended, 2000-2015

Dependent Variables:	Leave_Ext	EEA_Ext	Anglosax_Ext
Model:	(1)	(2)	(3)
female	0.0490***	0.0273*	0.0201
	(0.0150)	(0.0140)	(0.0120)
age	-0.0048***	-0.0017*	-0.0030***
	(0.0014)	(0.0008)	(0.0006)
pre_PhD_publications	-0.0005	0.0010	-0.0011
	(0.0025)	(0.0016)	(0.0015)
PhD_publications	-0.0015	-0.0015**	6.51×10^{-6}
	(0.0009)	(0.0006)	(0.0007)
prior_foreign_affiliation	0.1146*	0.0674	0.0394
	(0.0633)	(0.0463)	(0.0241)
supervisor_productivity	-0.0002*	-0.0003**	2.85×10^{-5}
	(9.15×10^{-5})	(8.52×10^{-5})	(0.0001)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	1,233	1,233	1,233
R^2	0.07030	0.05660	0.04548
Within R^2	0.02421	0.01441	0.01276

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave_Ext* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication in the period $[g+2, g+7]$, where g is the graduation year. *Anglosax_Ext* is equal to 1 if *Leave_Ext*=1 and the destination is in the US, UK, Canada, Australia, New Zealand, or Ireland. *EEA_Ext* equals 1 if *Leave_Ext*=1 and the destination is in the EEA (excluding UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate had a foreign affiliation in any publication in $[g-4, g]$. *supervisor_productivity* measures the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* captures broad disciplinary categories.

Table 4.17: Migration by destination type and origin affinity - Extended, 2000-2015

Dependent Variables:	Leave	EEA	Anglosax		Origin	
Sample:			Full		Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.1214**	-0.0350	0.0305	0.0993**	0.0825*	0.0963**
	(0.0443)	(0.0494)	(0.0238)	(0.0437)	(0.0424)	(0.0408)
female	-0.0413	0.0059	0.0098	-0.0419	-0.0513	-0.0474
	(0.0398)	(0.0277)	(0.0181)	(0.0457)	(0.0440)	(0.0461)
age	-0.0017	-0.0039**	-0.0042*	0.0060*	0.0062**	0.0058*
	(0.0035)	(0.0015)	(0.0022)	(0.0030)	(0.0027)	(0.0028)
pre_PhD_publications	0.0014	-0.0071*	0.0042	0.0063*	0.0030	0.0026
	(0.0024)	(0.0034)	(0.0031)	(0.0029)	(0.0059)	(0.0060)
PhD_publications	8.77×10^{-5}	0.0027*	-2.21×10^{-6}	-0.0028**	-0.0010	-0.0010
	(0.0011)	(0.0014)	(0.0008)	(0.0012)	(0.0029)	(0.0029)
prior_foreign_affiliation	0.3236***	-0.0119	0.0171	0.3078***	0.2906***	0.2958***
	(0.0394)	(0.0253)	(0.0172)	(0.0257)	(0.0375)	(0.0295)
supervisor_productivity	0.0002	0.0002	-3.96×10^{-5}	-2.01×10^{-5}	0.0002	0.0001
	(0.0003)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	685	685	685	685	609	624
R^2	0.20899	0.06741	0.04340	0.18310	0.17078	0.17212
Within R^2	0.12548	0.01798	0.01519	0.13909	0.12985	0.12958

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave_Ext*, *EEA_Ext*, and *Anglosax_Ext* are binary variables equal to 1 if the graduate has a post-PhD foreign affiliation in the respective destination type, excluding their country of origin. *Origin_Ext* equals 1 if the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* in $[g-2, g+1]$, *prior_foreign_affiliation* indicates at least one foreign publication in $[g-4, g]$, and *supervisor_productivity* counts supervisor publications in $[g-4, g+1]$.

4.3 Citation-weighted Publications instead of Publications

4.3.1 Graduates Citation-weighted Publications

Table 4.18: Probability to Take a Postdoctoral Position by Foreign Status, 2000-2018

Dependent Variable:	Post_Doc				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	-0.0370	-0.0194	0.0068		
	(0.0212)	(0.0183)	(0.0215)		
non_EEA				0.0154	0.0273
				(0.0321)	(0.0299)
EEA				-0.0065	-0.0019
				(0.0146)	(0.0149)
female	-0.0438***	-0.0451***	-0.0437*	-0.0428*	-0.0491**
	(0.0100)	(0.0125)	(0.0214)	(0.0209)	(0.0222)
age	-0.0002	-0.0010	-0.0021	-0.0021	-0.0019
	(0.0012)	(0.0012)	(0.0015)	(0.0016)	(0.0015)
cit_pre_PhD_publications	0.0018**	0.0018***	0.0024***	0.0024***	0.0022**
	(0.0007)	(0.0006)	(0.0007)	(0.0007)	(0.0008)
cit_PhD_publications	0.0102***	0.0098***	0.0098***	0.0098***	0.0096***
	(0.0009)	(0.0011)	(0.0013)	(0.0013)	(0.0013)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	6,639	6,639	6,258
R^2	0.18339	0.19233	0.48857	0.48868	0.49825
Within R^2	0.16233	0.15742	0.15647	0.15665	0.15283

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *macro_field* refers to broad disciplinary groupings. *cit_pre_PhD_publications* counts citations to publications in $[g-4, g-3]$, *cit_PhD_publications* those in $[g-2, g+1]$. Column (5) excludes Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates is not statistically significant in columns (4) and (5).

Table 4.19: Postdoctoral Productivity by Foreign Status (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications (IRRs)				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
foreign	1.171	1.253	1.261***		
	(1.077)	(1.071)	(1.085)		
non_EEA				1.296*	1.115
				(1.151)	(1.085)
EEA				1.208***	1.239***
				(1.072)	(1.086)
female	0.726***	0.716***	0.701***	0.703***	0.721***
	(1.044)	(1.036)	(1.071)	(1.071)	(1.082)
age	1.012***	1.009***	1.013**	1.014***	1.014***
	(1.004)	(1.003)	(1.005)	(1.005)	(1.004)
cit_pre_PhD_publications	1.0039**	1.0037**	1.0043	1.0043	1.0039
	(1.002)	(1.002)	(1.003)	(1.003)	(1.003)
cit_PhD_publications	1.011***	1.012***	1.016***	1.016***	1.014***
	(1.001)	(1.001)	(1.005)	(1.005)	(1.004)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	5,902	5,902	5,510
Squared Correlation	0.02511	0.03341	0.34480	0.34615	0.37995
Pseudo R^2	0.08858	0.10348	0.36488	0.36498	0.36342
BIC	74,540.5	61,079.0	53,959.4	53,961.7	48,792.9

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *macro_fields* are broad disciplinary categories. *cit_pre_PhD_publications* counts citations to publications in $[g-4, g-3]$, *cit_PhD_publications* those in $[g-2, g+1]$. Fewer observations in columns with supervisor fixed effects are due to Poisson dropping groups with no within-group variation in the outcome. Column (5) excludes Chinese graduates. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (4) and (5) is not statistically significant.

Table 4.20: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
foreign	0.2315***	0.2114***	0.1579***			
	(0.0159)	(0.0149)	(0.0284)			
non_EEA				0.1866***	0.1912***	0.1906***
				(0.0294)	(0.0296)	(0.0288)
EEA				0.1180***	0.1191**	0.1224***
				(0.0365)	(0.0440)	(0.0368)
female	-0.0257***	-0.0098	-0.0071	-0.0049	0.0014	0.0031
	(0.0071)	(0.0104)	(0.0108)	(0.0107)	(0.0087)	(0.0106)
age	-0.0018*	-0.0008	0.0003	0.0004	3.8×10^{-5}	5.49×10^{-5}
	(0.0009)	(0.0011)	(0.0019)	(0.0019)	(0.0020)	(0.0019)
cit_pre_PhD_publications	-3.06×10^{-5}	-0.0001	-0.0003	-0.0003	-0.0001	-0.0002
	(0.0003)	(0.0004)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
cit_PhD_publications	0.0005	0.0004	0.0009*	0.0009*	0.0008	0.0008
	(0.0003)	(0.0003)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
prior_foreign_affiliation	0.2899***	0.2980***	0.2889***	0.2885***	0.2655***	0.2751***
	(0.0236)	(0.0282)	(0.0230)	(0.0232)	(0.0258)	(0.0277)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21006	0.20567	0.58891	0.58993	0.59350	0.59429
Within R^2	0.17217	0.15895	0.12315	0.12531	0.11262	0.11917

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *cit_pre_PhD_publications* counts citations to publications in $[g-4, g-3]$, *cit_PhD_publications* those in $[g-2, g+1]$. . *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically not statistically significant in columns (4) and (5) while significant at the 10% level in column (6).

