# 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
[*]Contact author: andriy.romanyuk@u-bordeaux.fr

August 25, 2025

**Provisional draft, not to quote nor circulate**

**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). 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 significant increase in the stay rates of non-EU graduates in the years immediately following graduation. However, this result is sensitive to the stringency of the criteria used to assign graduates to the treated–non-EU–and control–EU–groups.

**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 either the French graduates or the graduates from selected EU regions as the control ones; 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 5 and 8 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, contextual evidence from reports by the *Confédération des jeunes chercheurs* (CJC),

based on interviews and surveys, suggests that the debate over visa policies for early-career researchers remained contentious until the 2006 reform (CJC, 2010). This indicates that the 1998 policy may not have fully resolved the underlying issues. We plan to extend our analysis to the 2006 reform in future drafts of this same paper.

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 B.2 in Appendix B.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*.

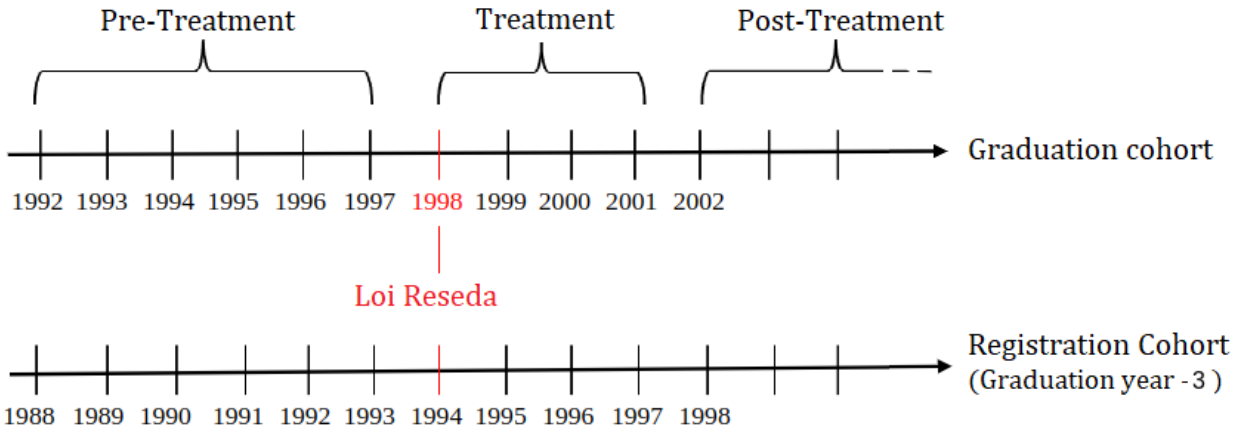



**Figure 1:** Experiment Timeline

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.

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

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, $\varepsilon$ 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.

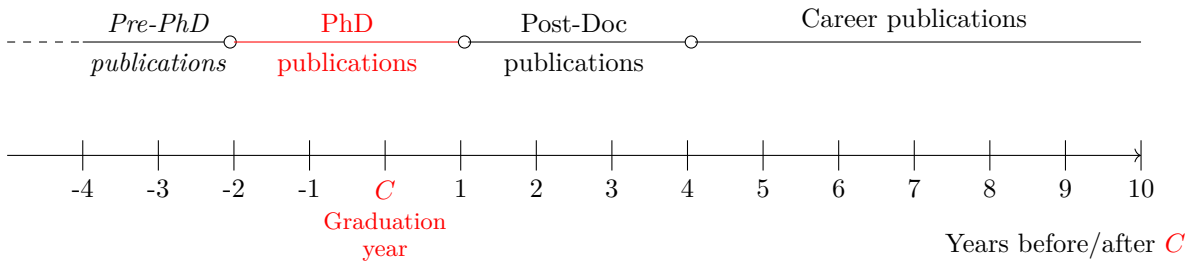



**Figure 2:** PhD students' publication timeline and classification

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 A). 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).

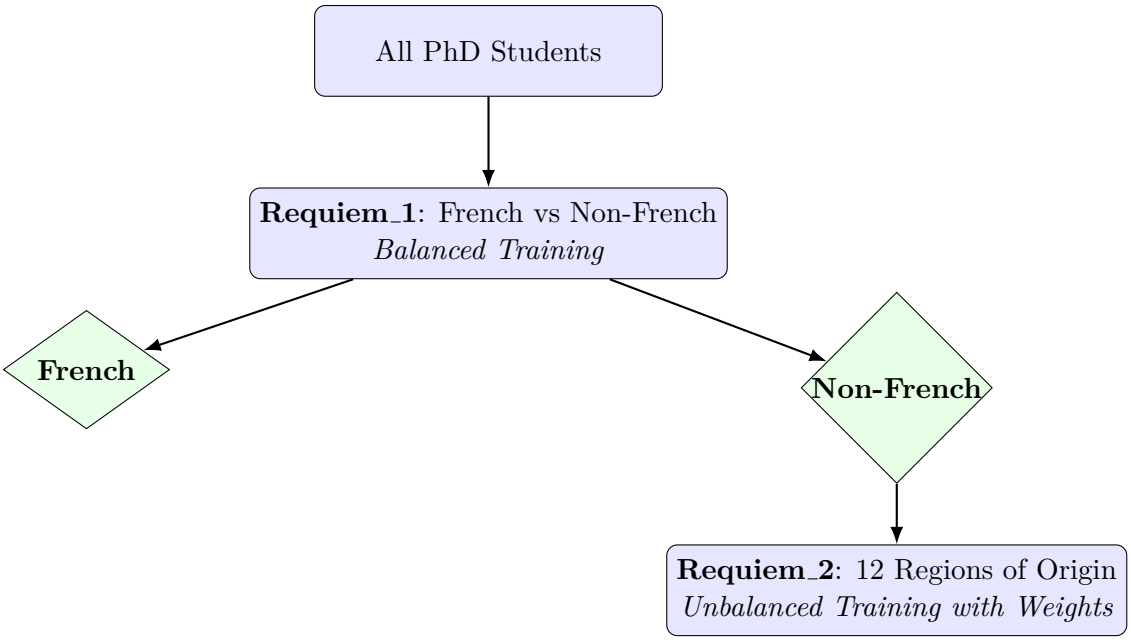
[14]For a summary introduction to LSTM models, see dedicated chapters in Subasi (2020) and Theodoridis (2020). For a more detailed one, see Hochreiter (1997).

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 B.1 in Appendix B.) 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 B.2 and B.3 in Appendix B.2). Robustness checks using more conservative thresholds are presented in Appendix F.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 B.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.

