NBER WORKING PAPER SERIES

# THE ROLE OF INDUSTRY, OCCUPATION, AND LOCATION-SPECIFIC KNOWLEDGE IN THE SURVIVAL OF NEW FIRMS

C. Jara Figueroa
Bogang Jun
Edward L. Glaeser
César Hidalgo

Working Paper 24868
http://www.nber.org/papers/w24868

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

C.J.F., B.J., and C.A.H. acknowledge support from the MIT Media Lab Consortia, the MIT Skoltech Program, the Masdar Institute of Science and Technology (Masdar Institute), Abu Dhabi, USA—Reference 02/MI/MIT/CP/11/07633/GEN/G/ 00 and the Center for Complex Engineering Systems (CCES) at King Abdulaziz City for Science and Technology (KACST). We thank comments from Kerstin Enflo, Jian Gao, Tarik Roukny, and Siqi Zheng, as well as data assistance from Manuel Aristarán and Elton Freitas. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w24868.ack

The Role of Industry, Occupation, and Location-Specific Knowledge in the Survival of New Firms

C. Jara Figueroa, Bogang Jun, Edward L. Glaeser, and César Hidalgo

NBER Working Paper No. 24868
July 2018
JEL No. D22,J24,N1,N16,O1,O14,O15,O5,O54,R12

## ABSTRACT

How do regions acquire the knowledge they need to diversify their economic activities? How does the migration of workers among firms and industries contribute to the diffusion of that knowledge? Here we measure the industry, occupation, and location specific knowledge carried by workers from one establishment to the next using a dataset summarizing the individual work history for an entire country. We study pioneer firms–firms operating in an industry that was not present in a region–because the success of pioneers is the basic unit of regional economic diversification. We find that the growth and survival of pioneers increase significantly when their first hires are workers with experience in a related industry, and with work experience in the same location, but not with past experience in a related occupation. We compare these results with new firms that are not pioneers and find that industry specific knowledge is significantly more important for pioneer than non-pioneer firms. To address endogeneity we use Bartik instruments, which leverage national fluctuations in the demand for an activity as shocks for local labor supply. The instrumental variable estimates support the finding that industry related knowledge is a predictor of the survival and growth of pioneer firms. These findings expand our understanding of the micro-mechanisms underlying regional economic diversification events.

C. Jara Figueroa
Massachusetts Institute of Technology
MIT Medial Lab
75 Amherst St.
Cambridge, MA 02139
crisjf@mit.edu

Bogang Jun
Massachusetts Institute of Technology
MIT Medial Lab
75 Amherst St.
Cambridge, MA 02139
bjun@mit.edu

Edward L. Glaeser
Department of Economics
315A Littauer Center
Harvard University
Cambridge, MA 02138
and NBER
eglaeser@harvard.edu

César Hidalgo
Massachusetts Institute of Technology
MIT Medial Lab
75 Amherst St.
Cambridge, MA 02139
hidalgo@mit.edu

# The role of industry, occupation, and location specific knowledge in the survival of new firms

C. Jara-Figueroa[a], Bogang Jun[a], Edward Glaeser[b], and Cesar Hidalgo[a,1]

[a]MIT Media Lab, Massachusetts Institute of Technology
[b]Department of Economics, Harvard University
[1]To whom correspondence should be addressed. E-mail: hidalgo@mit.edu

July 17, 2018

**Abstract**

How do regions acquire the knowledge they need to diversify their economic activities? How does the migration of workers among firms and industries contribute to the diffusion of that knowledge? Here we measure the industry, occupation, and location specific knowledge carried by workers from one establishment to the next using a dataset summarizing the individual work history for an entire country. We study pioneer firms–firms operating in an industry that was not present in a region–because the success of pioneers is the basic unit of regional economic diversification. We find that the growth and survival of pioneers increase significantly when their first hires are workers with experience in a related industry, and with work experience in the same location, but not with past experience in a related occupation. We compare these results with new firms that are not pioneers and find that industry specific knowledge is significantly more important for pioneer than non-pioneer firms. To address endogeneity we use Bartik instruments, which leverage national fluctuations in the demand for an activity as shocks for local labor supply. The instrumental variable estimates support the finding that industry related knowledge is a predictor of the survival and growth of pioneer firms. These findings expand our understanding of the micro-mechanisms underlying regional economic diversification events.

Can developing countries and cities thrive through their own entrepreneurship, or must they attract external investment? What are the factors that influence the success of local ventures? Development depends on undertaking new tasks, which require knowledge. In this paper, we estimate the impact of a worker's knowledge about an industry, occupation, and location in the survival of pioneer firms [1]: firms that start operating in a region where their industry was not present.

Understanding the success of pioneer firms is key to understanding the mechanisms behind industrial diversification. When a pioneer firms succeeds, the region where this firm is now present will have successfully developed a new industry. Here, we use a large administrative data set with almost complete work histories for all the individual workers of a country, to measure the knowledge carried by workers from their previous jobs into pioneer firms. This dataset allows us to estimate the industry specific knowledge, occupation specific knowledge, formal schooling, and knowledge about a location that each worker brings into a pioneer firm. We use this fine grained description to to test which type of knowledge matters most for the growth and survival of pioneer firms, and compare these results with new firms that are not pioneers; non-pioneer firms.

