

The Geography of Knowledge in Europe: 1000-1800

Chiara Zanardello

▶ **To cite this version:**

Chiara Zanardello. The Geography of Knowledge in Europe: 1000-1800. 2026. ⟨hal-05850532⟩

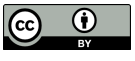
HAL Id: hal-05850532

https://hal.science/hal-05850532v1

Preprint submitted on 3 Nov 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

The Geography of Knowledge in Europe: 1000–1800

Chiara Zanardello*

October 28, 2026

Abstract

Between 1000 and 1800, the set of European scholars' birthplaces expanded geographically, even as their distribution became more concentrated among the leading birthplaces. At the same time, scholars' workplaces remained much more concentrated in a smaller number of cities, which is consistent with substantial lifetime mobility into academic institutions. Unlike the distribution of scholars' origins, the urban rank–size distribution remained relatively stable. I also show a strong correlation between the number of academic institutions and a higher concentration of the scholars' birthplace distribution, while no comparable relationship appears for urban population.

JEL classifications: N93, R12, I23

Keywords: scholars' birthplaces; human capital; rank–size distribution; academic institutions; urban population.

1 Introduction

Before the Industrial Revolution, the production of formal knowledge in Europe was concentrated in universities, academies, and other institutions located in a relatively small set of cities. But the concentration of scholarly activity does not tell us where scholars came from. Major intellectual centers could have drawn scholars from the same small set of established locations, or they could have concentrated scholarly activity while drawing talent from a much broader geography. This distinction has potential implications for the allocation of talent across space, particularly for questions of efficiency and lifetime mobility (Jaravel 2023). Modern evidence shows that geographic origins and exposure to knowledge-intensive environments matter for who becomes an inventor or scientist (Bell et al. 2019; Agarwal and Gaule 2020). More broadly, barriers to the allocation of talent can have important aggregate consequences (Hsieh et al. 2019). This question matters especially during the Scientific Revolution and the Enlightenment, when the link between knowledge and growth arguably became tighter (Mokyr 2005).

*Toulouse School of Economics, Université Toulouse Capitole, Toulouse, France.
Email address: chiara.zanardello@iast.fr

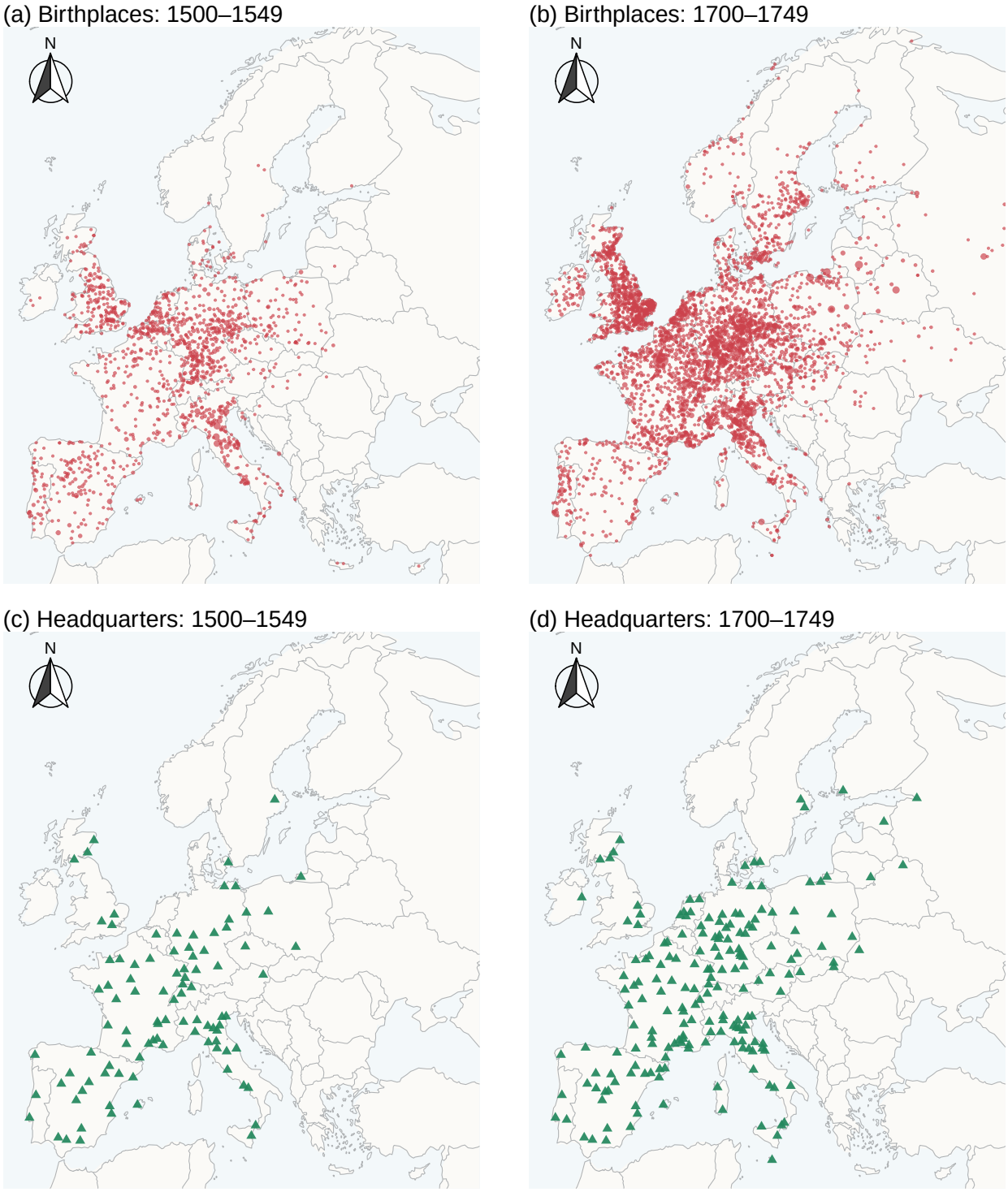


Figure 1: Geography of scholars' birthplaces and academic institutions in Europe, 1500–1749
Notes: Panels (a) and (b) show the locations where scholars were born in 1500–1549 and 1700–1749, respectively. Red circles indicate birthplaces, with their area proportional to the number of scholars born in each location during the corresponding half-century. Panels (c) and (d) show the headquarters of universities and academies active during the corresponding periods. Green triangles indicate locations hosting at least one active institution, independently of the number of institutions located there. The maps use present-day national borders.

Figure 1 makes this distinction visible. Between 1500 and 1700, both the geography of scholars' birthplaces (panels (a)–(b)) and the geography of academic institutions (panels (c)–(d)) expanded. However, there is also no mechanical prediction for how more academic institutions should affect the distribution of birthplaces. On the one hand, the expansion of universities and academies across different locations may create several regional points of attraction and lead to a flatter distribution of scholars' origins. On the other hand, having more academic institutions may disproportionately strengthen established scholarly regions and make the upper part of the birthplace distribution more important. This puzzle motivates my research question: as European knowledge became increasingly institutionalized within universities and academies, how did the geography of scholars' origins change, and did it simply mirror the geography of cities and institutions?

Hence, this paper documents how the geography of European scholars' origins evolved between 1000 and 1800 and how it differed from the geography of workplaces and urban population. I combine long-run urban population data from Buringh (2021) with individual-level data on European scholars from the UTHC-RETE project (de la Croix, Scebba, and Zanardello 2026; de la Croix 2021, 2026). The latter database tracks almost 90,000 individuals active in around 350 European academic institutions over eight centuries. I compare these origins with scholars' workplaces, which capture the geography of active scholarly life, and with urban population, which provides a benchmark for the broader distribution of European cities. I aggregate these data to the location–half-century level.

My paper also separates two margins that are often conflated. The number of scholars' birthplaces captures the geographic reach of scholarly origins, whereas the distribution of scholars across those places captures how concentrated those origins were. A broader *geographic reach* can therefore coexist with greater *concentration* among the leading birthplace locations. This paper documents both margins and compares them with the geography of scholars' workplaces and European urban population, and examines the relationship between the concentration of scholars' origins and the number of academic institutions.

To do this, I summarize each spatial distribution using a rank–size relationship. Locations are ranked by size, and the logarithm of rank is regressed on the logarithm of size. The estimated slope is the elasticity of rank with respect to size; equivalently, its inverse captures how quickly size declines as one moves down the ranking. For example, if the distribution follows the canonical Zipf's law (Zipf 1949), the slope equals -1, and the largest city is twice as large as the second, three times as large as the third, and so on.[1] I use the magnitude and direction of these slopes with respect to the Zipf benchmark, used as a common metric, to compare the concentration of different spatial distributions over time.[2]

I obtain three main findings. First, the geography of scholars' origins changed substan-

1. This is a regularity often found in the contemporary setting (Gabaix 1999). See Gabaix (2016) for a comprehensive review of contemporary Zipf-style distributions. For city-size distributions specifically, Arshad, Hu, and Ashraf (2018) review methodological issues and alternative formulations of Zipf-style regularities.

2. Another concept, often related to Zipf-style estimations, is Gibrat's law (Eeckhout 2004). It concerns the growth process rather than the static rank–size distribution: it states that proportional city growth is independent of initial size. In standard urban models, Gibrat-style growth can generate Zipf-like city-size distributions (Gabaix 1999). Whether a similar dynamic applies to the geography of scholarly human capital is an empirical question, which I examine as a complementary exercise (Appendix F).

tially between the medieval and early modern periods. In the Middle Ages, scholars came from a relatively small set of observed birthplaces, and the largest of these were not particularly dominant relative to other locations producing scholars. Between the late Middle Ages and the seventeenth century, this started to change with the largest birthplaces becoming increasingly important relative to smaller ones. At the same time, scholars originated from a rapidly expanding set of locations. Hence, the geographic reach of scholarly origins expanded even as their distribution became more concentrated at the top. Urban population followed a different pattern, with a rank–size distribution already much more concentrated in the medieval period and relatively stable over time.

Second, the geography of scholars' origins remained very different from that of scholarly activity. Workplaces were much more concentrated than birthplaces throughout the period, with a relatively small number of major institutional centers accounting for a disproportionately large share of scholarly activity. Scholars' birthplaces, instead, were spread across a much broader and expanding set of locations. The gap between the geographic reach of scholars' origins and their workplaces therefore widened over time. This suggests that the concentration of scholarly activity in major institutional centers increasingly coexisted with a broader geographic reach of scholars' origins, consistent with high lifetime mobility.