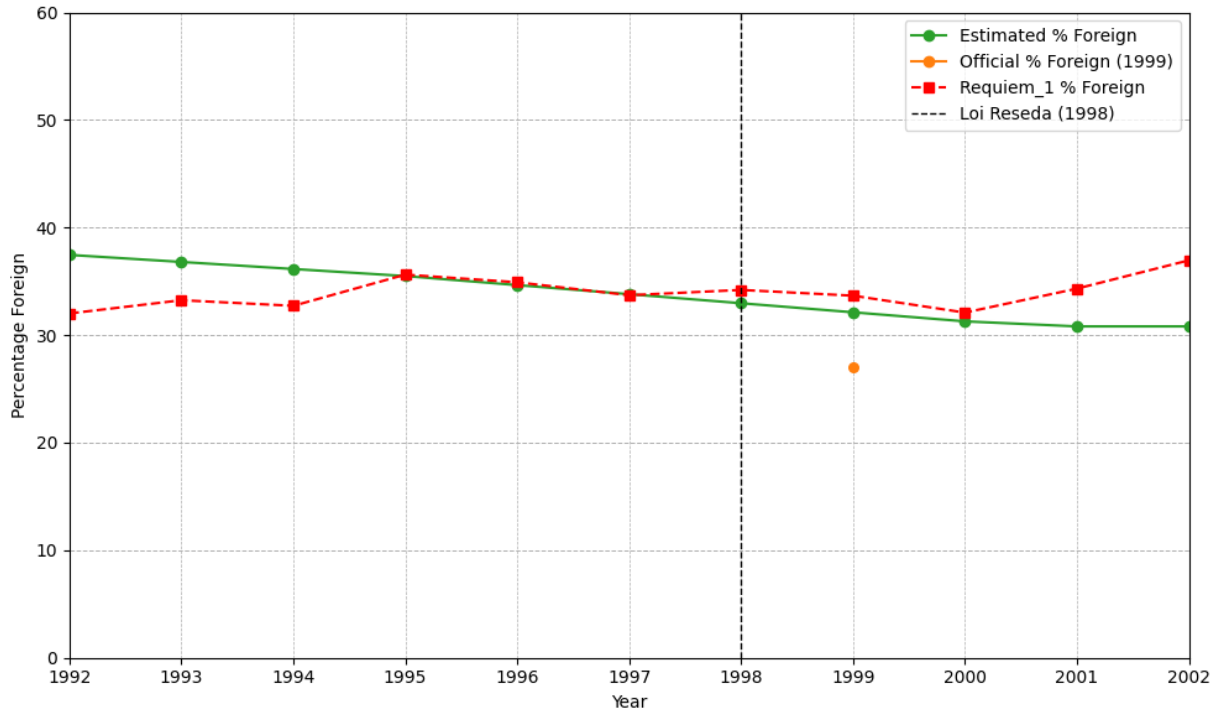



**Figure 4:** Percentage of foreign doctoral graduates in France from 1992 to 2002

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

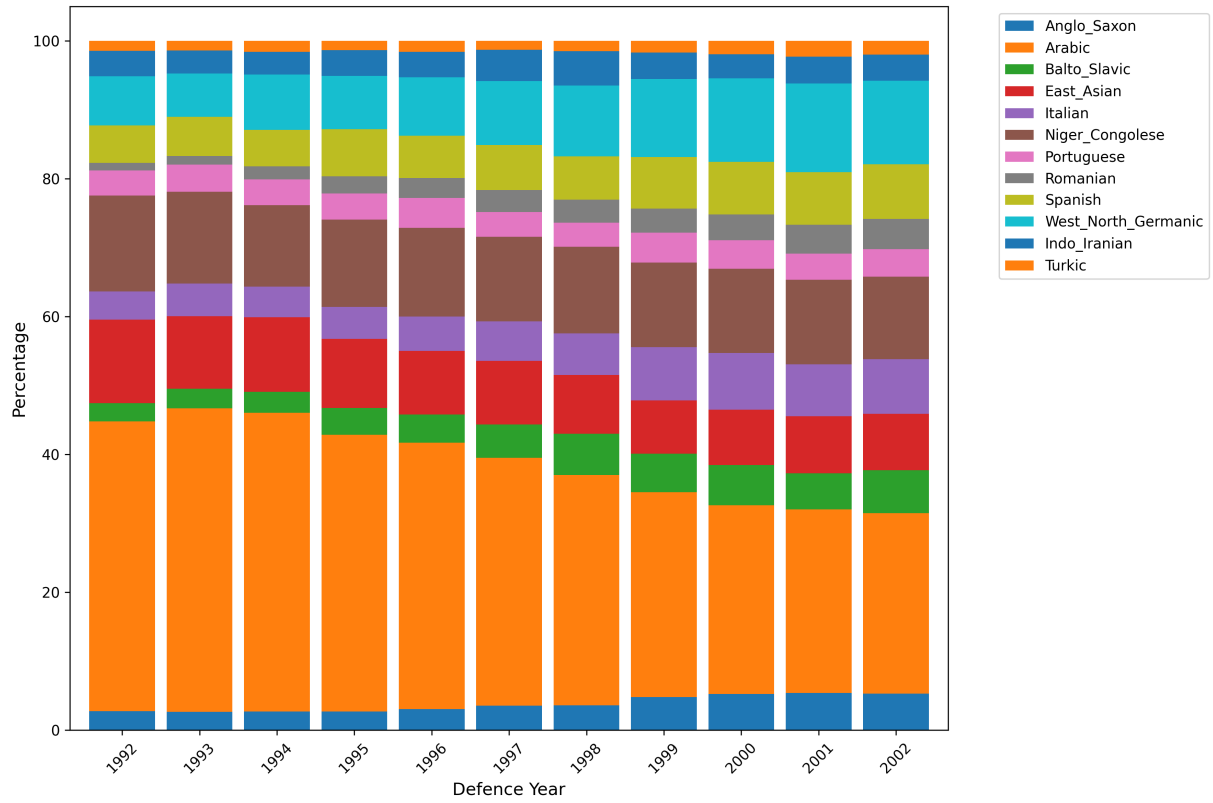



**Figure 5:** Proportion of regions of origin within foreign Doctoral Graduates in France, 1992-2002

**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 B.2 in Appendix B.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 B.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

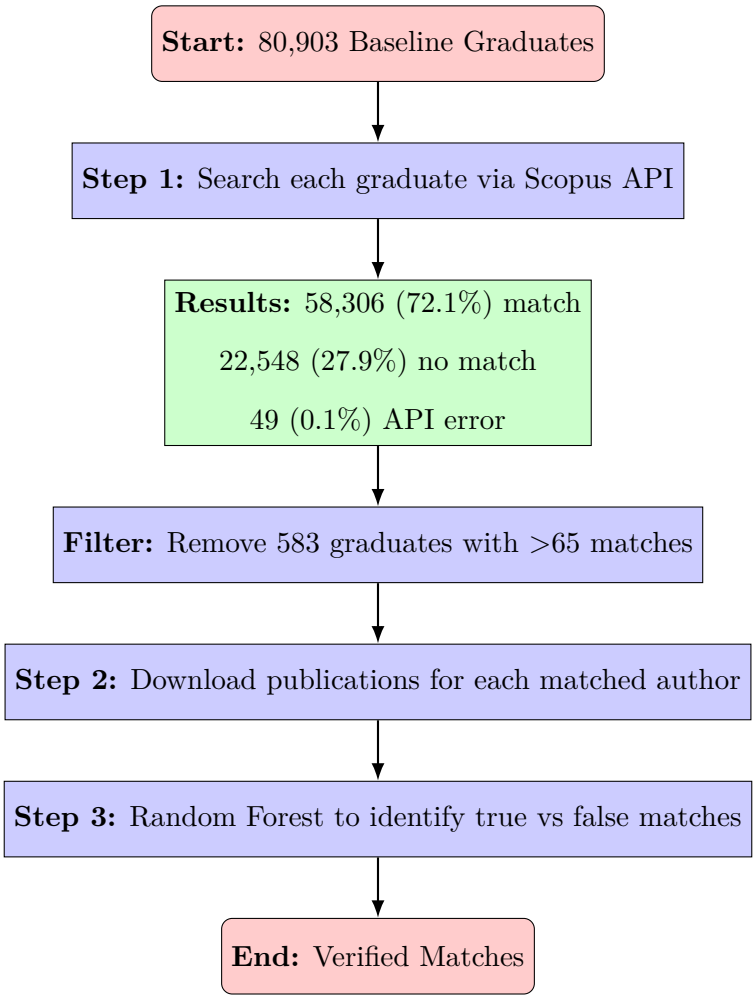
Compared to the attribution of a country or region of origin, the classification of the doctoral graduates by gender is a rather trivial task, due to the widespread availability of gender-name dictionaries covering multiple countries, as well as of well-tested methodologies. We follow Corsini et al. (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 B.3.1 reports a confusion matrix (Provost and Kohavi, 1998), detailing misclassifications between the *French* class and the various regions of origin (Table B.5), as well as among the regions themselves (Figures B.4 and B.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 link doctoral graduates from our *theses.fr* sample to author profiles in *Scopus*. Using the Scopus API, we retrieve potential matches for all 80,903 graduates and validate them using a Random Forest algorithm, as detailed in Appendix A. During this process, 632 graduates are excluded: 49 Chinese graduates with over 5,000 potential matches, and 583 individuals with more than 65 matches, removed to reduce computational burden.

In total, 59% of doctoral graduates are successfully matched with at least one Scopus author. This match rate is lower than the 68% reported by Corsini et al. (2022) for a similar *theses.fr–Scopus* linkage. The difference is primarily due to a time trend: the share of matched graduates increases from 40% in 1992 to 78% in 2002, reflecting growing pressure on recent cohorts to publish and to do so in English.

In our main analysis, we retain only the top-scoring match per graduate, as determined by the Random Forest model. Alternative matching strategies are discussed in the robustness checks in Appendix F.2.

### 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 6 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.

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, 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

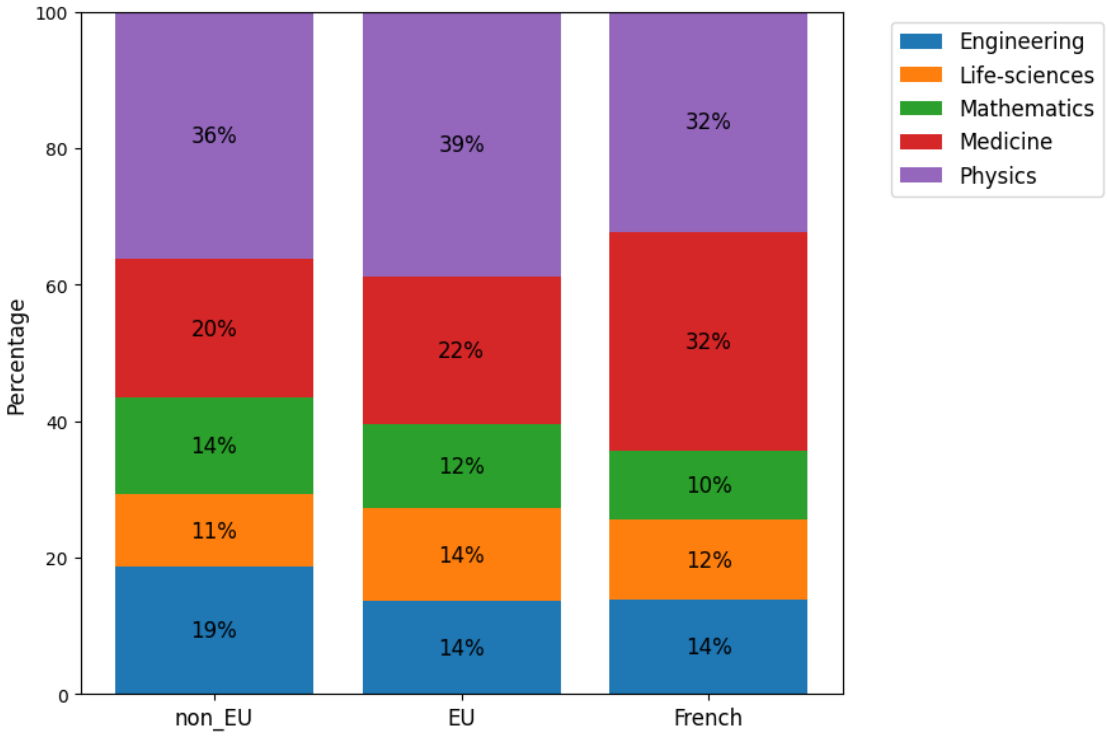



**Figure 6:** Doctoral graduates from Treated and Control groups by field, 1992 - 2002

**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*).

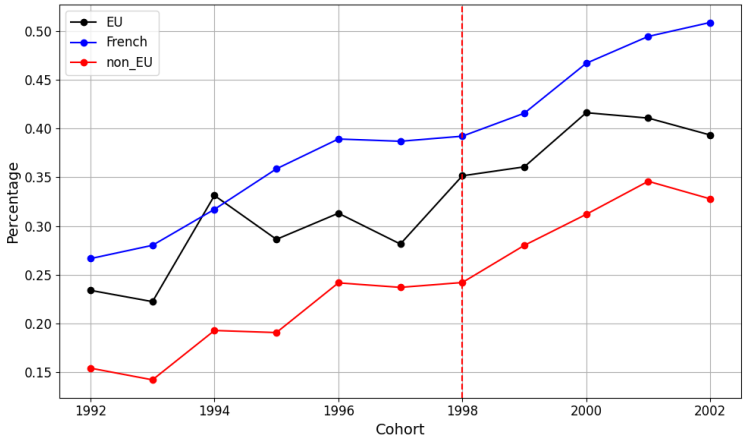
published prior to the likely start of their PhD (*Published_before_PhD*) and are more productive during their doctoral studies (*Nr_Publications_during_PhD*). However, the gap between EU and French graduates is smaller than that between the latter and the non-EU ones.