For decades, human capital has been recognized as an important determinant of economic growth [2–10]. But human capital is not just a worker's formal schooling. Workers acquire important skills, knowledge, and contacts at work. A forty-year-old worker brings, on average, more years of experience into a company than years of schooling. This work experience, which is specific to an industry, location, and occupation, should impact the growth and survival of the activities where these workers are involved.

The specificity of this knowledge pushes us to think of human capital not only in terms of intensity, but in terms of relatedness. Workers are not simply knowledgeable or skilled, but posses knowledge that is related to specific activities, even to new activities that have never before been present in a city or a country. In this paper, we test what type of related knowledge is a more critical ingredient in the success of new firms that lead to the development of new industries. While there is a long literature measuring relatedness between products [11, 12], industries [13, 14], technologies [15, 16], and even occupations [17], there is little work separating these relatedness measures into multiple forms of human capital.

In this paper we decompose knowledge into a two-dimensional representation, measuring how related the

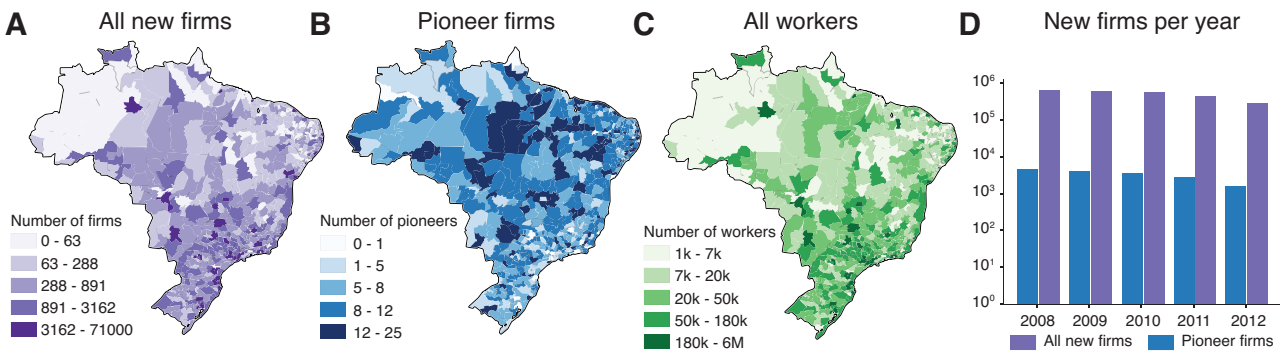


Figure 1: Spatial distribution of new firms in Brazil created between 2008 and 2012. **A** all firms, **B** only pioneer firms, and **C** distribution of workers. **D** number of firms created each year.

previous experience of a worker is to the industry and to the occupation of their new job. A worker with abundant formal schooling and experience can be classified as someone with little related experience if her work history involves occupations and industries that are unrelated to her current employment. Conversely, a worker with low formal education can be classified as having high related experience if she moves into an industry and occupation that are related to the ones she has performed previously. The dimensions of industry and occupation knowledge are not necessarily tied together, since a worker can have abundant experience in the occupation of her new job, while having very little experience in a related industry. We test the relative importance of these dimensions of knowledge relatedness for the survival and growth of pioneer firms, and compare these results with their relative importance for new firms that are not pioneers.

The idea that workers bring in knowledge into the firms they participate in is an idea that has a long tradition in organizational learning. According to Herbert Simon, organizations acquire knowledge either by the learning of its members or by ingesting new members [18]. Because pioneer firms do not start with members that can learn, the knowledge this firm has needs to come from the workers that it hires. We find that the survival of pioneer firms increases significantly when their first hires are people with industry specific knowledge, and with experience in that location, but not with occupation specific knowledge. When comparing pioneers with non-pioneers, we find that industry knowledge is significantly more important for pioneers than for non-pioneers, and that occupation specific knowledge plays a relatively more important role for non-pioneers.

There are some serious concerns relating to the endogeneity of starting a firm and of hiring. For instance, firms with more social capital may be able to hire more people from related industries. We cannot address these concerns fully, but we can instrument for the number of workers from a related industry available in a labor market by looking at national industrial shifts using a Bartik-style instrument [19]. Intuitively, the supply of related workers is higher in areas with related local industries that have received adverse national or global shocks. Our results on the importance of related knowledge are similar when we use this instrument.

Together, our results show how work histories can be used to measure the types of knowledge brought by workers into pioneer firms, and also, help uncover the relative importance of industry and occupation specific knowledge in pioneering economic activities. These results tell us that the success of the pioneering activities that promote diversification depends strongly on the move of local workers with related knowledge into these new activities.

