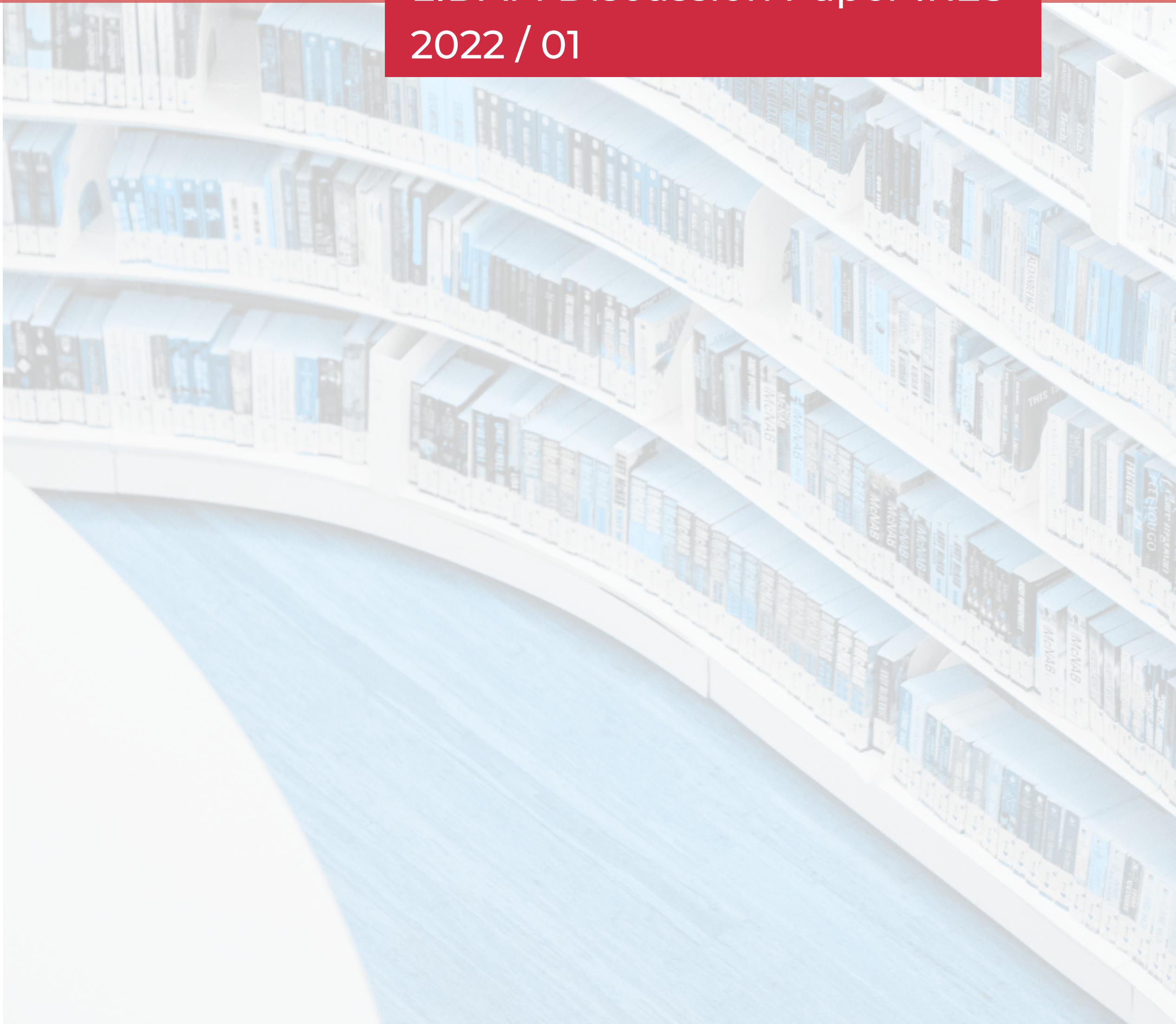
MARKET FORCES IN ITALIAN ACADEMIA TODAY (AND YESTERDAY)

Chiara Zanardello

LIDAM Discussion Paper IRES
2022 / 01



UCLouvain

LIDAM
IRES | Louvain Institute of Data Analysis and Modeling in economics and statistics

Market forces in Italian Academia today (and yesterday)[*]

Chiara Zanardello[†]

February 2022

Abstract

I investigate the operation of the academic market in Italy, mapping current scholars' location choices. I build a new dataset of current professors, associating each scholar with a composite indicator of their quality. The analysis includes the quality of the university and the features of the city where the institution is located. I estimate the strength of different factors: gravity (distance), agglomeration (scholars are attracted to higher quality universities), selection (better scholars travel longer distances), and sorting (the better the scholar, the more the quality of universities is weighted). I find that all of these factors have an effect, and do not vary according to scholars' gender. I find a greater expected utility for scholars in choosing private universities over public ones, through a consistent nesting procedure. Comparing these forces to historical trends in Italian academia, the sorting effect delineates a new momentum for the current academic market in Italy.

Keywords: Human capital, Academic market, Universities, Scholars, Sorting, Italy.

JEL Classification: I23, O15, N34, N33.

[*] I thank David de la Croix, Massimiliano Bratti, Daniele Checchi, and Fabrizio Iacone, who have contributed since the early draft of the project. I also thank all the scholars who replied to the survey. Simone Bertoli and Kenneth Train provided useful technical advice. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, under grant agreement No 883033 Did elite human capital trigger the rise of the West? Insights from a new database of European scholars.

[†] IRES/LIDAM, UCLouvain, Belgium

1 Introduction

Knowledge and knowledge mobility have enormous social impact. The location choice of scholars, especially those in the upper tail of the human capital distribution, has a measurable effect on innovation and knowledge formation. In the literature, the connection between human capital and economic growth is confirmed in theoretical and empirical results. The essential role of education in adaptation to continuously changing contexts and for driving modernization was first claimed by Nelson and Phelps (1966). Lucas (1988) modeled the positive spillover effects of education, and more recently many empirical studies have presented clear evidence about the essential role of human capital in improving current societies (Barro, 1991; Barro, 2001; Hanushek and Woessmann, 2008; Cohen and Soto, 2007; Barra and Zotti, 2018).

Notwithstanding its fundamental role in socio-economic progress, investments in knowledge are not linear over time. There is a non-steady process of human capital accumulation, with periods of growth, decline, and recovery that seeds a new cycle of expansion (Artige et al., 2004). Italy is a clear example of this fluctuating process: it dominated intellectual activities until the late Renaissance, but its contemporary cultural and educational system is weak. Italian universities outperformed learning institutions in other countries from the 14th century to the first part of the 16th century (de la Croix et al., 2020), but only three universities count among the 200 top institutions in the 2021 QS World University Rankings[1] (Polytechnic of Milan - 137th, University of Bologna - 160th and Sapienza University of Rome - 171st).

I study the mobility of modern day Italian academics, and how they choose where to develop their career, to learn more about the modern university system of the peninsula. I take several factors into account: distance is increasing the cost of travelling, but it might be offset by the skills and knowledge acquired by the scholar and by the prestige of the university. I estimate professors' location choice as a function of distance and quality, given the location of universities. I map the current academic market with its embodied human capital and compare this to previous eras of Italian academia. By comparing how scholars are moving nowadays to how they moved in the past, I locate Italy in the fluctuating cycle, estimate the path of Italian academia, and map out future directions.

[1] 2021 QS World University Rankings are at: https://www.topuniversities.com/university-rankings/world-university-rankings/2021.

For my research I have built a new dataset of contemporary Italian professors, capturing information about their origins and their individual quality. To collect information about the birthplaces of live persons I had to secure privacy authorizations, which could have hindered data collection. To overcome this missing data issue, I used a more accessible proxy: the location of professors' lowest level of education. Once I had a value for the birth and/or education location, I used a Principal Component Analysis (PCA) to build a composite indicator of individual quality out of eight bibliometric indexes.

I use a Random Utility Model (RUM), and specifically, a multinomial logistic regression to compare scholars' utility of living in a region other than their birthplace. I limit the analysis to choices made within the academic world, given the impossibility to consider the choice faced by academics when they decide whether to become a professor or to follow other career paths. I use the approach developed by de la Croix et al. (2020), who study the European Academic Market from 1000 CE to 1800 CE, to compare past and present outcomes in academia. My main estimations rely on information about geographical distance, individual quality of current professors (*human capital* hereafter), and aggregate quality of Italian universities (*notability* hereafter). (1) *Agglomeration* investigates whether Italian scholars are attracted by universities with higher notability, (2) *positive sorting* tests whether scholars with higher human capital weight the notability of universities higher than do professors with lower individual quality, through the interaction term between human capital and notability, and (3) *positive selection* questions whether scholars with higher human capital move further, utilising the interaction between human capital and distance. There is an extensive literature showing that better-educated individuals are the most mobile portion of the population, with their higher growth perspectives giving them stronger incentives to move (Schiller and Cordes, 2016; Handler, 2018; Barrientos, 2007; Docquier and Marfouk, 2006; Faggian and McCann, 2009; Zhao et al., 2021). Grogger and Hanson (2011) showed precisely that highly educated people are more likely to move (positive selection) and how these highly-specialised migrants choose destinations that compensate knowledge better (positive sorting).

I estimate these effects for Italian academia, and find that the standard distance effect is negative and has a magnitude in line with migration literature (Beine et al., 2011). To study *agglomeration* I include attractive features of the city in which the university is located (size and wealth), in addition to notability. The latter is strong and positive, signalling that the quality of Italian universities is strongly attractive for contemporaneous scholars. Together with disposable household income in the city (i.e., city wealth), these two forces highlight the effect of *agglomeration*. However, the estimator for the size of the city is negative, implying a dispersion effect – although with a lower magnitude than agglomeration. This finding is crucial for understanding mobility patterns and policy directions: it is essential to attract high-skilled people to create a dynamic context and generate positive spillovers for society, which may lead the country into the virtuous part of the cycle (Kerr et al., 2016; Grogger and Hanson, 2015; Kerr et al., 2017; Stephan and Levin, 2001).

I also find evidence of positive selection and positive sorting. Indeed, *positive selection* (interaction between distance and individual quality) is a solid result, which confirms that the higher the individual quality, the stronger the incentives for the scholar to travel to progressively better destinations. *Positive sorting* (interaction between human capital and notability) has a weaker significance level than selection. The weakness of sorting is due to the structure of Italian higher education, which is still influenced by the traditional seniority-based system (Rebora and Turri, 2008; Capano, 2008; MacLeod and Urquiola, 2021). Reforms to increase the autonomy of universities,[2] permitting greater investment in local excellence, are too recent to be strongly detected in the current project, and so sorting only reaches a low level of significance. The seniority-based system may explain why Italian universities lost their leading position: there is evidence of a highly significant positive sorting only until 1526, which fades towards 1800. The sorting effect only regains power in the sample of present-day scholars, but the significance of current sorting is not as strong as at the birth of Italian universities. This is probably due to the very recent academic reforms. Either way, these results are key to understanding Italy's current position in the cycle, where there is a new momentum for Italian universities.

In addition to the main regressions, I test for gender differences and find no significant outcomes. Men and women have similar patterns of mobility in Italy, but women represent only 30% of the sample. I do find important differences between public and private univer-

[2] In particular, DPR n. 390/1998 and law n. 210/1998.

sities. A variant of the standard logit model shows a greater expected utility for scholars in choosing private universities over public ones. This bolsters the argument in favour of a more autonomous, excellence-driven academic apparatus.

My analysis contributes to the migration and knowledge-based mobility literature. To the best of my knowledge, much of this literature deals with more general samples of high-educated/high-skilled people (Beine et al., 2011; Grogger and Hanson, 2011; Kerr et al., 2016; Kerr et al., 2017; Handler, 2018; Docquier and Marfouk, 2006). Only a few articles investigate the mobility of academics or scientists. Stephan and Levin (2001) find evidence of the extra vitality brought by foreign scientists (foreign-born and foreign-educated) to the U.S. in the fields of Science and Engineering (S&E). Grogger and Hanson (2015) study the mobility of foreign-born students in S&E after earning an American Ph.D. degree, claiming positive spillover effects for destination countries. The migration of German-affiliated researchers is addressed in Zhao et al. (2021), who find a net outflow of researchers from Germany.

The current research keeps the focus on the academia and aims to integrate the knowledge-based mobility literature about the Italian university system. To the best of my knowledge, published papers on Italian scholars have studied the role of individual quality on selection processes (Checchi and Verzillo, 2014; Checchi et al., 2014a) or its link with the competition and incentives generated within the Italian scientific sector (Checchi et al., 2014b). There have been no studies connecting human capital and the mobility of professors. Within the Italian university system, only student mobility has been analysed (Agasisti and Bianco, 2007; Triventi and Trivellato, 2008; Bratti and Verzillo, 2019), and no previous works have investigated the drivers of scholars' location choice in Italy.

2 Data sampling

2.1 Institutional context

Italy is home to the oldest University in Europe[3] and has a long tradition of literates and scholars such as Giovanni Boccaccio, Leonardo da Vinci and Galileo Galilei, who belong in

[3] University of Bologna, founded as a university in 1088.

the upper tail of the human capital distribution. The Italian academic system has interesting peculiarities, which are worth mentioning before the empirical analysis. Italy's education system was centralized for a long time, make it subject to the whims of the governing body. This increased the importance of hierarchy within academia, based on informal relationships between the most important chaired scholars and government ministers (Rebora and Turri, 2008; Capano, 2008). In the 20th century, this centralization of the system was intended to reduce the inequalities in the Italian education system (Triventi and Trivellato, 2008) and there were some positive outcomes. Social mobility improved (Barone and Guetto, 2016) and performance among geographical areas converged (Baldissera and Cornali, 2020), but academia remains seniority-based (Rebora and Turri, 2008), not only in Italy but throughout Europe (MacLeod and Urquiola, 2021). In 1946, to improve the functionality and equality of the system, the the universities' autonomy principle (art. 33 paragraph 6) was defined in the Italian Constitution. This precept aimed to underline local excellence (Checchi and Verzillo, 2014), giving each university the autonomy to hire eligible professors. However, this Constitutional principle entered into force only at the end of the 90s, due to the lack of technical standards. The actual implementation of the reforms[4] fragmented the Italian academic market, and it retained some elements of the seniority-based apparatus (Rebora and Turri, 2008; Bertola and Sestino, 2011; Perotti, 2008; Bini and Chiandotto, 2003). Among the other modifications, it is important to note that Berlinguer's decree (DPR n. 390/1998) shifted recruitment from a national to a local process[5]. In 2010, the selection procedure was modified again and became a two-stage process. Nowadays, a scholar has to pass a national open competitive exam to be eligible, and then must win a local contest to be hired by a university (Rossi, 2016; Checchi and Verzillo, 2014; Durante et al., 2011).