Table 4.21: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
female	0.0214*	0.0246**	0.0011
	(0.0106)	(0.0094)	(0.0077)
age	-0.0033***	-0.0012**	-0.0020***
	(0.0005)	(0.0004)	(0.0005)
cit_pre_PhD_publications	-0.0002	0.0002	-0.0003
	(0.0004)	(0.0003)	(0.0002)
cit_PhD_publications	0.0005	0.0001	0.0004
	(0.0004)	(0.0002)	(0.0003)
prior_foreign_affiliation	0.2152***	0.1030*	0.0961***
	(0.0537)	(0.0473)	(0.0153)
supervisor_productivity	-0.0002*	-0.0002**	0.0000
	(0.0001)	(0.0001)	(0.0001)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.07098	0.04047	0.04016
Within R^2	0.04676	0.02033	0.02321

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *cit_pre_PhD_publications* counts citations to publications in $[g-4, g-3]$, *cit_PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.22: Migration of foreign doctoral recipients by destination type and origin, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax		Origin	
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.0595**	-0.0272	0.0069	0.0614*	0.0387	0.0555
	(0.0238)	(0.0301)	(0.0150)	(0.0280)	(0.0347)	(0.0322)
female	-0.0542***	0.0090	-0.0066	-0.0437**	-0.0443**	-0.0464**
	(0.0149)	(0.0100)	(0.0172)	(0.0155)	(0.0171)	(0.0155)
age	0.0042	-0.0005	-0.0030	0.0064**	0.0068**	0.0062**
	(0.0030)	(0.0016)	(0.0019)	(0.0027)	(0.0027)	(0.0027)
cit_pre_PhD_publications	7.95×10^{-5}	-0.0001	0.0005*	-0.0001	0.0001	-8.96×10^{-5}
	(0.0003)	(0.0004)	(0.0003)	(0.0006)	(0.0006)	(0.0006)
cit_PhD_publications	0.0008	0.0007	0.0009	-0.0014**	-0.0015*	-0.0016*
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0008)	(0.0009)
prior_foreign_affiliation	0.3415***	0.0207	0.0637***	0.2506***	0.2102***	0.2293***
	(0.0378)	(0.0131)	(0.0150)	(0.0167)	(0.0275)	(0.0248)
supervisor_productivity	-0.0004	1.52×10^{-5}	-0.0001	-0.0003*	-0.0003*	-0.0003
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0001)	(0.0002)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.15649	0.02524	0.04257	0.14818	0.13218	0.13869
Within R^2	0.11953	0.00406	0.02595	0.11828	0.10035	0.10709

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication in $[g+2, g+4]$, with g the graduation year. *EEA* equals 1 if *Leave*=1 and the destination is in the EEA (excluding the UK and Ireland) but not the graduate's country of origin. *Anglosax* is 1 if *Leave*=1 and the destination is in the US, UK, Canada, Australia, New Zealand, or Ireland, and not the graduate's country of origin. *Origin* equals 1 if *Leave*=1 and the destination is the graduate's country of origin. *cit_pre_PhD_publications* is the count of cited publications in $[g-4, g-3]$; *cit_PhD_publications* in $[g-2, g+1]$. *prior_foreign_affiliation* is 1 if a publication before or during the PhD lists a foreign institution. *supervisor_productivity* counts the supervisor's publications in $[g-4, g+1]$. All models include fixed effects for year, university, and macro field.

4.3.2 Supervisors Citation-weighted Publications

Table 4.23: Migration of Dutch doctoral recipients by destination type, 2000-2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
female	0.0216*	0.0239**	0.0020
	(0.0110)	(0.0097)	(0.0082)
age	-0.0030***	-0.0013**	-0.0017***
	(0.0006)	(0.0004)	(0.0004)
pre_PhD_publications	-0.0046***	0.0010	-0.0052***
	(0.0013)	(0.0016)	(0.0012)
PhD_publications	0.0009	-0.0006	0.0014***
	(0.0009)	(0.0007)	(0.0003)
prior_foreign_affiliation	0.2152***	0.1048*	0.0948***
	(0.0567)	(0.0479)	(0.0158)
supervisor_productivity_cit	-2.91×10^{-6}**	-2.05×10^{-6}**	-8×10^{-7}
	(1×10^{-6})	(8.73×10^{-7})	(9.45×10^{-7})
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.07050	0.03912	0.04163
Within R^2	0.04627	0.01896	0.02470

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity_cit* is the number of citations up to 5 years of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.24: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax		Origin	
Sample:			Full		Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.0585**	-0.0298	0.0046	0.0652**	0.0430	0.0602*
	(0.0241)	(0.0306)	(0.0150)	(0.0269)	(0.0330)	(0.0305)
female	-0.0508***	0.0105	-0.0048	-0.0436**	-0.0463**	-0.0482**
	(0.0155)	(0.0095)	(0.0174)	(0.0167)	(0.0194)	(0.0177)
age	0.0040	-0.0003	-0.0035*	0.0066**	0.0071**	0.0066**
	(0.0030)	(0.0015)	(0.0019)	(0.0027)	(0.0027)	(0.0027)
pre_PhD_publications	-0.0020	-0.0068**	0.0019	0.0029	0.0008	-0.0006
	(0.0027)	(0.0031)	(0.0020)	(0.0041)	(0.0056)	(0.0054)
PhD_publications	0.0025**	0.0023	0.0002	0.0002	0.0009	0.0011
	(0.0010)	(0.0014)	(0.0003)	(0.0014)	(0.0031)	(0.0030)
prior_foreign_affiliation	0.3319***	0.0250	0.0613***	0.2379***	0.2020***	0.2218***
	(0.0392)	(0.0148)	(0.0189)	(0.0178)	(0.0248)	(0.0225)
supervisor_productivity_cit	-3.17×10^{-6}	1.48×10^{-6}	1.2×10^{-6}	-9.32×10^{-6}**	-8.6×10^{-6}**	-8.91×10^{-6}**
	(6.12×10^{-6})	(2.75×10^{-6})	(2.67×10^{-6})	(3.06×10^{-6})	(3.16×10^{-6})	(3.16×10^{-6})
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.15664	0.02920	0.03705	0.14534	0.12777	0.13447
Within R^2	0.11968	0.00810	0.02033	0.11534	0.09578	0.10272

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity_cit* is the number of citations up to 5 years of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.4 Alternative Fixed Effects

4.4.1 Field Fixed Effects

Table 4.25: Probability to Take a Postdoctoral Position by Foreign Status, 2000-2018

Dependent Variable:	Post_Doc				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
foreign	-0.0575***	-0.0454**	-0.0256		
	(0.0183)	(0.0176)	(0.0218)		
non_EEA				-0.0166	0.0070
				(0.0292)	(0.0234)
EEA				-0.0396**	-0.0379*
				(0.0157)	(0.0182)
female	-0.0172*	-0.0201*	-0.0257	-0.0248	-0.0276
	(0.0094)	(0.0108)	(0.0160)	(0.0155)	(0.0184)
age	-0.0017**	-0.0021**	-0.0034***	-0.0034***	-0.0029**
	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0009)
pre_PhD_publications	-0.0066	-0.0061	-0.0061	-0.0060	-0.0148
	(0.0109)	(0.0110)	(0.0139)	(0.0138)	(0.0123)
PhD_publications	0.0291***	0.0263***	0.0232***	0.0232***	0.0334***
	(0.0027)	(0.0025)	(0.0039)	(0.0039)	(0.0051)
Field dummies					
AGRI		0.0099	1.534	1.397	0.0244
ARTS		0.0059	-0.4252	-0.3079	1.351
BIOC		-0.0343			
BUSI		-0.0084			0.0626
CENG		0.0022			-0.0130
CHEM		-0.0091			0.1111
COMP		-0.0182			
DECI		-0.0040			0.0035
DENT		0.0144			
EART		0.0614***			0.0472
ECON		0.0172	0.0007	0.0089	
ENER		-0.0285**	-0.0003	0.0102	
ENGI		0.0143			
ENVI		-0.0005			
HEAL		0.0418**			
IMMU		0.0111			
MATE		0.0027			
MATH		-0.0255**			
MULT		0.0484***			
NEUR		-0.0036			
NURS		0.0375*	-0.0009	-0.0066	
PHAR		-0.0505**			
PHYS		0.0022			
PSYC		0.0125			
SOCI		-0.0056	-0.0106	0.0101	
VETE		0.0005			
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	6,639	6,639	6,258
R^2	0.22026	0.22211	0.48533	0.48545	0.52666
Within R^2	0.20015	0.19597	0.15112	0.15132	0.20081

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. Each field dummy has value 1 if the supervisor has at least one publication in that field and 0 otherwise. The baseline field is Medicine as the most frequent one. Standard errors for field dummies are not included for readability. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant, while significant at 5% in column (5).

Table 4.26: Postdoctoral Productivity by Foreign Status (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications (IRRs)				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
foreign	0.9912	1.0090	1.1003**		
	(0.0832)	(0.0606)	(0.0462)		
non_EEA				1.1406*	1.0703
				(0.0720)	(0.0499)
EEA				1.0466	1.0733
				(0.0536)	(0.0620)
female	0.7860***	0.7838***	0.8301***	0.8324***	0.8371***
	(0.0408)	(0.0274)	(0.0603)	(0.0603)	(0.0625)
age	1.0060*	1.0046	1.0029	1.0028	1.0029
	(0.0034)	(0.0037)	(0.0040)	(0.0040)	(0.0046)
pre_PhD_publications	1.0082	1.0076	0.9951	0.9958	0.9933
	(0.0096)	(0.0089)	(0.0200)	(0.0197)	(0.0197)
PhD_publications	1.0230***	1.0231***	1.0742***	1.0743***	1.0789***
	(0.0027)	(0.0028)	(0.0056)	(0.0057)	(0.0097)
Field dummies					
AGRI		0.8337*	78.18	$1.42 \cdot 10^{-6}$	10237.73
ARTS		0.9892	$1.03 \cdot 10^{9}$	$1.87 \cdot 10^{-9}$	1.1994
BIOC		0.9584	0.3629	$6.70 \cdot 10^{-5}$	$1.44 \cdot 10^{-4}$
BUSI		0.9976	1.3835		10.3521
CENG		0.9623	$7.46 \cdot 10^{14}$		3.7621
CHEM		1.0974	$1.20 \cdot 10^{-7}$		
COMP		1.0322	$2.50 \cdot 10^{-6}$	0.0328	$6.99 \cdot 10^{-6}$
DECI		1.0449			0.2010
DENT		1.0217		43.5953	0.0219
EART		1.1466	0.1445		
ECON		1.2403***			
ENER		0.9879			
ENGI		0.9077		125.1656	0.9141
ENVI		1.0383		9.1273	
HEAL		1.1610*		9.2410	
IMMU		0.9233			
MATE		0.9749			
MATH		1.0182	0.2504	4.4304	
MULT		0.9812			
NEUR		1.0343			
NURS		1.1962**			2.2152
PHAR		0.8851			
PHYS		1.1265**			
PSYC		1.0049			
SOCI		1.0166			
VETE		0.9373			
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	5,902	5,902	5,510
Squared Correlation	0.16951	0.19456	0.62006	0.62042	0.67116
Pseudo R^2	0.18042	0.20341	0.48212	0.48225	0.48228
BIC	67,061.9	54,518.1	46,747.9	46,748.0	42,387.6

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. Each field dummy has value 1 if the supervisor has at least one publication in that field and 0 otherwise. The baseline field is Medicine as the most frequent one. Standard errors for field dummies are not included for readability. Fewer observations in columns with supervisor fixed effects are due to Poisson dropping groups with no within-group variation in the outcome. Column (5) excludes Chinese graduates. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (4) and (5) is not statistically significant.