# 1 Data

We use Brazil's RAIS (Annual Social Security Information Report) compiled by the Ministry of Labor and Employment (MET) of Brazil between 2002 and 2013. The RAIS dataset uses the National Classification of Economic Activities (CNAE) for industries, and the Brazilian Occupations Classification (CBO) for occupations, both revised by the Brazilian Institute of Geography and Statistics (IBGE).

The RAIS dataset covers about 97% of the Brazilian formal labor market [20] and contains fine-grained information about individual workers, including 5,570 municipalities (which are grouped by the IBGE into 558 microregions based on similar productive structure and spatial interaction [21]), 501 occupations, and 284 industries for more than 30 million workers each year. Location information is provided at the discrete level of each municipality, so a continuous treatment is not possible. Municipalities in Brazil are grouped by IBGE into microregions based on similar productive structure and spatial interaction [21]. Microregions are grouped into 137 mesoregions, which are grouped into 27 states, and states are grouped into 5 macroregions. All the results presented in the main text use the 3-digit level for industries, the 4-digit level for occupations, and microregions as the spatial unit of analysis. We use microregions because they provide a more stringent criteria than municipalities for identifying pioneer firms; it is easier to

be the first firm to operate in an industry inside a small municipality than inside a much larger microregion. The supplementary information provides an alternative operational definition of pioneer firms based on microregions plus their neighborhood.

One of the key characteristics of RAIS that make it so useful for research is its granularity. The variables in RAIS can be tracked down to the individual level, which makes it the most important source of information on the formal labor market dynamics in the country. The classification of industries went through a major revision between 2005 and 2006, which we solve by splitting the analysis into before and after 2006.

Unfortunately, a firm that does not declare RAIS in a particular year may not be necessarily "dead," but just facing economic problems that make it rational not to pay taxes in that year or not to appear in any official control mechanism. In fact, many firms simply freeze their activities awaiting better economic events. This will lead us to underestimate the survival rate of firms, although the exit from RAIS is surely itself an important event. Because Brazilian legislation makes it relatively easy to open a company, but relatively difficult to close one, many firms, especially small firms, often close without informing official authorities, suggesting that the exit from RAIS might be a better expression of a company's status than the official closing of the firm. Studies conducted by the IBGE and MTE estimate that the rate of underreport of firms' death range from 14% to 20% of actually closed firms. To partially address these issues, we will consider firms to be "dead" when they stop reporting for at least two consecutive years. Despite these limitations, RAIS is the main source of information on the rate of firm creation and destruction at the municipal level [20]. In fact, the Central Registry of Firms (CEMPRE) is built by IBGE and MTE based on the information available in RAIS.

## 2 Results

Pioneer firms are the basic units of economic diversification. Here, we define a pioneer firm as a firm that is new (no record of it for at least 6 years), and that operates in an industry that is new to its region (no record of the industry in the region for at least two years before the pioneer). For companies starting after 2006 we will add the extra condition that they operate for at least two consecutive years, so as to filter out small short lived firms. Because we need at least two years of work history of the a pioneer's first hires, and because CNAE went through a major revision between 2005 and 2006, we analyze only firms created either in 2005, or after 2008 (for more information see SI Appendix).

Figure 1 shows the spatial distribution for all new firms (A), pioneer firms (B), and workers (C), across Brazilian microregions between 2008 and 2012. During the observation period, Brazil produced roughly 500,000 new firms a year, of which only about 3,000 to 4,000 (less than 1%) were pioneers (Figure 1 D). For information about the industries of pioneer firms see SI Appendix.

For pioneers, all their employees are new hires, so all their initial stock of knowledge is connected to their initial workforce [18]. We base our measure of the knowledge brought in by a company's new hire on the industry and the occupation of their previous job. Because of the limited time range of the data, we consider only jobs performed during the two years before the creation of the pioneer firm. For instance, if a worker was a teller (occupation) for a telecommunication company (industry), we assume that she brings two types of knowledge to the pioneer firm: industry specific knowledge about the telecommunication industry and occupation specific knowledge about being a teller. Because different industries and different occupations vary along a continuum, we abandon the view of industry and occupation knowledge as two binary variables [22]. We instead use a continuous approach building on the literature on relatedness. For example, the industries of shoe manufacturing and shirt manufacturing are different industries, but they are similar enough that a worker moving from shoe manufacturing to shirt manufacturing should be regarded as having some industry specific knowledge about shirt manufacturing, relative to workers coming from a less related industry such as animal agriculture. The diagram presented in Figure 2 A shows a pioneer firm made of three workers: the first and third come from the same occupation, but an unrelated industry, and the second comes from a different occupation, but a related industry.

To measure the relatedness between the industry of a pioneer firm and the work histories of that firm's workers we follow the literature on relatedness and use labor flows between pairs of industries at the national level [13, 14]. Similarly, we measure relatedness for each pair of occupations by looking at labor flows among occupations across the entire Brazilian economy. Unfortunately, the CBO classification has not been successfully linked to skill compositions, so we cannot use direct measure of skill similarity. Logically, labor should flow freely between industries and occupations that require similar knowledge and not between industries and occupations that require wildly different knowledge. In fact, the relatedness measure based on labor mobility has been termed "skill relatedness" by some authors [14, 23], because individuals changing jobs will likely remain in activities that value the skills associated with their previous work.