Figure 7 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

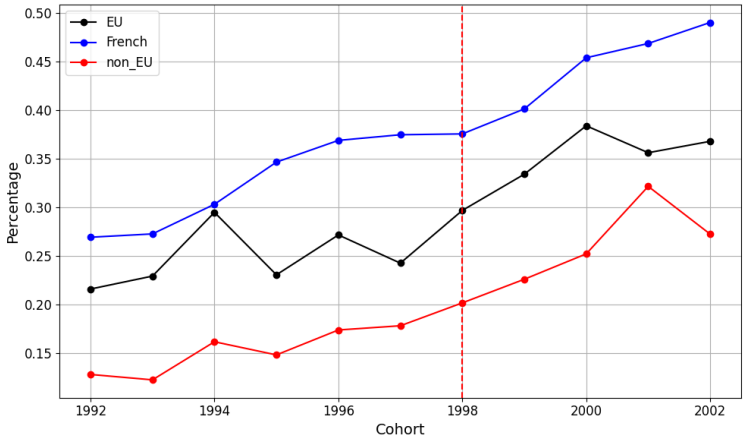
EU and France.

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 7a). 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 7b). 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 7c).

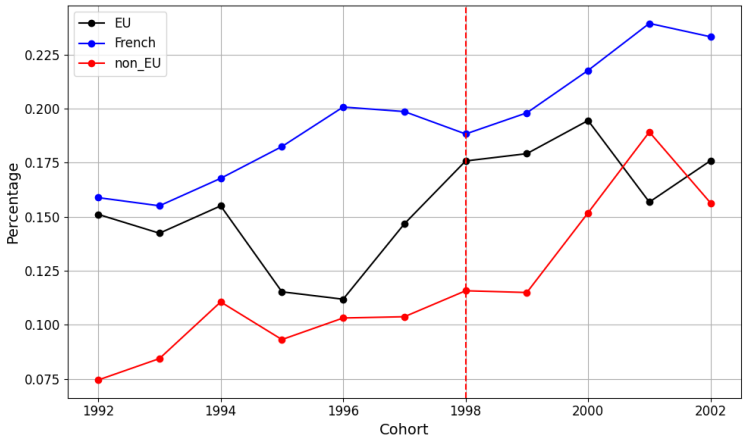
For all indicators, pre-treament trends for non-EU and French graduates are rather parallel. Instead, the lines for EU ones are more irregular, especially for *Supervisor*. This is possibly due to the relatively small sample for this group.




**(a)** Post Doc Stay




**(b)** Career Stay



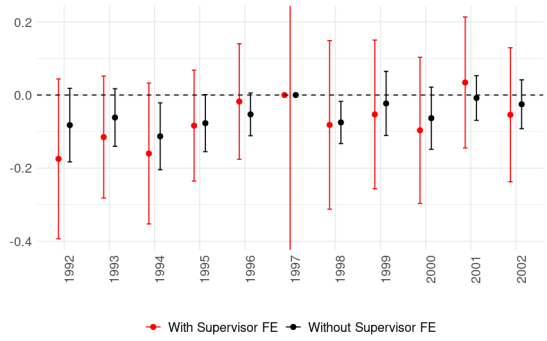

**(c)** Supervisor

**Figure 7:** Proportion of Stayers, by EU, non-EU and French status, and graduation cohort (1992–2002)

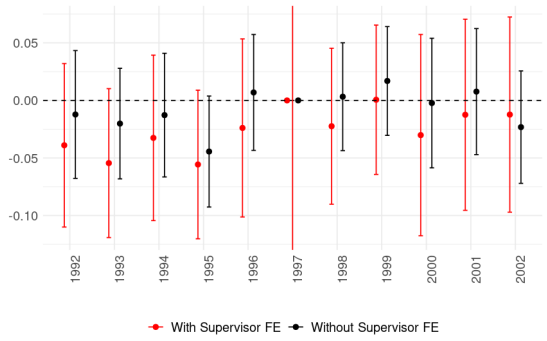
Post Doc Stay: share of students in cohort $c$ with $Post\_Doc\_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort $c$ with $Career\_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort $c$ with $Supervisor_{i,c+t} = 1$

# 5 Results

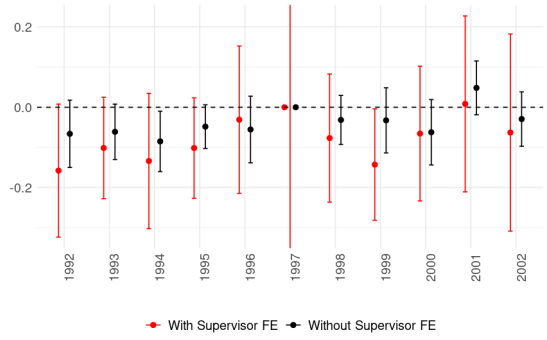
We estimate Equation 1 by means of a linear probability model, with either *Post_Doc_Stay*, *Career_Stay* or *Supervisor* as the binary outcome variables. We run two different regressions with bo alternative control samples, one containing the EU graduates and the French ones.



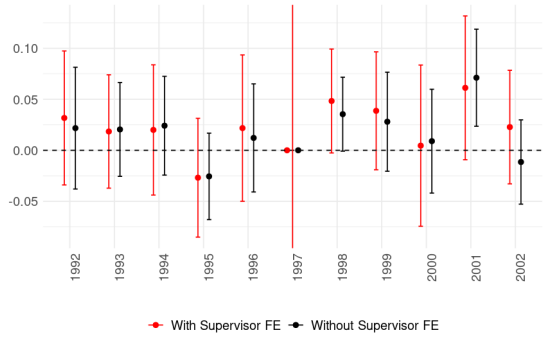

**(a)** *Post_Doc_Stay* - non-EU vs EU



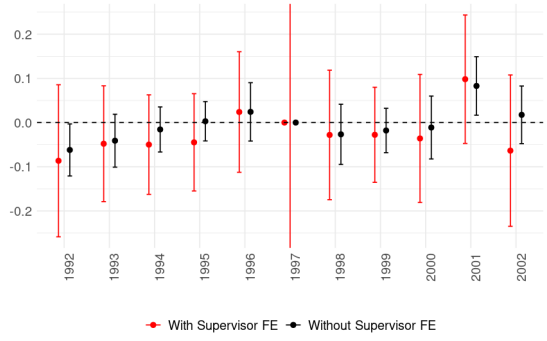

**(b)** *Post_Doc_Stay* - non-EU vs French



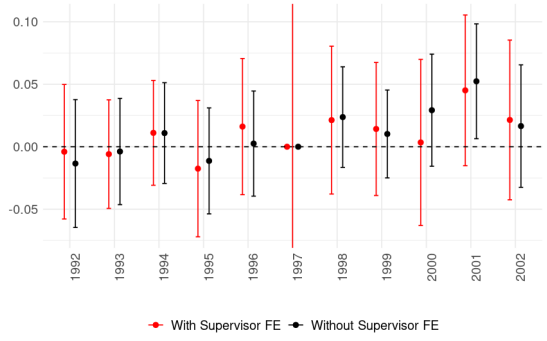

**(c)** *Career_Stay* - non-EU vs EU




**(d)** *Career_Stay* - non-EU vs French




**(e)** *Supervisor* - non-EU vs EU




**(f)** *Supervisor* - non-EU vs French

**Figure 9:** Estimated $\delta$ Coefficients (95% C.I.) by Cohorts: All outcome variables and treated vs control groups, 1992 – 2002

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

Our primary focus is on the value of the various $\delta$ coefficients in Equation 1, one for each interaction term between cohorts $c$ and the treatment groups ($Non\_EU_i \times Cohort_c$). Figure 9 plots the estimated values of such $\delta$ coefficients in six different regressions, which differ for the choice of the dependent variable (*Post_Doc_Stay*, *Career_Stay* and *Supervisor*) and the sample used (non-EU:EU versus non-EU:France samples). Tables 5and 6 report all the other estimated coefficients, respectively for the regressions with EU and French graduates as control groups.[18]

As displayed in Table 5, 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. If we try to add *Spanish* and *Portuguese*, 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 D).

We also notice that the parallel trend assumption underlying this causal interpretation of our results holds better in the non-EU vs French regression, with pre-treatment coefficients always very close to zero. At the same time, this sample is much bigger than the non-EU vs EU sample, which may explain the lower significance of the estimated coefficients in the treatment period.

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, results for the non-EU versus EU comparison are not interpretable because the pre-policy coefficients differ significantly from zero, violating the parallel trends assumption. No effect is observed when French graduates are used as the control group.