Third, the concentration of scholars' origins is systematically associated with the expansion of universities and academies. Using a country–half-century panel for 1500–1749, I find that countries with more academic institutions tend to exhibit a more concentrated distribution of scholars across their birthplaces. This suggests that the distribution of birthplaces becomes more top-heavy. Importantly, the same relationship does not appear for urban population, suggesting that the geography of scholarly origins evolved differently from the general urban system. These results are robust to a series of alternative specifications.

This paper contributes to the economic history of knowledge and elite human capital by documenting the changing geography of scholars' origins in Europe. A large literature emphasizes the importance of knowledge and the intellectual foundations of modern growth (Mokyr 2005, 2011, 2016). Universities and academies have been studied for their role in human-capital formation, knowledge production, and economic development (Cantoni and Yuchtman 2014; Bosker, Buringh, and Van Zanden 2013; Benos et al. 2024; Cinnirella, Hornung, and Koschnick 2025; Zanardello 2024), while others examine the emergence of an integrated academic market (de la Croix et al. 2024) and the contribution of knowledge elites to long-run growth (Squicciarini and Voigtländer 2016; Serafinelli and Tabellini 2022; Laouenan et al. 2022). I contribute to this literature by examining a different aspect of Europe's knowledge system: how the geographic origins of scholars evolved as universities and academies expanded.

More specifically, this paper relates to the literature on the geography of creative and intellectual activity. Existing studies have examined the concentration of creative individuals in major centers, their mobility, and the benefits of geographical clustering (O'Hagan and Borowiecki 2010; Borowiecki 2013; Cox and Figueroa 2025; Kuld and Mitchell 2023; Serafinelli and Tabellini 2022). In particular, Kuld and Mitchell (2023) show how German literary clusters increasingly developed through migration into creative centers rather than through a comparable concentration of authors' birthplaces. While these contributions high-

light the importance of distinguishing where intellectuals came from and where they worked, my focus is on the long-run spatial distribution of scholars' origins across Europe as a whole. Most closely related in terms of data, Curtis and de la Croix (2026) use RETE scholars' birthplaces to study regional development, and Curtis et al. (2026) construct measures of academic human capital across European cities and regions using publication outcomes from RETE. I instead study the distribution of scholars' origins, distinguishing the geography of birthplaces from the geography of scholarly activity.

Finally, I build on Dittmar (2011), who documents the emergence of Zipf's law in the distribution of West European city populations from the Middle Ages onward. Using a similar rank–size approach, I contribute by looking at the geography of scholarly human capital while keeping the historical perspective.[3] Comparing scholars' birthplaces with both urban population and scholarly workplaces documents whether the origins of knowledge producers followed the evolution of the broader urban system or developed a distinct spatial distribution. The findings show that, while scholarly activity remained concentrated in a relatively small number of institutional centers, scholars came from an expanding set of locations, and the concentration of their origins was systematically associated with the institutional environment.

Section 2 presents the scholarly and urban population data and discusses the distinction between birthplaces and workplaces. Section 3 documents the evolution of their rank–size distributions. Section 4 studies the relationship between university and academy networks and the rank–size relationship of scholars' origins at the country–half-century level. Section 4.1 reports robustness checks. Section 5 concludes.

2 Data: Scholarly and Urban Population

This section describes the urban and scholarly population data. The focus is mostly on the latter, as it represents the most original source of information used in the paper.

Urban population is taken from Buringh (2021), who provides updated population estimates for 2,262 European settlements that reached at least 5,000 inhabitants at least once between 700 and 2000 (Bairoch 1988). Historical population is available once every century until 1500, and once every half-century from 1500 onwards.[4]

Scholarly population is taken from de la Croix (2021), who provides individual-level information on scholars active in European universities and academies between 1000 and 1800.[5] The version used in this paper is from February 2026 and includes 88,863 unique scholars, as in de la Croix, Scebba, and Zanardello (2026). For these scholars, we observe biographical information such as name, surname, place and year of birth and death. We can

3. The methods used in this paper build on the literature on rank–size distributions and Zipf's law (Zipf 1949; Rosen and Resnick 1980; Soo 2005; Ioannides and Overman 2003). See Gabaix (2016), Arshad, Hu, and Ashraf (2018), and Hanlon and Heblich (2022) for broader discussions of the theoretical and empirical literature.

4. As historical urban populations are reconstructed from surviving sources, these figures should be interpreted as estimates rather than exact population counts, particularly for earlier periods.

5. I contributed to the project by enhancing the coverage of scientific academies; more details are provided in Zanardello (2024).

also track scholars over their active life through information on their institutional affiliations: headquarters of the institutions where they were active, appointment dates, and field of study.

An important feature of the database is that inclusion is institution-based rather than based on retrospective prominence. Scholars enter the database because they are recorded as teachers or members of the universities, academies, and other related institutions covered by the project, rather than because they appear in modern biographical dictionaries or are otherwise classified as famous.[6] The data are manually collected from institution-specific secondary sources, including membership and professor lists, modern prosopographical databases, and, where more recent sources are unavailable, older institutional histories. The database therefore aims at broad coverage of the scholarly population attached to these institutions, including relatively obscure individuals whenever the sources allow it. Moreover, the quality and completeness of the underlying sources vary across institutions, periods, and regions (de la Croix 2021, 2026).[7]

For the empirical analyses, I aggregate this micro-level information at the city level. The database allows for several possible aggregations: by birthplace, death place, or workplace.[8] The main analysis uses scholars' birthplaces, because they capture the spatial origins of scholarly human capital and are less mechanically linked to the location of universities and academies than workplaces. Death places are less attractive for this purpose, since they may reflect cumulative mobility, late-career sorting, and survival selection. Workplaces are instead informative about the geography of active scholarly life, but they are obtained from the same underlying affiliation records used to construct institutional exposure. For this reason, they are more exposed to mechanical endogeneity in the institutional analysis of Section 4. I therefore use birthplaces as the primary sample, and workplaces as comparison in Section 3.[9]

Table 2 (Appendix A) summarizes the construction of the analytical samples. Of the 88,863 scholars in the February 2026 database, 57,065, or 64.2%, have a known birthplace, while 39,142, or 44.0%, have a known birth year. Requiring both pieces of information leaves 33,267 scholars. To facilitate the comparison with urban population, I further restrict the sample to Europe, based on a European-country ISO whitelist, leading to a final birthplace sample of 33,076 scholars, corresponding to 37.2% of the full database. By comparison,

6. For datasets on notable individuals, see, for example, Laouenan et al. (2022) and Serafinelli and Tabellini (2022).

7. The institutional universe is built from established lists of European universities (Frijhoff 1996) and academies (McClellan 1985) and is supplemented by language academies, important Renaissance academies, and other higher-education institutions that conferred academic degrees or operated on the same academic market. The RETE institution-level publications document the specific sources used for each institution; see de la Croix (2026).

8. I rely on the geolocated birthplaces provided in RETE and interpret them as recorded origins. This distinction matters because birthplaces, workplaces, and death places capture different moments in scholars' life cycles (de la Croix and Scebba 2025).

9. The unit of observation differs slightly across the two distributions. Each scholar in the birthplace sample contributes exactly once to one birthplace and one birth half-century. For workplaces, a scholar observed in several headquarters cities within the same half-century contributes once to each of these cities. This concerns 10.6% of workplace scholar-periods. Workplace counts should therefore be interpreted as city-level affiliation counts rather than as a partition of scholars across cities.

88,792 scholars have at least one usable European workplace/reference-date record.

This sample construction raises a potentially important selection concern. Both inclusion in the surviving historical sources and the availability of birthplace and birth-year information may depend on how well an individual's life is documented. Changes in documentation over time can therefore mechanically affect both the number of observed birthplaces and the estimated shape of their distribution. Appendix B documents these patterns directly using the same February 2026 snapshot employed in the empirical analysis.

Biographical coverage varies substantially over time. Figure 3 (Appendix B) shows that coverage is strongly non-monotonic: the earliest records are unusually complete, coverage falls sharply during the later medieval period, and then improves during the early modern period. From 1450 onward, birthplace availability is comparatively stable, while the availability of birth years increases much more strongly. The unusually high completeness of the earliest observations is consistent with positive selection toward prominent and better-documented individuals; Appendix B provides additional evidence using modern external-visibility indicators.

Coverage also differs substantially across Europe. Table 3 (Appendix B) shows that over 1500–1749 the share of scholars for whom both birthplace and birth year are known ranges from around 23% in Southern Europe and 40% in Western Europe to more than 60% in Central Europe and around 80% in Northern and Eastern Europe. Table 4 (Appendix B) reports the resulting geographic composition of the analytical sample: Western and Central Europe account for approximately 40% and 32% of the final birthplace sample, respectively, Southern Europe for 22%, Northern Europe for 5%, and Eastern Europe for 2%.

These diagnostics do not eliminate historical source selection. In particular, the increase in the number of observed scholars and birthplaces over time partly reflects improvements in biographical documentation, especially in the availability of birth dates. I therefore do not interpret changes in the number of observed birthplace cities alone as evidence of geographic diffusion, nor do I interpret the early sample as a representative census of European scholarly origins. Instead, the rank–size estimates in Section 3 describe the shape of the distribution among observed scholars and locations. Early estimates should consequently be read with particular caution. Restricting the institutional analysis to the period after 1500 reduces the most severe early selection problem, although it does not make coverage uniform across either time or space. The corresponding estimates are therefore interpreted as conditional associations rather than causal effects.[10]

Finally, it is also clear that the distribution of 9,209 unique birthplaces is extremely skewed. Around 91% of observed birthplace-year cells contain a single scholar, while only about 2.4% contain three or more scholars and less than 0.5% contain five or more. This strong right-skewness is precisely why a power-law exercise is informative in this setting. To implement the analysis, I build a city × half-century panel, where cities are defined by

10. de la Croix (2026) provides additional period-specific statistics and a detailed discussion of the underlying source-selection process. The February 2026 snapshot used in this paper predates the subsequent update of the RETE database described in de la Croix (2026). I therefore use the latter to document the data-collection process and its limitations, while all sample counts and coverage statistics reported here are calculated directly from the February 2026 snapshot used in the empirical analysis.

scholars' birthplaces and half-centuries aggregate all scholars born in the corresponding 50-year interval indexed by their starting year. Thus, period t covers $[t, t+50)$: for example, the period labelled 1500 includes observations from 1500 through 1549. One data constraint is important here: the last observed birth year in the scholarly sample is 1786, so the last complete scholarly half-century is 1700–1749 rather than 1750–1799. I therefore treat the latter period as right-truncated throughout the descriptive analysis.