In a system where seniority is the main driver of an academic career, quality and individual ability may be irrelevant. However, Checchi et al. (2014b) found evidence of the opposite. They showed how better scholars were those who responded best to an increase in the level of competition within the university sector, even in the presence of weak incentives.

The literature about mobility in the Italian academic world is thin and mostly focuses on

[4] Among the others, the most important ones are laws n. 168/1989, n. 210/1998 and the DPR n. 390/1998.

[5] For a detailed explanation of the recruitment process in Italy see the following web page of the Ministry of Education, Universities and Research (only in Italian): https://www.miur.gov.it/reclutamento-nelle-universita.

student mobility (Agasisti and Bianco, 2007; Bratti and Verzillo, 2019; Triventi and Trivellato, 2008); I have not found any literature on drivers of professors' location choices. Insight into scholars' mobility within the country, and a characterization of the forces that attract them to an institution, can inform public policy.

2.2 Professors and Universities

This research is based on a new dataset. The data collection started with RePEc's[6] ranking of the "Top 25% Institutions and Economists in Italy". It uses the EDIRC database (Economics Departments, Institutes, and Research Centers in the World), which includes universities, public agencies, central banks, independent research centres, and associations (for more details see section 2.3 in Zimmermann, 2012). Each institution gains from every author's affiliation RePEc collates, implying an advantage for more populous entities (section 6 in Zimmermann, 2012; Seiler and Wohlrabe, 2011).

Universities. For the present work, only the following universities will be taken into account:[7] University of Bologna (UNIBO), Catholic University of the Sacred Heart (CATT), University of Verona (UNIVR), University of Catania (UNICT), University of Milan (UNIMI), University of Rome - Tor Vergata (UNIROMA2), University of Florence (UNIFI), University of Venice (UNIVE), Polytechnic University of the Marches (UNIVPM), Sapienza University of Rome (UNIROMA1), University of Turin (UNITO), University of Trento (UNITN), University of Naples Federico II (UNINA), University of Padua (UNIPD), Bocconi University (BOCCONI), University of Genoa (UNIGE), University of Palermo (UNIPA), Free University of Bozen (FUB), University of Bari (UNIBA), University of Milan-Bicocca (BICOCCA), Luiss University in Rome (LUISS).

In this list there are 16 public universities, one polytechnic (UNIVPM), and four privately founded universities (BOCCONI, CATT, FUB and LUISS).

Scholars. Each institution includes a list of members (registered in the RePEc Author

[6] Research Papers in Economics (RePEc) is a project collecting bibliographic data about papers in economics and similar fields, aiming to spread and enhance relative researches.

[7] From here onwards the words 'university' and 'institution' are treated as synonyms.

Service) and I include these observations in the dataset.[8] The people registered on the server have different roles inside academia. In this study I only include professors—full, associate, adjunct and assistant—and research fellows (also postdoctoral).[9] I include a few emeritus professors who are still teaching. Only scholars who are active in teaching are included in the sample: I call this a *”teaching disclaimer”* and it captures emeritus professors and academics taking part in visiting programs or national/international collaborations. Hence, a visiting professor is only included in the sample if she explicitly mentions her teaching activity at the host university (more on multiple affiliations later). Scholars "on leave" were not considered part of the sample, given the absence of the *teaching disclaimer*. This rule excluded research centres like CEPR, IZA, CESifo, given the honorific nature of their appointments. Table 1 presents the precise taxonomy for the scholars included in the dataset, with quantities and percentages.

Table 1: Taxonomy of scholars

Categories	**Quantity**	**Percentage**
Full professors	420	39%
Associate professors	303	28.13%
Assistant professors	104	9.66%
Adjunct professors	51	4.74%
Research fellows	116	10.77%
Post-doctoral fellows	30	2.79%
Emeritus professors	8	0.74%
Visiting professors	45	4.18%
Total	**1077**	**100%**

Affiliations. Once a scholar is identified, they are associated with their university. This process required a careful investigation for each academic. The Curriculum Vitae (CV) was the main source, but where it was out of date or incomplete I used LinkedIn[10] and personal web pages (institutional and/or private). I used the most updated affiliation at the moment of consultation.[11] Affiliations to telematic universities were not taken into account and research centres were excluded. Only the European University Institute (EUI)[12] and the University School for Advanced Studies in Pavia (IUSSPAVIA)[13] have been considered because they have

[8] The ranking is updated month by month; hence the names collected (and the status granted to them) can change with respect to the period of data collection, which is approximately December 2020 – September 2021.

[9] After determining the status of each member, doctoral students are excluded from the dataset.

[10] Professional social network: linkedin.com

[11] Consultation period: December 2020 - September 2021.

[12] https://www.eui.eu/en/home

[13] http://www.iusspavia.it/en/web/guest/university-school-for-advanced-studies

characteristics of actual universities.

For those universities with multiple locations, I counted the main location, assuming that the majority of the scholars teaching in one location are also teaching in the other(s). This can generate some bias when locations are far away from each other as in the case of Catholic University, with four locations, in Milan (main building), Brescia, Piacenza, and Rome. I discuss the robustness check for this in subsection 6.3.

Multiple affiliations. Some scholars are associated with more than one university, in Italy or abroad. Multiple affiliations comprise 7.06% of the sample, with a maximum of four affiliations. In the past, academics linked with multiple institutions were associated with high-quality scores (de la Croix et al., 2020), whereas nowadays it is more common to encounter multiple-affiliated scholars with low bibliometric indicators. Usually, these academics are younger and have a postdoc position in a university while teaching in another institution.

Empirically, each affiliation of the same scholar is treated as if it was chosen by different individuals, leading to their overestimation with respect to unique-affiliated scholars (see section 3.4). In the following part of the paper, the former will be called repeated movers (RM), and the latter single movers (SM).

Treating multiple affiliations in this way, the initial sample counts 1440 observations. A cleaning process removed from the sample scholars who are no longer members of the Italian academy, Ph.D. students and non-teaching emeritus and visiting professors, and those who are on leave.[14] The cleaning process reduced the sample to 1077 names.

Dataset. This procedure identified 76 universities, of which 39 are foreign universities and 37 are Italian. From this set, universities with fewer than 20 scholars have been excluded, given their minor relevance for academics' choice. The resulting list is the set of choices each professor faces when maximising their location decision, which now stands at 17 universities, all of which are Italian. The number of scholars in the database decreased to 936 observations, the percentage of multiple affiliations is now only 3.10% and the maximum number of associations decreased to three. From here onwards, this is the subset for analysis. Table 2 summarises

[14] One of the main difficulties was understanding the meaning of the various roles and titles indicated by each scholar. The final dataset was built to the best of available knowledge, however, minor errors may still be present.

the differences between the original dataset and the subset obtained after dropping universities with fewer than 20 scholars.

Table 2: Comparison between datasets

	Original	**Subset**
Tot. observations	1440	936
N. Universities	76	17
Obs. after cleaning process	1077	936
Obs. with known birthplaces	936 (86.91%)	815 (87.07%)
Obs. with known education	1044 (96.94%)	904 (96.58%)
Obs. with not-known education	33	32
Multiple affiliations	7.06%	3.10%
Max n. affiliations	4	3

2.3 Data on locations

In my analysis I study the distance a scholar is willing to travel to a given university to develop her career. I treat distance as an increasing cost for the individual. The further she is from her point of origin, the greater the distance and the higher the cost (Schwartz, 1976), also in terms of family attachment.

I collect the birthplace for each observation, and treat this as an observable proxy of scholars' usual life context. Other variables are not observable. For instance, where academics' families live is non-observable, as is the location of their partner's employment, even if these may be relevant drivers of professors' decisions. Other sources could be used to detect these determinants (i.e., surveys), but these were beyond the scope of the current project.

CVs and personal webpages were the main sources for affiliations, given that neither LinkedIn nor RePEc provide birthplace information.[15] Only about 30% of the sample indicated their place of birth somewhere in their public profile. Although information about living persons is abundant and often easy to access, bureaucratic and privacy authorizations, which are essential to protect personal information, slow the data collection process. Instead I sent direct emails requesting this information, increasing by about 55% the number of birthplaces collected.[16] This gives me a known birthplace for 87.07% of the academics (815 observations).

I included in the dataset the location of the institution where each scholar obtained her lowest, publicly-stated degree of education. I consider this another proxy for birthplace, given

[15] My thanks to RePEc administrators for their prompt answers.

[16] I express my heartfelt thanks to those who answered in so a interested manner.

that the two are likely to coincide or be reasonably close. This measure increased the dataset to 904 observations, reaching coverage of 96.58%. For the majority of the sample, the lowest level of education is the bachelor's degree, but some academics mentioned also the high school. Only for a few observations, the lowest education level available was the master's degree, while for five scholars only information about the Ph.D. is known. Given their small number (only 5 out of 936 observations), and given the location of their Ph.Ds: four of them received their doctorate from the same university (or a close one) that they teach at, and only one obtained their title abroad, an ad-hoc robustness check was not necessary. For most observations I found educational information in CVs or LinkedIn profiles, and if I could not find it online I requested it with a direct email.[17] However, education information is missing for 32 observations (3.42% of the sample) and they will be excluded. I implement two different regressions: birthplaces and the locations of the lowest level of education (see section 3.3).

I match decimal coordinates to location data,[18] giving me a dataset with i observations associated with a geo-localized birth and/or education site, and k geo-localized universities.

2.4 Data on quality

For quality indicators I collect *aggregate* quality scores (notability) and *individual* bibliometric indexes (human capital). The former are at the university level, while the latter are associated with each scholar.

Aggregate quality. Notability indicators may suffer from endogeneity, because university scores are related to the quality and quantity of scholars. I address this by using past indicators. The average age for Italian academics is 48 years (Elaborazioni su banche dati MIUR, DGCASIS – Ufficio VI Gestione patrimonio informativo e statistica, Morana, 2020) and careers usually begin at around 30 years, so I look for quality indicators from 20 years earlier. The RePEc archives provide aggregate quality scores for top institutions, organised by country, going back as far as 2007. Prior to that, only simple/ordinal rankings are available, and there is no institutional score. I elected to consider scores from 10 years ago, as of December 2010.[19,20]

[17] Usually in the case of emeritus professors, or persons not in the Italian academia anymore. Thus, this procedure helped also to determine the status of these exceptions.

[18] The websites used are: latitudelongitude.org, tuttitalia.it for Italy, while latlong.net for foreign cities.

[19] The research started in December 2020.

[20] The ranking is available here: https://ideas.repec.org/top/old/1012/

Scores for top universities were collected from country rankings. The scores in these rankings are weighted averages of the credit brought by each affiliated scholar: the highest portion (0.5) of affiliation is given to the scholar's main university and the remainder is a weighted average of the other appointments (for the specific formula see section 6, Zimmermann, 2012). This can generate some biases, for example decreasing the relevance of the main affiliation as more associations are added, as pointed out by Seiler and Wohlrabe (2011).

It was possible to assign a quality score to all 17 universities in the sample. RePEc uses reversed indexes in which lower scores indicate higher quality; I convert them to have a direct relation between indexes and quality. The notability ($\ln Q$) linked to each university ($k \in K$) can be visualized in Figure 1.

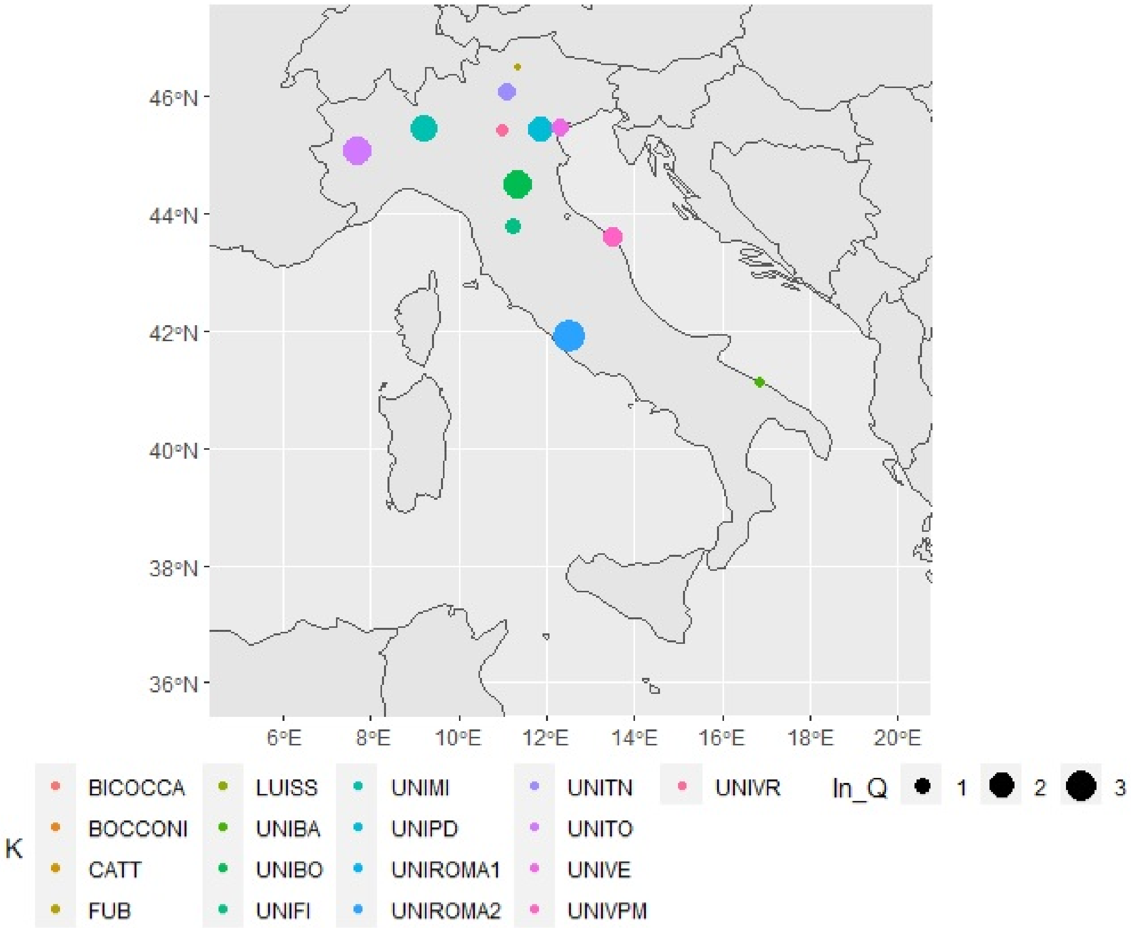


Figure 1: Bubble Plot on Italy map. Showing notability indexes associated to each university: the higher the $\ln Q$, the bigger the bubble on the map. Each point represents the location of $k \in K$ institutions.
Note: BICOCCA, BOCCONI and CATT are overwritten by UNIMI, which has the highest notability in Milan. UNIROMA1 and LUISS are overwritten by UNIROMA2, which has the highest notability in Rome.