Table 4.27: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
foreign	0.2314***	0.2119***	0.1584***			
	(0.0160)	(0.0135)	(0.0274)			
non_EEA				0.1878***	0.1934***	0.1930***
				(0.0290)	(0.0296)	(0.0286)
EEA				0.1173***	0.1175**	0.1208***
				(0.0354)	(0.0432)	(0.0361)
female	-0.0242***	-0.0049	-0.0011	0.0012	0.0056	0.0072
	(0.0075)	(0.0108)	(0.0114)	(0.0114)	(0.0090)	(0.0100)
age	-0.0019**	-0.0010	-0.0003	-0.0002	-0.0007	-0.0006
	(0.0009)	(0.0010)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0013	-0.0009	0.0017	0.0019	0.0043	0.0036
	(0.0017)	(0.0018)	(0.0036)	(0.0036)	(0.0043)	(0.0041)
PhD_publications	0.0012**	0.0017***	0.0026*	0.0026*	0.0016	0.0017
	(0.0005)	(0.0005)	(0.0012)	(0.0012)	(0.0025)	(0.0024)
prior_foreign_affiliation	0.2871***	0.2906***	0.2706***	0.2700***	0.2510***	0.2619***
	(0.0238)	(0.0283)	(0.0245)	(0.0248)	(0.0240)	(0.0261)
Field dummies						
AGRI		0.0033	0.2642	0.4943	2.596	-0.0017
ARTS		-0.0004			0.0647	-0.1569
BIOC		-0.0111			-2.846	
BUSI		0.0008			1.829	-0.0853
CENG		-0.0248	-0.0548	-0.0931		
CHEM		-0.0012				0.1125
COMP		0.0135				
DECI		0.0047	-1.820	-1.806		
DENT		-0.0246				
EART		0.0052			0.0131	
ECON		0.0415	0.7651	0.2247		
ENER		-0.0126				
ENGI		0.0235				
ENVI		0.0143				
HEAL		-0.0162				
IMMU		0.0019				
MATE		0.0148				
MATH		-0.0128				
MULT		0.0345*			0.0181	
NEUR		-0.0004				
NURS		-0.0524***				
PHAR		-0.0457				
PHYS		-0.0147				
PSYC		0.0054				
SOCI		0.0021				-0.0253
VETE		-0.0006				
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21012	0.21356	0.59089	0.59196	0.59484	0.59528
Within R^2	0.17223	0.17583	0.12736	0.12964	0.11556	0.12132

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. Each field dummy has value 1 if the supervisor has at least one publication in that field and 0 otherwise. The baseline field is Medicine as the most frequent one. Standard errors for field dummies are not included for readability. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 10% level in column (4) and at the 10% level in column (6), but not significant in column (5).

Table 4.28: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
Variables			
female	0.0231*	0.0256**	0.0012
	(0.0115)	(0.0084)	(0.0087)
age	-0.0029***	-0.0011**	-0.0018***
	(0.0008)	(0.0005)	(0.0006)
pre_PhD_publications	-0.0043**	0.0012	-0.0051***
	(0.0015)	(0.0016)	(0.0010)
PhD_publications	0.0008	-0.0006	0.0013***
	(0.0010)	(0.0008)	(0.0003)
prior_foreign_affiliation	0.2069***	0.1015*	0.0898***
	(0.0546)	(0.0467)	(0.0149)
supervisor_productivity	-0.0002*	-0.0001	-4.1×10^{-5}
	(0.0001)	(8.01×10^{-5})	(3.79×10^{-5})
Field dummies			
AGRI	-0.0149	-0.0363*	0.0207
ARTS	-0.0001	-0.0047	0.0078
BIOC	-0.0151	0.0100	-0.0285
BUSI	0.0040	-0.0247	0.0209*
CENG	-0.0175	0.0047	-0.0240**
CHEM	0.0155	0.0075	0.0023
COMP	0.0133	0.0057	0.0070
DECI	-0.0221	-0.0065	-0.0162
DENT	-0.0077	-0.0197***	0.0169
EART	0.0122	0.0017	0.0127
ECON	0.0221	0.0212	-0.0140
ENER	0.0151	0.0117	0.0038
ENGI	0.0209	0.0120	0.0076
ENVI	0.0007	0.0124	-0.0044
HEAL	0.0031	0.0018	0.0031
IMMU	-0.0150	-0.0054	-0.0074
MATE	0.0144	-0.0018	0.0174
MATH	0.0039	-0.0004	0.0087
MULT	0.0392	0.0259*	0.0084
NEUR	0.0338*	0.0208	0.0183
NURS	-0.0608***	-0.0377	-0.0228**
PHAR	-0.0002	-0.0302*	0.0248**
PHYS	0.0035	0.0041	0.0024
PSYC	0.0063	0.0067	-0.0006
SOCI	-0.0170	-0.0019	-0.0126
VETE	-0.0106	0.0124	-0.0209
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.08131	0.05086	0.05103
Within R^2	0.06258	0.03567	0.03691

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. Each field dummy has value 1 if the supervisor has at least one publication in that field and 0 otherwise. The baseline field is Medicine as the most frequent one. Standard errors for field dummies are not included for readability.

Table 4.29: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax		Origin	
Sample:			Full		Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
non_EEA	0.0437*	-0.0370	0.0039	0.0600*	0.0401	0.0551
	(0.0206)	(0.0301)	(0.0162)	(0.0283)	(0.0342)	(0.0317)
female	-0.0437**	0.0071	-0.0034	-0.0304	-0.0340	-0.0350*
	(0.0178)	(0.0091)	(0.0186)	(0.0174)	(0.0196)	(0.0176)
age	0.0031	-0.0006	-0.0031	0.0057*	0.0064**	0.0057**
	(0.0033)	(0.0017)	(0.0018)	(0.0026)	(0.0026)	(0.0026)
pre_PhD_publications	-0.0019	-0.0066*	0.0017	0.0028	-0.0001	-0.0014
	(0.0029)	(0.0031)	(0.0017)	(0.0041)	(0.0060)	(0.0056)
PhD_publications	0.0031**	0.0025	0.0003	0.0005	0.0017	0.0019
	(0.0010)	(0.0015)	(0.0003)	(0.0014)	(0.0031)	(0.0029)
prior_foreign_affiliation	0.3321***	0.0204	0.0609***	0.2461***	0.2109***	0.2308***
	(0.0394)	(0.0172)	(0.0188)	(0.0155)	(0.0235)	(0.0202)
supervisor_productivity	-0.0003	-7.67×10^{-5}	-4.85×10^{-6}	-0.0003	-0.0003	-0.0003
	(0.0002)	(7.87×10^{-5})	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Field dummies						
AGRI	0.0166	-0.0002	-0.0219	0.0579***	0.0412	0.0275
ARTS	0.0130	0.0033	-0.0124	0.0185	0.0115	0.0197
BIOC	-0.0282	-0.0187	-0.0178	-0.0034	-0.0143	-0.0033
BUSI	-0.0059	-0.0628***	0.0009	0.0391	0.0213	0.0241
CENG	-0.0374	-0.0096	-0.0350	0.0027	0.0009	0.0054
CHEM	-0.0143	-0.0330	0.0272	-0.0133	0.0060	-0.0113
COMP	-0.0025	-0.0038	0.0032	-0.0033	-0.0016	-0.0054
DECI	0.0446*	0.0095	0.0034	0.0356	0.0142	0.0210
DENT	-0.0334	0.0096	-0.0145	-0.0075	-0.0138	-0.0178
EART	-0.0129	0.0123	3.41×10^{-5}	-0.0407	-0.0415	-0.0389
ECON	0.0615	0.0273	0.0237	0.0247	0.0307	0.0276
ENER	-0.0203	0.0106	0.0104	-0.0389*	-0.0109	-0.0334*
ENGI	0.0300	-0.0094	-0.0018	0.0305	0.0214	0.0305
ENVI	0.0472*	0.0339	-0.0175	0.0301	0.0323	0.0342
HEAL	-0.0365	-0.0347	0.0063	0.0125	0.0309	0.0322
IMMU	0.0305	0.0352	0.0274*	-0.0522*	-0.0413	-0.0344
MATE	0.0221	0.0056	-0.0172	0.0359	0.0236	0.0288
MATH	-0.0256	-0.0068	0.0058	-0.0389	-0.0390	-0.0219
MULT	0.0450	0.0554*	-0.0110	-0.0067	-0.0131	-0.0087
NEUR	-0.0598	0.0218	-0.0153	-0.0403**	-0.0314*	-0.0421*
NURS	-0.0170	-0.0284	0.0097	-0.0099	-0.0040	-0.0083
PHAR	-0.0919**	-0.0064	-0.0181	-0.0847**	-0.0695*	-0.0823**
PHYS	-0.0512	-0.0117	0.0106	-0.0140	-0.0454	-0.0372
PSYC	0.0089	-0.0175	-0.0243	0.0331	0.0286	0.0355
SOCI	0.0351	0.0095	0.0089	0.0246	0.0378	0.0290
VETE	0.0377	0.0199	-0.0117	0.0355	0.0456	0.0455
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.18306	0.04474	0.05010	0.17367	0.15362	0.16090
Within R^2	0.14811	0.02859	0.03373	0.14841	0.12663	0.13353

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. Each field dummy has value 1 if the supervisor has at least one publication in that field and 0 otherwise. The baseline field is Medicine as the most frequent one. Standard errors for field dummies are not included for readability.