3 Power laws in scholarly and urban populations

I begin by estimating a standard rank–size specification, separately for scholars' birthplaces, workplaces, and for urban population. In the benchmark specification, I keep only places with positive counts in each half-century. This choice is important because zero scholarly counts are harder to interpret than zero urban population: in the scholarly data, they may reflect either a true absence of scholars or incomplete source coverage, especially in earlier periods. Urban population, as in Buringh (2021), may also contain measurement error, especially in earlier periods, even though this database updates previous population data (Bairoch 1988) and is broadly accepted by the literature. I keep the same positive-city convention for symmetry and comparability across the two datasets. Nevertheless, this implies an unbalanced panel with a varying number of places over time. I assess the relevance of this sample selection criterion in Section 4.1, where I also report a zero-inclusive robustness check.[11]

I estimate the following specification:

$$\log(R_{it}) = \alpha_t + \beta_t \log(S_{it}) + \varepsilon_{it}, \tag{1}$$

where R_{it} is the rank of city i in half-century t, and S_{it} is either the number of scholars born or active in that place over the corresponding 50-year window, or again the urban population of that city in that half-century. Throughout this section, the reported coefficients are the raw estimated slopes $\hat{\beta}_t$, which are negative in the distributions considered here. The canonical Zipf benchmark corresponds to $\hat{\beta} = -1$ (Zipf 1949). To better understand the dynamics, it is of interest to compare deviations from this benchmark coefficient. Indeed, deviations from -1 have different interpretations depending on their direction. Coefficients below -1, for example $\hat{\beta} = -2$, indicate that the largest places are smaller relative to lower-ranked locations than under Zipf; the distribution is therefore less top-heavy and with a thinner upper tail (or less concentrated) than the Zipf benchmark. Coefficients above -1, for example $\hat{\beta} = -0.6$, indicate the opposite: the largest places are disproportionately large relative to the rest, and the distribution is more top-heavy (or more concentrated) than Zipf. I therefore interpret movements toward -1 as 'convergence' toward the Zipf benchmark, while keeping track of whether these movements come from below or from above.[12] Appendix Figure 5 summarizes

11. The balanced panel contains 174,971 birthplace × half-century cells (9,209 locations × 19 period labels, from 900 through 1800). Because observed birth years range from 940 to 1786, the first and last non-empty periods are incomplete, while the 1800 endpoint is empty. The resulting panel contains 15,188 positive cells across 18 non-empty half-centuries.

12. In the exact rank–size relationship implied by the regression, city size can be written as proportional to $R^{1/\beta}$. Hence, when $\hat{\beta} = -1$, the largest place is twice the size of the second. When $\hat{\beta} = -2$, it is only $\sqrt{2}$

how movements in the raw rank–size coefficient translate into changes in top-heaviness and distance from the Zipf benchmark.

An important feature of rank–size data is the presence of ties: locations with identical observed sizes cannot be meaningfully ordered among themselves. I therefore use average ranks (i.e., midranks) throughout the analysis. Specifically, all locations within a tied group are assigned the midpoint of the ordinal positions occupied by that group. This convention preserves the multiplicity of locations in the distribution while making the resulting ranks invariant to arbitrary ordering within ties. The issue is particularly relevant for the scholarly distributions, where many locations have the same number of observed scholars, especially at the lower end of the distribution.[13]

Figure 4 (Appendix C) summarizes the main descriptive comparison between scholars' birthplaces, workplaces and urban population. Each panel reports the empirical rank–size relationship, shown by the red fitted line, together with the theoretical Zipf benchmark, shown by the dashed line. The benchmark is normalized to the largest observed place in each panel and imposes a slope of $\hat{\beta} = -1$. Hence, it represents the distribution that would arise if the second-ranked location were half as large as the first, the third-ranked city one third as large, and so on.

In the earlier periods, scholars' birthplaces display coefficients that are much more negative than those for urban population and workplaces, with $\hat{\beta}_t = -1.95$ in 1300. Given the rank–size specification used here, this means that the scholarly distribution initially lies below the Zipf benchmark: among positive scholars' birthplaces, the largest centers were not large enough relative to lower-ranked places to generate a Zipf-like concentration. Over time, however, the scholarly distribution moves toward the Zipf benchmark, reaching $\hat{\beta}_t = -1.3$ in 1700. The visual comparison between the slopes of empirical fitted lines and the theoretical benchmark conveys the same pattern: the rank–size distribution of scholars' birthplaces initially differs more from the Zipf benchmark than urban population, but gradually moves closer to it. By contrast, the urban system is already much closer to Zipf in 1300, with $\hat{\beta}_t = -1.15$, and evolves within a comparatively narrow range. Finally, workplace distributions show coefficients well above the canonical -1 already in 1100—even though this coefficient relies on a very small sample of only 16 locations—before stabilizing roughly between -0.4 and -0.6 for the rest of the time horizon. This implies a higher concentration of workplaces' distribution than under the canonical Zipf's law.

The distribution in 1500–1549 illustrates what this means. Bologna is the largest birthplace of scholars, with 41 scholars, followed by Florence with 35 and Siena with 31. Under Zipf's law, Florence's 35 scholars would imply a first-ranked location with approximately 70 scholars, while Siena's 31 would imply approximately 93. Bologna is therefore sizeable, but not nearly large enough relative to the next-ranked birthplaces to generate a Zipf-like distribution. This pattern extends further down the distribution. For example, the tenth-ranked birthplace still contains 18 scholars: under Zipf, a tenth-ranked location of this size would

times the size of the second; when $\hat{\beta} = -0.5$, it is four times the size of the second.

13. Rank–size coefficients are estimated only for samples containing at least five positive locations and at least three distinct positive sizes. These restrictions avoid estimating rank–size relationships from extremely sparse samples or from distributions with too little size variation to identify a meaningful slope.

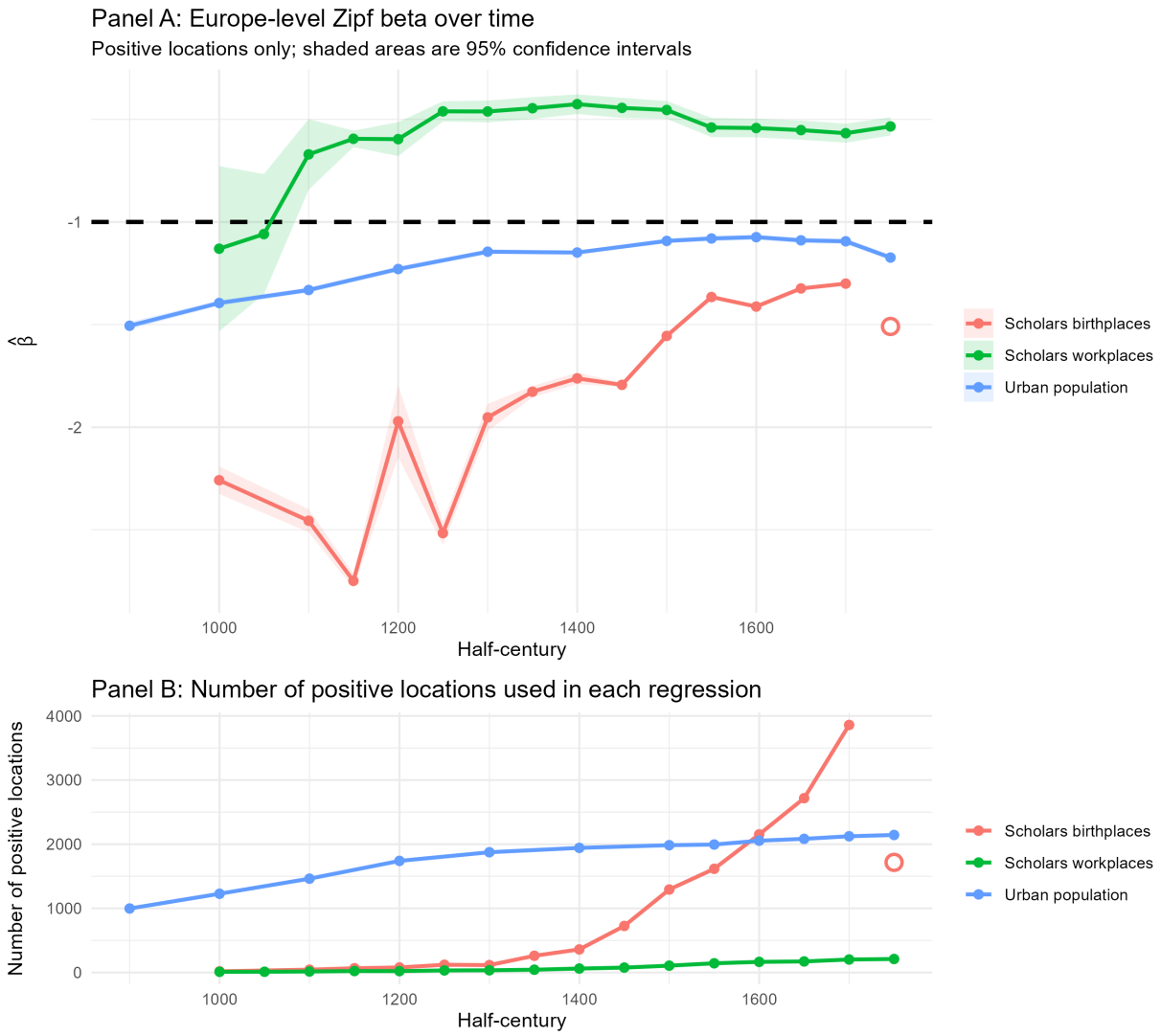


Figure 2: Rank-size coefficients over time.

Notes: Europe-level Zipf coefficients over time, with 95% confidence intervals and number of positive cities. Zipf baseline marked with a thicker dashed line at -1. Red lines correspond to birthplaces of scholarly population, blue lines to urban population, and green lines to scholars' workplaces. The dashed final birthplace segment denotes right truncation.

correspond to a first-ranked location with approximately 180 scholars, far above Bologna's observed 41. Thus, birthplace size declines relatively slowly down the ranking, with many sizeable locations but no correspondingly dominant upper tail. This is what a coefficient below -1, such as $\hat{\beta} = -1.55$ in 1500, captures: the largest birthplaces are not large enough relative to lower-ranked locations, and the upper tail is therefore thinner than the Zipf benchmark.