Formally, we define the relatedness between industry $i$ and industry $i'$ as the residual of a regression explaining labor flows as a function of the size of industries and their growth rates [14]. That is, we consider a pair of industries (occupations) to be related when the labor flows

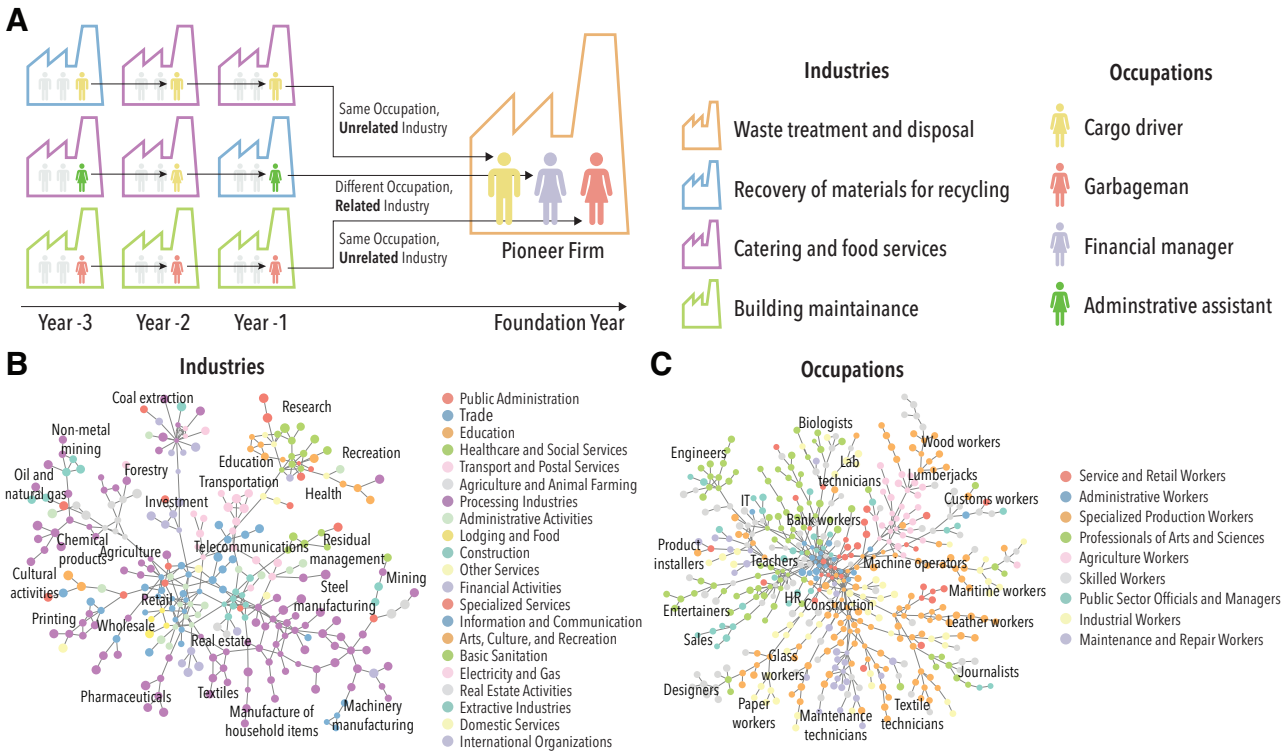


Figure 2: Work histories and networks of related activities. The diagram in **A** shows how individual work histories are used to infer the knowledge brought into the pioneer firm by its first hires. The color of each worker represents their occupation, while the color of the bounding box represents the industry. The yellow worker, for example, has experience as a cargo driver, the same occupation he was hired to perform in the pioneer firm, but comes from a very unrelated industry. The light blue worker has experience in a different occupation, but in a related industry. **B** shows the network of related industries and **C** shows the network of related occupations **C**. Node colors correspond to the highest level of the classification for occupations and industries. This figure only shows the most important edges for each network, selected based on a trimming algorithm that starts with the maximum spanning tree and then adds all edges above a threshold (see SI Appendix for details).

between them is higher than what we would expect based on the size and growth of a pair of industries. In other words, we take the residuals of the regression from Eq. 1, where $F^{(t)}_{i\leftrightarrow i'}$ is the total flow of workers in log-scale going from $i$ to $i'$ and from $i'$ to $i$ between year $t-1$ and $t$. $g^{(t)}_{ii'} = \max\{g^{(t)}_i, g^{(t)}_{i'}\}$ is the maximum growth rate in the number of employees $g^{(t)}_i = \ln L^{(t)}_i - \ln L^{(t-1)}_i$ between both industries, $\tilde{L}^{(t)}_{ii'} = \max\{L^{(t)}_i, L^{(t)}_{i'}\}$ is the maximum number of employees between both industries, in log-scale, and $L^{(t)}_i$ is the number of employees of industry $i$ in year $t$, also in log-scale. We normalize the residuals $\hat{\gamma}^{(t)}_{ii'}$ to keep them between zero and one (see Eq. 2). We measure relatedness between occupations $o$ and $o'$ in an analogous way (see Eqs. 3 and 4).