4.4.2 Supervisor Fixed Effects

Table 4.30: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
Variables			
female	0.0345	0.0237	0.0132
	(0.0199)	(0.0167)	(0.0118)
age	-0.0018	0.0006	-0.0024
	(0.0032)	(0.0016)	(0.0019)
pre_PhD_publications	-0.0044	0.0043	-0.0094***
	(0.0031)	(0.0028)	(0.0026)
PhD_publications	0.0021	-0.0011	0.0030
	(0.0028)	(0.0019)	(0.0023)
prior_foreign_affiliation	0.1971**	0.0983	0.0886**
	(0.0755)	(0.0729)	(0.0323)
supervisor_productivity	0.0003	0.0002	0.0002
	(0.0004)	(0.0003)	(0.0001)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	Yes	Yes	Yes
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.60820	0.63156	0.53747
Within R^2	0.04148	0.01925	0.03215

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.31: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax	Origin		
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
non_EEA	-0.0161	-0.0343	0.0383	-0.0009	-0.0338	-0.0237
	(0.0864)	(0.0478)	(0.0321)	(0.0708)	(0.1070)	(0.0946)
female	0.0234	0.0373	0.0209	-0.0297	-0.0019	-0.0121
	(0.0680)	(0.0378)	(0.0379)	(0.0818)	(0.0730)	(0.0574)
age	0.0034	0.0054	-0.0032	-0.0003	0.0005	0.0009
	(0.0103)	(0.0047)	(0.0050)	(0.0047)	(0.0045)	(0.0040)
pre_PhD_publications	-0.0009	-0.0095	0.0062	0.0003	-0.0075	-0.0085
	(0.0075)	(0.0061)	(0.0056)	(0.0087)	(0.0175)	(0.0169)
PhD_publications	0.0033	0.0035	-0.0009	0.0014	0.0055	0.0052
	(0.0028)	(0.0022)	(0.0015)	(0.0028)	(0.0091)	(0.0086)
prior_foreign_affiliation	0.3017**	0.0114	0.0927	0.1736**	0.1454	0.1781
	(0.1180)	(0.0265)	(0.0640)	(0.0635)	(0.0975)	(0.1008)
supervisor_productivity	0.0009	0.0010	-2.6×10^{-6}	-0.0001	-0.0005	-3.27×10^{-5}
	(0.0022)	(0.0009)	(0.0011)	(0.0011)	(0.0015)	(0.0013)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.70914	0.66171	0.67776	0.72642	0.75958	0.74856
Within R^2	0.10205	0.02596	0.04514	0.05753	0.04997	0.05715

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.5 EEA Enlargement

Table 4.32: Probability to Take a Postdoctoral Position & Postdoctoral Productivity by Foreign Status, 2000-2018

Dependent Variables:	Post_Doc		Post_Doc_Publications	
Sample:	Full	Chinese Excluded	Full	Chinese Excluded
Model:	(1)	(2)	(3)	(4)
	OLS	OLS	Poisson	Poisson
Variables				
non_EEA_ENL	-0.0127	0.0124	1.1394*	1.0684
	(0.0303)	(0.0250)	(0.0727)	(0.0526)
EEA_ENL	-0.0446**	-0.0427**	1.0496	1.0752
	(0.0144)	(0.0170)	(0.0541)	(0.0646)
female	-0.0242	-0.0268	0.8329***	0.8368***
	(0.0153)	(0.0182)	(0.0605)	(0.0631)
age	-0.0034***	-0.0029**	1.0028	1.0029
	(0.0008)	(0.0009)	(0.0040)	(0.0046)
pre_PhD_publications	-0.0060	-0.0148	0.9958	0.9933
	(0.0138)	(0.0123)	(0.0198)	(0.0197)
PhD_publications	0.0232***	0.0334***	1.0742***	1.0790***
	(0.0039)	(0.0051)	(0.0057)	(0.0097)
Fixed-effects				
year	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes
supervisor	Yes	Yes	Yes	Yes
Fit statistics				
Observations	6,639	6,258	5,902	5,510
Squared Correlation	0.48557	0.52689	0.62033	0.67118
Pseudo R^2	0.46125	0.52007	0.48224	0.48228
BIC	24,568.7	23,335.1	46,687.9	42,318.6

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates: *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. *EEA_ENL* and *non_EEA_ENL* are alternative definitions of the EEA group with countries entering the group at the year of the enlargement rather then when the restrictions were lifted for that country. Columns (2) and (3) exclude Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (1), (3) and (4) is not statistically significant, while significant at 5% level in column (2).

Table 4.33: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave		
Sample:	Full	Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)
Variables			
non_EEA_ENL	0.1928***	0.2001***	0.1993***
	(0.0279)	(0.0294)	(0.0280)
EEA_ENL	0.1114***	0.1110**	0.1145***
	(0.0319)	(0.0400)	(0.0331)
female	0.0020	0.0066	0.0082
	(0.0112)	(0.0087)	(0.0100)
age	-0.0002	-0.0007	-0.0006
	(0.0017)	(0.0017)	(0.0017)
pre_PhD_publications	0.0019	0.0043	0.0036
	(0.0035)	(0.0042)	(0.0041)
PhD_publications	0.0026*	0.0016	0.0017
	(0.0012)	(0.0025)	(0.0024)
prior_foreign_affiliation	0.2700***	0.2510***	0.2618***
	(0.0250)	(0.0240)	(0.0264)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	Yes	Yes	Yes
Fit statistics			
Observations	3,303	3,141	3,187
R^2	0.59231	0.59529	0.59569
Within R^2	0.13039	0.11652	0.12221

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. *EEA_ENL* and *non_EEA_ENL* are alternative definitions of the EEA group with countries entering the group at the year of the enlargement rather then when the restrictions were lifted for that country.Column (2) excludes Chinese graduates. Column (3) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 5% level in columns (1) and (3) and at the 10% level in column (3).

Table 4.34: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax	Origin		
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
non_EEA_ENL	0.0654**	-0.0315	0.0078	0.0700**	0.0464	0.0643*
	(0.0237)	(0.0320)	(0.0152)	(0.0283)	(0.0344)	(0.0318)
female	-0.0484***	0.0099	-0.0042	-0.0414**	-0.0436**	-0.0458**
	(0.0149)	(0.0096)	(0.0173)	(0.0165)	(0.0198)	(0.0179)
age	0.0039	-0.0003	-0.0036*	0.0066**	0.0071**	0.0067**
	(0.0031)	(0.0015)	(0.0020)	(0.0027)	(0.0027)	(0.0027)
pre_PhD_publications	-0.0023	-0.0068**	0.0018	0.0027	0.0005	-0.0008
	(0.0026)	(0.0030)	(0.0020)	(0.0041)	(0.0056)	(0.0053)
PhD_publications	0.0026**	0.0023	0.0003	0.0003	0.0010	0.0012
	(0.0010)	(0.0014)	(0.0003)	(0.0014)	(0.0029)	(0.0029)
prior_foreign_affiliation	0.3343***	0.0246	0.0619***	0.2401***	0.2039***	0.2241***
	(0.0392)	(0.0152)	(0.0188)	(0.0176)	(0.0249)	(0.0222)
supervisor_productivity	-0.0004	2.14×10^{-5}	-6.99×10^{-5}	-0.0004**	-0.0004**	-0.0004*
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.15947	0.02926	0.03733	0.14644	0.12888	0.13481
Within R^2	0.12264	0.00816	0.02061	0.11648	0.09693	0.10307

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories. *EEA_ENL* and *non_EEA_ENL* are alternative definitions of the EEA group with countries entering the group at the year of the enlargement rather then when the restrictions were lifted for that country.

4.6 Age Squared

Table 4.35: Probability to Take a Postdoctoral Position by Foreign Status, 2000-2018

Dependent Variable:	Post_Doc				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
foreign	-0.0573**	-0.0458**	-0.0253		
	(0.0185)	(0.0175)	(0.0216)		
non_EEA				-0.0163	0.0072
				(0.0289)	(0.0233)
EEA				-0.0392**	-0.0377*
				(0.0157)	(0.0184)
female	-0.0172*	-0.0196*	-0.0259	-0.0249	-0.0277
	(0.0094)	(0.0102)	(0.0162)	(0.0157)	(0.0186)
age	-0.0029	-0.0015	-0.0074**	-0.0074**	-0.0050
	(0.0023)	(0.0028)	(0.0033)	(0.0033)	(0.0038)
age2	1.58×10^{-5}	-4.85×10^{-6}	5.12×10^{-5}	5.08×10^{-5}	2.61×10^{-5}
	(2.58×10^{-5})	(2.95×10^{-5})	(3.29×10^{-5})	(3.3×10^{-5})	(4.08×10^{-5})
pre_PhD_publications	-0.0065	-0.0062	-0.0059	-0.0058	-0.0146
	(0.0108)	(0.0108)	(0.0138)	(0.0138)	(0.0121)
PhD_publications	0.0291***	0.0262***	0.0232***	0.0232***	0.0333***
	(0.0027)	(0.0025)	(0.0038)	(0.0039)	(0.0051)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	6,639	6,639	6,258
R^2	0.22028	0.21504	0.48550	0.48562	0.52671
Within R^2	0.20017	0.18111	0.15140	0.15160	0.20089

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Estimates. Linear Probability Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. Column (5) excludes Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant, while significant at 5% level in column (5).