Individual quality. There are many individual bibliometric indicators to choose from. RePEc has the top authors per country ranking (i.e. "Top 25% Institutions and Economists in Italy"). These human capital scores are the harmonic mean of various rankings based on different factors (section 5, Zimmermann, 2012) and more than 800 scholars are ranked. I use the December 2020 ranking (see below for missing data).[21]

In the literature, academic quality is measured by indicators provided by *Web of Science* (*WoS* – with its three subject specific ISI citation databases; Yang and Meho, 2006). The *WoS* social science indicator goes back to 1956.[22] For a long time it has been one of the few multidisciplinary databases to assign authors' scores based on citations from an original set of sources (Neuhaus and Hans-Dieter, 2008; Jacso, 2005). The main issue with Web of Science measures is the relative coverage: only a fraction of sources are considered, although those that are considered (i.e. journal literature) are significant (Norris and Oppenheim, 2007). However, for economics and social science, this literature is not the main way that knowledge is disseminated (Neuhaus and Hans-Dieter, 2008).

Quality-evaluation possibilities are now augmented with the automated databases *Scopus*, from Elsevier, and Google Scholar. The former covers a wider range of sources than *WoS*: it starts with an Elsevier database and it goes back to 1996 for social science[23] (Jacso, 2005; Yang and Meho, 2006; Norris and Oppenheim, 2007). *Google Scholar* is a free Google database that uses a wide range of sources, but does not identify clearly what those sources are. This gives it low reliability, which is added to weak, imprecise performance, as pointed out by Neuhaus and Hans-Dieter (2008). However, because it is free and has some of the widest coverage among bibliographic indicators, Google scholar still has value as a measure of quality (Neuhaus and Hans-Dieter, 2008).

I add to the comparison the *WorldCat identities* index. This database has measures for works (*Worldcat Works*) and library holdings (*Worldcat Library*) for each scholar (and organization) found in WorldCat.org and OCLC sources (OCLC Research, WorldCat identities).[24]

Because no single indicator is perfect, I create a composite indicator of: RePEc score, Worldcat works and library holdings,[25] Google Scholar citations, H-index and i10-index,[26] WoS

[21] Given that the ranking is updated every month, the current online score could present some differences.

[22] *"Coverage in Web of Science goes back to 1945 for Science Citation Index, 1956 for Social Sciences Citation Index, and 1975 for Arts & Humanities Citation Index."* (Yang and Meho, 2006)

[23] Scopus goes back at maximum to 1966. (Yang and Meho, 2006)

[24] https://www.oclc.org/research/areas/data-science/identities.html

[25] https://www.worldcat.org/identities/

[26] https://scholar.google.com/

H-index,[27] and Scopus H-index.[28] To understand the information added by each indicator I use a *Principal Component Analysis* (PCA) to reduce the number of variables, without losing too much accuracy and information. Once the correlation between the variables is computed (Figure 2) and their standardization is completed, the PCA compresses most of the information among the first principal components, which are new uncorrelated variables. For this research, I take the first two components into consideration, because their standard deviation is greater than one. Moreover, the cumulative information explained by these two components is 78.09% of the total (Table 3). Hence, considering the two components together, the analysis gains simplicity while losing only a little portion of its accuracy; from here on I use them to represent the new individual quality index.

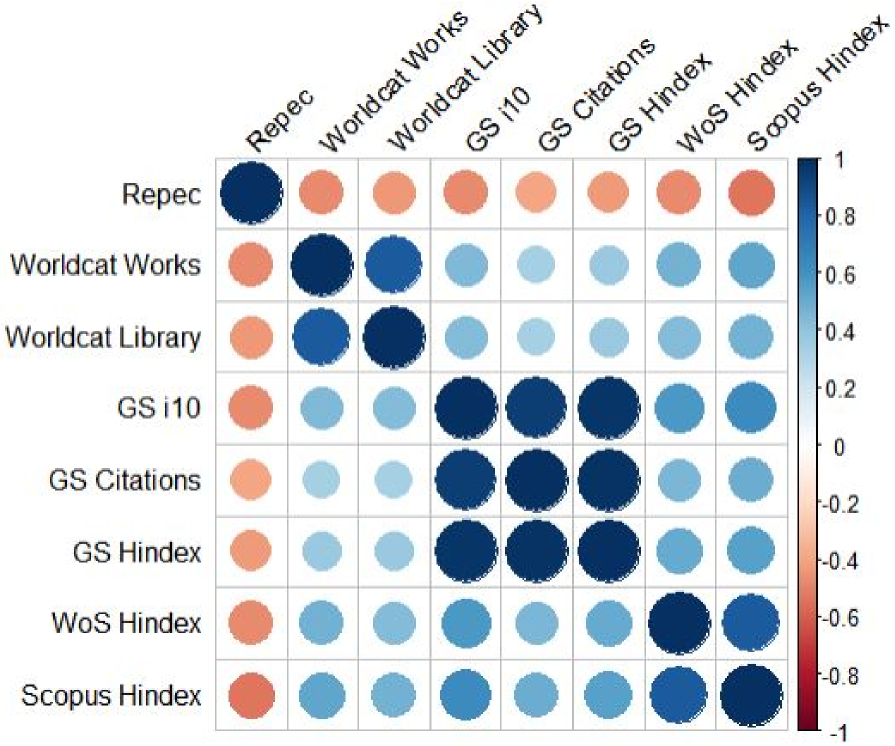


Figure 2: Correlation Matrix Plot showing the correlations between the eight different bibiliomeric indicators included in the analysis.

Table 3: Principal Components table. Showing the standard deviation (St.dv.), the proportion of variance (Pr.Var.) and the cumulative proportion (Cum.Pr.) for each principal component.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**
St.dv.	2.2051	1.1768	0.9033	0.7604	0.4096	0.3861	0.1879	8e−02
Pr.Var.	0.6078	0.1731	0.1020	0.0723	0.0210	0.0186	0.0044	8e−04
Cum.Pr.	0.6078	0.7809	0.8829	0.9552	0.9762	0.9948	0.9992	1.0000

[27] https://app-webofknowledge-com.pros.lib.unimi.it/author/search?lang=en_US&SID=C45nUucTHDOd2z2VUPs

[28] https://www.scopus.com/freelookup/form/author.uri?zone=TopNavBar&origin=NO\%20ORIGIN\%20DEFINED

3 Methodology

3.1 Main hypotheses

In Section 2.1 I described some interesting features of Italian universities. One of the main aim of this paper is to understand how these features have changed over time. In order to achieve this objective I compare my results with de la Croix et al. (2020). The authors tested the following hypotheses in the whole Europe for the period between 1000 CE and 1800 CE, while I study them for contemporary Italy. The role of the location of higher education institutions (Barra and Zotti, 2017; Audretsch, 1998; Drucker and Goldstein, 2007; Agasisti et al., 2019; Cottini et al., 2019) is considered to be exogenous.

Hypothesis 1: *Agglomeration: scholars are attracted by universities with higher notability.*

I expect to find agglomeration (Kerr et al., 2016; Kerr et al., 2017; Grogger and Hanson, 2015) although the distance covered by academics could appear shorter than in the past, with a lower magnitude of the coefficient. In Italy, the local appointment of professors may have increased the probability of finding local excellence (Checchi and Verzillo, 2014), and the importance of networks and nepotism (Durante et al., 2011). With this hypothesis I test for agglomeration forces, such as notability of the university, and the attractiveness of the city in which the institution is located, measured by the size of the population (istat.it) and the local disposable income of private households (finanze.gov.it).

Hypothesis 2: *Positive sorting: scholars with higher human capital weight the notability of universities higher than scholars with lower human capital do.*

I hypothesise that better scholars have better career prospects, and their expected gains are higher in high-quality environments (Grogger and Hanson, 2015; Docquier and Marfouk, 2006). Thus better professors would assign higher weight to the notability of the university.[29]

Hypothesis 3: *Positive selection: scholars with higher human capital move over greater distances than scholars with lower human capital.*

[29] I intend the term 'professor' as a synonym of 'scholar/academic'.

The literature shows that better-educated people are more mobile (Schiller and Cordes, 2016; Grogger and Hanson, 2011; Handler, 2018; Barrientos, 2007), hence my hypothesis that better professors travel further.

3.2 The model

I use a Random Utility Model (RUM), a gravity model widely used in migration analysis (Grogger and Hanson, 2011; Ortega and Peri, 2013). It determines the individual utility of living in a certain region and compares it to the expected utility from moving to alternative locations (Ramos, 2016).

I implement a standard multinomial logit model (Akcigit et al., 2016; Ortega and Peri, 2013), which is a specification of the RUM and requires perfect elasticity of demand in the academic market i.e., that there is a position available for every scholar. In Italian academia, there is a two-step hiring procedure: scholars are filtered at the national level and then at the local level. The assumption of perfectly elastic demand implies that each professor who succeeds at the national level will succeed in finding a chair that she prefers at the local level. This is a reasonable assumption, because the reforms of the university system (in 1998 and in 2010) simplified bureaucratic processes and increased the opening of vacancies (Checchi and Verzillo, 2014, Rossi, 2016). However, in practice only professors with higher individual quality can freely choose the location of their career. To account for this I include the individual human capital score in the analysis. Keeping the perspective of partial equilibrium analyses, I introduce competition variables as demand-side factors, i.e. universities' notability, desirability of the city and individual human capital.

A multinomial logit model allows us to compute the probability that a k university, belonging to K set of choices, is maximising an i scholar's utility (McFadden, 1974). The first step is to define the utility function for each i scholar. It is defined by a deterministic component $V_{ik} = \beta x_{ik}$, capturing average benefits and costs of each location choice, and by a random component ϵ_{ik} orthogonal to βx_{ik}, which describes unobservable factors that may influence the utility.

The utility function can be written as follows:

$$U_{ik} = V_{ik} + \epsilon_{ik} = \beta x_{ik} + \epsilon_{ik} \tag{1}$$

The standard logit model relies on the assumption of independent individual choices, which requires ϵ_{ik} be independent and identically distributed. Under this assumption, the main equation of the multinomial logit model defines the probability of choosing a university k, which depends on the specificities of that institution compared to the specificities of the remaining set of available choices:

$$p_{ik} \equiv Prob[U_{ik} = \max_{k' \in K} U_{ik'}] = \frac{\exp(\beta x_{ik})}{\sum_{k' \in K} \exp(\beta x_{ik'})} \tag{2}$$

Another important assumption of the logit model is the Independence of Irrelevant Alternatives (IIA). With independent and identically distributed error terms, the IIA assumption implies that the choice between two specific alternatives should depend only on their own features, without any influence from a third feature (McFadden, 1974). This means that the choice between two universities depends only on the two institutions considered and not on other alternatives. In Subsections 3.4 and 3.6, I relax this assumption.

In the next step, I explicit the deterministic component, which captures the difference between average benefits and average costs of choosing $k \in K$. The benefits are an increasing function of the university's notability Q_k (as defined in the previous subsection), and of the attractiveness of the city. Hence, I include the variables P_k and Y_k, representing respectively cities' total population (capturing the size of the city) and households' disposable income[30] (capturing the wealth status) of the city in which k university is located. In addition, I include an interaction term ($q_i Q_k$) to capture the fact that better scholars (with a high individual quality index q_i) gain more from a welcoming environment (i.e., a university with high Q_k).

[30] Italy (2018, average disposable income IRPEF, city-level): https://www1.finanze.gov.it/finanze3/analisi_stat/index.php?search_class[0]=cCOMUNE&opendata=yes#

Therefore, the benefits equation is:

$$B_{ik} = a_0 + a_1 Q_k + a_2 P_k + a_3 Y_k + a_4 q_i Q_k \tag{3}$$

where $\forall a \in \{a_1, a_2, a_3, a_4\}$ greater than zero.

The costs are competition costs: the greater the distance from the birthplace (or education-place) the higher is the burden of travel, and hence the lower the competition among scholars. However, the better an academic (i.e., the higher q_i), the more she has to gain in a certain university environment Q_k, implying a reduction in competition costs ($q_i Q_k$). In addition, as shown by the literature, a better scholar should be willing to move longer distances, so I include the interaction term between distance and individual quality ($d_{ik} q_i$), which may increase competition costs. Finally, the higher the attractiveness of a city (P_k and Y_k) and of a university (Q_k), the higher the competition. Therefore, the costs equation is:

$$C_{ik} = b_0 + b_1 Q_k + b_2 P_k + b_3 Y_k - b_4 q_i Q_k - b_5 d_{ik} + b_6 d_{ik} q_i \tag{4}$$

where $\forall b \in \{b_1, b_2, b_3, b_4, b_5, b_6\}$ greater than zero.

Having defined the equations for the benefits and the costs, I specify the net benefit for each dyadic match (i.e., the association of scholar i with university k) and explicit the deterministic component of the utility function. This is done by subtracting (4) to (3), with the addition of a fixed effect γ_k. The subscript k suggests that fixed effects refer to time-invariant, non-measurable universities' (cities') characteristics which may influence their ability to attract human capital.[31] These fixed effects perfectly identify the agglomeration variables (Q_k, P_k, Y_k) included in the model, given that both are destination-specific and time-invariant. For this reason, I exclude fixed effects when agglomeration forces are considered (more on this later). The final expression is:

$$\beta x_{ik} \equiv V_{ik} \equiv B_{ik} - C_{ik} = \beta_0 + \beta_1 Q_k + \beta_2 P_k + \beta_3 Y_k + \beta_4 q_i Q_k + \beta_5 d_{ik} + \beta_6 d_{ik} q_i + \gamma_k \tag{5}$$

β is a vector, whose parameters are common to each scholar. Specifically, the constant

[31] Fixed effects are used to overcome omitted variables biases, they control for unobserved variables which do not change over time. If there is a change over time, they can be inefficient, with large standard errors. (Williams, 2019)

β_0 in equation (5) is the difference between the two constants in (3) and (4). To define the *agglomeration effect*, I look at $\beta_1, \beta_2, \beta_3$ computed as $a_j - b_j$, with $j = \{1, 2, 3\}$, which represent notability of universities and attractiveness of cities. A positive sign indicates the presence of agglomeration and a negative sign evidences dispersion. I measure the *sorting effect* with $\beta_4 \equiv a_4 + b_4$. A positive coefficient means that the higher the individual quality, the smaller the cost (or the higher the gain) to travel to better universities. The coefficient $\beta_5 \equiv -b_5$ captures the expected effect of distance, considered as a cost, as previously mentioned. $\beta_6 \equiv b_6$ underlines the *selection effect*: when there is a positive sign, better scholars move further.