$$F^{(t)}_{i\leftrightarrow i'} = \beta_0 + \beta_1 g^{(t)}_{ii'} + \beta_2 \tilde{L}^{(t)}_{ii'} + \gamma^{(t)}_{ii'}, \tag{1}$$

$$\phi^{(t)}_{ii'} = \begin{cases} \frac{\hat{\gamma}^{(t)}_{ii'} - \min_{ii'}\{\hat{\gamma}^{(t)}_{ii'}\}}{\max_{ii'}\{\hat{\gamma}^{(t)}_{ii'}\} - \min_{ii'}\{\hat{\gamma}^{(t)}_{ii'}\}} & , \quad i \neq i' \\ 1 & , \quad i = i' \end{cases} \tag{2}$$

$$F^{(t)}_{o\leftrightarrow o'} = \beta_0 + \beta_1 g^{(t)}_{oo'} + \beta_2 \tilde{L}^{(t)}_{oo'} + \theta^{(t)}_{oo'}, \tag{3}$$

$$\psi^{(t)}_{oo'} = \begin{cases} \frac{\hat{\theta}^{(t)}_{oo'} - \min_{oo'}\{\hat{\theta}^{(t)}_{oo'}\}}{\max_{oo'}\{\hat{\theta}^{(t)}_{oo'}\} - \min_{oo'}\{\hat{\theta}^{(t)}_{oo'}\}} & , \quad o \neq o' \\ 1 & , \quad o = o' \end{cases} \tag{4}$$

Relatedness among industries and among occupations define two weighted undirected networks for each year. Figures 2 B and C show the networks of related industries and occupations for 2008, after selecting the most important edges for purpose of the visualization (see SI Appendix for details). All of our analysis are conducted with the full, time dependent, weighted networks.

Next, we use these measures of relatedness to create indicators of the stock of related knowledge that workers bring into pioneer firms. For each pioneer firm, we measure the amount of industry and occupation specific knowledge brought into it by its workers by aggregating relatedness across all its workers:

$$\Phi^{(t)}_{f,i,r} = \sum_{i'} s_{f,i'} \phi^{(t)}_{ii'} \tag{5}$$

$$\Psi^{(t)}_{f,i,r} = \sum_{o'} s_{f,o,o'} \psi^{(t)}_{oo'}, \tag{6}$$

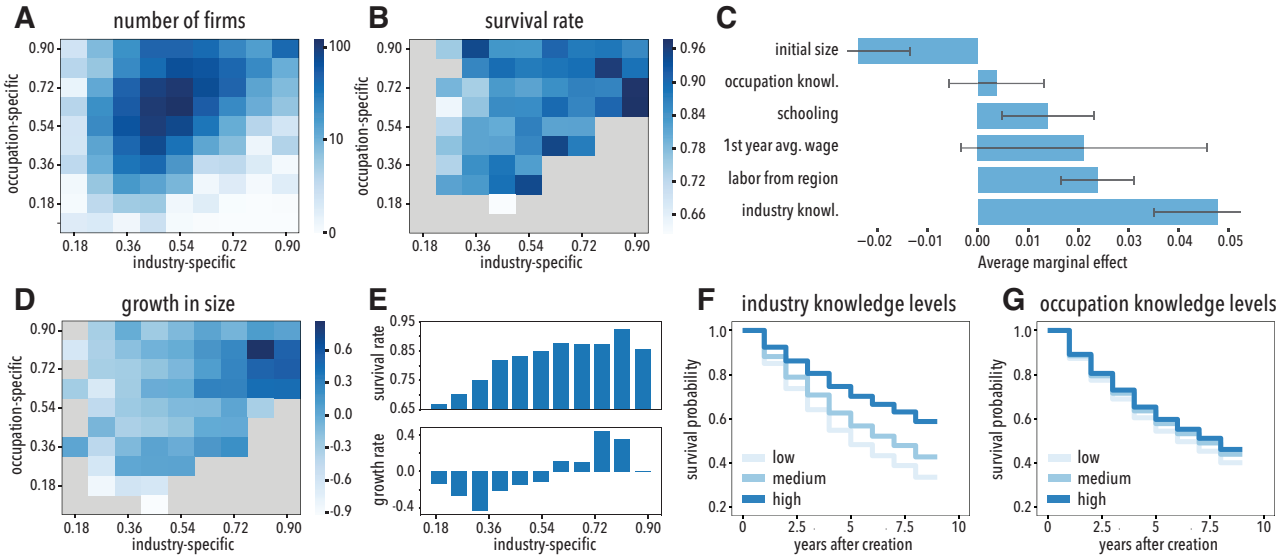


Figure 3: Characteristics of pioneer firms that started after 2008, as a function of the industry and occupation specific knowledge brought by their workers: **A** shows the number of firms observed in the data, **B** shows the empirical survival rate at the third year, **D** shows the empirical employment growth rate at the third year of firms that survived, and **E** shows survival rate and growth rate as a function of industry specific knowledge only. The gray color represents situations with not enough data points. **C** Shows the average Marginal Effect on survival of each variable from model (6) in Table 1, for firms that started after 2008. **F** Shows the predicted values for model (5) from Table 3 for firms that started in 2005, for different levels of industry knowledge: low, medium, and high. **G** is similar to F, but for different levels of occupation knowledge. In both F and G, *low* means the smallest observed value among pioneers, *medium* means the median of the observed values, and *high* means the maximum observed value.