Table 4.36: Postdoctoral Productivity by Foreign Status (Exponentiated Coefficients), 2000-2018

Dependent Variable:	Post_Doc_Publications (IRRs)				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
foreign	0.9935	1.0445	1.1006**		
	(0.0824)	(0.0799)	(0.0454)		
non_EEA				1.1402*	1.0699
				(0.0715)	(0.0491)
EEA				1.0477	1.0744
				(0.0525)	(0.0617)
female	0.7852***	0.7809***	0.8293***	0.8318***	0.8362***
	(0.0399)	(0.0295)	(0.0601)	(0.0602)	(0.0633)
age	0.9916	0.9909	0.9972	0.9976	0.9986
	(0.0086)	(0.0078)	(0.0214)	(0.0214)	(0.0228)
age2	1.0002**	1.0001**	1.00006	1.00006	1.00005
	(7.22×10^{-5})	(6.75×10^{-5})	(0.0002)	(0.0002)	(0.0002)
pre_PhD_publications	1.0083	1.0062	0.9958	0.9964	0.9938
	(0.0096)	(0.0083)	(0.0182)	(0.0180)	(0.0175)
PhD_publications	1.0230***	1.0230***	1.0739***	1.0739***	1.0788***
	(0.0026)	(0.0023)	(0.0059)	(0.0060)	(0.0100)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	5,902	5,902	5,510
Squared Correlation	0.16951	0.19513	0.61979	0.62015	0.67094
Pseudo R^2	0.18081	0.19951	0.48215	0.48229	0.48231
BIC	67,039.0	54,580.5	46,693.4	46,693.8	42,326.0

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Poisson Regression: coefficients are exponentiated (e^{β}) and interpreted as Incidence Rate Ratios (IRRs), which reflect multiplicative effects on the number of postdoctoral publications. For instance, an IRR of 1.1 implies a 10% increase in expected publication count for foreign graduates relative to natives. *Post_Doc_Publications* is the number of publications in years $[g+2, g+4]$, with g as the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_fields* are broad disciplinary categories. Fewer observations in columns with supervisor fixed effects are due to Poisson dropping groups with no within-group variation in the outcome. Column (5) excludes Chinese graduates. The difference between the estimated coefficients for EEA and non-EEA graduates in columns (4) and (5) is not statistically significant.

Table 4.37: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
foreign	0.2307***	0.2101***	0.1578***			
	(0.0159)	(0.0145)	(0.0275)			
non_EEA				0.1880***	0.1933***	0.1930***
				(0.0288)	(0.0291)	(0.0283)
EEA				0.1155***	0.1160**	0.1193***
				(0.0357)	(0.0432)	(0.0361)
female	-0.0242***	-0.0071	-0.0007	0.0017	0.0061	0.0078
	(0.0075)	(0.0103)	(0.0111)	(0.0110)	(0.0087)	(0.0101)
age	0.0034	0.0042	0.0043	0.0048	0.0035	0.0037
	(0.0028)	(0.0032)	(0.0047)	(0.0048)	(0.0048)	(0.0047)
age2	-6.8×10^{-5}**	-6.25×10^{-5}*	-5.47×10^{-5}	-5.94×10^{-5}	-4.91×10^{-5}	-5.1×10^{-5}
	(2.91×10^{-5})	(3.15×10^{-5})	(4.08×10^{-5})	(4.14×10^{-5})	(4.09×10^{-5})	(4.06×10^{-5})
pre_PhD_publications	-0.0015	-0.0018	0.0014	0.0015	0.0039	0.0031
	(0.0016)	(0.0017)	(0.0035)	(0.0034)	(0.0042)	(0.0041)
PhD_publications	0.0013**	0.0017***	0.0027**	0.0026*	0.0017	0.0018
	(0.0005)	(0.0004)	(0.0012)	(0.0012)	(0.0025)	(0.0024)
prior_foreign_affiliation	0.2866***	0.2926***	0.2708***	0.2702***	0.2512***	0.2622***
	(0.0238)	(0.0295)	(0.0243)	(0.0246)	(0.0238)	(0.0259)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21085	0.20690	0.59121	0.59234	0.59513	0.59558
Within R^2	0.17300	0.16026	0.12805	0.13045	0.11617	0.12196

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 10% level in columns (4) and (6), but not significant in column (5).

Table 4.38: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
Variables			
female	0.0213*	0.0238**	0.0016
	(0.0106)	(0.0093)	(0.0080)
age	-0.0017	-6.37×10^{-5}	-0.0017
	(0.0025)	(0.0015)	(0.0012)
age2	-1.59×10^{-5}	-1.55×10^{-5}	3.62×10^{-7}
	(2.65×10^{-5})	(1.74×10^{-5})	(1.06×10^{-5})
pre_PhD_publications	-0.0046***	0.0009	-0.0051***
	(0.0013)	(0.0016)	(0.0012)
PhD_publications	0.0010	-0.0004	0.0013***
	(0.0009)	(0.0007)	(0.0003)
prior_foreign_affiliation	0.2157***	0.1052*	0.0949***
	(0.0570)	(0.0483)	(0.0158)
supervisor_productivity	-0.0002	-0.0001**	1.08×10^{-5}
	(0.0001)	(5.81×10^{-5})	(5.84×10^{-5})
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
R^2	0.07118	0.04026	0.04154
Within R^2	0.04698	0.02012	0.02461

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.39: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax	Origin		
Sample:		Full		Full	Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.0588**	-0.0295	0.0045	0.0659**	0.0439	0.0611*
	(0.0233)	(0.0306)	(0.0153)	(0.0268)	(0.0328)	(0.0303)
female	-0.0475**	0.0114	-0.0032	-0.0406**	-0.0422*	-0.0448**
	(0.0157)	(0.0094)	(0.0172)	(0.0170)	(0.0204)	(0.0181)
age	0.0124*	0.0034		0.0142***	0.0140***	0.0137***
	(0.0064)	(0.0030)		(0.0038)	(0.0041)	(0.0039)
age2	-0.0001*	-4.85×10^{-5}	-3.28×10^{-5}	-9.84×10^{-5}**	-8.79×10^{-5}*	-9.06×10^{-5}**
	(5.78×10^{-5})	(3.66×10^{-5})	(1.92×10^{-5})	(3.38×10^{-5})	(4.37×10^{-5})	(3.8×10^{-5})
pre_PhD_publications	-0.0026	-0.0070**	0.0015	0.0024	3.76×10^{-5}	-0.0013
	(0.0022)	(0.0029)	(0.0021)	(0.0041)	(0.0055)	(0.0053)
PhD_publications	0.0027***	0.0024	0.0004	0.0004	0.0012	0.0014
	(0.0009)	(0.0013)	(0.0004)	(0.0015)	(0.0030)	(0.0030)
prior_foreign_affiliation	0.3330***	0.0248	0.0609***	0.2388***	0.2028***	0.2230***
	(0.0391)	(0.0150)	(0.0186)	(0.0173)	(0.0252)	(0.0225)
supervisor_productivity	-0.0004	2.06×10^{-5}	-6.74×10^{-5}	-0.0004**	-0.0004**	-0.0004*
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.16007	0.02964	0.03469	0.14725	0.13010	0.13569
Within R^2	0.12326	0.00854	0.01792	0.11732	0.09820	0.10398

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.7 prior_foreign_affiliation x foreign/non_EEA interactions

Table 4.40: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
foreign	0.1994***	0.1851***	0.1289**			
	(0.0224)	(0.0261)	(0.0495)			
non_EEA				0.1789***	0.1962***	0.1907***
				(0.0474)	(0.0484)	(0.0464)
EEA				0.1204***	0.1167***	0.1215***
				(0.0312)	(0.0367)	(0.0322)
female	-0.0237***	-0.0074	0.0017	0.0013	0.0054	0.0073
	(0.0074)	(0.0108)	(0.0113)	(0.0114)	(0.0095)	(0.0097)
age	-0.0021**	-0.0010	-0.0005	-0.0002	-0.0007	-0.0006
	(0.0009)	(0.0010)	(0.0018)	(0.0017)	(0.0017)	(0.0016)
pre_PhD_publications	-0.0021	-0.0022	0.0009	0.0018	0.0044	0.0035
	(0.0016)	(0.0018)	(0.0037)	(0.0036)	(0.0049)	(0.0045)
PhD_publications	0.0013**	0.0017***	0.0026**	0.0025*	0.0016	0.0017
	(0.0005)	(0.0004)	(0.0011)	(0.0012)	(0.0026)	(0.0025)
prior_foreign_affiliation	0.1967***	0.2139***	0.1865**	0.2571***	0.2544***	0.2591***
	(0.0455)	(0.0536)	(0.0834)	(0.0518)	(0.0544)	(0.0524)
foreign × prior_foreign_affiliation	0.1609**	0.1403*	0.1512			
	(0.0686)	(0.0779)	(0.1469)			
non_EEA × prior_foreign_affiliation				0.0374	-0.0118	0.0091
				(0.1319)	(0.1598)	(0.1419)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field		Yes	Yes	Yes	Yes	Yes
supervisor			Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21478	0.20980	0.59295	0.59208	0.59486	0.59529
Within R^2	0.17712	0.16333	0.13176	0.12990	0.11558	0.12134

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients in columns (4), (5) and (6) is not statistically significant.

4.8 PhD_publications x foreign/non_EEA interactions

Table 4.41: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
foreign	0.1995***	0.1840***	0.1371**			
	(0.0188)	(0.0174)	(0.0460)			
non_EEA				0.1876***	0.2075***	0.2076***
				(0.0333)	(0.0343)	(0.0283)
EEA				0.1173***	0.1165**	0.1197***
				(0.0351)	(0.0430)	(0.0358)
female	-0.0257***	-0.0088	-0.0016	0.0012	0.0061	0.0078
	(0.0077)	(0.0105)	(0.0111)	(0.0116)	(0.0088)	(0.0103)
age	-0.0019*	-0.0008	-0.0003	-0.0002	-0.0007	-0.0007
	(0.0009)	(0.0010)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0022	-0.0024	0.0010	0.0019	0.0050	0.0043
	(0.0017)	(0.0016)	(0.0039)	(0.0036)	(0.0050)	(0.0049)
PhD_publications	-0.0009	-0.0002	0.0010	0.0025*	0.0020	0.0021
	(0.0010)	(0.0006)	(0.0016)	(0.0014)	(0.0020)	(0.0020)
prior_foreign_affiliation	0.2875***	0.2936***	0.2717***	0.2700***	0.2503***	0.2613***
	(0.0233)	(0.0291)	(0.0234)	(0.0248)	(0.0239)	(0.0260)
foreign × PhD_publications	0.0036**	0.0030***	0.0024			
	(0.0013)	(0.0009)	(0.0029)			
non_EEA × PhD_publications				2.6×10^{-5}	-0.0016	-0.0017
				(0.0011)	(0.0017)	(0.0015)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
R^2	0.21150	0.20731	0.59125	0.59196	0.59496	0.59541
Within R^2	0.17368	0.16069	0.12813	0.12964	0.11582	0.12160

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 5% level in column (6) and not significant in all the other columns column.