Equation (5) includes only destination-specific regressors. Human capital (q_i) is always interacted with university-specific variables (i.e., notability and distance), because it influences all dyadic matches in a symmetric manner.

3.3 Main results

In this section, I use the multinomial logit model described above to estimate the main regression of the research. First, I consider scholars for whom the place of birth is known (815 observations - 87.07% of the sample). Second, as a robustness check, I use the site of their lowest level of education (904 observations - 96.58% of the sample). I link each site with its geographic coordinates and each academic with a unique individual quality index, computed with a PCA (see section 2.4). I discarded universities with fewer than 20 professors from the database, assuming that they have minor relevance in the total set of choices.[32] The university set counts 17 geo-localized institutions linked to their RePEc quality score (see section 2.4). Because I work in logarithm terms, the estimation does not allow for zero indexes at aggregate or individual level. If a scholar does not have a positive score, I fill this gap with the lowest human capital index of the sample (794,82 for RePEc, 1 for all the other indicators). It is reasonable to assume that such a scholar does not publish as much as her peers with a positive score. However, it is possible that the sources used to compute bibliometric indicators do not accurately reflect her work, which is a known flaw in quality evaluations. I apply the same reasoning for universities with indexes at zero and link them with the lowest positive score of notability. Finally, I also

[32] In the Appendix I show the results of the two main regressions, considering also a less stringent threshold of 5 scholars per university - the major difference is in the sorting effect, which totally disappears in the complete model of birthplaces analysis.

take the logarithm of the measure of distance which raises the issue of zero distances, affecting scholars born in the same city where they teach. These academics bear the minimum cost of distance, which I assume to be the same as in de la Croix et al. (2020): 3,5 km, the walking distance from the Vatican city to the Colosseum, in the old city of Rome.

In the following part of the section, I describe the results of the main regression which considers scholars' locations of birth. I use the package called "mlogit", written by Croissant (2020). I focus the evaluation on the sign and on the significance of the coefficients of distance, agglomeration, selection, and sorting effect. I control for unobserved characteristics of universities with fixed effects in each regression, except for when I introduce agglomeration effects. In this case, I include the variables which capture the observed characteristics of the city where the university is located (P_k and Y_k - see section 3.2) and represent the reputation of the institution (Q_k). Table 4 presents some descriptive statistics.

Table 4: Descriptive statistics

Variables	**Obs***	**Mean**	**St.Dv.**	**Min**	**Max**
Birthplace analysis:					
ln of distance	13855	5.4535	1.3096	1.2529	9.1979
ln of human capital	13855	0.1747	2.1573	−4.4680	7.6000
ln of notability	13855	1.5438	0.8581	0.5391	3.5405
ln of population	13855	12.9904	1.7133	7.8579	14.8786
ln of income	13855	10.1586	0.1378	9.9396	10.3854
Lowest level of education analysis:					
ln of distance	15368	5.2210	1.5693	1.2529	9.1979
ln of human capital	15368	0.0259	2.2034	−4.4680	7.6000
ln of notability	15368	1.5438	0.8581	0.5391	3.5405
ln of population	15368	12.9904	1.7133	7.8579	14.8786
ln of income	15368	10.1586	0.1378	9.9396	10.3854

Note: *Obs counts the number of possible dyadic matches in each analysis.

Table 5 shows the results of the multinomial logit estimations with known birthplaces. The dataset counts 815 scholars (87.07% of the sample) who choose among 17 universities, resulting in 13855 possible dyadic matches.

The first column contains the basic gravity equation and highlights the negative sign of *distance* coefficients, $\ln d$. This means that the greater the distance between the birthplace and the location of the university, the higher the costs and the lower the probability of finding a dyadic match. Distance coefficients remain highly significant in every specification. The magnitude is consistent with the contemporary migration literature; for example, in *"Diasporas"* (by Beine et al., 2011) they also find distance coefficients of around 0.7 when migrants are not divided into low- and high-skilled categories. However, this coefficient is lower than in analyses of past periods (de la Croix et al., 2020).

I add *selection* effect in the second column, defined by the interaction term between human capital and distance, $\ln q \ln d$. As expected the sign is positive, which means that scholars with higher human capital are less affected by distance than scholars with lower human capital. The high significance of the coefficient (at 1%) confirms the third hypothesis of *positive selection* in every specification of the model.

Column (3) shows the effect of *sorting*, through the interaction between individual human capital and university notability ($\ln q \ln Q$). The positive sign of the coefficient is evidence for *positive sorting*, as expected from the second hypothesis. Despite this, the significance of sorting appears weaker than selection. The sorting effect is non-significant when considered alone in column (3), but it becomes slightly significant (at 10%) in column (4) when I include selection. Sorting maintains the level of significance at 10% in the complete model (column (6)). Finally, I compare log-likelihood (LL) values in order to compute a likelihood-ratio (LR) test: considering column (4) over column (1), the null hypothesis of no selection and no sorting is rejected at any conventional significance level.

To investigate *agglomeration*, I exclude university fixed effects from the regression (columns (5) and (6)), otherwise the effect of agglomeration variables cannot be identified (see Subsection 3.2). Without fixed effects, I can study the relevance of the attractiveness of cities where universities are located. All three included variables are highly significant. The coefficient of the logarithm of population ($\ln P_k$) is negative, which preludes the presence of dispersion: the probability that a scholar chooses university k decreases as the city size increases. The coefficient of the logarithm of disposable income ($\ln Y_k$) is positive, which implies that the variable has a strong attractive force: the richer the city, the greater the likelihood a professor develops her career at that institution. The coefficient of the logarithm of university notability ($\ln Q$) is also significant at 1% and positive, which means that the better the reputation of the university, the higher the possibility that a scholar moves there. Therefore, the first hypothesis about agglomeration holds: although from $\ln P_k$ there is a tendency for dispersion (given its

Table 5: Multinomial logit regressions: standard logit model - birthplaces analysis, threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.709^{***}	-0.723^{***}	-0.709^{***}	-0.723^{***}	-0.710^{***}	-0.723^{***}
	(0.029)	(0.029)	(0.029)	(0.029)	(0.028)	(0.029)
Selection:						
$\ln q \ln d$		0.043^{***}		0.043^{***}		0.042^{***}
		(0.013)		(0.013)		(0.013)
Sorting:						
$\ln q \ln Q$			0.031	0.032^{*}		0.034^{*}
			(0.019)	(0.019)		(0.020)
Agglomeration:						
$\ln P_k$					-0.122^{***}	-0.127^{***}
					(0.029)	(0.029)
$\ln Y_k$					2.076^{***}	2.179^{***}
					(0.355)	(0.356)
$\ln Q$					0.171^{***}	0.168^{***}
					(0.044)	(0.044)
k FE	YES	YES	YES	YES	NO	NO
Obs	815	815	815	815	815	815
R^2	0.157	0.160	0.158	0.161		
LL	−1,867.206	−1,861.567	−1,865.936	−1,860.156	−1,916.167	−1,909.365

Note: *p<0.1; **p<0.05; ***p<0.01

negative sign), it is more than compensated by the attractive force of the other two variables ($\ln Y_k$ and $\ln Q$).

Finally, when I consider all the coefficients together (column (6)), they still maintain their signs, significance levels, and magnitudes.

3.4 Robustness checks

In this section, I implement three additional checks to test the robustness of the main results. Firstly, I substitute the data on scholars' birthplaces with data on their lowest level of education, which covers 96.58% of the sample. Secondly, I correct the human capital index by scholars' age: younger professors with similar bibliometric indicators of senior ones should receive more credit in the computation of their human capital index. Finally, I check the overestimation of repeat movers with different strategies.

Lowest level of education. Table 10 (in the Appendix) presents the results of multinomial logit estimations when I study the location of the lowest level of scholars' education. Now the dataset counts 15368 dyadic matches, which associate 904 observations with 17 universities.

Only the *distance* and *agglomeration* coefficients remain significant at 1%. The sign of the former is still negative and each specification confirms the magnitude of about 0.7, although it slightly decreases compared to the birthplaces analysis. From the models without fixed effects (columns (5) and (6)), agglomeration variables ($\ln P_k$, $\ln Y_k$, $\ln Q$) confirm again the first hypothesis, with the same signs as in the birthplaces analysis.

The coefficients of *selection effect* ($\ln q \ln d$) are still positive, but not significant anymore. I find similar evidence for *sorting* ($\ln q \ln Q$), which has positive signs but not significant coefficients. These results prove that the second and the third hypotheses are confirmed only when I take into account the actual location of birth; indeed the LR test between (4) and (1) now fails to reject the null hypothesis of no effects. On the other hand, for the standard effect of distance, I can consider the lowest level of education as a proxy for birthplace. This reasoning holds also for the agglomeration effect, given that in this case the analysis focuses on features of the universities (reputation/quality) and cities (population size and income level); aspects that do not vary compared to the previous analysis. The change of dataset affects more the individual level of quality, which appears in both selection and sorting.

Given the results of both regressions, I consider the birthplaces analysis more relevant for the project. I use this as the benchmark model in the following part of the paper, where I develop further robustness checks.

Age. In the benchmark model, I do not consider the age of scholars to compute the human capital index. I can define this as the *current* human capital indicator, computed as today. However, younger professors with the same level of human capital index as senior ones ought to receive more credit. I include professors' age in the dataset, to compute *age-expected* human capital index. CVs are the main source for this data: most of the scholars disclose their year of birth. However, for 28.85% of the dataset this information is missing and for these cases, I look to the final year of their Ph.D. and assume that scholars at the end of their doctorate are 30 years old. This leaves 29 observations for which I do not have any age reference, neither the year of birth nor the last year of their Ph.D. I exclude them from the analysis, bringing the number of dyadic matches to 13362, which corresponds to 786 academics. Figure 3 presents the age distribution of scholars where the average age is 50 years and a few months.

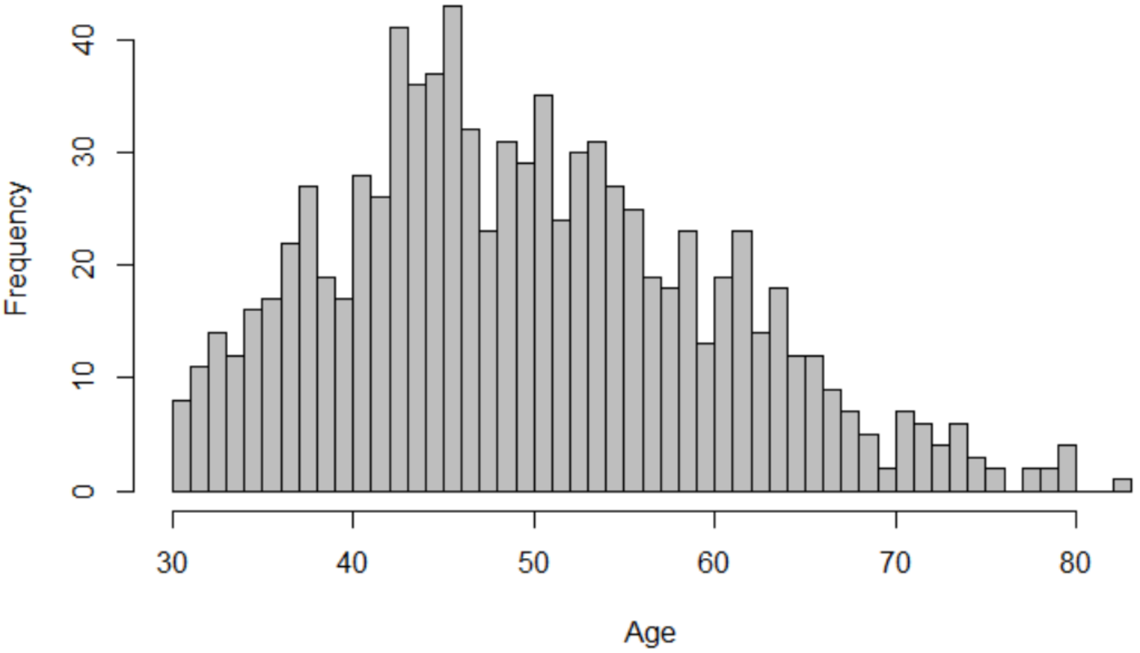


Figure 3: Histogram showing the age distribution of scholars: mean 50.0751, median 49, variance 108.505 (std. dev. 10.4166), skewness 0.4387 (right-skewed distribution) , kurtosis -0.2510 (platykurtic distribution), min 30, max 83.

I expect age to be a crucial factor in explaining individual human capital: as age increases, the probability of publishing great researches increases, augmenting in turn the individual quality index. I run an ad-hoc regression, which confirms this expectation at the 1% significance level (Table 11 - Appendix). I estimate the *age-expected* human capital index at 40 years old for each scholar using the coefficients of the age regression. I re-run the benchmark model with this new indicator and show the results in Table 12 (in the Appendix).

Distance ($\ln d$) and *agglomeration* ($\ln P_k$, $\ln Y_k$, $\ln Q$) coefficients have the same sign as in the benchmark model and are still highly significant: the first hypothesis holds also in the age-adjusted model. Both *selection* ($\ln q \ln d$) and *sorting* ($\ln q \ln Q$) coefficients are positive but not significant, with highly decreased magnitudes. The second and third hypotheses do not hold anymore: I cannot claim that scholars with a higher human capital index move over longer distances or weight more the quality of universities than those with lower individual quality.

In conclusion, when I introduce age, the model confirms only standard results concerning distance and agglomeration, but not more sophisticated ones, like selection and sorting. These findings indicate that the *current* human capital index employed in the main regression was already able to capture age specificities of Italian scholars. Given this, I retain the original model as the benchmark (Table 5).

Multiple affiliations. After I eliminate universities with fewer than 20 scholars, multiple affiliations count for 3.10% of the sample. Given how they enter the dataset (see section 2.2), the choices of these scholars[33] are overweight with respect to those of single movers.[34] However, there should not be an over-representation of better scholars against worse ones as in de la Croix et al. (2020), because my dataset associates repeat movers with a low-quality score, with an average human capital of -0.431 against 0.0318 of single movers.[35] To confirm that repeat movers do not influence benchmark results, I exclude them from the dataset in columns (3) and (4) of Table 15 (in the Appendix) and then I associate them randomly with one of their affiliations in columns (5) and (6). Neither modification alters any sign. Magnitudes increase, but notability slightly decreases when single movers are considered (column (4)). Significance levels remain almost as in the benchmark model, but in both variations (column (3),(4) and (6)) the sorting effect is now significant at 5% and gains some relevance.

So far, I assumed independent career choices, as required by the IIA assumption in standard logistic models. This is violated when individuals choose more than one alternative at the same time, which is the case when scholars are affiliated to more than one university. I develop a mixed logit model to test for correlated preferences. This version of logistic regression allows me to consider the presence of heterogeneous agents. It is similar to the standard model but more flexible: the coefficients are scholar-specific and the utility function includes an additional term which permits correlated choices and the relaxation of the IIA assumption (Ye et al., 2020; Train, 2009). I illustrate the results of the mixed logit estimation (columns (7) and (8) in Table 15); technical detail is in the Appendix.