Some positive effects are observed when using *Career_Stay* as the dependent variable and French graduates as the control group for the 2001 cohort, 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 8 percentage points increase for the non-EU versus EU comparison, and a 5 percentage points increase for the non-EU versus France comparison 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

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

excluding graduates with multiple matches entirely (see Appendix F.2).

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 F.1). In the non-EU versus EU comparison, the interaction term for the 2001 cohort in the *Supervisors* outcome loses statistical significance at the 5 percentage points level but remains significant at the 10% level (see Figure F.3). In the non-EU versus France comparison, the *Career_Stay* effect for the 2001 cohort becomes statistically insignificant, while the *Supervisors* effect similarly loses significance at the 5% level but remains significant at the 10% level (see Figure F.5). 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.

Results for *Arabic* are displayed in Appendix E. 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 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 x Cohort | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* | *Yes* |
| 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.

**Table 6:** Dynamic Difference-in-Differences: non-EU vs France, LPM, With and Without Supervisor FE, 1992-2002

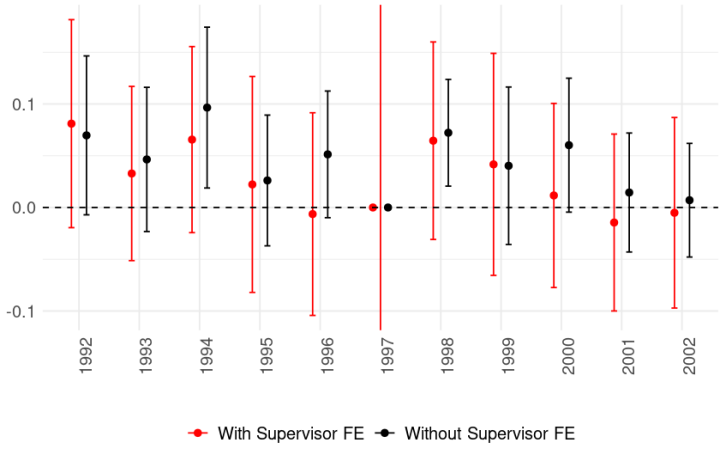
| | Post Doc Stay | | Career Stay | | Supervisor | |
|---|---|---|---|---|---|---|
| non_EU | −0.109*** | −0.078** | −0.165*** | −0.157*** | −0.076*** | −0.068*** |
| | (0.019) | (0.026) | (0.018) | (0.023) | (0.016) | (0.019) |
| Male | 0.028*** | 0.039*** | 0.066*** | 0.068*** | 0.060*** | 0.055*** |
| | (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) |
| 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.

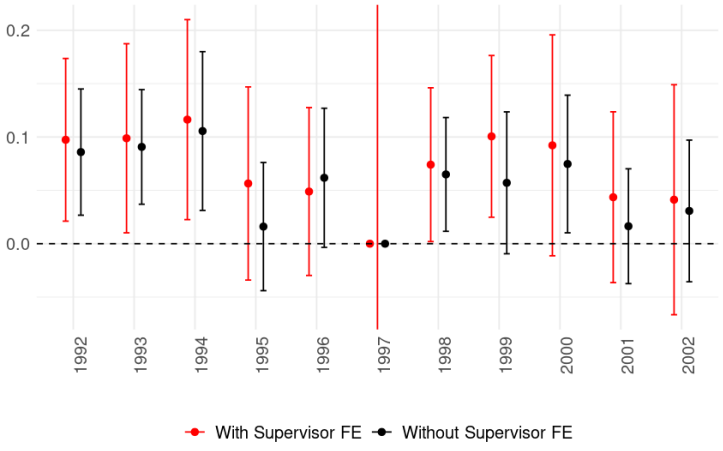
Tables C.1 and 6 reports all the estimated coefficients besides $\delta$, from the same regressions underlying Figure 9. The coefficient for non-EU in Table C.2 indicates that, on average, PhD students from non-EU are less likely to remain in postdoctoral or career positions compared to their French 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.

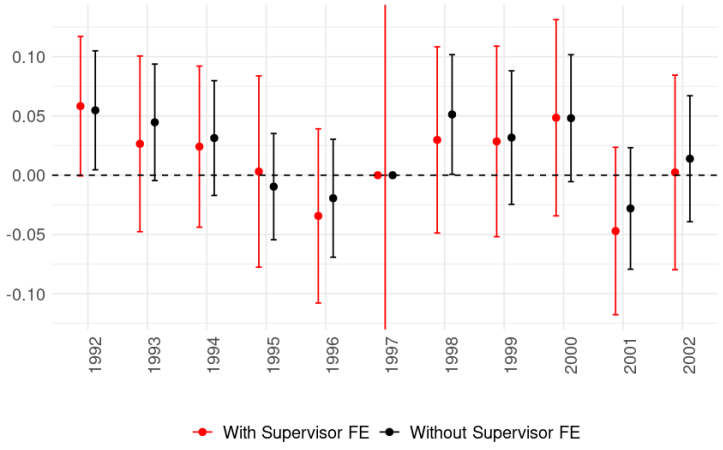
In Figure 11 we report the results of a simple placebo test, performed on sample where we include only the French and EU graduates, with the latter taking the place of the non-EU. We run regressions with *Post_Doc_Stay*, *Career_Stay* and *Supervisor* as dependent variables. We obtain some positive coefficients 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 11:** 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 country 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*, pre-policy differences violate the parallel trends assumption, preventing interpretation for the non-EU versus EU comparison, and no effect is found when using French graduates as the control group.

Some evidence of positive effects emerges for the 2001 cohort, particularly in *Career_Stay* (non-EU versus France) and *Supervisors* (non-EU versus EU and non-EU versus France), 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.

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 Graduates Scopus Matching

### A.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 A.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 A.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 A.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

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

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

associated Author ID in the `authors_id` field, due to Scopus's limit of listing only the first 100 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.

## A.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 our application, the algorithm predicts whether a match is correct using features derived from publication records, including supervisor co-authorship, institutional or national affiliation, as well as similarity in title, discipline, and author names.

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[19]. As shown in Figure A.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 ROBUSTNESS CHECKS).

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

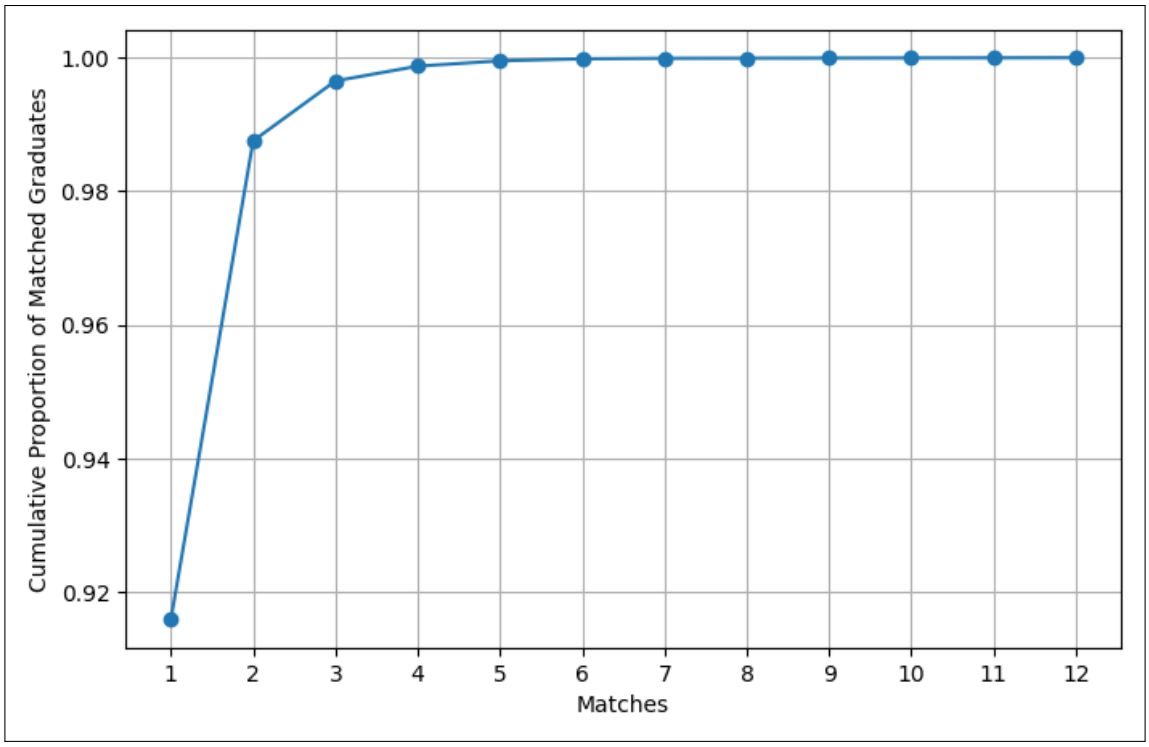



**Figure A.1:** Cumulative Distribution Function of the Number of Matches of the Matched Graduates

# B Region of origin assignment

## B.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 B.1. For that cohort, 33% of PhD students were foreign, a share that increased steadily to 41% by 2010 (MESR, 2020)[20]. 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*[21], 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

[20] 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)

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

**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 |

(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[22], 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

[22]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).

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.