Figure 2 shows the same broad movement toward -1 from below between the late Middle Ages and the seventeenth century. As the number of locations with at least one observed scholar born there increases, the confidence intervals narrow, making the later coefficients

considerably more precise.[14] One caveat concerns the estimate for 1750, which is right-truncated because the last observed birth year is 1786. Because of this, I exclude the right-truncated $[1750, 1800)$ half-century from the institutional analyses in Section 4, and focus on the period 1500-1749, which includes the five complete half-centuries labelled 1500, 1550, 1600, 1650, and 1700. This window also provides a more stable coverage of locations over time and reduces concerns about selection arising from the data-collection process. Hence, I interpret the main descriptive result as the long-run convergence of the distribution of scholars' birthplaces toward the Zipf benchmark between the late Middle Ages and the seventeenth century.

Furthermore, in Figure 2 we can also see that the coefficients associated with the workplace distribution are substantially less negative than the birthplace coefficients and lie above the Zipf benchmark. This indicates a considerably more top-heavy rank–size distribution: scholarly activity was concentrated in an upper tier of large institutional centers, followed by a much longer tail of small workplace locations. The distribution of scholar–city affiliation counts in 1500–1549 illustrates this pattern. Cambridge is the largest workplace, with 966 scholars, followed by Padova with 840. A second tier of large institutional centers follows: Rome has 349 scholars, Oxford 334, Krakow 318, Bologna 313, and Leipzig 307. This upper tier remains large relative to the Zipf benchmark well beyond the first few ranks. For example, the twentieth-ranked workplace contains 129 scholars, compared with about 48 under a Zipf distribution normalized on Cambridge, and even the fiftieth-ranked location contains 33 scholars rather than about 19. Further down the distribution, however, workplace size falls sharply: the seventy-fifth-ranked location contains only 7 scholars, compared with about 13 under Zipf. Thus, the relatively flat upper tier is followed by a steep lower tail, producing a much more top-heavy overall rank–size distribution than the Zipf benchmark.

Figure 2 also reveals a widening gap in observed geographic reach. Between 1100 and 1700, the number of observed birthplace locations rose from 45 to 3,859, whereas the number of workplace locations increased from 16 to 203. As discussed in Section 2, part of the expansion in observed birthplaces may reflect improving biographical coverage. Even so, the much broader geography of recorded origins relative to scholarly activity is consistent with an important role for mobility into academic institutions. This interpretation concerns the relationship between the two aggregate distributions; it does not constitute a direct estimate of individual origin–destination moves.[15] Nevertheless, this complements the maps in Figure 1, showing that the broader reach of scholars' origins was also accompanied by a more concentrated distribution of scholarly workplaces.

Country-level regressions broadly reinforce these Europe-level patterns, although they are substantially noisier in the earlier periods, when many countries contain only a small number of positive scholarly locations. For scholars' birthplaces, the movement toward the

14. Additional quantile regressions and Gabaix curvature tests yield similar qualitative conclusions, although the early scholarly estimates remain noisy because of sparse support and the large number of tied location sizes.

15. The underlying database contains individual affiliation histories, and related work uses these records to study academic mobility and network formation directly (de la Croix, Scebba, and Zanardello 2026; de la Croix et al. 2024; Zanardello 2023). The present paper instead focuses on the changing aggregate geography of scholars' origins and activity.

Zipf benchmark is particularly visible in large countries such as France, Germany, and Great Britain. Between 1500 and 1700, for example, the estimated coefficient moves from -1.76 to -1.21 in France, from -1.58 to -1.28 in Germany, and from -1.86 to -1.53 in Great Britain. Italy is already close to the Zipf benchmark by 1500–1600, whereas the pattern is less systematic in Spain. By contrast, workplace coefficients remain well above -1 across countries and periods, generally in the range of roughly -0.3 to -0.6, confirming that scholarly activity remained considerably more top-heavy than scholars' origins. Urban coefficients remain comparatively close to -1 and vary within a substantially narrower range across countries and periods. Country-level urban regressions are also estimated on much larger and more stable samples because the Buringh (2021) data provide wider spatial support, although small countries can still have limited city counts in the earliest periods.

4 Institutions and rank–size distributions

The second part of the analysis studies whether universities and academies are associated with the spatial distribution of scholars' origins. I examine whether country-periods with more active institutions display systematically different rank–size distributions of scholars' birthplaces. I therefore move from the Europe-level descriptive analysis to an unbalanced country $\times$ half-century panel and estimate specifications of the form:

$$\hat{\beta}_{ct}^{\text{birth}} = \theta_0 + \theta_1 \log(1 + \text{Institutions}_{ct}) + \mu_c + \tau_t + \varepsilon_{ct}, \tag{2}$$

where $\hat{\beta}_{ct}^{\text{birth}}$ is the rank–size coefficient estimated for scholars' birthplaces in country c and half-century t (Soo 2005). Since the rank–size coefficients are negative, a positive value of θ_1 indicates that the coefficient becomes less negative as the number of institutions increases. As discussed in Section 3, this corresponds to a more top-heavy distribution among locations producing at least one scholar: the largest birthplaces become relatively larger compared with lower-ranked locations. For country-periods whose rank–size coefficients lie below -1, a higher coefficient also corresponds to movement toward the Zipf benchmark from below (see Appendix D for more technical details on the interpretation of the coefficients).

It is also important to note that this intensive-margin relationship is distinct from the expansion in the number of scholars' birthplaces documented in Section 3. A larger number of locations may produce at least one scholar while, at the same time, the distribution among those positive locations becomes more top-heavy. Hence, an expansion of the geographical reach can coexist with an increasing concentration at the top.

As already mentioned, there are two possible hypotheses for how more academic institutions across different locations should affect the birthplace distribution:

H1: More academic institutions may create several regional points of attraction, resulting in a flatter distribution of scholars' origins.

H2: Alternatively, more academic institutions may reinforce established scholarly regions, resulting in a more top-heavy distribution of scholars' origins.

This section therefore asks which of these hypotheses characterizes the historical data. The baseline institutional measure is the number of active institutions, combining universities and academies, in each country-half-century. I construct this measure from the lists of McClellan (1985) and Frijhoff (1996), as summarized by de la Croix (2021). An institution is counted as active in a half-century whenever its period of activity overlaps the corresponding 50-year interval; institution-level observations are then aggregated to the country–half-century level. Specifically, $\log(1 + \text{Institutions}_{ct})$ is the logarithm of the number of active institutions in country c and half-century t, with one added to accommodate country-periods with no active institution, which is more common for earlier periods. I interpret these estimates as conditional associations rather than causal effects, since institutional systems and the geography of scholars may jointly reflect unobserved characteristics or may reinforce one another. Finally, the panel is unbalanced because some country–half-century cells do not contain enough positive locations or distinct sizes to estimate a meaningful rank–size coefficient.

Table 1 builds four specifications progressively. Column (1) reports a simple pooled regression. Column (2) adds country and half-century fixed effects, so that the coefficient is identified from within-country changes in institutional network size over time rather than from permanent differences across countries. This distinction matters empirically. In the pooled specification, the coefficient for scholars' birthplaces is small, negative, and statistically insignificant, whereas after the inclusion of country and half-century fixed effects it becomes positive and statistically significant. Thus, the relationship documented below is a within-country relationship rather than one driven by cross-sectional differences between countries with larger and smaller institutional systems.[16] Column (3) adds precision weights proportional to the inverse estimated variance of the country-period rank–size coefficient, $1/\widehat{\text{SE}}(\hat{\beta}_{ct})^2$, rescaled to have mean one. This weighting scheme accounts for the fact that the dependent variable is itself an estimated coefficient: a slope estimated more precisely should receive greater weight than a much noisier estimate (Soo 2005). Column (4) keeps the same weighted specification but clusters standard errors at the country level to allow for serial correlation within countries over time. The final birthplace sample contains 93 country–half-century observations from 23 countries, while the corresponding urban sample contains 136 observations from 28 countries. Since the number of country clusters is moderate, especially for the scholarly sample, inference should be interpreted with some caution (Soo 2005; Ioannides and Overman 2003).

Table 1 yields a clear result for scholars' birthplaces. In the preferred specification with fixed effects, precision weights, and country-clustered standard errors, the coefficient on $\log(1 + \text{Institutions}_{ct})$ is 0.467, positive and highly significant (s.e. 0.100). Within countries, periods with larger institutional environments therefore tend to display less negative rank–size coefficients for scholars' birthplaces. This means that the distribution among positive

16. As a diagnostic, I also estimate a specification including only country and half-century fixed effects. Fixed effects alone explain a substantial share of the overall variation in the birthplace rank–size coefficient with an R^2 equal to 0.704, compared with $R^2 = 0.758$ when institutional count is added in column (2). The within R^2 reported in Table 1 provides a complementary measure: it captures the explanatory power of the institutional count for the variation that remains after country and half-century fixed effects have been removed.

Table 1: Institutions and rank–size coefficients, 1500–1749

	(1) Simple	(2) + FE	(3) + Precision weights	(4) + Clustered SE
Panel A. Scholars' birthplaces: dependent variable $\hat{\beta}_{ct}^{\text{birth}}$				
$\log(1 + \text{institutions}_{ct})$	-0.051	0.364***	0.467***	0.467***
	(0.031)	(0.096)	(0.079)	(0.100)
Observations	93	93	93	93
R^2	0.029	0.758	0.894	0.894
Within R^2		0.180	0.352	0.352
Panel B. Urban population: dependent variable $\hat{\beta}_{ct}^{\text{urban}}$				
$\log(1 + \text{institutions}_{ct})$	0.017	-0.043	0.034	0.034
	(0.030)	(0.081)	(0.042)	(0.055)
Observations	136	136	136	136
R^2	0.002	0.783	0.899	0.899
Within R^2		0.003	0.007	0.007
Country FE	✗	✓	✓	✓
Half-century FE	✗	✓	✓	✓
Precision weights	✗	✗	✓	✓
Country-Clustered SE	✗	✗	✗	✓

Significance codes: *** 0.01, ** 0.05, * 0.1

Notes: Standard errors in parentheses. The sample covers the five complete half-centuries from 1500 through 1749. Precision weights are proportional to $1/\widehat{\text{SE}}(\hat{\beta}_{ct})^2$ and are rescaled to have mean one. A constant term is included in column (1). Positive coefficients indicate that a larger institutional environment is associated with a higher, less negative rank–size coefficient and therefore with a more top-heavy distribution; negative coefficients indicate the opposite. The birthplace sample contains 93 country–half-century observations from 23 countries, while the urban-population sample contains 136 observations from 28 countries.

birthplaces becomes more top-heavy: the largest scholar-producing locations become relatively larger compared with lower-ranked locations, consistent with the pattern described in *H2*. For birthplace distributions lying below the Zipf benchmark, this same movement also brings the coefficient closer to -1 from below (see Appendix Figure 5 for the technical interpretation).