With respect to the benchmark model, the magnitude varies for every coefficient. Signs and significance levels of *distance* coefficients confirm the gravity literature and magnitudes increase by almost 0.4. However, all the other results lose their significance and decrease in magnitude. The signs of *agglomeration* are all the opposite compared to the benchmark, but no coefficient is significant. *Selection* turns negative with a magnitude close to zero; the magnitude of *sorting* drastically decreases but remains positive. The LR test between columns (7) and (1) rejects the benchmark version, although there are six additional parameters estimated with the mixed

[33] NB: I call these scholars 'repeated movers', which means that they are associated with more than one university.

[34] NB: by 'single mover' I mean a scholar associated with only one university.

[35] To compute the mean of $\ln q$, I consider 29 observations for repeat movers and 877 observations for single movers.

logit (not reported in Table 15). Nevertheless, the mixed logit is weaker than the benchmark since it involves simulations and not a maximization. Moreover, the assumption on parameters' distribution is essential to obtain these results, which may change when considering another assumption. The original model remains the benchmark.

3.5 Gender analysis

In the dataset, women are about one-third of the sample (30.24%). I test for gender differences in the effects found in the benchmark model.

I estimate the same regression with the addition of an interaction term: each categorical variable interacts with a gender dummy (1 if male, 0 if female). As shown in Table 14 (in the Appendix) none of the interaction terms with the gender dummy ($\ln dxM$, $\ln P_k xM$, $\ln Y_k xM$, $\ln QxM$, $\ln q \ln dxM$, $\ln q \ln QxM$) are significant, revealing no evidence of gender differences. Negative *distance* coefficients show lower magnitudes when only women are considered. These coefficients increase in magnitude only when models include men, except in the last column where the men's distance coefficient is positive. Similarly, all three agglomeration coefficients reinforce if I analyse male scholars. Selection and sorting effects are lower when the male portion of the sample is involved. The coefficients for women are almost always in line with the benchmark model, which confirms that there are no significant gender differences.

3.6 Private/Public universities

As mentioned in the description of the sample (Section 2.2), four of the universities originally considered are private: Bocconi University, Catholic University, Free University of Bozen and LUISS University. Private universities have more hiring autonomy and discretion around remuneration (Trivellato et al., 2016), making them more attractive to better scholars. To understand how private institutions influence the benchmark estimation, I run a regression excluding them from the sample. The total of dyadic matches is now 7774, with 598 observations and 13 universities. Table 17 (in the Appendix) presents the results.

Distance coefficients confirm previous findings, each of them is negative, highly significant and with a magnitude a little larger than the benchmark, but still in line with the literature.

Agglomeration variables (columns (5) and (6)), represented by population size ($\ln P_k$) and university notability ($\ln Q$), maintain their signs and significance levels. The magnitude of notability is greater than $\ln P_k$, confirming agglomeration. The income coefficient ($\ln Y_k$) is not significant, although its sign is still positive. Its magnitude decreases significantly, likely because the private universities excluded are located in rich cities: two of the four are in Milan, the city with the highest disposable income. *Selection* ($\ln q \ln d$) has a positive sign, its significance level is at 1% and its magnitude is similar to the benchmark model. *Sorting* coefficients are positive and highly significant as well, their magnitude almost double. Table 17 confirms *positive selection* and *positive sorting* and brings evidence for a reinforcement of the latter effect. Hence, I can claim that all the hypotheses hold also when only public universities are included in the model.

I develop a nested logit model to investigate further. It enables more appropriate comparison of the two systems and permits the relaxation of the IIA assumption. This method still denies correlation of error terms between the two sectors (private and public), but now there is the possibility of error terms dependency within a nest (McFadden, 1978; Train, 2003). Hence, the IIA assumption holds within a nest, where the unobserved portions of utility still have the same mean, while the assumption does not hold between nests, where means of the error terms can now differ (Train, 2003; Heiss, 2002). With a nested logit model, it is possible to test whether one type of university implies systematically higher utility.

In general, the nested logit model permits grouping of alternatives in nests with similar characteristics, with a certain degree of correlation λ_s. I divide the university set K per status s: private and public. Thus, the utility function of scholar i is decomposed into two parts, plus a random component $\epsilon_{ik'}$. The first portion H_{is} depends only on the nest s, and the other portion $M_{ik'}$ depends on a specific alternative k' within nest s (Train, 2003). The new utility function is defined as follows:

$$U_{ik'} = H_{is} + M_{ik'} + \epsilon_{ik'} \tag{6}$$

for $k' \in K_s$.

Starting from this decomposed utility function, it is possible to describe the probability of choosing $k \in K_s$ as the product between two probabilities: the conditional probability of choosing k given that the choice of nest K_s has been made (i.e., a standard logit model between alternatives in nest K_s) and the marginal probability of choosing universities in nest K_s (i.e., a standard logit model between nests). The probability of the final choice k for scholar i is the product of two standard logit models:

$$p_{ik} = P_{ik|K_s} P_{iK_s} \tag{7}$$

where

$$P_{ik|K_s} = \frac{\exp(M_{ik'}/\lambda_s)}{\sum_{k' \in K_s} \exp(M_{ik'}/\lambda_s)} \tag{8}$$

with

$$I_{is} = \ln \sum_{k' \in K_s} \exp(M_{ik'}/\lambda_s) \tag{9}$$

and where

$$P_{iK_s} = \frac{\exp(H_{is} + \lambda_s I_{is})}{\sum_{l=1}^{S} \exp(H_{is} + \lambda_l I_{il})} \tag{10}$$

The quantity I_{is} is called *inclusive value* or *log-sum term*. It is essential for connecting information in the upper model (marginal probability) with information in the lower model (conditional probability) and it is defined by the logarithm of the lower model denominator - equation (8) (Train, 2003). λ_s is called *log-sum coefficient* or *dissimilarity parameter* and it reveals informations about the degree of error terms correlation: the higher the λ_s, the higher the independence (or the lower the correlation) of the unobserved portion of utility. The standard multinomial logit model requires λ_s be equal to 1, which implies complete independent error terms (i.e., zero correlation of error terms) (Train, 2003; Heiss, 2002). λ_s captures the substitutability of alternatives: if there is more substitution within than between nests, then λ_s is lower than one, while if substitution is greater between rather than within nests, then λ_s is greater that one (Train et al., 1987).

Once I_{is} multiplies λ_s, their product $\lambda_s I_{is}$ represents the extra expected utility of scholar i from choosing the best university in nest K_s. This extra expected utility is added to H_{is}, which defines the expected utility of choosing whatever alternative is in the nest. H_{is} depends on nest-specific variables, which are not present in my analysis. Hence, only the product $\lambda_s I_{is}$ tells the difference in the expected utility of choosing a private or a public university: the higher the $\lambda_s I_{is}$, the higher the gain for the scholar (Train, 2003). For a nested logit model to be globally consistent with Random Utility Models, the density function must be non-negative; this condition is always met for dissimilarity parameters within the unite interval (Borsch-Supan, 1990; Kling and Herriges, 1995). When λ_s are larger than one (Train et al., 1987), the consistency condition may still hold locally, i.e. for some value of the explanatory variables (Borsch-Supan, 1990; Kling and Herriges, 1995).

Table 7 presents the results of the simultaneous nested logit model.[36] All results still hold qualitatively when compared to the benchmark regression. One cannot compare magnitudes between the benchmark and the nested logit model directly, given the presence of the additional parameters λ_s, but it is possible to analyze meaningful ratios. With *selection* effect in column (6) of Table 7, when a scholar has a human capital index of 5, her distance costs decrease by more that 20% with respect to a scholar with a human capital indicator of 2. Once I compute this percentage using benchmark coefficients, I can claim that there is no relevant difference between the two models in terms of *selection*: the percentage of cost reduction is almost the same (−20,60% for the nested model,[37] −19,72% for the standard model.)[38] The *sorting* effect presents some difference, with more inequalities among scholars in the nested than in the standard specification. In column (6) of Table 7, when a scholar has a human capital index of 5, with the sorting effect her gains are 55,08%[39] higher than the gains of a scholar with an individual quality indicator of 2. On the other hand, in the benchmark model (Table 5), gains for better scholars are 43,22%[40] higher than for scholars with lower human capital.[41]

[36] A consistent nested logit model can be computed also sequentially, but this latter method is less efficient than the simultaneous approach currently employed (Heiss, 2002; Train, 2003).

[37] The cost for a scholar with a human capital index of 2 is: $-1.143 + 2{\cdot}0.069 = -1.005$
The cost for a scholar with a human capital index of 5 is: $-1.143 + 5{\cdot}0.069 = -0.798$
The cost reduction for a better scholar is: $(-0.798 + 1.005)/-1.005 = -0.2060 = \mathbf{-20{,}60\%}$

[38] The cost for a scholar with a human capital index of 2 is: $-0.723 + 2{\cdot}0.042 = -0.639$
The cost for a scholar with a human capital index of 5 is: $-0.723 + 5{\cdot}0.042 = -0.513$
The cost reduction for a better scholar is: $(-0.513 + 0.639)/-0.639 = -0.1972 = \mathbf{-19{,}72\%}$

[39] The gain for a scholar with a human capital index of 2 is: $0.193 + 2{\cdot}0{,}056 = 0.305$
The gain for a scholar with a human capital index of 5 is: $0.193 + 5{\cdot}0{,}056 = 0.473$
The increase of gains for a better scholar is: $(0.473 - 0.305)/0.305 = 0.5508 = \mathbf{55{,}08\%}$

[40] The gain for a scholar with a human capital index of 2 is: $0.168 + 2{\cdot}0{,}034 = 0.236$
The gain for a scholar with a human capital index of 5 is: $0.168 + 5{\cdot}0{,}034 = 0.338$
The gain increase for a better scholar is: $(0.338 - 0.236)/0.338 = 0.4322 = \mathbf{43{,}22\%}$

[41] Here, better scholars have a human capital index of 5, while scholars with a lower human capital index have an indicator of 2.

I also compare two opposite situations to compute the gain percentage variation: the gains of a better scholar (i.e., with a human capital index of 5) who teaches in a better university (i.e., with a notability index of 3) to the gains of a scholar with a lower individual quality (i.e., with a human capital index of 2) who teaches in a worse university (i.e., with a notability index of 1). In the nested logit model (Table 7), the gains for the better scholar are 365%[42] higher than those of her peer with a lower human capital indicator. When I compare these two opposite situations in the benchmark model, the gains for the better scholar who teaches in a better university are 330%[43] higher than for a scholar with a lower human capital index who teaches in a university with lower notability. In general terms, the *sorting* effect is stronger in the nested than the standard logit model.

This nesting procedure seems to be justified: the null hypothesis of no nests is rejected through a log-likelihood ratio test (LL = -1909.4, p-value = 0.000), and the correlation within nests is different from zero (Wald test = 21.636, p-value = 0.000), but the null hypothesis of unique nest elasticity cannot be rejected (Wald test = 0.3037, p-value = 0.5816; LL = -1894.8, p-value = 0.507). This raises questions about the applicability of the grouping strategy I employ here, although dividing private and public universities appears reasonable. To clearly define the pertinence of this nested logit model, it is necessary to look at the additional parameters in the last two rows of the output, the λ_s. Firstly, all dissimilarity parameters exceed the unity, which pose another question on the global consistency of this nested model with utility maximization. Following consistency tests proposed by Kling and Herriges (1995), I check the first and the second-order partial consistency conditions: they do not reject the null hypothesis of utility maximization compatibility. I present the technical details in the Appendix.

As the nested logit model is consistent with random utility maximization, I can define the expected gain each scholar obtains from choosing either a private or a public university. Given the lack of nest-specific variables, this utility is only given by the product $\lambda_s I_{is}$, which varies for every scholar. Among the 815 professors considered, 127 have greater expected utility (EU)

[42] The gain for a scholar with a human capital index of 2 in a university with notability of 1 is: 0.193 + 2·0,056 = 0.305
The gain for a scholar with a human capital index of 5 in a university with notability of 3 is: 0.193·3 + 5·3·0,056 = 1.419
The gain increase for a better scholar in a better university is: (1.419 − 0.305)/0.305 = 3.65 = **365%**

[43] The gain for a scholar with a human capital index of 2 in a university with notability of 1 is: 0.168 + 2·0,034 = 0.236
The gain for a scholar with a human capital index of 5 in a university with notability of 3 is: 0.168·3 + 5·3·0,034 = 1.014
The gain increase for a better scholar in a better university is: (1.014 − 0.236)/0.236 = 3.30 = **330%**

from teaching in public universities than in private ones, while 688 realize higher expected gains by affiliating to private institutions. I compare these two groups (Table 6) and the mean of the individual quality for those who prefer public universities is negative (-0.444), against a positive for those who prefer private institutions (0.289). *Sorting* effect is evident: better professors prefer more favourable environments. Private institutions, with more available resources, create better contexts to attract more relevant human capital.

I offer more insights into the influence of Bocconi and Catholic University in the Appendix (see Subsection 6.3).

Table 6: Descriptive statistics: groups of scholars preferring public or private universities

Variables	**Obs**	**Mean**	**St.Dv.**	**Min**	**Max**
EU Public > EU Private:					
ln of human capital	127	-0.4435	1.8589	-4.1965	3.4072
EU Private > EU Public:					
ln of human capital	688	0.2888	2.1916	-4.4680	7.6000

Table 7: Multinomial logit regressions: nested logit model - birthplaces analysis - threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.860^{***}	-0.900^{***}	-0.874^{***}	-0.923^{***}	-1.085^{***}	-1.143^{***}
	(0.170)	(0.178)	(0.172)	(0.182)	(0.144)	(0.155)
Selection:						
$\ln q \ln d$		0.053^{***}		0.055^{***}		0.069^{***}
		(0.019)		(0.019)		(0.022)
Sorting:						
$\ln q \ln Q$			0.041	0.046^{*}		0.056^{*}
			(0.026)	(0.028)		(0.033)
Agglomeration:						
$\ln P_k$					-0.167^{***}	-0.180^{***}
					(0.050)	(0.053)
$\ln Y_k$					4.093^{***}	4.413^{***}
					(0.827)	(0.893)
$\ln Q$					0.191^{***}	0.193^{**}
					(0.073)	(0.076)
$\lambda_{private}$	1.821^{***}	1.871^{***}	1.886^{***}	1.968^{***}	1.635^{***}	1.730^{***}
	(0.553)	(0.552)	(0.579)	(0.590)	(0.354)	(0.378)
λ_{public}	1.185^{***}	1.218^{***}	1.204^{***}	1.249^{***}	1.582^{***}	1.637^{***}
	(0.240)	(0.247)	(0.243)	(0.253)	(0.210)	(0.223)
k FE	YES	YES	YES	YES	NO	NO
Obs	815	815	815	815	815	815
R^2	0.159	0.161	0.159	0.162		
LL	$-1,864.157$	$-1,858.387$	$-1,862.610$	$-1,856.624$	$-1,902.012$	$-1,894.613$

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

4 Comparison between the present and the past

It is interesting to compare features of the contemporaneous academic market in Italy with those of the past. I run the same logistic regression as before but I use a sample of professors who worked in Italy from 1000 to 1800,[44] the whole considered by period de la Croix et al. (2020). Agglomeration variables are not fully comparable; I cannot test the first hypothesis with the updated dataset of de la Croix et al. (2020) because the authors consider the level of city democracy instead of the average disposable income of the households ($\ln Y_k$).