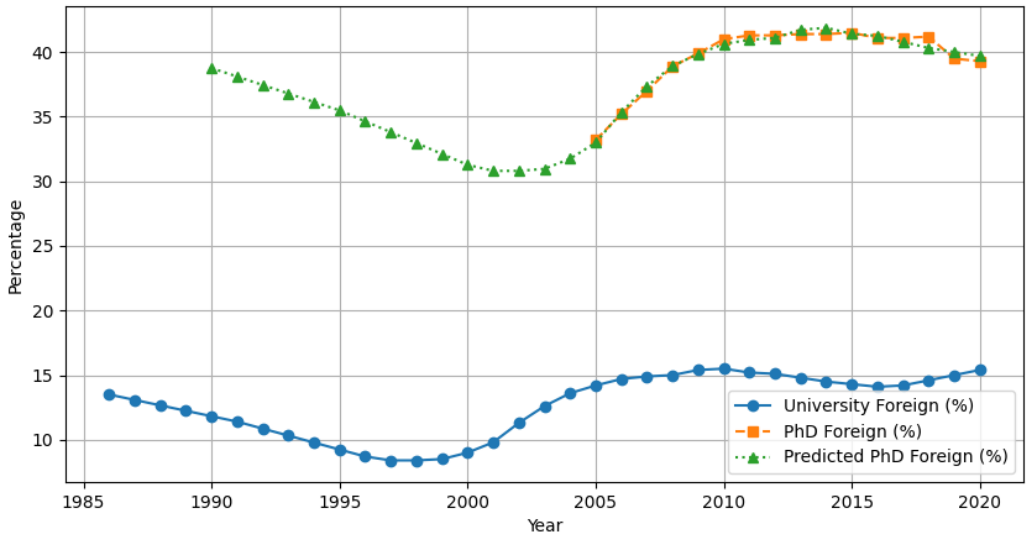



**Figure B.1:** Proportion foreign PhD students in France 1990-2020

*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 B.2.

**Table B.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.

## B.2 Threshold Analysis

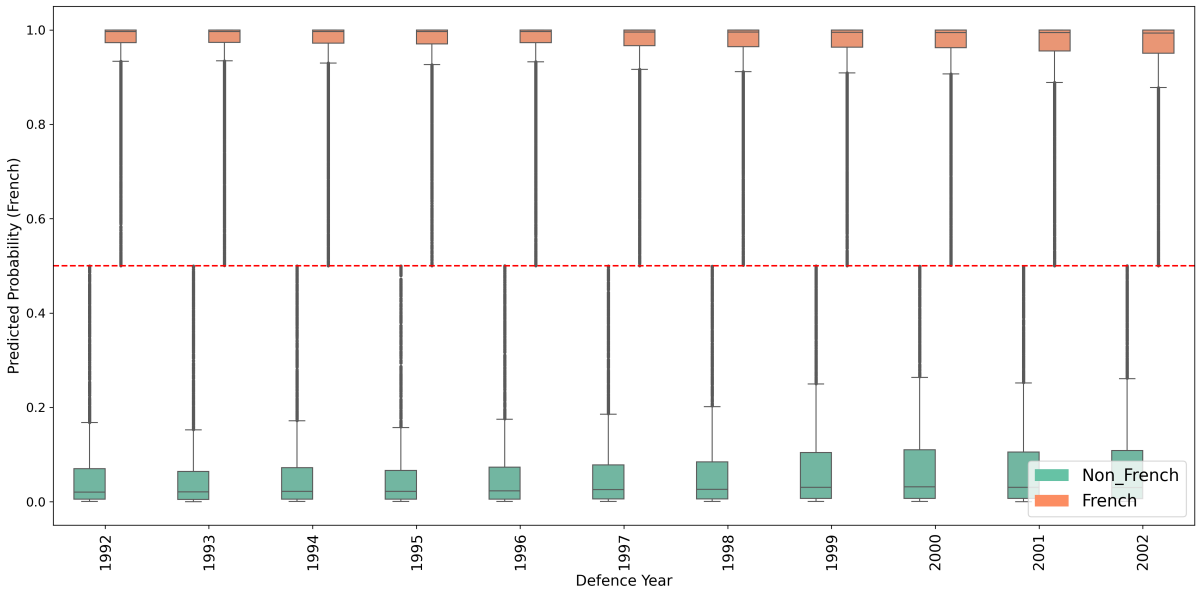



**Figure B.2:** Distribution of the Highest Predicted Class Probability Across French and non-French, 1992-2002

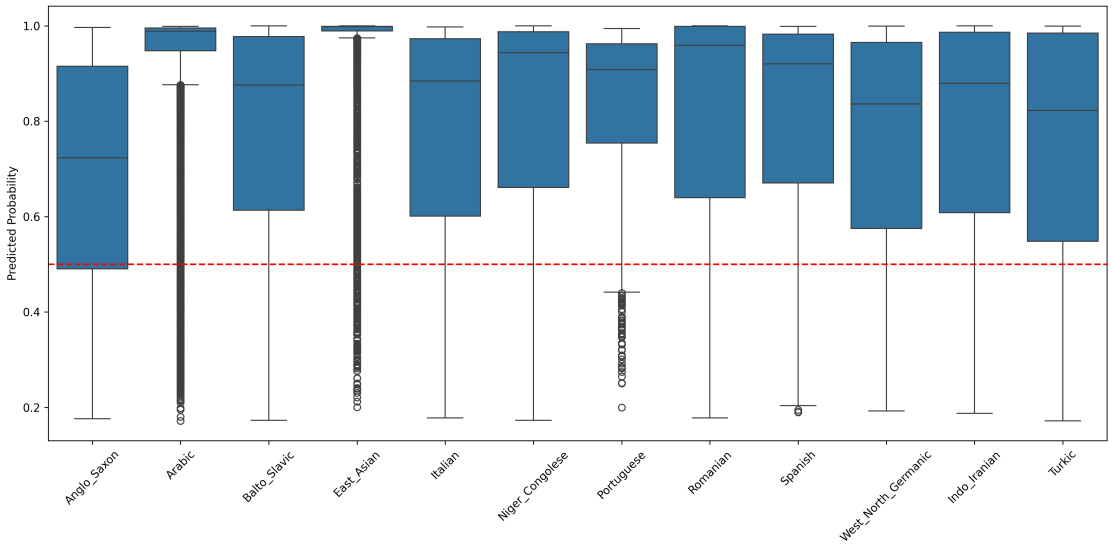



**Figure B.3:** Distribution of the Highest Predicted Class Probability Across Regions of Origin, 1992-2002

## B.3 Model's performance

### B.3.1 In-Sample Performance

The classification report for **Requiem_1** in Table B.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.[23]

[23] 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 B.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 B.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 B.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 B.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**.[24] 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,

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

the *West_North_Germanic* class (8.45%) again likely includes individuals from Belgium.

**Table B.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 B.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 B.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 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 B.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.