where $s_{f,i'}$ is the fraction of workers in firm $f$ with experience on industry $i'$, and $s_{f,o,o'}$ is the fraction of workers in firm $f$ performing occupation $o$ with experience in occupation $o'$.

These two aggregate variables quantify, respectively, the industry and occupation specific knowledge that workers bring–based on their previous experience–into a pioneer firm $f$.

Figure 3 A shows a bi-variate histogram of the number of pioneer firms starting with a certain stock of industry and occupation specific knowledge. We note that the median relatedness between a pair of industries or a pair of occupations is about 0.4, so most pioneer firms hire workers with a level of industry and occupation relatedness that is much higher than if they would be hiring those workers at random. The best interpretation of this fact is that the firms and workers recognize the importance of related knowledge and search and hire accordingly. When we study the histogram we observe that pioneer firms tend to hire workers with occupation specific knowledge (top rows) but only with an intermediate level of industry specific knowledge (middle columns).

Next, we look at the pioneer firms that survive. Figure 3 B shows a bi-variate histogram for the average three year survival rate of pioneer firms. Surprisingly, the distribution of surviving firms is quite different from the distribution of all pioneer firms. While pioneer firms tend to hire workers with occupation specific knowledge, surviving pioneer firms tend to be those that hired workers with high levels of industry specific knowledge (Figure 3 B). In fact, the three year survival rate of pioneer firms increases from about 60% when workers do not have industry specific knowledge, to more than 85% when workers bring an average industrial relatedness of more than $\Phi_f > 0.5$ (Figure 3 E). Figure 3 D shows the growth in employment of surviving pioneer firms. Here we see that pioneer firms with high stocks of industry specific knowledge also grow much faster than those lacking industry specific knowledge (Figure 3 E).

We formalize these results using multivariate regression analysis that predicts the three year survival rate $S_{f,i,r}^{(t+3)}$ and employment growth $G_{f,i,r}^{(t+3)}$ of pioneer firm $f$, operating in industry $i$ and region $r$. We use logistic regression to predict the three year survival rate and OLS to predict growth. We focus on the three year survival rate as a simple way to address right censoring of our data (companies that outlive our observation period). If we were to study survival at longer time periods using a logistic model, we would have to shrink the pool of pioneer firms we can track (for alternative models see SI Appendix).

Our models for survival and growth are a function of the firm's stock of industry specific knowledge ($\Phi$), occupation specific knowledge ($\Psi$), average years of schooling of its workers ($edu$), number of initial workers ($n_0$), average wage ($w$), and local knowledge ($\rho$), which we define as the fraction of workers with work experience in the same region. In all of our models, the four knowledge

| | *Dependent variable:* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Survival rate at third year, $S^{(t+3)}$ | | | | | | Three year growth rate, $G^{(t+3)}$ | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Industry knowl. | | $0.466^{***}$ | | | | $0.457^{***}$ | | $0.174^{***}$ | | | | $0.185^{***}$ |
| ($\Phi$) | | (0.114) | | | | (0.123) | | (0.029) | | | | (0.031) |
| Occupation knowl. | | | $0.184^{**}$ | | | 0.035 | | | 0.033 | | | $-0.029$ |
| ($\Psi$) | | | (0.085) | | | (0.092) | | | (0.022) | | | (0.023) |
| Years of schooling | | | | $0.163^{*}$ | | 0.134 | | | | 0.023 | | 0.012 |
| ($edu$) | | | | (0.086) | | (0.091) | | | | (0.025) | | (0.025) |
| Local knowledge | | | | | $0.238^{***}$ | $0.228^{***}$ | | | | | 0.014 | 0.007 |
| ($\rho$) | | | | | (0.071) | (0.072) | | | | | (0.019) | (0.019) |
| Initial size | $-0.246^{***}$ | $-0.251^{***}$ | $-0.261^{***}$ | $-0.226^{**}$ | $-0.235^{**}$ | $-0.227^{**}$ | $-0.393^{***}$ | $-0.394^{***}$ | $-0.395^{***}$ | $-0.391^{***}$ | $-0.393^{***}$ | $-0.391^{***}$ |
| ($\log(n_0)$) | (0.093) | (0.095) | (0.094) | (0.092) | (0.093) | (0.096) | (0.031) | (0.030) | (0.031) | (0.031) | (0.031) | (0.030) |
| Average wage | 0.208 | 0.136 | 0.188 | 0.137 | 0.342 | 0.202 | $0.231^{***}$ | $0.209^{***}$ | $0.228^{***}$ | $0.221^{***}$ | $0.238^{***}$ | $0.208^{***}$ |
| ($\log(w)$) | (0.220) | (0.233) | (0.221) | (0.224) | (0.235) | (0.257) | (0.071) | (0.069) | (0.071) | (0.072) | (0.072) | (0.071) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 |
| McFadden | 0.2128 | 0.2265 | 0.2161 | 0.2153 | 0.2212 | 0.2367 | | | | | | |
| AICc | 1,635.9 | 1,619.1 | 1,633.9 | 1,635.0 | 1,626.6 | 1,612.7 | | | | | | |
| Log Likelihood | $-558.1$ | $-548.4$ | $-555.8$ | $-556.3$ | $-552.1$ | $-541.1$ | | | | | | |
| $R^2$ | | | | | | | 0.324 | 0.343 | 0.325 | 0.324 | 0.324 | 0.344 |
| Adjusted $R^2$ | | | | | | | 0.194 | 0.216 | 0.194 | 0.194 | 0.194 | 0.215 |
| F Statistic | | | | | | | $2.490^{***}$ (df = 222) | $2.699^{***}$ (df = 223) | $2.487^{***}$ (df = 223) | $2.481^{***}$ (df = 223) | $2.480^{***}$ (df = 223) | $2.665^{***}$ (df = 226) |
| *Note:* | | | | | | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and standard errors are in parentheses. |