Because the regressor is logarithmic, the association with one additional institution depends on the initial number of institutions. In the preferred specification, moving from zero to one active institution is associated with an increase of approximately 0.32 in the rank–size coefficient, whereas moving from four to five institutions is associated with a smaller increase of approximately 0.09.[17]

The same relationship is not present for urban population. In the preferred specification, the coefficient is only 0.034 (s.e. 0.055) and is statistically indistinguishable from zero. The

17. More technically, considering column (4), adding one institution is associated with a change in the rank–size coefficient of $0.467 \times \log((I + 2)/(I + 1))$, where I is the initial number of institutions.

contrast with the scholarly regressions is informative: institutional expansion is systematically associated with changes in the distribution of scholars' origins, but not with comparable changes in the overall urban rank–size distribution. This contrast also appears in the within R^2. The number of active institutions accounts for a substantial share of the remaining within-country variation in the birthplace coefficient in the preferred specification (within $R^2 = 0.352$), whereas its explanatory power for within-country changes in the urban coefficient is negligible (within $R^2 = 0.007$). This comparison should nevertheless be interpreted with some caution, since the urban distribution is already much closer to the Zipf benchmark at the beginning of the period and evolves within a narrower range.

Taken together, these estimates suggest that the geographical expansion of academic institutions was associated with changes in the concentration of scholars' origins, rather than simply reflecting changes in the broader urban distribution. This relationship is consistent with institutions playing a role in shaping the geography of scholarly human capital, although the estimates do not directly identify whether it operated through local access to scholarly careers, recruitment by institutional centers, or geographic mobility (Jaravel 2023).

This relationship is also remarkably stable over time. Allowing the institutional slope to vary by half-century shows that, for scholars' birthplaces, the coefficient is positive and statistically significant in every period from 1500 to 1700, ranging only from 0.434 to 0.489. This suggests that the main relationship is not driven by a single half-century. For urban population, by contrast, all period-specific coefficients are small and statistically indistinguishable from zero. Overall, the dynamic specification therefore reinforces the contrast between a stable institutional association with the distribution of scholars' origins and the absence of a comparable relationship for the urban distribution. Coefficients by half-century are summarized in Table 5 (Appendix E).

4.1 Robustness checks

In this section, I report five robustness checks. First, I replace the institutional count with a dummy for the presence of at least one institution. Second, I weight observations by the number of places used to estimate each rank–size coefficient rather than by coefficient precision. Third, I use the institutional count lagged by one half-century. Fourth, I construct a zero-inclusive city panel and re-estimate the rank–size coefficient using $\log(1 + \text{pop})$, where pop is either scholarly population or urban population. This exercise should not be interpreted as the benchmark estimate, but as a robustness check that keeps the city universe fixed over time and explicitly includes zero-population settlements. The fifth robustness check addresses a different and particularly important sample-comparability concern. The universe of scholars' birthplaces is much broader than the urban universe in Buringh (2021), which contains settlements that reached at least 5,000 inhabitants at some point. Differences between the scholarly and urban rank–size distributions could therefore partly reflect differences in the underlying set of locations. I address this issue by matching scholarly birthplaces to Buringh settlements and re-estimating the two distributions on a common set of locations. Birthplaces are matched using place names, including alternative historical names, and geographical coordinates, with ambiguous cases reviewed manually. I then retain, within each country and half-century, locations with both a positive scholarly count

and a positive urban population. The scholarly and urban rank–size coefficients in this exercise are therefore estimated on exactly the same locations. Results are shown in Appendix Table 6.

The robustness results broadly reinforce the main result for scholars' birthplaces, although they also clarify which aspect of institutional development matters. The dummy for the presence of at least one institution is close to zero and statistically insignificant. Thus, the main association does not appear to come simply from the transition between having no institution and having at least one. By contrast, when I retain the institutional count but replace precision weights with weights based on the number of locations used to estimate each rank–size coefficient, the estimate is 0.480 (s.e. 0.104), very close to the benchmark result and highly statistically significant. The result is therefore not driven by the particular choice of precision weights.[18]

The lagged specification in Column (3) also provides a useful result. When institutional count is measured one half-century earlier, the coefficient remains positive and statistically significant, at 0.197 (s.e. 0.074). Hence, the relationship is not exclusively contemporaneous: higher institutional count is also associated with a higher, less negative birthplace rank–size coefficient fifty years later. This should not be interpreted causally, since institutional systems are highly persistent and the lag does not eliminate the possibility of common underlying determinants. It nevertheless shows that the benchmark relationship does not disappear as soon as contemporaneous institutional counts are replaced by predetermined ones.

Column (4) shows that the result is also robust to explicitly including zero-population locations. I construct a fixed location universe, use a $\log(1+\text{pop})$ transformation, and retain the same country–half-century support as in the benchmark specification. The coefficient remains positive and is statistically significant at the 10% level. Its magnitude is not directly comparable with the benchmark estimate, because the zero-inclusive exercise changes the scale and shape of the first-stage rank–size distribution. The relevant comparison is therefore primarily the direction and persistence of the institutional association rather than the magnitude of the coefficient.[19]

Column (5) provides an especially important robustness check. The benchmark scholarly and urban distributions are based on different location universes: the birthplace data include

18. As an additional robustness check inspired by Dittmar (2011), I re-estimate the country–half-century rank–size coefficients using Theil regressions, which are based on the median of pairwise slopes and are therefore less sensitive than OLS to individual leverage points. I use the same 93 country–half-century observations as in the benchmark analysis and weight them by the number of positive locations, since no first-stage precision estimate is constructed for the Theil coefficient. The institutional coefficient remains positive and highly statistically significant, at 0.425 (s.e. 0.107), compared with 0.467 (s.e. 0.100) in the benchmark OLS specification. The stepwise specification also displays a similar pattern: the pooled relationship is slightly negative, while the coefficient becomes positive once country and half-century fixed effects are introduced.

19. Allowing the zero-inclusive specification to use all country–half-century cells for which a valid rank–size coefficient can be estimated expands the birthplace sample from 93 to 121 observations and the urban sample from 136 to 139 observations. The institutional coefficient remains positive and significant at the 10% level for scholars' birthplaces, at 0.505 (s.e. 0.292), and remains small and statistically insignificant for urban population, at 0.060 (s.e. 0.056).

many small places that do not enter the Buringh urban sample. Restricting the analysis to matched Buringh settlements therefore changes the scholarly sample substantially and directly addresses the possibility that the contrast between scholarly and urban distributions is mechanically generated by their different spatial support. Despite this much more restrictive common support, the institutional coefficient for scholars' birthplaces remains positive and statistically significant, at 0.197 (s.e. 0.085). In contrast, when urban population is estimated on the same matched location universe, the coefficient is -0.089 (s.e. 0.102) and statistically indistinguishable from zero. This is particularly reassuring because the main contrast survives even when the scholarly and urban rank–size distributions are constructed from the same places. The relationship between the institutional network size and the distribution of scholars' origins therefore does not appear to be an artifact of the much broader spatial coverage of the birthplace data.

An important qualification is that the matched-city coefficient should not be interpreted mechanically as convergence toward Zipf. Restricting the scholarly sample to Buringh settlements disproportionately retains established urban locations and removes many smaller birthplaces, thereby changing the shape of the distribution itself. The relevant result in column (5) is instead that the same positive relationship between institutional count and the raw rank–size coefficient survives: within this very different city universe, countries with larger institutional environments still tend to display a higher, less negative, and hence more top-heavy birthplace distribution.

The urban robustness checks remain less significant, as in the benchmark results. None of the alternative specifications produces a statistically significant relationship between the institutional network size and the urban rank–size coefficient, and the estimated coefficients remain small. This is also true when columns (1), (2), and (4) retain the full 136 country–half-century observations used in the benchmark urban specification; only the lagged specification has a smaller sample because the lagged institutional measure is not available for all country–periods. The within R^2 also remains very small across these specifications. This reinforces the contrast found in Table 1: institutional count is systematically associated with changes in the distribution of scholars' origins, whereas no comparable within-country relationship emerges for the urban distribution.

Taken together, the results reveal two distinct features of the changing geography of scholars' origins. Descriptively, the number of locations producing scholars expanded over time, while the rank–size relationship among positive birthplaces became more top-heavy and moved toward the Zipf benchmark from below. Separately, within countries, larger institutional environments are systematically associated with higher, less negative rank–size coefficients for scholars' birthplaces. This association survives alternative weighting schemes, a Theil estimator, a lagged measure of institutions, the inclusion of zero-population locations, and, importantly, the restriction to a common Buringh location universe. The same systematic relationship is not observed for urban population.

These results should not be interpreted as showing that institutional expansion caused the geographic broadening of scholars' origins, nor as evidence that institutions were irrelevant for urban development. Rather, they show that the number of active institutions is closely associated with the concentration of observed scholars' birthplaces, whereas no comparable

within-country relationship appears for the urban rank–size distribution.

Appendix F reports a related Gibrat-style exercise. The results do not support size-independent proportional growth among existing scholars' birthplaces: from 1500 onward, initially larger scholar-producing locations tend to grow more slowly than smaller ones.

5 Conclusions

This paper shows that the geography of European scholars' origins evolved differently from both scholarly activity and urban population. Between the late Middle Ages and the seventeenth century, scholars came from an expanding set of locations, while the largest birthplaces became increasingly important relative to smaller ones. Hence, a broader geographic reach coexisted with greater concentration at the top of the birthplace distribution. At the same time, scholarly activity remained concentrated in a much smaller set of institutional centers, widening the gap between where scholars came from and where they worked. This pattern is consistent with an important role for mobility into intellectual centers, although individual movements are not directly measured. Within countries, periods with more universities and academies were also associated with a more concentrated distribution of scholars' origins, while no comparable relationship appears for urban population. Together, these findings suggest that the geography of scholarly human capital was closely associated with Europe's academic institutions rather than simply following broader urban development.

These findings also raise broader questions about the allocation of talent across space. The fact that scholars increasingly originated from a larger set of locations while scholarly activity remained concentrated in a few institutional centers suggests that the geography of knowledge production alone provides an incomplete picture of the European scholarly system. Understanding whether this broader geographic reach reflected greater access to scholarly careers, the attraction of talent toward intellectual centers, or changes in the local conditions producing scholars would require further research. In particular, the historical record allows me to observe individuals who became scholars, but not those who could have entered scholarly careers and did not. This distinction is important for understanding the implications of the geography of knowledge for the efficiency of talent allocation (Jaravel 2023).