Table 4.42: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax	Origin		
Sample:		Full		Full	Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
non_EEA	0.0907*	-0.0358	-0.0071	0.1079**	0.1023*	0.1194**
	(0.0430)	(0.0371)	(0.0222)	(0.0416)	(0.0469)	(0.0424)
female	-0.0483***	0.0102	-0.0051	-0.0407**	-0.0427*	-0.0455**
	(0.0148)	(0.0099)	(0.0174)	(0.0173)	(0.0202)	(0.0183)
age	0.0038	-0.0003	-0.0035*	0.0065**	0.0069**	0.0064**
	(0.0031)	(0.0014)	(0.0020)	(0.0027)	(0.0026)	(0.0026)
pre_PhD_publications	-0.0026	-0.0068**	0.0019	0.0023	0.0020	0.0006
	(0.0029)	(0.0029)	(0.0019)	(0.0040)	(0.0053)	(0.0052)
PhD_publications	0.0055**	0.0018	-0.0007	0.0040	0.0045	0.0048*
	(0.0024)	(0.0023)	(0.0014)	(0.0023)	(0.0027)	(0.0026)
prior_foreign_affiliation	0.3309***	0.0255	0.0628***	0.2357***	0.1974***	0.2178***
	(0.0394)	(0.0155)	(0.0187)	(0.0177)	(0.0252)	(0.0223)
supervisor_productivity	-0.0004	2.48×10^{-5}	-6.53×10^{-5}	-0.0004**	-0.0004**	-0.0004*
	(0.0002)	(9.62×10^{-5})	(0.0001)	(0.0002)	(0.0001)	(0.0002)
non_EEA × PhD_publications	-0.0034	0.0006	0.0012	-0.0044	-0.0063**	-0.0063**
	(0.0022)	(0.0025)	(0.0015)	(0.0025)	(0.0022)	(0.0021)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
R^2	0.15985	0.02915	0.03772	0.14869	0.13483	0.13992
Within R^2	0.12304	0.00805	0.02101	0.11881	0.10309	0.10837

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Linear Probability Model estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.9 Alternative Model Specifications

4.9.1 Logit Regressions

Table 4.43: Probability to Take a Postdoctoral Position by Foreign Status, 2000-2018

Dependent Variable:	Post_Doc				
Sample:	Full				Chinese Excluded
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
foreign	-0.0575***	-0.0457**	-0.0256		
	(0.0183)	(0.0174)	(0.0218)		
non_EEA				-0.0166	0.0070
				(0.0291)	(0.0234)
EEA				-0.0396**	-0.0379*
				(0.0157)	(0.0182)
female	-0.0172*	-0.0196*	-0.0257	-0.0248	-0.0276
	(0.0094)	(0.0102)	(0.0160)	(0.0155)	(0.0184)
age	-0.0017**	-0.0019*	-0.0034***	-0.0034***	-0.0029**
	(0.0008)	(0.0009)	(0.0008)	(0.0008)	(0.0009)
pre_PhD_publications	-0.0066	-0.0061	-0.0061	-0.0060	-0.0148
	(0.0109)	(0.0108)	(0.0139)	(0.0138)	(0.0123)
PhD_publications	0.0291***	0.0262***	0.0232***	0.0232***	0.0334***
	(0.0027)	(0.0025)	(0.0039)	(0.0039)	(0.0051)
Fixed-effects					
year	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes
Fit statistics					
Observations	8,339	6,639	6,639	6,639	6,258
Squared Correlation	0.22026	0.21504	0.48533	0.48545	0.52666
Pseudo R^2	0.17219	0.16798	0.46083	0.46100	0.51963
BIC	10,290.3	8,287.4	24,565.0	24,572.2	23,340.0

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Logit Estimates. *Post_Doc* is a binary variable equal to 1 if the graduate publishes in years $[g+2, g+4]$, where g is the graduation year, and 0 otherwise. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *macro_field* refers to broad disciplinary groupings. Column (5) excludes Chinese graduates from the sample. The difference between the estimated coefficients for EEA and non-EEA graduates in column (4) is not statistically significant, while significant at 5% level in column (5).

Table 4.44: Probability to Migrate by Foreign Status, 2000–2018

Dependent Variable:	Leave					
Sample:	Full				Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
foreign	0.2314***	0.2110***	0.1584***			
	(0.0160)	(0.0146)	(0.0273)			
non_EEA				0.1878***	0.1934***	0.1930***
				(0.0290)	(0.0296)	(0.0286)
EEA				0.1173***	0.1175**	0.1208***
				(0.0354)	(0.0432)	(0.0360)
female	-0.0242***	-0.0075	-0.0011	0.0012	0.0056	0.0072
	(0.0075)	(0.0105)	(0.0114)	(0.0114)	(0.0090)	(0.0100)
age	-0.0019**	-0.0008	-0.0003	-0.0002	-0.0007	-0.0006
	(0.0009)	(0.0010)	(0.0017)	(0.0017)	(0.0017)	(0.0017)
pre_PhD_publications	-0.0013	-0.0015	0.0017	0.0019	0.0043	0.0036
	(0.0017)	(0.0018)	(0.0036)	(0.0036)	(0.0042)	(0.0041)
PhD_publications	0.0012**	0.0016***	0.0026*	0.0026*	0.0016	0.0017
	(0.0005)	(0.0004)	(0.0012)	(0.0012)	(0.0025)	(0.0024)
prior_foreign_affiliation	0.2871***	0.2929***	0.2706***	0.2700***	0.2510***	0.2619***
	(0.0238)	(0.0296)	(0.0245)	(0.0248)	(0.0239)	(0.0261)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	No	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	Yes	Yes	Yes	Yes
Fit statistics						
Observations	4,080	3,303	3,303	3,303	3,141	3,187
Squared Correlation	0.21012	0.20627	0.59089	0.59196	0.59484	0.59528
Pseudo R^2	0.20576	0.20376	0.78835	0.79066	0.82871	0.81680
BIC	4,014.1	3,289.5	13,521.7	13,521.2	12,938.8	13,093.7

Clustered (university) standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Logit estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *macro_fields* are broad disciplinary categories. Column (5) excludes Chinese graduates. Column (6) excludes only Chinese graduates without CSC (Chinese Scholarship Council) funding. The difference between EEA and non-EEA coefficients is statistically significant at the 10% level in column (4) and (6), but not significant in column (5).

Table 4.45: Migration of Dutch doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax
Model:	(1)	(2)	(3)
Variables			
female	0.0213*	0.0238**	0.0016
	(0.0106)	(0.0092)	(0.0080)
age	-0.0030***	-0.0013***	-0.0016***
	(0.0006)	(0.0004)	(0.0004)
pre_PhD_publications	-0.0045***	0.0010	-0.0051***
	(0.0013)	(0.0016)	(0.0012)
PhD_publications	0.0010	-0.0004	0.0013***
	(0.0009)	(0.0007)	(0.0003)
prior_foreign_affiliation	0.2157***	0.1052*	0.0949***
	(0.0570)	(0.0483)	(0.0157)
supervisor_productivity	-0.0002	-0.0001**	1.09×10^{-5}
	(0.0001)	(5.82×10^{-5})	(5.79×10^{-5})
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations	2,104	2,104	2,104
Squared Correlation	0.07112	0.04015	0.04154
Pseudo R^2	0.11435	0.38941	-0.26950
BIC	1,492.9	425.96	-129.76

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Logit estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, or Ireland. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in an EEA country (excluding the UK and Ireland). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.46: Migration of foreign doctoral recipients by destination type, 2000–2018

Dependent Variables:	Leave	EEA	Anglosax		Origin	
Sample:			Full		Chinese Excluded	Chinese with no CSC Excl
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
non_EEA	0.0580**	-0.0299	0.0044	0.0652**	0.0430	0.0602*
	(0.0238)	(0.0306)	(0.0152)	(0.0275)	(0.0336)	(0.0312)
female	-0.0498***	0.0104	-0.0046	-0.0426**	-0.0447**	-0.0471**
	(0.0150)	(0.0094)	(0.0173)	(0.0168)	(0.0199)	(0.0181)
age	0.0039	-0.0003	-0.0036*	0.0066**	0.0072**	0.0067**
	(0.0031)	(0.0015)	(0.0020)	(0.0027)	(0.0027)	(0.0027)
pre_PhD_publications	-0.0023	-0.0068**	0.0018	0.0027	0.0006	-0.0007
	(0.0026)	(0.0030)	(0.0020)	(0.0041)	(0.0056)	(0.0054)
PhD_publications	0.0026**	0.0023	0.0003	0.0002	0.0009	0.0011
	(0.0010)	(0.0014)	(0.0003)	(0.0014)	(0.0030)	(0.0029)
prior_foreign_affiliation	0.3336***	0.0250	0.0618***	0.2393***	0.2031***	0.2231***
	(0.0392)	(0.0150)	(0.0188)	(0.0179)	(0.0253)	(0.0227)
supervisor_productivity	-0.0004	2.22×10^{-5}	-7.04×10^{-5}	-0.0004**	-0.0004**	-0.0004*
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Fixed-effects						
year	Yes	Yes	Yes	Yes	Yes	Yes
university	Yes	Yes	Yes	Yes	Yes	Yes
macro_field	Yes	Yes	Yes	Yes	Yes	Yes
supervisor	No	No	No	No	No	No
Fit statistics						
Observations	1,199	1,199	1,199	1,199	1,037	1,083
Squared Correlation	0.15865	0.02907	0.03720	0.14552	0.12839	0.13404
Pseudo R^2	0.12034	0.04226	0.22029	0.15755	0.15228	0.14966
BIC	1,790.5	1,077.9	437.36	1,284.7	1,064.1	1,158.1