Table 1: Estimates of the effect of different types of knowledge on the survival rate (models 1-6, logistic regressions) and growth rate (models 7-12, OLS) at the third year for pioneer firms. For all models reported standard errors are robust and clustered by region, and the four knowledge variables are expressed in standard deviation units.

variables ($\Phi, \Psi, edu, \rho$) are measured in units of standard deviations from their respective means, to make their coefficients more easily interpretable and comparable. Formally, our models take the form defined in Eqs. 7 and 8. The model in Eq. 7 is a logistic regression, and $\mu_i$, $\lambda^{(t)}$, and $\eta_r$ from Eqs. 7 and 8 are industry, year, and region fixed effect, respectively. Because we control for these fixed effects, our model can capture the effect of different types of human capital on firms' survival and growth, while controlling for time-invariant characteristics of industries and regions (such as the life cycle of an industry), as well as nation wide trends. Moreover, by adding the initial number of workers and the average wage of each firm, we are controlling for size effects and for the other effects regarding how attractive the jobs at each firm are.

Table 1 presents the results for both models for pioneer firms, with $\Phi$, $\Psi$, $edu$, and $\rho$ measured in standard deviation units. Across all specifications the effects of industry specific knowledge ($\Phi$) in the survival and growth of firms remains strong, whereas the effects of occupation specific knowledge ($\Psi$) and schooling ($edu$), are weak when considered in isolation, and insignificant after controlling for industry specific knowledge ($\Phi$). Figure 3 C shows the average marginal effects for model (6) from Table 1. An increase in one unit of standard deviation of industry knowledge leads to an average $\sim 5\%$ increase in the firm's probability of survival.

$$
\begin{aligned}
S_{f,i,r}^{(t+3)} &= \beta_0 + \beta_1 \Phi_{f,i,r}^{(t)} + \beta_2 \Psi_{f,i,r}^{(t)} + \beta_3 edu_{f,i,r}^{(t)} + \beta_4 \rho_{f,i,r}^{(t)} \\
&\quad + \beta_5 \log(n_{0\,f,i,r}^{(t)}) + \beta_6 \log(w_{f,i,r}^{(t)}) \\
&\quad + \mu_i + \lambda^{(t)} + \eta_r + \varepsilon_{f,i,r}^{(t)} \qquad (7) \\
G_{f,i,r}^{(t+3)} &= \beta_0 + \beta_1 \Phi_{f,i,r}^{(t)} + \beta_2 \Psi_{f,i,r}^{(t)} + \beta_3 edu_{f,i,r}^{(t)} + \beta_4 \rho_{f,i,r}^{(t)} \\
&\quad + \beta_5 \log(n_{0\,f,i,r}^{(t)}) + \beta_6 \log(w_{f,i,r}^{(t)}) \\
&\quad + \mu_i + \lambda^{(t)} + \eta_r + \varepsilon_{f,i,r}^{(t)} \qquad (8)
\end{aligned}
$$

Is industry knowledge important only for pioneer firms, or for all new firms? Table 2 shows a comparison between pioneers and other non-pioneer new firms. The industry knowledge coefficient for non-pioneers is significantly lower than for pioneers (the interaction term in model (3) is positive and significant), and for non-pioneers the occupation knowledge coefficient remains significant even when we consider it together with industry specific knowledge. Although we cannot reject the view that general knowledge and occupation related knowledge matter for both pioneers and for all firms, our results show that their effect is small compared to industry specific knowledge. In fact, the point estimate for schooling is actually larger for pioneers than for all new firms. These results suggest that industry specific knowledge is more important for pioneer firms than for new firms.