Acknowledgements and Funding

I am grateful to Victor Gay for guidance. This project benefited from discussions with David de la Croix, Antoine Germain, and Arthur Silve. I thank the RAs working on the ERC project 883033 and on the PDR A2.11903.007-F for the help with the data collection. 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". This work was also supported by the Fonds de la Recherche Scientifique-FNRS under Grant n° A2.11903.007-F "Human capital and the rise of the West: the key role of scientific academies". Funding from the French Agence Nationale de la Recherche (under the Investissement d'Avenir programme, ANR-17-EURE-0010) is gratefully acknowledged.

Declaration of AI use

During the preparation of this work the author used ChatGPT to improve grammar and style and to assist with coding and debugging. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Agarwal, Ruchir, and Patrick Gaule. 2020. Invisible geniuses: Could the knowledge frontier advance faster? *American Economic Review: Insights* 2 (4): 409–424.

Arshad, Sidra, Shougeng Hu, and Badar Nadeem Ashraf. 2018. Zipf's law and city size distribution: A survey of the literature and future research agenda. *Physica A: Statistical mechanics and its applications* 492:75–92.

Bairoch, Paul. 1988. *Cities and economic development: from the dawn of history to the present.* University of Chicago Press.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. Who becomes an inventor in America? The importance of exposure to innovation. *The Quarterly Journal of Economics* 134 (2): 647–713.

Benos, Nikolaos (Nikos), Maurizio Conti, Michail Papazoglou, and Stamatis Tsoumaris. 2024. The Rise of Universities and Economic Development in Medieval and Early Modern Europe (700–1800). *SSRN http://dx.doi.org/10.2139/ssrn.4967505.*

Borowiecki, Karol Jan. 2013. Geographic Clustering and Productivity: An Instrumental Variable Approach for Classical Composers. *Journal of Urban Economics* 73 (1): 94–110. https://doi.org/10.1016/j.jue.2012.07.004.

Bosker, Maarten, Eltjo Buringh, and Jan Luiten Van Zanden. 2013. From baghdad to london: unraveling urban development in europe, the middle east, and north africa, 800–1800. *Review of Economics and Statistics* 95 (4): 1418–1437.

Buringh, Eltjo. 2021. The population of european cities from 700 to 2000: social and economic history. *Research Data Journal for the Humanities and Social Sciences* 6 (1): 1–18.

Cantoni, Davide, and Noam Yuchtman. 2014. Medieval universities, legal institutions, and the commercial revolution. *The Quarterly Journal of Economics* 129 (2): 823–887.

Cinnirella, Francesco, Erik Hornung, and Julius Koschnick. 2025. Flow of ideas: Economic societies and the rise of useful knowledge. *The Economic Journal* 135 (669): 1496–1535.

Cox, Gary W., and Valentin Figueroa. 2025. Agglomeration and Creativity in Early Modern Britain. *Explorations in Economic History* 95:101644. https://doi.org/10.1016/j.eeh.2024.101644.

Curtis, Matthew, and David de la Croix. 2026. Seeds of knowledge: Premodern scholarship, academic fields, and European growth. *European Economic Review,* 105299.

Curtis, Matthew, David de la Croix, Filippo Manfredini, and Mara Vitale. 2026. Academic human capital in european countries and regions, 1200–1793. *Explorations in Economic History* 101:101756. ISSN: 0014-4983. https://doi.org/10.1016/j.eeh.2026.101756.

de la Croix, David. 2021. Scholars and Literati in European Academia before 1800. *Repertorium Eruditorum Totius Europae* 5:35–41.

———. 2026. Scholars and Literati in European Academia before 1800 – an update. *Repertorium Eruditorum Totius Europae* 17:13–20. https://doi.org/10.14428/rete.v17i0/Global26.

de la Croix, David, Frédéric Docquier, Alice Fabre, and Robert Stelter. 2024. The academic market and the rise of universities in medieval and early modern Europe (1000–1800). *Journal of the European Economic Association* 22 (4): 1541–1589.

de la Croix, David, and Rossana Scebba. 2025. Geolocalization and the Birth-to-Death Distance. *Repertorium Eruditorum Totius Europae* 14:37–42. https://doi.org/10.14428/rete.v14i0/Locations.

de la Croix, David, Rossana Scebba, and Chiara Zanardello. 2026. *Flora, cosmos, salvatio: pre-modern academic institutions and the spread of ideas.* Technical report. CEPR Discussion Paper 20569.

Dittmar, Jeremiah. 2011. *The Emergence of Zipf's Law.* Technical report.

Eeckhout, Jan. 2004. Gibrat's law for (all) cities. *American Economic Review* 94, no. 5 (December): 1429–1451. https://doi.org/10.1257/0002828043052303.

Frijhoff, Willem. 1996. Patterns. Chap. 2 in *A history of the University in Europe. vol. ii: Universities in Early Modern Europe (1500–1800),* edited by Hilde de Ridder-Symoens. Cambridge University Press.

Gabaix, Xavier. 1999. Zipf's law and the growth of cities. *American Economic Review* 89 (2): 129–132.

———. 2016. Power laws in economics: an introduction. *Journal of Economic Perspectives* 30 (1): 185–206.

Hanlon, W. Walker, and Stephan Heblich. 2022. History and urban economics. Urban Economics and History, *Regional Science and Urban Economics* 94:103751. ISSN: 0166-0462. https://doi.org/10.1016/j.regsciurbeco.2021.103751.

Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow. 2019. The Allocation of Talent and U.S. Economic Growth. *Econometrica* 87 (5): 1439–1474. https://doi.org/10.3982/ECTA11427.

Ioannides, Yannis M, and Henry G Overman. 2003. Zipf's law for cities: an empirical examination. *Regional science and urban economics* 33 (2): 127–137.

Jaravel, Xavier. 2023. *Marie Curie habite dans le Morbihan: démocratiser l'innovation.* Seuil.

Kuld, Lukas, and Sara Mitchell. 2023. Market Structure and Creative Cluster Formation: The Origins of Urban Clusters in German Literature, 1700–1932. *European Review of Economic History* 27 (3): 380–411. https://doi.org/10.1093/ereh/head003.

Laouenan, Morgane, Palaash Bhargava, Jean-Benoit Eyméoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. A cross-verified database of notable people, 3500bc-2018ad. *Scientific Data* 9 (1): 290.

McClellan, James E. 1985. *Science reorganized: Scientific societies in the eighteenth century.* Columbia University Press.

Mokyr, Joel. 2005. The intellectual origins of modern economic growth. *The Journal of Economic History* 65 (2): 285–351.

———. 2011. The gifts of Athena: Historical origins of the knowledge economy. Princeton University Press.

———. 2016. A Culture of Growth. Princeton University Press.

O'Hagan, John, and Karol Jan Borowiecki. 2010. Birth Location, Migration, and Clustering of Important Composers: Historical Patterns. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 43 (2): 81–90. https://doi.org/10.1080/01615441003729945.

Rosen, Kenneth T, and Mitchel Resnick. 1980. The size distribution of cities: an examination of the pareto law and primacy. *Journal of urban economics* 8 (2): 165–186.

Serafinelli, Michel, and Guido Tabellini. 2022. Creativity over time and space: A historical analysis of European cities. *Journal of Economic Growth* 27 (1): 1–43.

Soo, Kwok Tong. 2005. Zipf's law for cities: a cross-country investigation. *Regional science and urban Economics* 35 (3): 239–263.

Squicciarini, Mara P, and Nico Voigtländer. 2016. *Knowledge elites and modernization: evidence from revolutionary france.* Technical report. National Bureau of Economic Research.

Zanardello, Chiara. 2023. Market forces in italian academia today (and yesterday). *Scientometrics* 128 (1): 651–698.

———. 2024. Early modern academies and growth. LIDAM Discussion Paper - 2024/12.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: an introduction to human ecology.* Reading, MA: Addison-Wesley.

A Descriptive statistics

Table 2: Construction of the scholarly samples

	Scholars	Share of full database
Full scholar database	88,863	100.0%
Known birthplace	57,065	64.2%
Known birth year	39,142	44.0%
Known birthplace and birth year	33,267	37.4%
Known birthplace and birth year, European birthplace	33,076	37.2%
Usable European workplace/reference-date sample	88,792	99.9%

Notes: The first five rows describe the construction of the main birthplace sample. The last row reports the comparison sample used to construct workplace distributions. Percentages are relative to the 88,863 scholars in the February 2026 database.

B Sample composition and coverage diagnostics

This appendix provides additional information on the composition and coverage of the scholarly sample. The purpose is to assess whether changes in the availability of biographical information may affect the spatial patterns documented in the main text. In particular, the construction of the birthplace sample requires both a known birthplace and a known birth year, and the availability of these variables varies over time and across regions.

For these diagnostics only, scholars are assigned to half-centuries using the reference date `refdatel`, rather than their year of birth. The variable `refdatel` is designed to approximate the beginning of a scholar's active academic period (de la Croix, Scebba, and Zanardello 2026). When an exact affiliation date is available, it corresponds to the earliest known year of affiliation with a formal educational institution: for university professors, this is generally the year in which they began teaching, while for academy members it is the year in which they were elected to the academy. When exact dates are unavailable, approximate affiliation dates are used. In more extreme cases, the reference date is constructed from the earliest available date among 30 years after the scholar's birth year, the year of death, the institution's closing date, or 1793 (e.g., our sample end point).

Although birth year can therefore enter the construction of `refdatel` as a fallback when it is available, `refdatel` does not mechanically require a known birth year. It can consequently be used to organize the missingness diagnostic without restricting the denominator to scholars for whom `YearBorn` is observed. Here, `refdatel` is used only to assign scholars to broad reference periods; it is not interpreted as a birth date or as an indication that the scholar was located at his or her birthplace in that period. The main birthplace analysis remains assigned by scholars' actual birth years.

Figure 3 reports the availability of birthplace and birth-year information by half-century. Using `refdatel` rather than birth year is important because using birth year to study missing birth-year information would mechanically exclude precisely the observations whose missingness I want to measure.

Coverage is strongly non-monotonic. The earliest observations are unusually well documented: for example, 57.5% of the 40 scholars assigned to the reference half-century 1000 have both birthplace and birth year recorded. Joint coverage then falls below 12% between 1250 and 1450 before increasing steadily during the early modern period. From 1450 onward, birthplace availability itself is comparatively stable, at around 60–68%, while birth-year coverage rises much more strongly, from 13.0% in 1450 to 24.9% in 1500, 45.6% in 1600, and 61.9% in 1700. Consequently, the share for which both birthplace and birth year are observed rises from 11.3% in 1450 to 21.3% in 1500, 38.9% in 1600, and 52.6% in 1700.