Clustered (university) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Logit estimates. *Leave* is a binary variable equal to 1 if the graduate has a foreign affiliation in her last publication among those produced in years $[g+2, g+4]$, where g is the graduation year. *Anglosax* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in one of the following countries: US, UK, Canada, Australia, New Zealand, and Ireland, and none of these is the graduate's country of origin. *EEA* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in a EEA country (except the UK or Ireland), and none of these is the graduate's country of origin. *Origin* is a binary variable equal to 1 if *Leave*=1 and the foreign affiliation is in the graduate's country of origin. *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.9.2 Multinomial Logistic Regressions

Table 4.47: Multinomial Logistic Model: Migration of Dutch doctoral recipients by destination type, 2000–2018

Destination	EEA	Anglosax
female	0.381.	0.061
	(0.194)	(0.218)
age	-0.033.	-0.056**
	(0.018)	(0.022)
pre_PhD_publications	0.007	-0.152*
	(0.040)	(0.060)
PhD_publications	-0.004	0.023.
	(0.015)	(0.014)
prior_foreign_affiliation	1.406***	1.504***
	(0.231)	(0.251)
supervisor_productivity	-0.003*	0.000
	(0.002)	(0.002)
(Intercept)	-1.920	-1.381
	(1.247)	(1.309)
Fixed-effects		
year	Yes	Yes
university	Yes	Yes
macro_field	Yes	Yes
supervisor	No	No
Fit statistics		
Observations	4676	4676
Log-Likelihood	-1140.912	
AIC	2515.825	
BIC	3108.302	
McFadden's pseudo R^2	0.122	

Standard errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: Coefficients are reported as odds ratios from a multinomial logistic model. The reference category is having a Dutch affiliation in the last publication within the period $[g+2, g+4]$. Destination categories include: *Anglosax*: affiliation in an Anglophone country (United States, United Kingdom, Canada, Australia, New Zealand, or Ireland); *EEA*: affiliation in a country within the European Economic Area (excluding the Netherlands). *pre_PhD_publications* counts publications in $[g-4, g-3]$, *PhD_publications* those in $[g-2, g+1]$. *prior_foreign_affiliation* equals 1 if the graduate has a publication with a foreign affiliation in $[g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_field* denotes broad disciplinary categories.

Table 4.48: Multinomial Logistic Model: Migration of foreign doctoral recipients by destination type, 2000–2018

Destination	EEA	Anglosax	Origin
non_EEA	-0.098	0.211	0.559**
	(0.193)	(0.257)	(0.191)
PhD_publications	0.029*	0.018	0.019
	(0.012)	(0.014)	(0.012)
age	0.004	-0.040.	0.040**
	(0.017)	(0.021)	(0.014)
female	-0.013	-0.173	-0.428*
	(0.189)	(0.247)	(0.184)
pre_PhD_publications	-0.068.	0.007	-0.007
	(0.036)	(0.033)	(0.027)
prior_foreign_affiliation	0.912***	1.480***	1.895***
	(0.208)	(0.251)	(0.184)
supervisor_productivity	-0.001	-0.002	-0.004**
	(0.001)	(0.002)	(0.001)
(Intercept)	-1.953	-10.151***	-12.799***
	(1.384)	(0.881)	(0.625)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Fit statistics			
Observations		4676	
Log-Likelihood		-1140.912	
AIC		2515.825	
BIC		3108.302	
McFadden's pseudo R^2		0.122	

Standard errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: Coefficients are reported as odds ratios from a multinomial logistic model. The reference category is having a Dutch affiliation in the last publication within the period $[g+2, g+4]$. Destination categories include: *Anglosax*: the graduate has a foreign affiliation in the US, UK, Canada, Australia, New Zealand, or Ireland, and this country is not the graduate's country of origin; *EEA*: the graduate has a foreign affiliation in an EEA country (excluding the UK and Ireland), and this country is not the graduate's country of origin; *Origin*: the graduate has a foreign affiliation in her own country of origin (outside the Netherlands). *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

Table 4.49: Multinomial Logistic Model: Migration of foreign doctoral recipients by destination type - Chinese Gradates Excluded, 2000–2018

Term	EEA	Anglosax	Origin
non_EEA	0.054	0.234	0.43*
	(0.199)	(0.271)	(0.202)
female	-0.028	-0.126	-0.457*
	(0.200)	(0.267)	(0.208)
age	-0.001	-0.046*	0.044**
	(0.018)	(0.022)	(0.015)
PhD_publications	0.019	0.019	0.015
	(0.016)	(0.018)	(0.015)
pre_PhD_publications	-0.037	0.009	0.001
	(0.040)	(0.039)	(0.031)
prior_foreign_affiliation	0.887***	1.426***	1.751***
	(0.218)	(0.269)	(0.203)
supervisor_productivity	-0.001	-0.003	-0.004*
	(0.002)	(0.002)	(0.002)
(Intercept)	-1.763	-9.368***	-12.132***
	(1.398)	(0.920)	(0.659)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Model Fit			
Log-Likelihood		-980.0143	
AIC		2194.029	
BIC		2769.4	
Observations		4040	
McFadden's pseudo R^2		0.1163	

Standard errors in parentheses

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: Coefficients are reported as odds ratios from a multinomial logistic model.
The reference category is having a Dutch affiliation in the last publication within the period $[g+2, g+4]$. Destination categories include: *Anglosax*: the graduate has a foreign affiliation in the US, UK, Canada, Australia, New Zealand, or Ireland, and this country is not the graduate's country of origin; *EEA*: the graduate has a foreign affiliation in an EEA country (excluding the UK and Ireland), and this country is not the graduate's country of origin; *Origin*: the graduate has a foreign affiliation in her own country of origin (outside the Netherlands).
pre_PhD_publications is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

Table 4.50: Multinomial Logistic Model: Migration of foreign doctoral recipients by destination type - Chinese Graduates without CSC Excluded, 2000–2018

Term	EEA	Anglosax	Origin
non_EEA	-0.002	0.224	0.545**
	(0.197)	(0.267)	(0.197)
female	0.012	-0.106	-0.440*
	(0.197)	(0.262)	(0.198)
age	0.002	-0.044*	0.040**
	(0.018)	(0.022)	(0.014)
PhD_publications	0.019	0.020	0.018
	(0.016)	(0.018)	(0.014)
pre_PhD_publications	-0.036	0.007	-0.008
	(0.040)	(0.038)	(0.030)
prior_foreign_affiliation	0.874***	1.498***	1.820***
	(0.216)	(0.265)	(0.195)
supervisor_productivity	-0.001	-0.003	-0.003*
	(0.002)	(0.002)	(0.002)
(Intercept)	-1.943	-9.627***	-12.547***
	(1.400)	(0.918)	(0.643)
Fixed-effects			
year	Yes	Yes	Yes
university	Yes	Yes	Yes
macro_field	Yes	Yes	Yes
supervisor	No	No	No
Model Fit			
Log-Likelihood		-1030.034	
AIC		2294.069	
BIC		2874.6	
Observations		4224	
McFadden's pseudo R^2		0.1186	

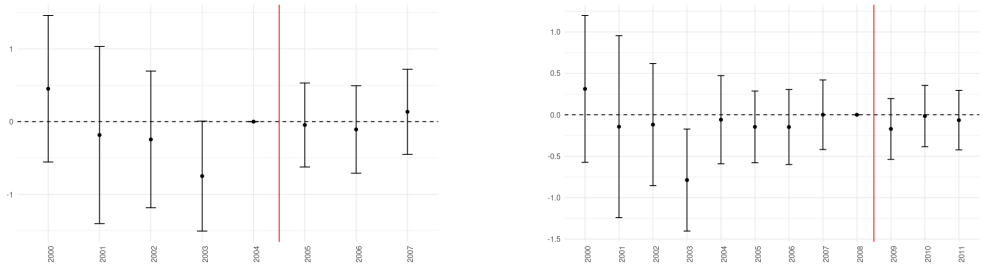
Standard errors in parentheses

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Notes: Coefficients are reported as odds ratios from a multinomial logistic model. The reference category is having a Dutch affiliation in the last publication within the period $[g+2, g+4]$. Destination categories include: *Anglosax*: the graduate has a foreign affiliation in the US, UK, Canada, Australia, New Zealand, or Ireland, and this country is not the graduate's country of origin; *EEA*: the graduate has a foreign affiliation in an EEA country (excluding the UK and Ireland), and this country is not the graduate's country of origin; *Origin*: the graduate has a foreign affiliation in her own country of origin (outside the Netherlands). *pre_PhD_publications* is the number of publications in years $[g-4, g-3]$, *PhD_publications* that in years $[g-2, g+1]$. *prior_foreign_affiliation* is a binary variable equal to 1 if the graduate has a publication with a foreign affiliation in years $\in [g-4, g]$. *supervisor_productivity* is the number of the supervisor's publications in $[g-4, g+1]$. *macro_fields* are broad disciplinary categories.

4.10 Dynamic DID Robustness Checks

4.10.1 Chinese Graduates Excluded

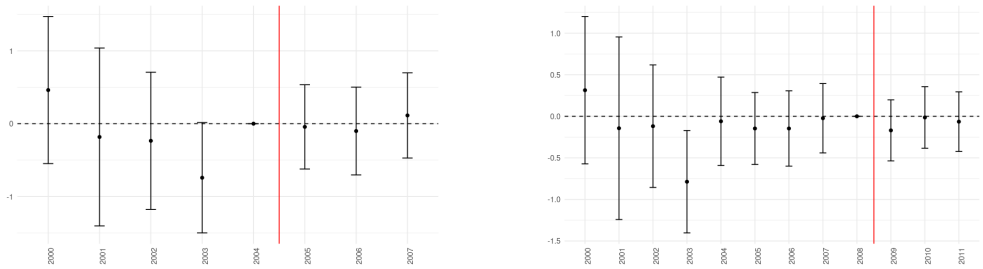


(a) Knowledge Migrants Scheme (2004) (b) Researcher EU 2005/71 (2008)

Figure 4.1: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact - Chinese Graduates Excluded

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates.