In Table 8, column (1) summarises the other findings using present professors (i.e., professors who currently work in Italy), while column (2) involves past scholars (i.e., professors who worked in Italy between 1000 and 1800).

In both cases, I confirm standard results for *distance* of gravity models: the greater the distance, the lower the probability a scholar chooses to travel that route. From Table 8, the difference in magnitude between these coefficients is evident but both are still in line with the literature, which provides greater magnitude for past periods than for current times. Furthermore, the distance in column (1) is the Euclidean distance, while in de la Croix et al. (2020) it is the cost distance. However, the Euclidean distance increases linearly with the cost distance, which limits the relevance of this computational difference. The magnitude of *selection* effects halves in current times with respect to the past, due to changes in individual quality measures - the human quality indexes are both the result of a *PCA* but they consider different bibliometric indicators.[45] In the Appendix, Table 13 compares the total effect of distance now and in the past for different levels of human capital: the effect is almost the same for top scholars in both columns, which confirms the comparability of the results. There is another important difference when I consider *sorting.* To compute the notability of the university, shown in the second column, de la Croix et al. (2020) aggregate the 5 highest human capital indexes associated with scholars active in that institution during the preceding 25 years (for more technical details see de la Croix and Stelter, 2021). In the first column, I link

[44] Professor David de la Croix provided this sample to this project

[45] Present indicators: RePEc score, Worldcat works and library holdings, Google Scholar citations, H-index and i10-index, WoS H-index, Scopus H-index.
Past indicators: number of characters of the longest Wikipedia page, number of Wikipedia pages in different languages, Worldcat works, library holdings, and publication languages.

the notability index to the RePEc score of each university as of 10 years ago (see section 2.4). Further developments of the project could implement a version of notability index similar to de la Croix and Stelter (2021), to avoid the disadvantages of RePEc indicators (i.e., institutions with many registered affiliated scholars are advantaged, Zimmermann, 2012) and exclude other possible endogeneity problems. Finally, the significance of sorting coefficients in Table 8 is weaker for current times than for the past, when the same effect had a high relevance.

Table 8: Multinomial logit regressions: standard logit model, birthplaces analysis - comparison of results from the present and from the past without agglomeration variables

	(1)	(2)
	PRESENT[1]	PAST[2]
Distance:		
$\ln d$	-0.723^{***}	-1.455^{***}
	(0.029)	(0.013)
Selection:		
$\ln q \ln d$	0.043^{***}	0.081^{***}
	(0.013)	(0.006)
Sorting:		
$\ln q \ln Q$	0.032^{*}	0.014^{***}
	(0.019)	(0.002)
k FE	YES	YES
Obs	815	12,003
R^2	0.161	0.400
LL	−1,865.936	−17,813.470

Note: *p<0.1; **p<0.05; ***p<0.01

[1] Present professors, currently working in Italy

[2] Past professors, working in Italy from 1000 to 1800

The time horizon shown in the second column of Table 8 is too broad to freely compare it with the shorter time-span of column (1). Instead I exploit the division in periods developed by de la Croix et al. (2020) to seize more directly possible changes and fluctuations of the cycle that occurred in past centuries. In de la Croix et al. (2020) there are eight time segments with a different number of observations available for each period. I follow this two-by-two partition and I group together: the 2nd and 3rd period (1348-1449 / 1450-1526), the 4th and 5th (1527-1617 / 1618-1685), and the 6th and 7th (1686-1733 / 1734-1800). I exclude the first two periods from 1000 to 1347, because there are too few observations, which leads to a negligible empirical relevance and less comparable results with respect to the other segments.

Table 9 shows the results. *Distance* is negative and highly significant in every specification. Its magnitude reflects the corresponding literature and historical period: it is higher in columns

(2) and (3) than in column (1). This shift corresponds to the rise of national states and the increase in barriers and customs duties, leading to higher transportation costs. These burdens decrease only in recent times with technological improvements and transport innovations. *Selection* effect is always present, with its positive sign and high relevance. Its magnitude drastically decreases in column (3) and lowers even more when the model involves current scholars.[46] *Sorting* appears the most fluctuating effect across the time horizon, as expected. It is positive and highly significant in the first time range, when most of the major universities are already established (i.e., Bologna, Rome (Sapienza), Florence (Studium generale)) and are among the European top five institutions (de la Croix et al., 2020). These features allow me to position the 1348 - 1526 Italian academic market in the upward part of the aforementioned fluctuating cycle. However, sorting totally disappears in the second period I consider. Its sign is negative in both columns (2) and (3), but it is slightly significant only in the third one. These results might be due to the characteristics of the Italian academic world: its decline starts after the sixteenth century, a time of strict censorship of revolutionary concepts by the Catholic Church (Blassuto and de la Croix, 2021). Notable scholars were strongly attracted to the high quality of the first universities, but the sorting effect was diluted with the flow of time and with other universities entering the academic market. This decline in the sorting effect locates the 1527 - 1800 Italian university system in the downward portion of the cycle. Sorting regains its positive sign only when the model considers current scholars. In column (4), positive sorting is slightly significant, which may signal a new momentum for current Italian universities. With the local recruitment of professors and the greater autonomy of each university, quality should gain attention and importance. However, the current analysis cannot detect these reforms with confidence; they are too recent and the influence of the previous seniority-based apparatus persists. This explains the weak sorting effect in the sample of contemporaneous Italian scholars. The same structural explanation applies to the low significance level of sorting in the past (column (2) and (3)): the strong control of the powers in charge (i.e., Catholic Church) limited the relevance of university quality while favouring more denominational sorting, which relies on membership and networks rather than on meritocracy (MacLeod and Urquiola, 2021).

[46] When comparing past periods with the present, both human capital indexes are the results of a PCA but they involve different bibliometrics indicators.

Table 9: Multinomial logit regressions: standard logit model, birthplaces analysis - comparison of results from the present and from the past without agglomeration variables

	(1) (1348-1526)	(2) (1527-1685)	(3) (1686-1800)	(4) (PRESENT)
Distance:				
$\ln d$	-1.360^{***}	-1.602^{***}	-1.624^{***}	-0.723^{***}
	(0.022)	(0.023)	(0.033)	(0.029)
Selection:				
$\ln q \ln d$	0.096^{***}	0.097^{***}	0.056^{***}	0.043^{***}
	(0.010)	(0.009)	(0.013)	(0.013)
Sorting:				
$\ln q \ln Q$	0.017^{***}	-0.0005	-0.011^{*}	0.032^{*}
	(0.022)	(0.004)	(0.006)	(0.019)
k FE	YES	YES	YES	YES
Obs.	4412	4527	2217	815
R^2		0.416	0.440	0.161
LL	$-5{,}839.514$	$-6{,}584.718$	$-3{,}246.440$	$-1{,}865.936$
Note:				*p<0.1; **p<0.05; ***p<0.01

5 Conclusions

Using a new sample of contemporaneous scholars, this research confirms and discloses important features of the Italian academic market. Gravity highlights a recurrent effect widely explained in migration literature and agglomeration forces of Italian universities are present as well. The high significance of university notability is of particular interest, confirming that the quality of institutions is a strong factor for attracting relevant human capital. Selection effect is also remarkably strong in the benchmark model, which implies that contemporaneous professors travel longer distances when they have greater human capital indexes. Sorting is weaker in this specification, but still significant and positive, which means that notability is more relevant for scholars with a higher individual quality index. Although it is less clear than the others, this last effect might direct the position of Italy in the fluctuating cycle. The difference in current and past sorting represents an important initial step for Italian academia: implementing reforms may enhance Italian universities' notability. Policies to improve the quality of high-education institutions would stimulate excellence and in turn, would increase the attractiveness of Italian universities. This would trigger a virtuous circle for the whole economy - improving the sorting effect will feed the system with more resources, attract more remarkable scholars and increase the likelihood of innovations and economic enhancements. The United States, which has the top universities and research centers, has reaped the benefits of these positive spillovers. Since the early 1900s, sorting has been much stronger in America than it has been in Europe, where centralized systems favored equal growth of high-education institutions while simultaneously preventing the most promising ones from completely exploiting their potential (MacLeod and Urquiola, 2021). The recent reforms in the Italian system might be seen as a watershed moment: the positive achievements reached in terms of equality under a centralized system (Baldissera and Cornali, 2020; Barone and Guetto, 2016) can be bolstered by growing investments in excellence.

Future research can relax the assumption that demand in academia is totally elastic. This would necessitate the use of alternative gravity models, which do not impose the same stringent constraints as the multinomial logit. Gravity models that allow the consideration of both sides of the market can achieve a more complex general equilibrium analysis. Finally, the notability measure can be improved when using the conventional multinomial logit model. This could be accomplished by creating an index similar to de la Croix et al. (2020) to mitigate (if not eliminate) endogeneity issues with RePEc indicators.

Bibliography

Agasisti, T., Barra, C., & Zotti, R. (2019). *Research, knowledge transfer, and innovation: The effect of Italian universities' efficiency on local economic development 2006/2012.* (Vol. 59). Journal of Regional Science.

Agasisti, T., & Bianco, A. D. (2007). *Determinants of college student migration in italy: Empirical evidence from a gravity approach.*

Akcigit, U., Baslandze, S., & Stantcheva, S. (2016). *Taxation and the international mobility of inventors.* (Vol. 106/10). American Economic Review.

Artige, L., Camacho, C., & de La Croix, D. (2004). *Wealth Breeds Decline: Reversals of Leadership and Consumption Habits.* (Vol. 9). Journal of Economic Growth.

Audretsch, D. (1998). *Agglomeration and the location of innovative activity.* (Vol. 14/2). Oxford Review of Economic Policy.

Baldissera, A., & Cornali, F. (2020). *Geography of human capital in italy: A comparison between macro-regions.* (Vol. 25). Cambridge University Press.

Barone, C., & Guetto, R. (2016). *Verso una meritocrazia dell'istruzione? Inerzia e mutamento nei legami tra origini sociali, opportunità di studio e destini lavorativi in Italia (1920–2009).* (Vol. 1).

Barra, C., & Zotti, R. (2017). *Investigating the human capital development–growth nexus does the efficiency of universities matter?* (Vol. 40/06). International Regional Science Review.

Barra, C., & Zotti, R. (2018). *The contribution of university, private and public sector resources to italian regional innovation system (in)efficiency.* The Journal of Technology Transfer.

Barrientos, P. (2007). *Analysis of international migration and its impacts on developing countries.* (Vol. 12). Institute for Advanced Development Studies, Development Research Working Paper.

Barro, R. (1991). *Economic growth in a cross section of countries.* (Vol. 106/2). The Quarterly Journal of Economics, Oxford University Press.

Barro, R. (2001). *Human capital and growth.* (Vol. 91/2). American Economic Review, American Economic Association.

Beine, M., Docquier, F., & Özden, Ç. (2011). *Diasporas.* (Vol. 95).

Bertola, G., & Sestino, P. (2011). *A Comparative Perspective on Italy's Human Capital Accumulation.* (Vol. 6). Economic History Working Papers, Bank of Italy.

Bini, M., & Chiandotto, B. (2003). *La valutazione del sistema universitario italiano alla luce della riforma dei cicli e degli ordinamenti didattici.* (Vol. 2). Studi e note di economia.

Blassuto, F., & de la Croix, D. (2021). Catholic censorship and the demise of knowledge production in early modern italy. *Available at SSRN:* https://ssrn.com/abstract=3928688

Borsch-Supan, A. (1990). On the compatibility of nested logit models with utility maximization. *Journal of Econometrics*, *43*(3), 373–388.

Bratti, M., & Verzillo, S. (2019). *The 'gravity' of quality: Research quality and the attractiveness of universities in italy.* [Regional Studies].

Capano, G. (2008). *Looking for serendipity: The problematical reform of government within italy's universities.* (Vol. 55).

Checchi, D., Fraja, G. D., & Verzillo, S. (2014a). *And the winners are... an axiomatic approach to selection from a set.* (Vol. 14/05) [Discussion Paper]. University of Nottingham, School of Economics.

Checchi, D., Fraja, G. D., & Verzillo, S. (2014b). *Publish or perish: An analysis of the academic job market in italy.* (Vol. 14/04) [Discussion Paper]. University of Nottingham, School of Economics.

Checchi, D., & Verzillo, S. (2014). *Selecting university professors in italy: Much ado about nothing?* Mimeo, Milano Italy.

Cohen, D., & Soto, M. (2007). *Growth and human capital: Good data, good results.* (Vol. 46/3). Journal of Economic Growth 12.

Cottini, E., Ghinetti, P., & Moriconi, S. (2019). *Higher education supply, neighbourhood effects and economic welfare.* [DISCE - Working Papers del Dipartimento di Economia e Finanza]. Università Cattolica del Sacro Cuore, Dipartimenti e Istituti di Scienze Economiche (DISCE).

Croissant, Y. (2020). *Multinomial logit models.*

Daly, A., & Zachary, S. (1978). Improved multiple choice models.

de la Croix, D., Docquier, F., Fabre, A., & R., S. (2020). *The academic market and the rise of universities in medieval and early modern europe.*

de la Croix, D., & Stelter, R. (2021). *Scholars and Literati at the University of Göttingen (1734–1800).* Repertorium Eruditorum Totius Europae - RETE.

Docquier, F., & Marfouk, A. (2006). *International migration by educational attainment (1990–2000).* [In: Ozden, C., Schiff, M. (Eds.), International Migration, Remittances and Development.]. Palgrave Macmillan, New York. Chapter 5.

Drucker, J., & Goldstein, H. (2007). *Assessing the regional economic development impacts of universities: A review of current approaches.* (Vol. 30). International Regional Science Review.

Durante, R., Labartino, G., & Perotti, R. (2011). *Academic dynasties: Decentralization and familism in the italian academia.* [NBER Working Paper No. 17572].

Faggian, A., & McCann, P. (2009). *Universities, agglomerations and graduate human capital mobility.* (Vol. 100/2). Journal of Economic; Social Geography.