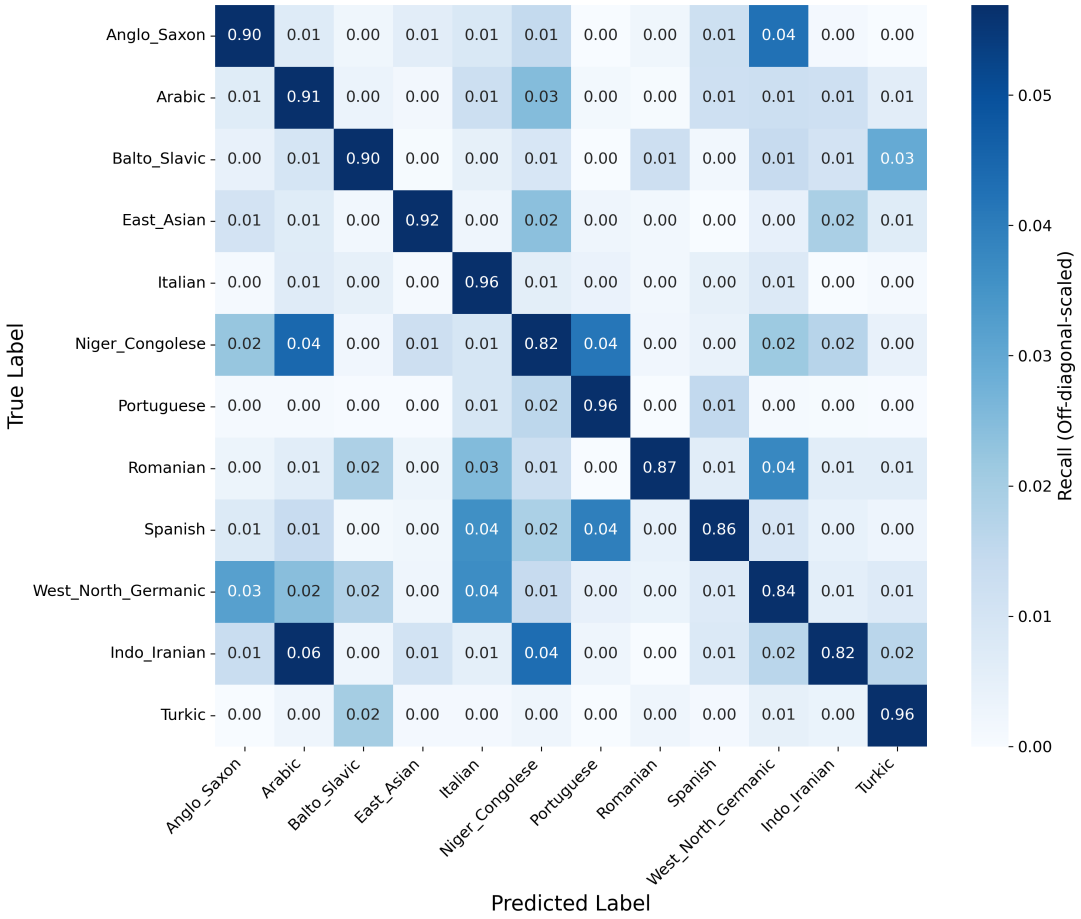



**Figure B.4: Requiem_2** - Confusion Matrix - Recall

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

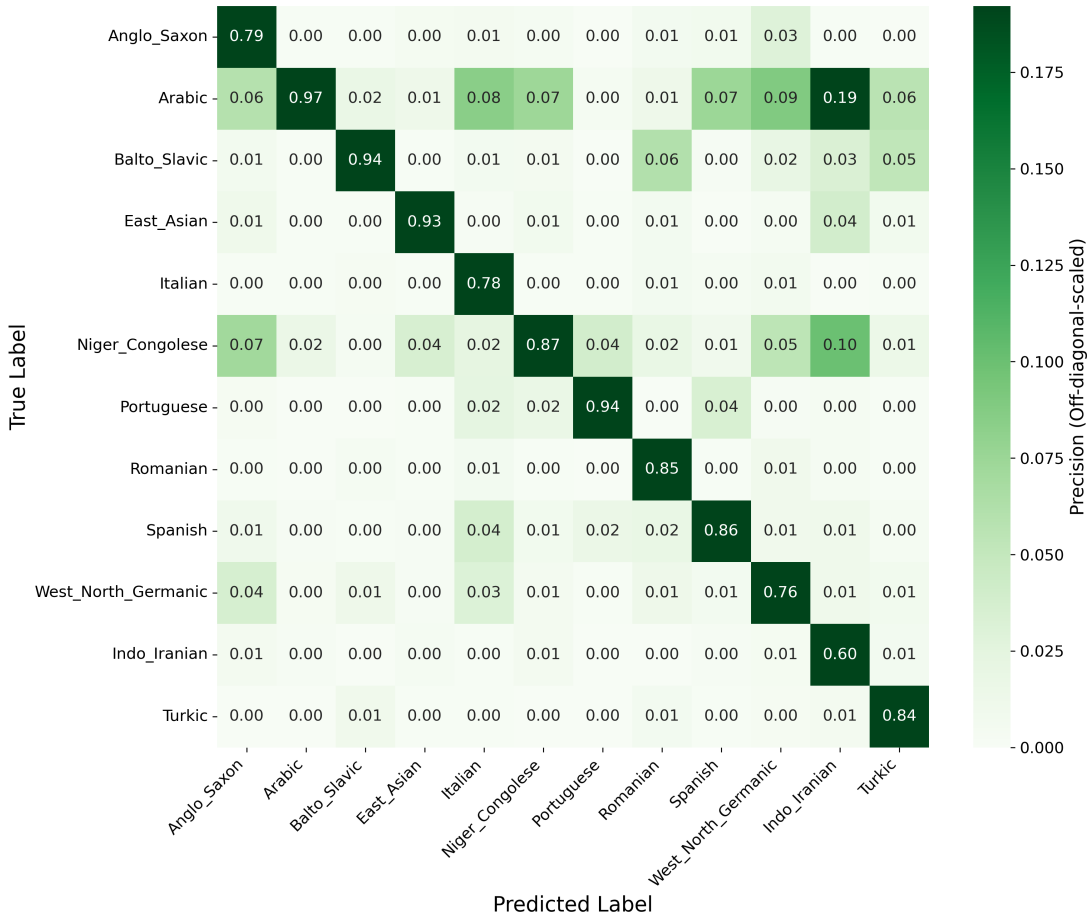



**Figure B.5: Requiem_2** - Confusion Matrix - Precision

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

### B.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 B.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 B.7 summarizes the distribution of these observations by the first available type of location information.

**Table B.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 B.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 B.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 B.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 |

# C Full Regressions Tables

**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 |

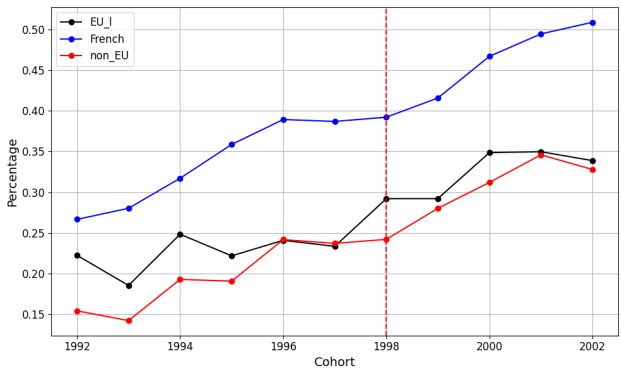
*p<0.1; **p<0.05; ***p<0.01. Standard errors in parentheses.

**Table C.2:** Dynamic Difference-in-Differences: non-EU vs France, LPM, With and Without Supervisor FE, 1992-2002

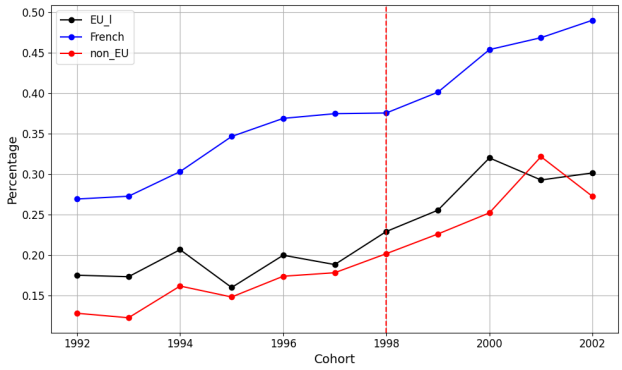
| | Post Doc Stay | | Career Stay | | Supervisor | |
|---|---|---|---|---|---|---|
| non_EU | −0.109*** | −0.078** | −0.165*** | −0.157*** | −0.076*** | −0.068*** |
| | (0.019) | (0.026) | (0.018) | (0.023) | (0.016) | (0.019) |
| Male | 0.028*** | 0.039*** | 0.066*** | 0.068*** | 0.060*** | 0.055*** |
| | (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.

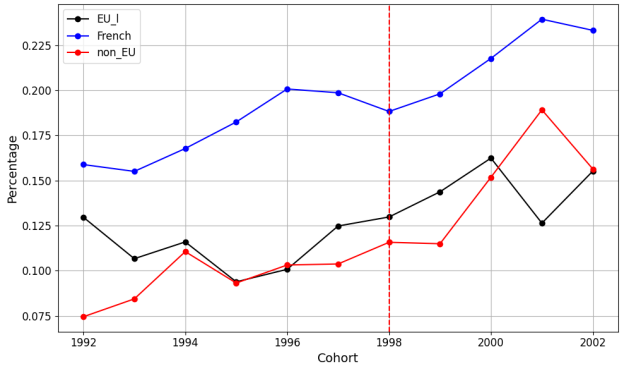
# D EU "Large" Graphs with Spanish and Portuguese




**(a)** Post Doc Stay




**(b)** Career Stay

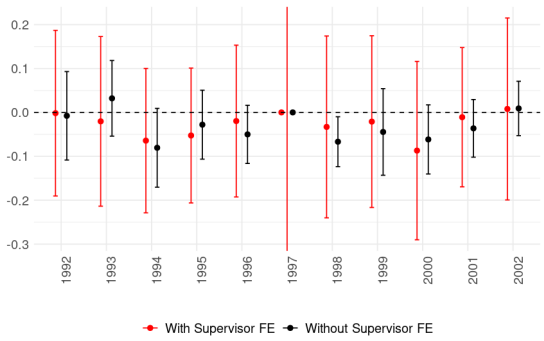



**(c)** Supervisor

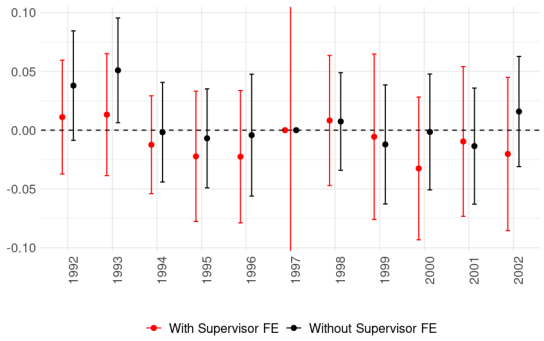
**Figure D.1:** Proportion of Stayers, by EU, non-EU and French status, and graduation cohort (1992–2002)

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.

# E Arabic Results

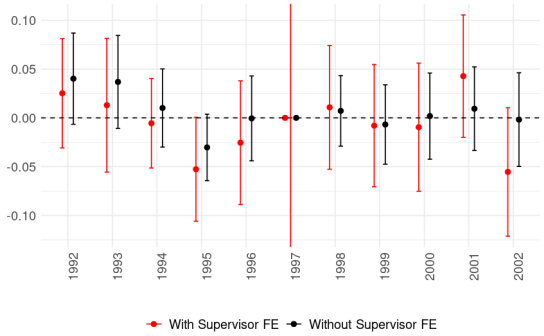



**(a)** Post Doc Stay - Arabic vs EU

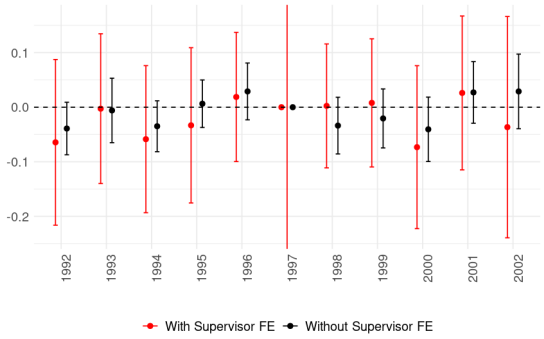



**(b)** Post Doc Stay - Arabic vs French

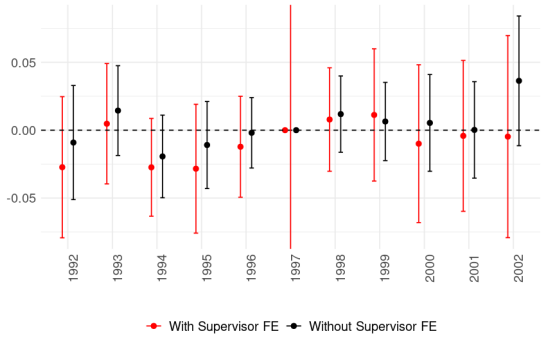

**(c)** Career Stay - Arabic vs EU




**(d)** Career Stay - Arabic vs French




**(e)** Supervisor - Arabic vs EU




**(f)** Supervisor - Arabic vs French

**Figure E.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 $Supervisor_{i,c+t} = 1$

# F Robustness Check

## F.1 More Stringent Thresholds

In our main analysis, given the distribution of predicted probabilities shown in the box plots in Figures B.2 and B.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 F.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 F.1, when we evaluate model performance on a validation sample with known location information (see Appendix B), 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 F.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. Figures F.3 and F.5 present the regression results from the main analysis under these stricter thresholds.