4.10.2 Chinese Graduates without CSC Excluded

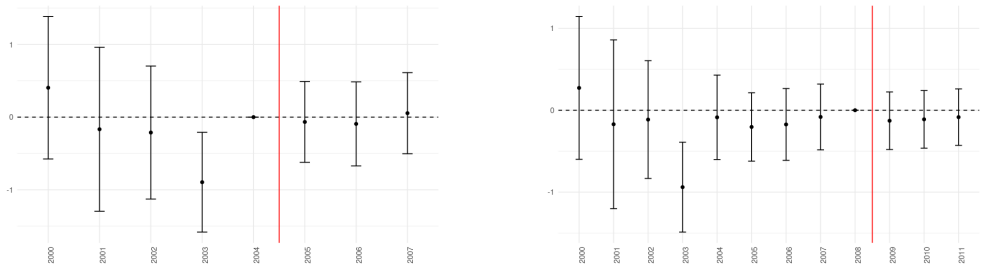


(a) Knowledge Migrants Scheme (2004) (b) Researcher EU 2005/71 (2008)

Figure 4.3: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact - Chinese Graduates without CSC Excluded

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates.

4.10.3 Chinese Graduates with CSC Excluded

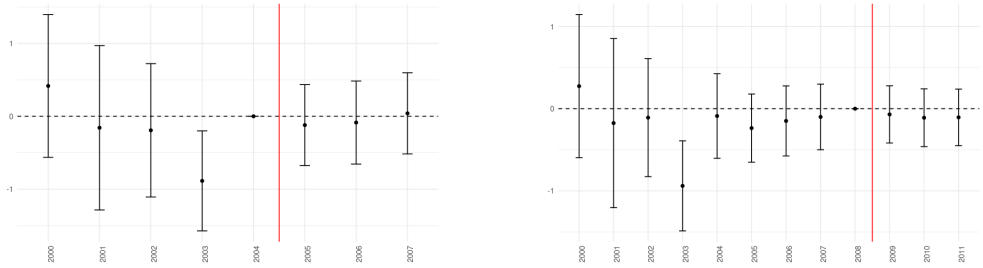


(a) Knowledge Migrants Scheme (2004) (b) Researcher EU 2005/71 (2008)

Figure 4.5: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact - Chinese Graduates with CSC Excluded

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates.

4.10.4 EEA Enlargement

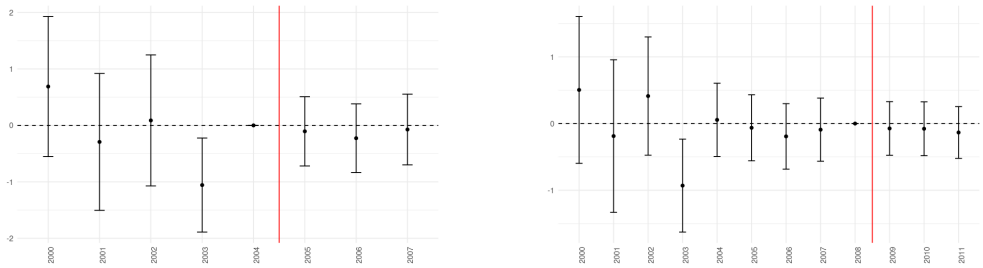


(a) Knowledge Migrants Scheme (2004) (b) Researcher EU 2005/71 (2008)

Figure 4.7: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact - EEA Enlarged

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates.

4.10.5 Extended Dependent Variable



(a) Knowledge Migrants Scheme (2004)

(b) Researcher EU 2005/71 (2008)

Figure 4.9: Knowledge Migrants Scheme (2004) and Researcher EU 2005/71 permit (2008): Estimated impact - Extended

Panels (a) and (b) report the estimated coefficients θ in equation 5, for the interaction of *non_EEA* with the year fixed effects respectively from 2000 to 2007 (2004 as reference year) and from 2000 to 2011 (2008 as reference year). The coefficients measure the impact on the probability to leave the Netherlands, conditional on having at least one *Post_Doc* publication, of the introduction of Knowledge Migrants Scheme in 2004 and the Researcher EU 2005/71 permit in 2008, respectively. The regression sample includes only foreign-born graduates.

General Conclusion

This dissertation has analyzed the post-graduation trajectories and careers of foreign doctoral graduates in Europe, focusing in particular on France and the Netherlands, while also developing an original methodological approach for inferring the graduates' countries of origin through names. The main finding of the empirical analysis is that the special visas introduced in both France and the Netherlands (respectively: the *carte de sejour "scientifique-chercheur"* and the *kennismigrant* permit) had no significant effect on the likelihood of foreign doctoral graduates to stay in the country after graduation. This implies that other, deeper structural factors may determine the foreign graduates' stay decision, which the lower immigration costs coming with cheaper and simpler immigration procedure cannot offset. This result raises broader reflections about the generalizability of the evidence to other European countries, the comparison with the United States, and the interaction with citizenship policies.

A first question is whether the absence of measurable effects can be extended to the entire European Union. Over the past two decades, the EU has promoted several directives to facilitate the mobility of highly skilled workers, among which the Blue Card, requiring member states to introduce national schemes or adapt pre-existing instruments (Cerna, 2013). The experiences of France and the Netherlands, combined with the literature on a similar policy in Germany (Wogart and Schüller, 2011), suggest that such measures may have had only limited success in retaining researchers. This indicates that the problem lies not in the absence of legal instruments but in their limited integration with long-term policies addressing career opportunities, funding, and social inclusion (Buenstorf et al., 2022).

At the international level, European evidence appears to contrast with evidence on the same topic emerging from the United States, where restrictive policies have been associated with a decline in stay rates (Ganguli and MacGarvie, 2025). One possible explanation is an asymmetry of effects: restrictive policies reduce retention, whereas permissive policies do not necessarily increase it. Moreover, in the United States there is a clear supply constraint: the demand for STEM workers from firms far exceeds the number of visas available, as shown by the need to assign H-1B visas through a lottery (only academic institutions are exempt from this cap (Amuedo-Dorantes and Furtado, 2019)). It remains unclear whether a similar pool of foreign graduates eager to stay in

Europe, but excluded for lack of visas, actually exists. This represents a promising area of future research, and the data assembled for this dissertation could serve as the basis for survey-based investigations of graduates' intentions to stay, in the spirit of earlier work for the United States (Finn, 2010).

Another difference between Europe and the United States concerns access to citizenship. In France, for example, years spent studying count toward the five-year requirement for naturalization. The same holds in the Netherlands, although only for doctoral studies (coming them along a work contract with the host institution). instead, in the United States, the year counting begins only after the acquisition of a Green Card. This difference means that many French doctoral graduates in our sample - and possibly some Dutch ones, too - might have reached citizenship before or soon after qualifying special visas. Furthermore, the French and Dutch policies under examination also apply to master's graduates who intend to continue with a PhD, suggesting that their effects may be more relevant for this group than for doctoral students already enrolled. This too represents an avenue for further research.

One additional results from our analysis, so far limited to the Netherlands case, is that foreign and native doctoral students appear to be similarly likely to engage in an academic career upon graduation, with similar productivity rates - after controlling for composition effects. This reminds of similar results for the United States by Hunt (2010), albeit referred to college graduates: foreigners contributed disproportionately to innovation inputs (publishing and, in Hunt's studies, patenting) due to their STEM specializations rather than higher skills. This implies that positive self-selection of foreign students may not be at work. Indeed, we also find no evidence of self-selection among stayers nor - similarly - among leavers towards the United States.

On top of these substantive findings, this dissertation also makes a methodological contribution by employing name-based classification to infer the origin of doctoral graduates. This approach, already used in the study of patents, demonstrates its value in contexts where administrative data are incomplete, as it enabled the reconstruction of mobility trajectories over several decades. At the same time, its potential can be further enhanced by richer and more balanced training data, for example by drawing on university registries or census records to improve classification accuracy. Alternative data sources such as alumni surveys, social media, or professional networking platforms may complement or partially substitute name-based approaches, though each carries its

own limitations in terms of coverage, selection bias, or privacy. The combination of multiple sources appears the most promising direction for future work.

In conclusion, this dissertation shows that selective migration policies directed at researchers in France and the Netherlands have not produced measurable effects on the retention of foreign doctoral graduates; and it casts doubt on their effects in other European countries, which have introduced similar policies. In the near future, we plan to investigate more structural factors such as access to citizenship, career opportunities, and existing international research networks. Methodologically, the innovative use of machine learning techniques on names illustrates both the promise and the limitations of this approach, highlighting the need for further development and integration with alternative sources of information.

References

Amuedo-Dorantes, C. and D. Furtado (2019): "Settling for academia?: H-1B visas and the career choices of international students in the United States," *Journal of Human Resources*, 54, 401–429.

Buenstorf, G., D. P. Heinisch, and M. Kapa (2022): "International doctoral graduates as inventors in the German innovation system," *Industry and Innovation*, 29, 488–510.

Cerna, L. (2013): "Understanding the diversity of EU migration policy in practice: the implementation of the Blue Card initiative," *Policy studies*, 34, 180–200.

Finn, M. G. (2010): "Stay rates of foreign doctorate recipients from US universities, 2007," Tech. rep., Oak Ridge Inst. for Science and Education (ORISE), Oak Ridge, TN (United States).

Ganguli, I. and M. MacGarvie (2025): "International students' migration: policy impacts and implications for innovation," in *A Research Agenda for Migration and Innovation*, ed. by F. Lissoni and A. Morrison, Edward Elgar Publishing, 129–150.

Hunt, J. (2010): "Skilled immigrants' contribution to innovation and entrepreneurship in the United States," in *Open for Business: Migrant Etrepreneurship in OECD Countries*, ed. by OECD, OECD Publishing, 257–272.

Wogart, J. P. and M. Schüller (2011): "The EU's Blue Card: will it attract Asia's highly skilled?" .