Grogger, J., & Hanson, G. (2011). *Income maximization and the selection and sorting of international migrants.* (Vol. 95/1). Journal of Development Economics.

Grogger, J., & Hanson, H. (2015). *Attracting talent: Location choices of foreign-born phds in the united states.* (Vol. 35(S1)). Journal of Labor Economics, University of Chicago Press.

Handler, H. (2018). *Economic links between education and migration: An overview.* (Vol. 4) [Flash Paper]. Policy Crossover Center Vienna - Europe.

Hanushek, E. A., & Woessmann, L. (2008). *The role of cognitive skills in economic development* (Vol. 46/3). Journal of Economic Literature.

Heiss, F. (2002). Structural choice analysis with nested logit models. *Stata Journal*, *2*(3), 227–252.

Jacso, P. (2005). *As we may search–comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases.* (Vol. 89). Current Science.

Kerr, S., Kerr, W., Özden, Ç., & Parsons, C. (2016). *Global talent flows.* (Vol. 30/4). Journal of Economic Perspectives, American Economic Association.

Kerr, S., Kerr, W., Özden, Ç., & Parsons, C. (2017). *High-skilled migration and agglomeration.* (Vol. 9). Annual Review of Economics.

Kling, C. L., & Herriges, J. A. (1995). An Empirical Investigation of the Consistency of Nested Logit Models with Utility Maximization. *American Journal of Agricultural Economics*, *77*(4), 875–884.

Leon, G. (2013). *Transportation choices and the value of statistical life.* (Vol. 9).

Lucas, R. J. (1988). *On the mechanics of economic development.* (Vol. 22/1). Journal of Monetary Economics, Elsevier.

MacLeod, W. B., & Urquiola, M. (2021). *Why does the united states have the best research universities? incentives, resources, and virtuous circles.* (Vol. 35/1).

McFadden, D. (1974). *Condition logit analysis of quantitative choice behavior.* University of California at Berkeley. Chapter 4.

McFadden, D. (1978). *Modeling the choice of residential location.* Transportation Research Record.

McFadden, D. (1979). *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments* (Behavioral Travel Modelling). David Hensher and P. Stopher. London : Croom Helm.

Morana, M. (2020). *Focus "il personale docente e non docente nel sistema universitario italiano - a.a 2019/2020.* Elaborazioni su banche dati MIUR, DGCASIS – Ufficio VI Gestione patrimonio informativo e statistica.

Nelson, R., & Phelps, E. (1966). *Investment in humans, technological diffusion, and economic growth.* (Vol. 56(1/2)). The American Economic Review.

Neuhaus, C., & Hans-Dieter, D. (2008). *Data sources for performing citation analysis: An overview.* (Vol. 64).

Norris, M., & Oppenheim, C. (2007). *Comparing alternatives to the web of science for coverage of the social sciences' literature.* Infometrics.

Ortega, F., & Peri, G. (2013). *The effect of income and immigration policies on international migration.* (Vol. 1). Migration Studies.

Perotti, R. (2008). *L'università truccata.* Einaudi. Torino, Italy.

Ramos, R. (2016). *Gravity models: A tool for migration analysis.* Institute of Labor Economics (IZA).

Rebora, G., & Turri, M. (2008). *La governance del sistema universitario in italia: 1989-2008.* (Vol. 32) [The World of Economics]. Liuc Papers Serie Economia Aziendale.

Rossi, P. (2016). *Stato giuridico, reclutamento ed evoluzione della docenza universitaria (1975–2015)* (Vol. 4/1). A Journal on Research Policy; Evaluation.

Schiller, D., & Cordes, A. (2016). *Measuring Researcher Mobility – A Comparison of Different Datasets and Methods with an Empirical Application of Micro-Data for the United States and Germany.* In OECD Blue Sky Forum.

Schwartz, A. (1976). *Migration, age and education.* (Vol. 24). Journal of Political Economy.

Seiler, C., & Wohlrabe, K. (2011). *Ranking economists and economic institutions using repec: Some remarks.*

Stephan, P., & Levin, S. (2001). *Exceptional contributions to us science by the foreign-born and foreign-educated.* (Vol. 20(1/2)). Population Research; Policy Review.

Train, K. (2003). *Discrete choice method with simulation.* Cambridge University press.

Train, K. (2009). *Discrete choice methods with simulation.* (2nd ed.). Cambridge University Press.

Train, K., McFadden, D., & Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *RAND Journal of Economics*, *18*(1), 109–123.

Trivellato, P., Triventi, M., & Traini, C. (2016). *Private higher education in italy.*

Triventi, M., & Trivellato, P. (2008). *Le onde lunghe dell'università italiana. partecipazione e risultati accademici degli studenti nel novecento.* (Vol. 22). Polis (Italy).

Williams, R. (2019). *Panel data 3: Conditional logit/ fixed effects logit models.* University of Notre Dame.

Yang, K., & Meho, L. (2006). *Citation analysis: A comparison of google scholar, scopus, and web of science.* (Vol. 43). Proc. Am. Soc. Info. Sci. Tech.

Ye, M., Chen, Y., Yang, G., Wang, B., & Hu, Q. (2020). *Mixed logit models for travelers' mode shifting considering bike-sharing.* (Vol. 12).

Zhao, X., Aref, S., Zagheni, E., & Stecklov, G. (2021). *International migration in academia and citation performance: An analysis of german-affiliated researchers by gender and discipline using scopus publications 1996-2020.*

Zimmermann, C. (2012). *Academic rankings with repec.* [Working Papers 2012-023]. Federal Reserve Bank of St. Louis.

6 Appendix

Table 10: Multinomial logit regressions: standard logit model - lowest level of education analysis - threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.679^{***}	-0.680^{***}	-0.679^{***}	-0.680^{***}	-0.680^{***}	-0.681^{***}
	(0.022)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)
Selection:						
$\ln q \ln d$		0.011		0.012		0.011
		(0.010)		(0.010)		(0.010)
Sorting:						
$\ln q \ln Q$			0.029	0.030		0.029
			(0.019)	(0.019)		(0.019)
Agglomeration:						
$\ln P_k$					-0.094^{***}	-0.096^{***}
					(0.028)	(0.029)
$\ln Y_k$					0.635^{*}	0.685^{*}
					(0.374)	(0.375)
$\ln Q$					0.137^{***}	0.141^{***}
					(0.043)	(0.043)
k FE	YES	YES	YES	YES	NO	NO
Obs	904	904	904	904	904	904
R^2	0.224	0.224	0.225	0.225		
LL	−1,911.623	−1,910.982	−1,910.463	−1,909.688	−1,947.317	−1,945.631

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Ordinary Least Square: age regression on human capital

	$\ln q$
AGE	0.485^{***}
	(0.015)
AGE^2	-0.004^{***}
	(0.0001)
Constant	-13.460^{***}
	(0.382)
Observations	13,632
R^2	0.119
Adjusted R^2	0.119
Residual Std. Error	2.026 (df = 13359)
F Statistic	900.433^{***} (df = 2; 13359)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Multinomial logit regressions: standard logit model, birthplaces age-adjusted analysis - threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.703^{***}	-0.703^{***}	-0.703^{***}	-0.703^{***}	-0.704^{***}	-0.704^{***}
	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)
Selection:						
$\ln q \ln d$		0.002		0.002		0.0004
		(0.015)		(0.015)		(0.015)
Sorting:						
$\ln q \ln Q$			0.007	0.007		0.004
			(0.021)	(0.021)		(0.021)
Agglomeration:						
$\ln P_k$					-0.117^{***}	-0.117^{***}
					(0.029)	(0.029)
$\ln Y_k$					2.093^{*}	2.093^{*}
					(0.360)	(0.360)
$\ln Q$					0.174^{***}	0.174^{***}
					(0.045)	(0.044)
k FE	YES	YES	YES	YES	NO	NO
Obs	786	786	786	786	786	786
R^2	0.154	0.154	0.154	0.154		
LL	−1,805.994	−1,805.983	−1,805.945	−1,805.933	−1,853.601	−1,853.584

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Effect of distance in the present and in the past without agglomeration variables

	(1) PRESENT[1]	(2) PAST[2]
$q_{min}{}^{*}$	−0.916	−1.455
$q_{75}{}^{**}$	−0.653	−1.426
$q_{max}{}^{***}$	−0.395	−0.428
Obs	815	12,003

[1] Present professors, currently working in Italy

[2] Past professors, working in Italy from 1000 to 1800

*Minimum q: −4.468 for present professors, 0 for past professors.

$^{**}75^{th}$ quantile of q: 1.611 for present professors, 0.359 for past professors.

***Maximum q: 7.600 for present professors, 12.604 for past professors.

Table 14: Multinomial logit regressions: standard logit model, birthplaces analysis - gender differences threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.709^{***}	-0.723^{***}	-0.709^{***}	-0.723^{***}	-0.709^{***}	-0.723^{***}
	(0.029)	(0.029)	(0.029)	(0.029)	(0.028)	(0.029)
$\ln dxM$	-0.019	-0.006	-0.019	-0.006	-0.008	0.004
	(0.039)	(0.041)	(0.039)	(0.041)	(0.040)	(0.041)
Selection:						
$\ln q \ln d$		0.043^{***}		0.043^{***}		0.042^{***}
		(0.013)		(0.013)		(0.013)
$\ln q \ln dxM$		-0.018		-0.018		-0.020
		(0.017)		(0.017)		(0.017)
Sorting:						
$\ln q \ln Q$			0.031	0.032^{*}		0.034^{*}
			(0.019)	(0.019)		(0.020)
$\ln q \ln QxM$			-0.005	-0.006		-0.010
			(0.023)	(0.023)		(0.023)
Agglomeration:						
$\ln P_k$					-0.122^{***}	-0.127^{***}
					(0.029)	(0.029)
$\ln P_k xM$					-0.034	-0.032
					(0.030)	(0.030)
$\ln Y_k$					2.085^{***}	2.188^{***}
					(0.355)	(0.357)
$\ln Y_k xM$					0.123	0.075
					(0.384)	(0.386)
$\ln Q$					0.169^{***}	0.166^{***}
					(0.044)	(0.044)
$\ln QxM$					0.041	0.046
					(0.054)	(0.055)
k FE	YES	YES	YES	YES	NO	NO
Obs.	815	815	815	815	815	815
R^2	0.157	0.160	0.158	0.161		
LL	$-1{,}867.089$	$-1{,}860.883$	$-1{,}865.803$	$-1{,}859.450$	$-1{,}915.353$	$-1{,}907.843$

Note: *p<0.1; **p<0.05; ***p<0.01

"xM" represents the relative effect when only Male scholars are considered.

Table 15: Repeat Movers' robustness checks and Mixed logit - birthplace analysis threshold at 20

	Benchmark		Removing RM		RM linked to 1uni.		Mixed Logit	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Distance:								
$\ln d$	-0.723^{***}	-0.723^{***}	-0.736^{***}	-0.736^{***}	-0.733^{***}	-0.733^{***}	-1.121^{***}	-0.956^{***}
	(0.029)	(0.029)	(0.030)	(0.030)	(0.030)	(0.030)	(0.082)	(0.062)
Agglomeration:								
$\ln P_k$		-0.127^{***}		-0.129^{***}		-0.135^{***}		0.021
		(0.029)		(0.029)		(0.029)		(0.029)
$\ln Y_k$		2.179^{***}		2.094^{***}		2.143^{***}		-0.600
		(0.356)		(0.371)		(0.365)		(0.397)
$\ln Q$		0.168^{***}		0.150^{***}		0.169^{***}		-0.067
		(0.044)		(0.046)		(0.045)		(0.054)
Selection:								
$\ln q \ln d$	0.043^{***}	0.042^{***}	0.051^{***}	0.050^{***}	0.050^{***}	0.049^{***}	-0.007	-0.008
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.018)	(0.017)
Sorting:								
$\ln q \ln Q$	0.033^{*}	0.034^{*}	0.042^{**}	0.044^{**}	0.038^{*}	0.041^{**}	0.019	0.020
	(0.019)	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.023)	(0.022)
k FE	YES	NO	YES	NO	YES	NO	YES	NO
Obs	815	815	763	763	789	789	815	815
R^2	0.161		0.169		0.166		0.172	
LL	$-1{,}860.156$	$-1{,}909.365$	$-1{,}724.414$	$-1{,}774.466$	$-1{,}786.364$	$-1{,}838.901$	$-1{,}834.740$	$-1{,}915.687$
Note:							*p<0.1; **p<0.05; ***p<0.01	

RM = Repeat Movers

Note: The mixed logit involves the six s.d. associated to each coefficient, only the s.d. linked to $\ln d$ is significantly different from zero.

6.1 Mixed Logit Model - technical details

The mixed logit model accounts for the heterogeneity of individuals and permits the relaxation of the IIA assumption. Specifically, coefficients are not fixed anymore but they vary across the population, and an additional term in the utility function allows for correlated choices (Train, 2009; Ye et al., 2020). Hence, the mixed logit model modifies scholar's utility function as follows:

$$U_{ik} = \beta_i x_{ik} + \eta_i x_{ik} + \epsilon_{ik} \tag{11}$$

Where the first term is the general deterministic component, which represents the utility of scholar i who chooses university k. The other two terms capture the unobservable part of the function: η_i is an individual deviation and ϵ_{ik} is a random term as before. I assume these two error terms to be normally distributed.

The mixed logit model, with the η term violating the IIA assumption, requires the integration of the conditional probability by using the joint probability density function, $f(\beta_i|\theta)$; where θ summarises the first and the second moment of the distribution. The vector of β coefficients is assumed to be independent and normally distributed and it is of length N. To obtain the unconditional probability of professor i choosing university k, the following formula applies:

$$P_{ik} = E(P_{ik}|\beta_i) = \int_{\beta} (P_{ik}|\beta_i) f(\beta_i|\theta) d\beta = \int_{\beta_1} \int_{\beta_2} \dots \int_{\beta_N} (P_{ik}|\beta_i) f(\beta_i|\theta) d\beta_1 d\beta_2 \dots d\beta_N \tag{12}$$

where

$$(P_{ik}|\beta_i) = \frac{\exp(\beta_i x_{ik})}{\sum_{k' \in K} \exp(\beta_i x_{ik'})} \tag{13}$$

is the conditional probability.

Simulations are used to draw the parameters from the β distribution: the unconditional probability is the average of the conditional probabilities computed for each scholar (Train, 2009; Ye et al., 2020; Leon, 2013). Table 16 presents all the specifications of the mixed logit regression, columns (4) and (6) are those reported in Table 15.