The unusually high completeness of the earliest records is consistent with positive selection toward prominent and better-documented individuals. This can also be seen directly in modern external-visibility indicators contained in the database. Across the reference periods 1000–1249, around 41% of scholars have a non-empty Wikipedia record and a similar share have a VIAF record. These shares fall to approximately 9% and 11%, respectively, over 1250–1499, as the database begins to include a much broader set of less prominent scholars, before increasing again in the early modern period. These indicators are not used as selection criteria or in the empirical analysis; they are reported only to characterize the changing composition of the surviving historical record.

Similarly, regional coverage is measured using scholars' modal workplace country, rather than their birthplace country, so that scholars with a missing birthplace are not mechanically excluded from the denominator. These alternative assignments are used only to describe sample coverage and do not change the construction of the main analytical sample, where scholars are assigned to periods according to their birth year and to locations according to their birthplace.

Table 3 shows that coverage is neither constant over time nor homogeneous across Europe. Table 4 then describes the geographic composition of the resulting final birthplace sample. The observed selection is an important limitation of the data, especially in the medieval period. The main analysis should therefore be interpreted as describing the spatial distribution of observed scholars' origins rather than the complete population of individuals who could potentially have become scholars.

Table 3: Biographical coverage by broad European region and period

	1000–1249		1250–1499		1500–1749	
Region	*N*	Both known	*N*	Both known	*N*	Both known
Central Europe	27	29.6%	3,369	13.3%	10,637	64.5%
Eastern Europe	–	–	–	–	605	80.5%
Northern Europe	–	–	10	40.0%	1,195	81.1%
Southern Europe	311	21.2%	6,074	11.0%	19,582	23.0%
Western Europe	485	31.5%	7,531	6.6%	19,760	40.2%

Notes: "Both known" is the share of scholars for whom both birthplace and birth year are observed. For this missingness diagnostic, scholars are assigned to a broad reference region using their modal European workplace country, so that geographic assignment does not condition on birthplace being observed. Periods are assigned using `refdatel`, as described above. Broad regions are fixed groupings of present-day ISO country codes and are used only for descriptive diagnostics. In 2.6% of cases, more than one workplace country is tied for the modal country; ties are resolved deterministically.

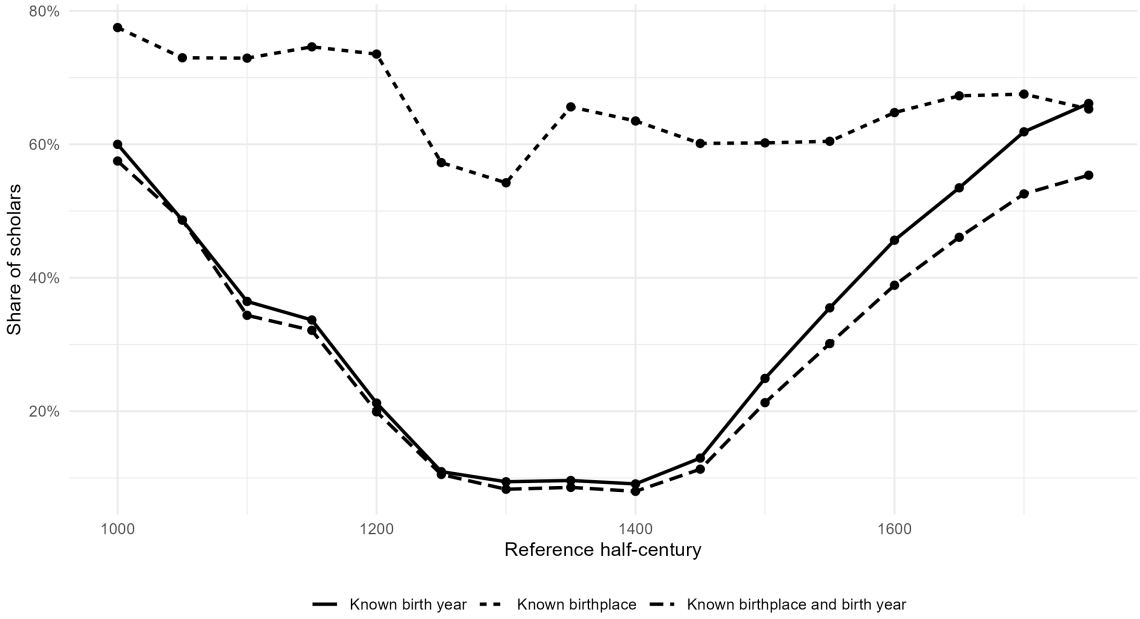


Figure 3: Biographical coverage of scholars over time.

Notes: The figure reports the share of scholars with a known birthplace, known birth year, and both pieces of information. For this missingness diagnostic only, periods are assigned using the independent reference date `refdatel`; the main birthplace analysis continues to assign scholars to half-centuries using their birth year.

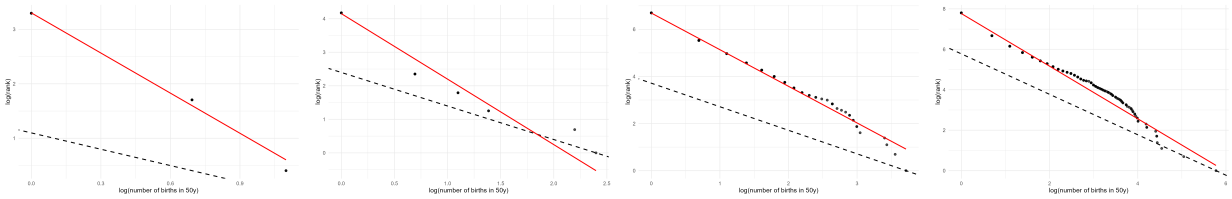
Table 4: Geographic composition of the final birthplace sample

Region	Scholars	Birthplace locations	Share of scholars
Western Europe	13,176	3,938	39.8%
Central Europe	10,553	2,888	31.9%
Southern Europe	7,119	1,603	21.5%
Northern Europe	1,628	633	4.9%
Eastern Europe	600	147	1.8%
Total	33,076	9,209	100.0%

Notes: The table describes the final analytical birthplace sample, restricted to scholars with a known birthplace and birth year and a birthplace in the European-country whitelist. Regions are fixed broad groupings of present-day country codes and are used for descriptive purposes only.

C Additional rank–size evidence

Figure 4 presents the empirical rank–size distributions of scholars' birthplaces, urban population, and scholars' workplaces for selected half-centuries. It complements Figure 2 in the main text by illustrating the underlying rank–size relationships and their deviations from the Zipf benchmark.

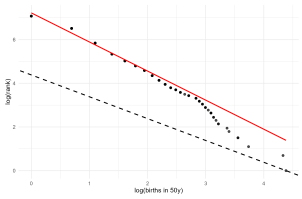


(a) Birthplaces in 1100
$\hat{\beta} = -2.46, N = 45$

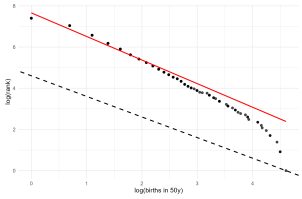
(b) Birthplaces in 1300
$\hat{\beta} = -1.95, N = 116$

(c) Birthplaces in 1500
$\hat{\beta} = -1.55, N = 1294$

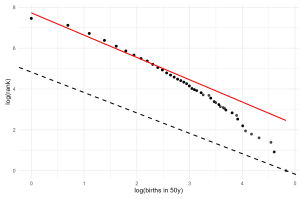
(d) Birthplaces in 1700
$\hat{\beta} = -1.3, N = 3859$



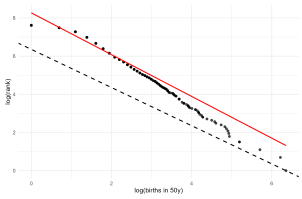
(e) Urban pop in 1100
$\hat{\beta} = -1.33, N = 1463$



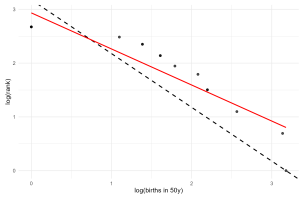
(f) Urban pop in 1300
$\hat{\beta} = -1.15, N = 1875$



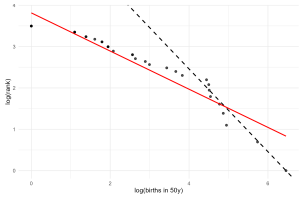
(g) Urban pop in 1500
$\hat{\beta} = -1.09, N = 1985$



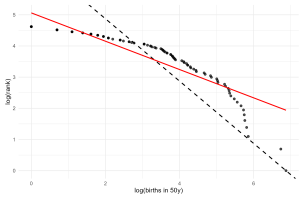
(h) Urban pop in 1700
$\hat{\beta} = -1.09, N = 2124$



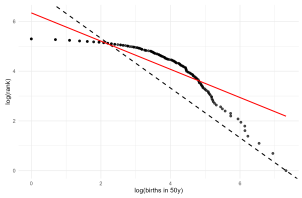
(i) Workplaces in 1100
$\hat{\beta} = -0.67, N = 16$



(j) Workplaces in 1300
$\hat{\beta} = -0.46, N = 35$



(k) Workplaces in 1500
$\hat{\beta} = -0.45, N = 107$



(l) Workplaces in 1700
$\hat{\beta} = -0.57, N = 203$

Figure 4: Rank–size plots for scholars' birthplaces, urban cities, and scholars' workplaces

Notes: Europe-level Zipf-style rank–size plots for scholars' birthplaces, urban cities, and scholars' workplaces; positive cities only. The dashed line reports the theoretical Zipf benchmark, normalized to the largest observed city in each panel and imposing $\hat{\beta} = -1$. The solid line reports the empirical fitted rank–size relationship. Ties handled by the average rank method.

D Interpreting the rank–size coefficient

Figure 5 summarizes the interpretation of movements in the raw rank–size coefficient. In the institutional regression of Equation (2), θ_1 captures the association between the number of active universities and academies and the estimated rank–size coefficient $\hat{\beta}$. A positive θ_1 implies a higher, less negative $\hat{\beta}$, corresponding to a more top-heavy distribution (as shown in the upper panels), while a negative θ_1 implies a lower, more negative $\hat{\beta}$ and a less top-heavy distribution (as shown in the lower panel). Whether these movements bring the distribution closer to or farther from the canonical Zipf benchmark, $\hat{\beta} = -1$, depends on which side of the benchmark the initial coefficient lies.