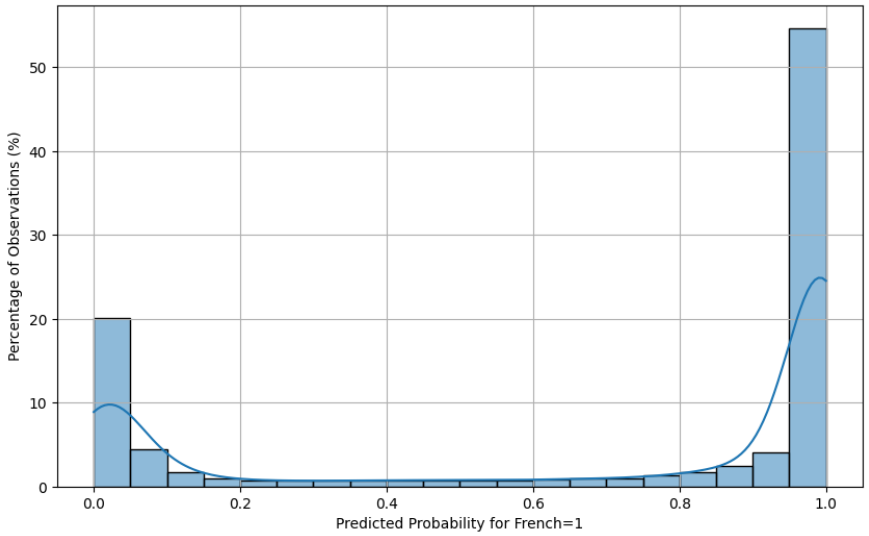



**Figure F.1:** Threshold Distribution **Requiem_1** Predictions of French equal 1, 1992-2002

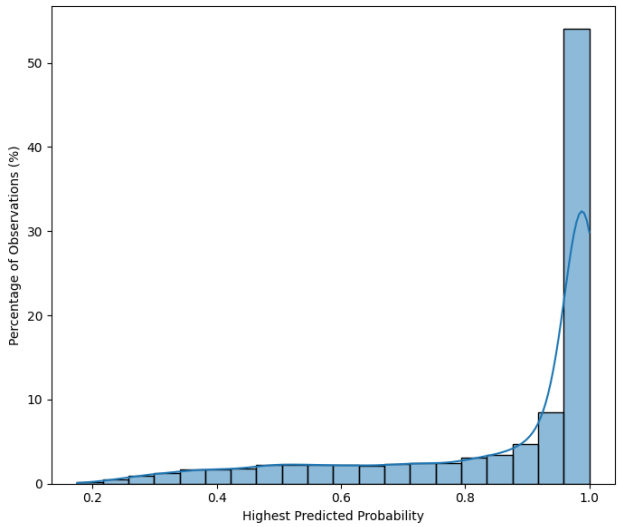
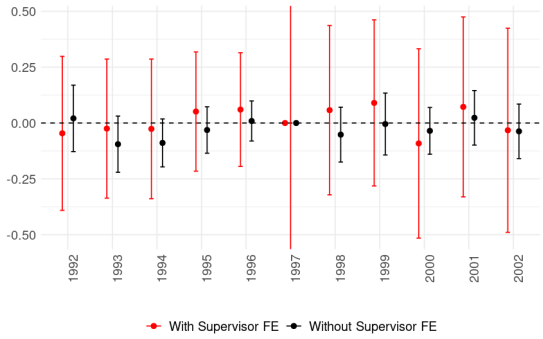


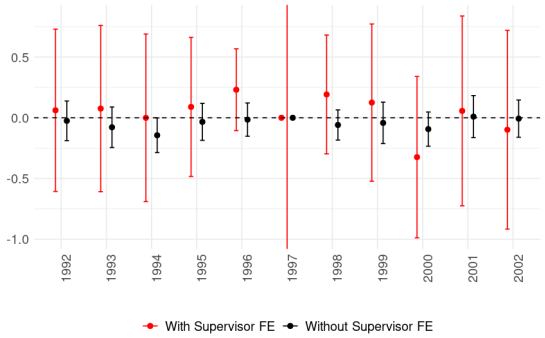

**Figure F.2:** Threshold Distribution **Requiem_2** Predictions of the Highest Region of Origin Value, 1992-2002

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

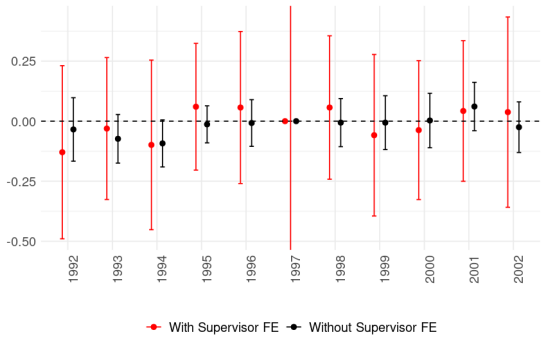
| Requiem_1 | Requiem_2 | Class | Precision | Recall | F1-score | Support | Accuracy |
|---|---|---|---|---|---|---|---|
| 0.50 | 0.50 | non_French | 0.61 | 0.89 | 0.73 | 149 | 0.85 |
| | | French | 0.96 | 0.84 | 0.90 | 513 | |
| 0.05/0.95 | 0.50 | non_French | 0.72 | 0.88 | 0.79 | 111 | 0.89 |
| | | 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 | |



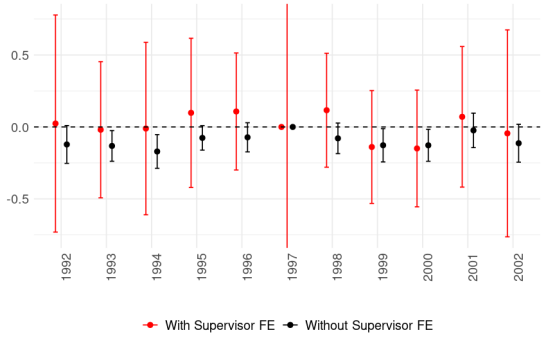

**(a)** Post Doc Stay - **Requiem_1** 05-95 Threshold



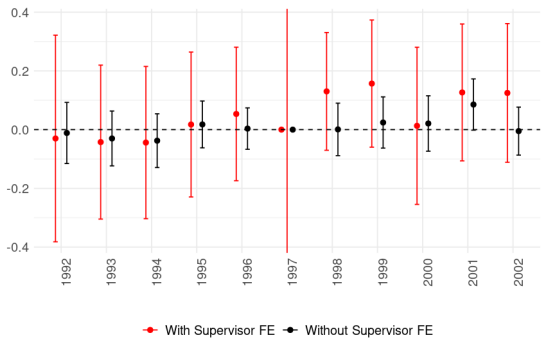

**(b)** Post Doc Stay - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold



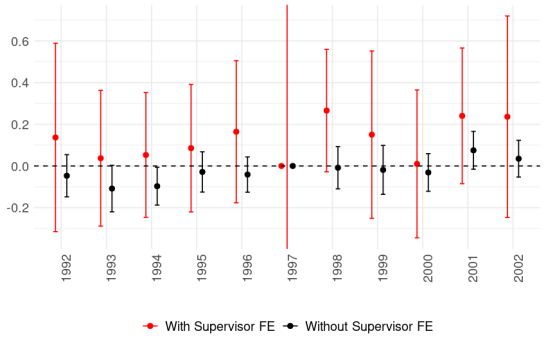

**(c)** Career Stay - **Requiem_1** 05-95 Threshold




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



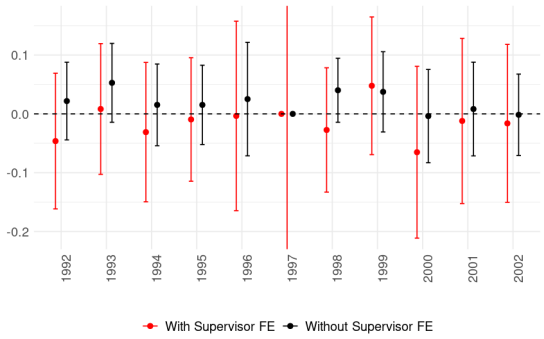

**(e)** Supervisor - **Requiem_1** 05-95 Threshold



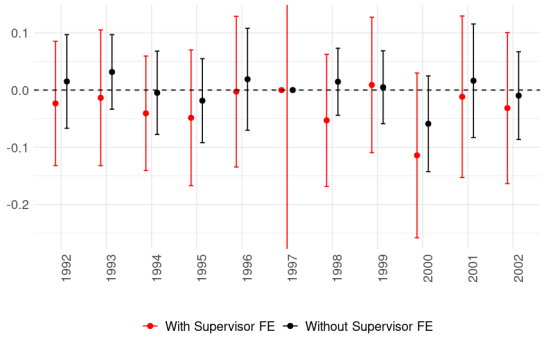

**(f)** Supervisor - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold

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

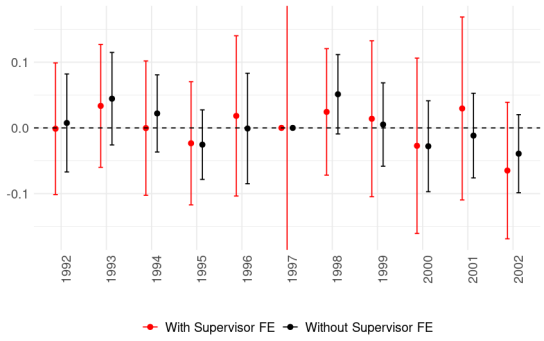
Post Doc Stay: share of students in cohort $c$ with $Post\_Doc\_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort $c$ with $Career\_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort $c$ with $Supervisor_{i,c+t} = 1$