Table 16: Multinomial logit regressions: mixed logit model - birthplaces analysis - threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance						
$\ln d$	-0.766^{***}	-1.123^{***}	-1.120^{***}	-1.121^{***}	-0.960^{***}	-0.956^{***}
	(0.047)	(0.082)	(0.082)	(0.082)	(0.062)	(0.062)
Selection						
$\ln q \ln d$		-0.008		-0.007		-0.008
		(0.017)		(0.018)		(0.017)
Sorting						
$\ln q \ln Q$			0.020	0.019		0.020
			(0.022)	(0.023)		(0.022)
Agglomeration						
$\ln P_k$					0.021	0.021
					(0.029)	(0.029)
$\ln Y_k$					-0.600	-0.600
					(0.398)	(0.397)
$\ln Q$					0.060	-0.067
					(0.053)	(0.054)
k FE	YES	YES	YES	YES	NO	NO
Obs	815	815	815	815	815	815
R^2	0.165	0.172	0.172	0.172		
LL	$-1{,}850.713$	$-1{,}835.489$	$-1{,}835.296$	$-1{,}834.740$	$-1{,}916.449$	$-1{,}915.687$

Note: *p<0.1; **p<0.05; ***p<0.01

Note: The mixed logit involves the six s.d. associated to each coefficient only the s.d. linked to $\ln d$ and to $\ln Y_k$ are significantly different from zero.

Table 17: Multinomial logit regressions: standard logit model, birthplaces analysis - private universities excluded - threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.735^{***}	-0.745^{***}	-0.735^{***}	-0.744^{***}	-0.732^{***}	-0.739^{***}
	(0.032)	(0.033)	(0.032)	(0.033)	(0.031)	(0.032)
Selection:						
$\ln q \ln d$		0.045^{***}		0.044^{***}		0.043^{***}
		(0.015)		(0.015)		(0.015)
Sorting:						
$\ln q \ln Q$			0.070^{***}	0.069^{***}		0.072^{***}
			(0.022)	(0.022)		(0.022)
Agglomeration:						
$\ln P_k$					-0.082^{***}	-0.092^{***}
					(0.031)	(0.031)
$\ln Y_k$					0.216	0.352
					(0.463)	(0.464)
$\ln Q$					0.209^{***}	0.216^{***}
					(0.046)	(0.047)
k FE	YES	YES	YES	YES	NO	NO
Obs.	598	598	598	598	598	598
R^2	0.208	0.211	0.211	0.214		
LL	$-1{,}179.174$	$-1{,}174.645$	$-1{,}174.030$	$-1{,}169.584$	$-1{,}199.210$	$-1{,}189.387$

Note: *p<0.1; **p<0.05; ***p<0.01

6.2 Nested Logit Model - Consistency Test

Nested logit models with dissimilarity parameters greater than one require additional tests to check for the consistency with utility maximization. Indeed, Daly and Zachary (1978) and McFadden (1979) show that obtaining λ_s inside the unit interval is essential for the model to be globally consistent. Nevertheless, when this consistency is relaxed to hold only locally, the dissimilarity parameters can exceed one; as showed by Kling and Herriges (1995). Specifically, two conditions must be checked. (i) The non-negativity of the first-order partial derivatives of the choice probabilities is the first necessary condition, and it is described as follows:

$$\lambda_s \leq U_{1s}(v) \equiv \frac{1}{1 - Q_s(v)} \quad s = 1, S \tag{14}$$

with v_k defined as the utility delivered by each alternative and $v \equiv (v_1, \ldots, v_K)$, and where $Q_s(v)$ is the upper model as in equation 10:[47]

$$Q_s(v) = P_{iK_s} = \frac{\exp(\lambda_s I_{is})}{\sum_{l=1}^{S} \exp(\lambda_l I_{il})} \tag{15}$$

For this first necessary condition to hold, Q_s must be sufficiently large. (ii) The second condition questions the non-positivity constraint on the mixed second-order of the choice probabilities, as follows:

$$\lambda_s \leq \mathrm{U}_{2s}(v) \equiv \quad \frac{4}{3\left[1 - Q_s(v)\right] + \sqrt{\left[1 + 7Q_s(v)\right]\left[1 - Q_s(v)\right]}} \tag{16}$$

To define these conditions, I compute Q_s from equation 15 and compare it with the λ_s, which are already in the output (Table 7). Kling and Herriges (1995) present different approaches to precisely test the consistency of nested models, I follow them and Table 18 shows my results.

One possible approach confronts $\hat{\lambda}_s$ with both $\hat{U}_{1s}(\bar{v})$ and $\hat{U}_{2s}(\bar{v})$, where $\bar{v}$ denotes the mean of the indirect utility function. Already this approach seems to highlight the consistency of my nested model, by finding that $\hat{\lambda}_s \leq \hat{U}_{1s}(\bar{v})$ and $\hat{\lambda}_s \leq \hat{U}_{2s}(\bar{v})$. Notwithstanding, another approach investigates at which level of precision the estimated coefficients (the λ_s) are able to reject - or not - the local consistency. Hence, I develop a one-tailed test for each condition. For

[47] The term H_{is} is not reported, because of the lack of nest-specific variables.

Table 18: Consistency tests of NLM with RUM

First-Order Conditions			
Nest	$\hat{\lambda}_s$	$\hat{U}_{1s}(\bar{v})$	**t-ratio**
Private	1.7300	18.4201	−44.1279
Public	1.6368	2.1878	−2.4687
Second-Order Conditions			
Nest	$\hat{\lambda}_s$	$\hat{U}_{2s}(\bar{v})$	**t-ratio**
Private	1.7300	3.8092	−5.49714
Public	1.6368	1.3067	1.4798

the first-order condition, I test the null hypothesis $H_{1O} : \lambda_s \leq U_{1s}(\bar{v})$ against the alternative $H_{1A} : \lambda_s > U_{1s}(\bar{v})$, and for the second-order condition I compare the null hypothesis $H_{1O} : \lambda_s \leq U_{2s}(\bar{v})$ to the alternative $H_{1A} : \lambda_s > U_{2s}(\bar{v})$. The last column of Table 18 reports the t-ratios of each test statistic ($t_{1,2} \equiv [\hat{\lambda}_s - \hat{U}_{1,2s}(\bar{v})]/SE$); negative coefficients immediately imply that the null hypothesis of local consistency cannot be rejected, which is almost always the case. Only the second-order consistency condition for public universities is rejected, as it was the case of the previous approach. Nevertheless, grouping private and public universities seems appropriate and the first-order consistency condition largely approves this nesting procedure. I consider these results consistent enough with utility maximisation models.

6.3 Private/Public universities - further insights

In addition to the regression in the main text, I develop other two estimations: one excludes only Catholic University while keeping Bocconi University, and another excludes Bocconi University while keeping Catholic University. With the former, I investigate the issue of secondary locations to understand whether it alters previous results. With the latter, I examine the position of excellence recently reached by Bocconi in several rankings - it has lately become the best university in Italy for economics and related fields and this may be due to its well-known ability to attract high-ranked personalities.

Table 19 presents the results obtained by excluding Catholic University: almost all coefficients are significant. *Distance* and *agglomeration* remain as in the benchmark (Table 5) for what concern signs and significance, the magnitude is similar as well, the income coefficient experiences the largest variation: a drop of 0.198 (from 2.179 to 1.981). *Positive selection* maintains its significance level of 1% in each specification, with a slight decrease in magnitude. *Positive sorting* is again weaker than selection but it gains significance in the third column when it is considered alone (with respect to zero significance level of the coefficient in the benchmark). This implies that there is no relevant bias due to the imprecise geographical coordinates associated with this university. Although I assume that every scholar teaches only in Milan, the benchmark model is not significantly influenced.

Excluding Bocconi from the set of choices allows me to find only significant coefficients which improves the solidity of the results (Table 20). *Distance* and *agglomeration* variables remain with the same sign and significance as in the benchmark model. The magnitude of the notability coefficient increases, while income's magnitude almost halves. I find *positive selection* as in the previous estimation without Catholic University. In this regression, there is strong evidence for *positive sorting* in each specification. This model reaches a high significance level, which means that private universities have different features not totally captured by the variables included in the benchmark: further investigations are necessary (i.e., applying different models, like the nested logit model - see Subsection 3.4). Despite this, I can claim that all the expected features of the contemporaneous academic world described in Section 3.1 still hold if I exclude Bocconi University.

Table 19: Multinomial logit regressions: standard logit model, birthplaces analysis - Catholic University excluded

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.705^{***}	-0.718^{***}	-0.705^{***}	-0.718^{***}	-0.708^{***}	-0.720^{***}
	(0.029)	(0.030)	(0.029)	(0.030)	(0.029)	(0.029)
Selection:						
$\ln q \ln d$		0.039^{***}		0.039^{***}		0.038^{***}
		(0.013)		(0.013)		(0.013)
Sorting:						
$\ln q \ln Q$			0.034^{*}	0.035^{*}		0.036^{*}
			(0.020)	(0.019)		(0.020)
Agglomeration:						
$\ln P_k$					-0.114^{***}	-0.119^{***}
					(0.029)	(0.029)
$\ln Y_k$					1.872^{***}	1.981^{***}
					(0.383)	(0.384)
$\ln Q$					0.171^{***}	0.166^{***}
					(0.044)	(0.045)
k FE	YES	YES	YES	YES	NO	NO
Obs	741	741	741	741	741	741
R^2	0.169	0.171	0.170	0.172		
LL	−1,635.306	−1,630.858	−1,633.796	−1,629.271	−1,683.633	−1,677.802
Note:				*p<0.1; **p<0.05; ***p<0.01		

Table 20: Multinomial logit regressions: standard logit model, birthplaces analysis - Bocconi University excluded, threshold at 20

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.708^{***}	-0.718^{***}	-0.709^{***}	-0.718^{***}	-0.710^{***}	-0.719^{***}
	(0.029)	(0.030)	(0.029)	(0.030)	(0.029)	(0.029)
Selection:						
$\ln q \ln d$		0.045^{***}		0.046^{***}		0.045^{***}
		(0.014)		(0.014)		(0.014)
Sorting:						
$\ln q \ln Q$			0.062^{***}	0.063^{***}		0.068^{***}
			(0.019)	(0.019)		(0.019)
Agglomeration:						
$\ln P_k$					-0.120^{***}	-0.128^{***}
					(0.029)	(0.029)
$\ln Y_k$					1.154^{***}	1.276^{***}
					(0.397)	(0.399)
$\ln Q$					0.230^{***}	0.234^{***}
					(0.046)	(0.046)
k FE	YES	YES	YES	YES	NO	NO
Obs	714	714	714	714	714	714
R^2	0.174	0.176	0.176	0.179		
LL	−1,578.996	−1,573.514	−1,574.347	−1,568.781	−1,612.901	−1,602.348
Note:				*p<0.1; **p<0.05; ***p<0.01		

6.4 Main regressions with the threshold at 5 scholars

I develop the two main regressions of the study with varied thresholds in order to further investigate the peculiarities of the current Italian academic market. Hence, rather than limiting myself to universities with more than 20 scholars, I now present the results considering institutions with more than 5 scholars. With this new threshold, I involve in the analysis also smaller universities - Genoa, Catania, Naples, and Palermo. On the one hand, this expands the scope of the research by allowing more universities in the southern part of Italy to participate (i.e., Catania, Naples, and Palermo). On the other hand, these small universities reduce the balance of the choice set, because they appear within the career possibilities of each scholar but they are rarely chosen.

Table 21 and Table 22 show the results of the birthplaces analysis and the lower level of education analysis, respectively. The latter is similar to the estimation in the main text, with *distance* and *agglomeration* maintaining their high significance levels and without evidence for selection and sorting effects. I find the main difference of the new threshold in the birthplace analysis. In Table 21 *distance*, *selection* and *agglomeration* correspond to those of the project, while *sorting* lose almost all the significance. It is slightly significant in column (3) when it is considered alone, but it is not significant when the model includes selection and/or agglomeration. This result is in favour of a more balanced choice set: the results gain clearness and precision when the threshold increased to 20 scholars. Keeping smaller universities in the dataset detracts attention from the real scholars' choices, although the inclusion of Southern universities would improve the reach of the project.

Table 21: Multinomial logit regressions: standard logit model - birthplaces analysis, threshold at 5

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.777^{***}	-0.793^{***}	-0.777^{***}	-0.791^{***}	-0.774^{***}	-0.787^{***}
	(0.026)	(0.027)	(0.026)	(0.027)	(0.025)	(0.026)
Selection:						
$\ln q \ln d$		0.046^{***}		0.045^{***}		0.044^{***}
		(0.012)		(0.012)		(0.012)
Sorting:						
$\ln q \ln Q$			0.032^{*}	0.028		0.029
			(0.018)	(0.018)		(0.018)
Agglomeration:						
$\ln P_k$					-0.155^{***}	-0.159^{***}
					(0.026)	(0.026)
$\ln Y_k$					2.535^{***}	2.619^{***}
					(0.310)	(0.311)
$\ln Q$					0.208^{***}	0.204^{***}
					(0.040)	(0.041)
k FE	YES	YES	YES	YES	NO	NO
Obs	876	876	876	876	876	876
R^2	0.200	0.203	0.200	0.203		
LL	$-2{,}017.696$	$-2{,}009.554$	$-2{,}016.089$	$-2{,}008.352$	$-2{,}070.630$	$-2{,}061.524$

Note: *p<0.1; **p<0.05; ***p<0.01

Table 22: Multinomial logit regressions: standard logit model - lowest level of education analysis - threshold at 5

	(1)	(2)	(3)	(4)	(5)	(6)
Distance:						
$\ln d$	-0.726^{***}	-0.727^{***}	-0.726^{***}	-0.726^{***}	-0.723^{***}	-0.723^{***}
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Selection:						
$\ln q \ln d$		0.014		0.015^{*}		0.015^{*}
		(0.009)		(0.009)		(0.009)
Sorting:						
$\ln q \ln Q$			0.026	0.027		0.026
			(0.018)	(0.017)		(0.017)
Agglomeration:						
$\ln P_k$					-0.127^{***}	-0.129^{***}
					(0.026)	(0.026)
$\ln Y_k$					1.086^{***}	1.135^{***}
					(0.317)	(0.318)
$\ln Q$					0.178^{***}	0.180^{***}
					(0.040)	(0.040)
k FE	YES	YES	YES	YES	NO	NO
Obs	969	969	969	969	969	969
R^2	0.259	0.259	0.259	0.260		
LL	$-2{,}069.106$	$-2{,}067.837$	$-2{,}068.037$	$-2{,}066.652$	$-2{,}108.792$	$-2{,}106.510$

Note: *p<0.1; **p<0.05; ***p<0.01

INSTITUT DE RECHERCHE
ÉCONOMIQUES ET SOCIALES

Place Montesquieu 3
1348 Louvain-la-Neuve

ISSN 1379-244X D/2022/3082/01

UCLouvain

LIDAM
IRES | Louvain Institute of Data Analysis and Modeling in economics and statistics