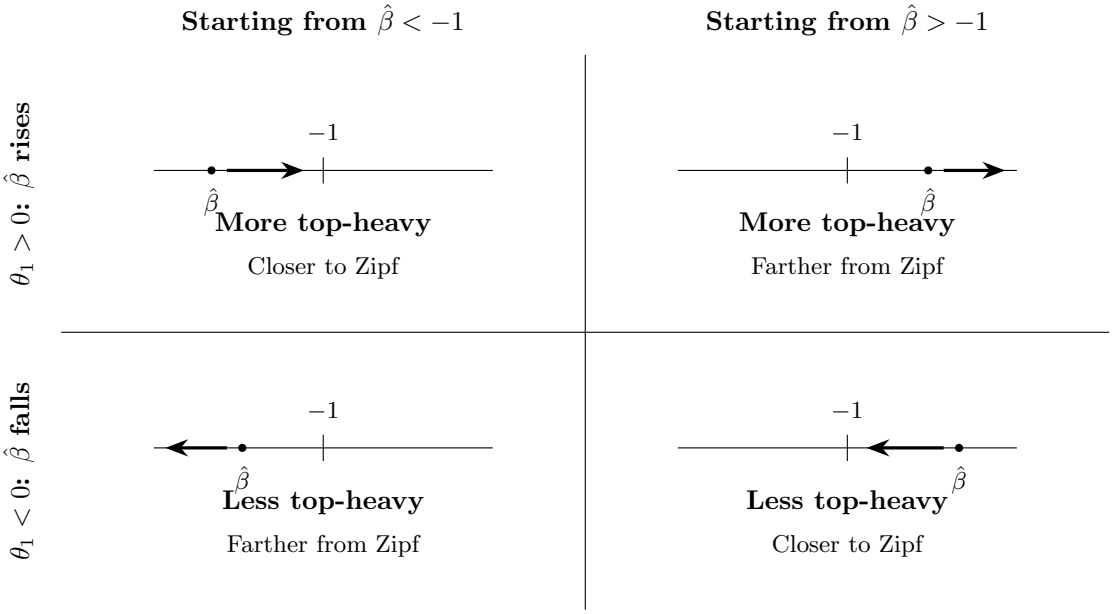


Figure 5: Interpreting movements in the raw rank–size coefficient.

Notes: A positive association, $\theta_1 > 0$, raises $\hat{\beta}$ and therefore makes the distribution more top-heavy, while a negative association lowers $\hat{\beta}$ and makes it less top-heavy. Whether the same movement represents convergence toward or divergence from the Zipf benchmark depends on the starting position relative to $\hat{\beta} = -1$. The figure is schematic and does not depict estimated magnitudes.

E Additional results and robustness checks

This appendix provides additional results supporting the institutional analysis in Section 4. Table 5 reports specifications allowing the association between institutional count and rank–size coefficients to vary across half-centuries. Table 6 presents robustness checks using alternative measures of institutional presence, weighting schemes, lagged institutional counts, zero-inclusive distributions, and a common set of locations matched to Buringh (2021).

Table 5: Institutions and rank–size coefficients by half-century, 1500–1749

	Scholars' birthplaces $\widehat{\beta}_{ct}^{\text{birth}}$	Urban population $\widehat{\beta}_{ct}^{\text{urban}}$
$\log(1 + \text{Institutions}_{ct}) \times 1500$	0.434***	0.057
	(0.096)	(0.052)
$\log(1 + \text{Institutions}_{ct}) \times 1550$	0.458***	0.048
	(0.104)	(0.048)
$\log(1 + \text{Institutions}_{ct}) \times 1600$	0.485***	0.050
	(0.097)	(0.052)
$\log(1 + \text{Institutions}_{ct}) \times 1650$	0.489***	0.037
	(0.087)	(0.049)
$\log(1 + \text{Institutions}_{ct}) \times 1700$	0.489***	0.004
	(0.093)	(0.051)
Country FE	✓	✓
Half-century FE	✓	✓
Precision weights	✓	✓
Country-Clustered SE	✓	✓
Observations	93	136
R^2	0.898	0.911
Within R^2	0.374	0.119

Significance codes: *** 0.01, ** 0.05, * 0.1

Notes: Each coefficient reports the half-century-specific association between $\log(1 + \text{Institutions}_{ct})$ and the estimated rank–size coefficient. Positive coefficients indicate a higher, less negative rank–size coefficient and therefore a more top-heavy distribution; negative coefficients indicate a less top-heavy distribution. All specifications include country and half-century fixed effects. Standard errors, clustered at the country level, are reported in parentheses. Observations are weighted proportionally to the inverse estimated variance of the rank–size coefficient, with weights rescaled to have mean one. The sample covers the five complete half-centuries from 1500 through 1749.

Table 6: Robustness checks: institutions and rank–size coefficients, 1500–1749

	(1) Dummy	(2) City weights	(3) Lagged count	(4) Zero-inclusive	(5) Buringh match
Panel A. Scholars' birthplaces: dependent variable $\hat{\beta}_{ct}^{\text{birth}}$					
Institution dummy	0.006				
	(0.336)				
$\log(1+\text{inst}_{ct})$		0.480***		0.599*	0.197**
		(0.104)		(0.328)	(0.085)
$\log(1+\text{inst}_{c,t-1})$			0.197**		
			(0.074)		
Observations	93	93	93	93	75
R^2	0.837	0.868	0.849	0.855	0.870
Within R^2	0.000	0.390	0.072	0.221	0.199
Panel B. Urban population: dependent variable $\hat{\beta}_{ct}^{\text{urban}}$					
Institution dummy	-0.024				
	(0.067)				
$\log(1+\text{inst}_{ct})$		0.042		0.050	-0.089
		(0.055)		(0.059)	(0.102)
$\log(1+\text{inst}_{c,t-1})$			0.011		
			(0.037)		
Observations	136	136	116	136	77
R^2	0.899	0.746	0.902	0.829	0.786
Within R^2	0.002	0.007	0.001	0.009	0.029
Country FE	✓	✓	✓	✓	✓
Half-century FE	✓	✓	✓	✓	✓
Country-clustered SE	✓	✓	✓	✓	✓

Significance codes: *** 0.01, ** 0.05, * 0.1

Notes: All columns include country and half-century fixed effects and country-clustered standard errors. In column (1), observations are weighted by the precision of the estimated rank–size coefficient. In column (2), observations are instead weighted by the number of locations used to estimate the corresponding coefficient. In column (3), the institutional variable is lagged by one half-century and observations are precision-weighted; the urban sample is smaller because the lagged institutional count is not available for all benchmark country–periods. In column (4), the city universe is held constant over time by including zero-population settlements and using a $\log(1+\text{pop})$ transformation; observations are precision-weighted, and the sample is restricted to country–periods entering the corresponding benchmark specification. The coefficient is therefore not directly comparable in magnitude with the benchmark estimate. In column (5), scholarly birthplace and urban rank–size coefficients are re-estimated using only locations matched between the scholarly and Buringh datasets and with positive values in both distributions in the corresponding half-century. Thus, within each country–period, the two coefficients are constructed from the same underlying locations. Observations are precision-weighted. The number of country–period observations can nevertheless differ between the two panels because the same matched locations may contain too few distinct scholarly counts to estimate a meaningful birthplace rank–size coefficient.

F Gibrat's Law

As a complementary exercise, I examine whether subsequent growth is independent of the initial size of scholars' birthplaces, as implied by Gibrat's law. This is not the main empirical framework of the paper, but it provides a useful alternative perspective on the evolution of the geography of scholars' origins, inspired by Gabaix (1999) and Eeckhout (2004). I estimate the following location-level specification for selected half-centuries:

$$g_{it} = \alpha_t + \gamma_t \log(S_{it}) + u_{it}, \tag{3}$$

where $g_{it} = \log(S_{i,t+1}/S_{it})$ is subsequent growth and S_{it} is the initial number of scholars born in location i. The sample includes locations with positive scholarly counts in both consecutive half-centuries. Under Gibrat's law, γ_t should be close to zero, indicating size-independent growth. If $\gamma_t < 0$, larger scholar-producing locations grow more slowly than smaller ones; if $\gamma_t > 0$, they grow faster.

The Europe-level results for scholars' birthplaces do not support convergence toward size-independent growth. In the early periods, the estimates are imprecise: the coefficient is -0.020 (s.e. 0.401) in 1100 and -0.121 (s.e. 0.190) in 1300. From 1500 onward, however, the relationship is clearly negative and statistically significant. The coefficient is -0.227 (s.e. 0.040) in 1500, -0.164 (s.e. 0.028) in 1600, and -0.171 (s.e. 0.024) in 1650. Hence, among locations that produce scholars in two consecutive periods, initially larger birthplaces subsequently grow more slowly than smaller ones. Rather than a Gibrat-like process, from 1500 onward, the data therefore show a persistent negative size–growth relationship, consistent with mean reversion. The final transition considered is from 1650–1699 to 1700–1749, so the right-truncated final birthplace half-century does not enter this exercise.

This result is interesting because, for urban population, Gibrat's law is often interpreted as a condition that can generate a Zipf-like city-size distribution (Gabaix 1999). For scholars' birthplaces, instead, the rank–size coefficient moves toward the Zipf benchmark from below even though growth among already-active locations remains systematically related to initial size. The Gibrat exercise therefore does not provide evidence that the evolution of the scholarly rank–size distribution resulted from size-independent proportional growth among existing scholar-producing locations.

At the same time, this exercise only considers locations with positive scholarly counts in two consecutive half-centuries and therefore does not directly capture entry into the scholarly distribution. Section 3 documents that the number of positive birthplaces expanded substantially over the same period. Taken together, the two results are consistent with an important extensive-margin component in the long-run evolution of scholars' origins: the set of locations producing scholars expanded even though growth among locations already present in the distribution was characterized by mean reversion.

Dittmar (2011) shows that urban population had a negative size–growth relationship before 1500 and moved closer to size-independent growth after 1500, implying that the growth distributions of large and small cities became more similar. As a comparison, I estimate the same Gibrat-type regression for urban population using Buringh (2021) for Europe as a whole. The urban results also reject strict size-independent growth. The coefficient is relatively close to zero in 1500, at -0.039 (s.e. 0.006), and remains fairly small in 1600, at -0.068 (s.e. 0.006). It subsequently becomes more negative, reaching -0.174 (s.e. 0.009) in 1700 and -0.135 (s.e. 0.009) in 1750. Thus, the urban distribution does not display a monotonic movement toward Gibrat's law either. The comparison

nevertheless shows that size dependence is substantially weaker for urban population than for scholars' birthplaces around 1500, when the corresponding scholarly coefficient is -0.227. The two distributions therefore need not approach a Zipf-like rank–size pattern through the same underlying growth process.