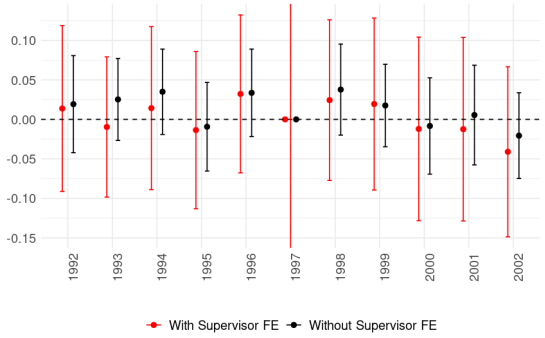

**(a)** Post Doc Stay - **Requiem_1** 05-95 Threshold



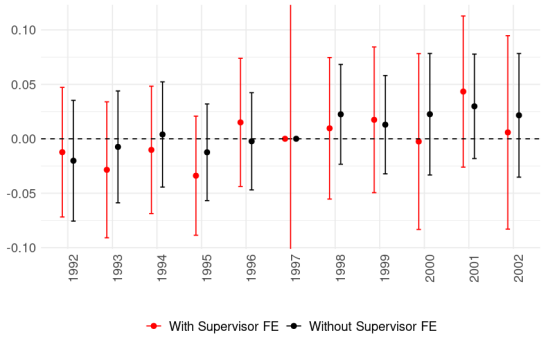

**(b)** Post Doc Stay - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold



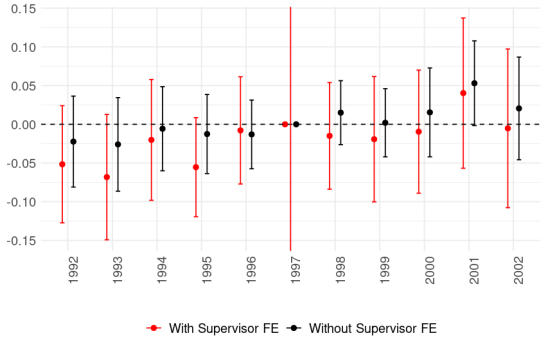

**(c)** Career Stay - **Requiem_1** 05-95 Threshold




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




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



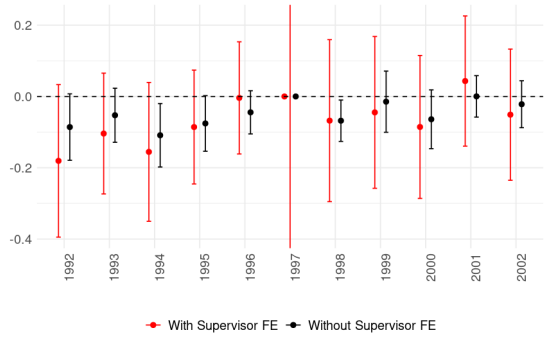

**(f)** Supervisor - **Requiem_1** 05-95 Threshold & **Requiem_2** 95 Threshold

**Figure F.5:** Estimated Coefficients (95% Confidence Interval) by Cohorts: non-EU vs French, 1992 – 2002

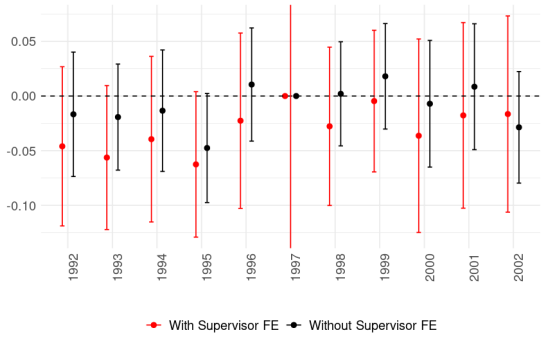
Post Doc Stay: share of students in cohort $c$ with $Post\_Doc\_Stay_{i,c+t} = 1$; Career Stay: share of students in cohort $c$ with $Career\_Stay_{i,c+t} = 1$; Supervisor: share of students in cohort $c$ with $Supervisor_{i,c+t} = 1$

## F.2 Multiple Matches Alternative Approaches

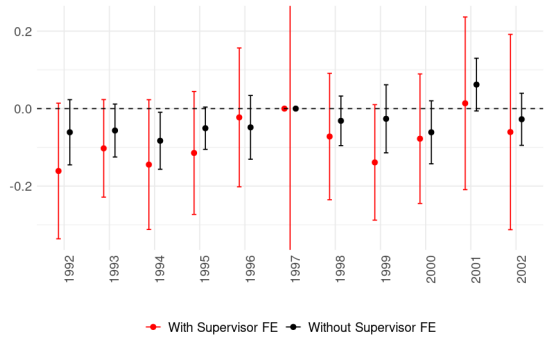
### F.2.1 Retention of the Top Three Matches per Graduate



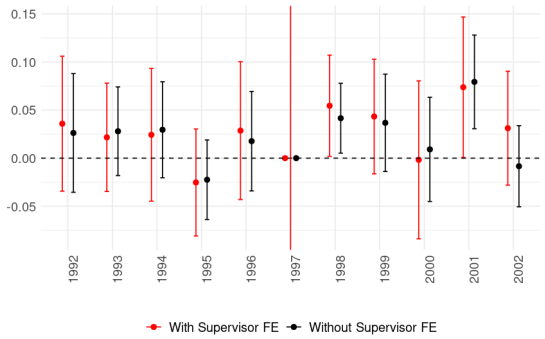

**(a)** Post Doc Stay 3 - non-EU vs EU




**(b)** Post Doc Stay 3 - non-EU vs French




**(c)** Career Stay 3 - non-EU vs EU

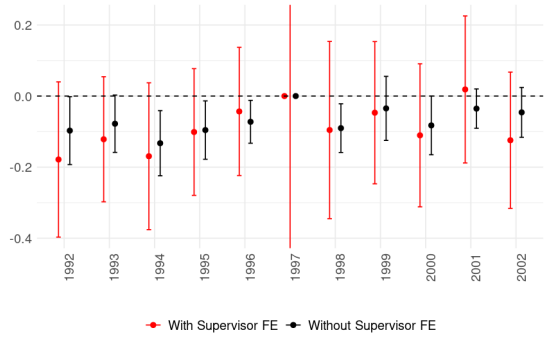



**(d)** Career Stay 3 - non-EU vs French

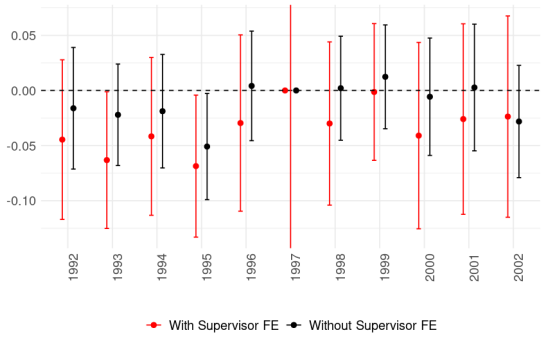
**Figure F.7:** Estimated Coefficients (95% Confidence Interval) by Cohorts: Treated vs Control, 1992 – 2002

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 $Supervisor_{i,c+t} = 1$. If a graduate has more than one match the top three matches are retained.

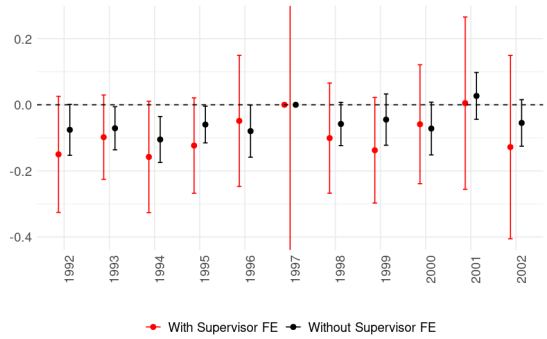
### F.2.2 Excluding Graduates with Multiple Scopus Author Matches (Post Random Forest)



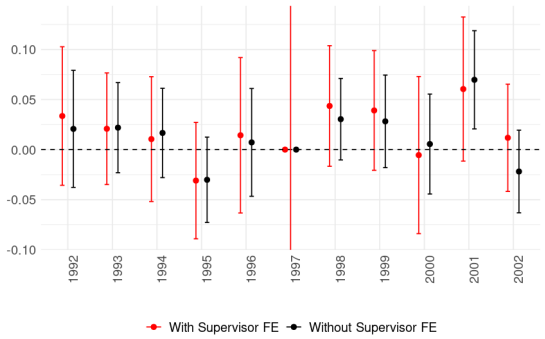

**(a)** Post Doc Stay - non-EU vs EU




**(b)** Post Doc Stay - non-EU vs French




**(c)** Career Stay - non-EU vs EU




**(d)** Career Stay - non-EU vs French

**Figure F.9:** 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.