To explore the long-run impact of knowledge on survival, we focus on firms that started operating in 2005 and use the Cox Proportional Ratios model [24, 25] with a similar specification as before (Eq. 7). Since we are only using pioneers from one year, a fixed effects model would lead to model overspecification. Instead, we control for region and firm characteristics as shown in Table 3. Figures 3 F and G show the predicted values for the survival rate of pioneer firms according to model (5) from Table

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Survival rate at third year, $S^{(t+3)}$ | | | Three year growth rate, $G^{(t+3)}$ | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Industry knowl. ($\Phi$) | $0.457^{***}$ (0.123) | $0.091^{***}$ (0.007) | $0.091^{***}$ (0.007) | $0.185^{***}$ (0.031) | $0.054^{***}$ (0.002) | $0.054^{***}$ (0.002) |
| Pioneer dummy | | | 0.156 (0.126) | | | $0.088^{**}$ (0.038) |
| Industry knowl.:pioneer dummy | | | $0.203^{**}$ (0.093) | | | $0.091^{***}$ (0.027) |
| Occupation knowl. ($\Psi$) | 0.035 (0.092) | $0.035^{***}$ (0.007) | $0.036^{***}$ (0.007) | $-0.029$ (0.023) | $0.012^{***}$ (0.002) | $0.011^{***}$ (0.002) |
| Years of schooling ($edu$) | 0.134 (0.091) | 0.008 (0.007) | 0.009 (0.007) | 0.012 (0.025) | 0.002 (0.002) | 0.002 (0.002) |
| Local knowledge ($\rho$) | $0.228^{***}$ (0.072) | $0.084^{***}$ (0.006) | $0.085^{***}$ (0.006) | 0.007 (0.019) | $-0.007^{***}$ (0.002) | $-0.007^{***}$ (0.002) |
| Firm controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Firm type | pioneers | non-pioneers | all | pioneers | non-pioneers | all |
| Observations | 1,632 | 284,369 | 286,001 | 1,376 | 242,192 | 243,568 |
| McFadden | 0.2367 | 0.0404 | 0.0404 | | | |
| AICc | 1,613 | 231,739 | 233,106 | | | |
| Log Likelihood | $-541$ | $-115,638$ | $-116,320$ | | | |
| $R^2$ | | | | 0.344 | 0.152 | 0.152 |
| Adjusted $R^2$ | | | | 0.215 | 0.151 | 0.151 |
| F Statistic | | | | $2.665^{***}$ (df = 226) | $188.475^{***}$ (df = 230) | $188.274^{***}$ (df = 232) |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 and standard errors are in parentheses. | | |

Table 2: Survival and growth at the third year for pioneer firms (models 1 and 4), non-pioneer firms (models 2 and 5), and for all new firms (models 3 and 6). The interaction between industry knowledge and a dummy for pioneers is positive and significant, meaning that the effect of industry specific knowledge is larger for pioneer companies. As before, all knowledge variables are expressed in standard deviation units. Firm controls include initial size and average wage.

3, for firms with low, medium, and high level of industry knowledge (Figure 3 F) and occupation knowledge (Figure 3 G). Industry knowledge has more distinctive effects on the survival rate than occupation specific knowledge (more details in SI Appendix).

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | death probability | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Industry knowl. ($\Phi$) | $-0.214^{**}$ (0.089) | | | | $-0.181^{**}$ (0.092) |
| Occupation knowl. ($\Psi$) | | $-0.107^{*}$ (0.059) | | | $-0.038$ (0.063) |
| Years of schooling ($edu$) | | | $-0.129^{**}$ (0.057) | | $-0.105^{*}$ (0.058) |
| local knowledge ($\rho$) | | | | $-0.145^{***}$ (0.047) | $-0.144^{***}$ (0.048) |
| Region controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| Firm controls | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 462 | 462 | 462 | 462 | 462 |
| $R^2$ | 0.026 | 0.019 | 0.023 | 0.032 | 0.054 |
| Wald Test | 11.580 (df = 8) | 9.070 (df = 8) | 10.790 (df = 8) | $15.660^{**}$ (df = 8) | $25.840^{***}$ (df = 11) |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 3: Cox proportional hazards model for pioneer firms that started on 2005. Firm controls include initial size and average wage, and region controls include population, GDP per capita, average schooling, available industry knowledge, and the survival rate of non-pioneer firms as a control for how competitive the region is. As before, all knowledge variables are expressed in standard deviation units.

The endogeneity of firm entry and hiring decisions both challenge these results. Perhaps, more productive firms just tend to hire related industry workers. Perhaps, occupation related knowledge does not matter, because firms only enter when they anticipate their ability to make up for any lack in occupation related skill. We cannot address all endogeneity concerns, but we use shocks to the supply of related human capital at the local level as an instrument of hiring such workers.

Here, we construct a Bartik labor supply shock $B_{ri}$ [19, 26, 27] using the demand shocks experienced by other related industries. In other words, we use the growth or decline of industry $i'$ at the national level, as a supply shock that respectively decreases or increases the availability of the workers with industry specific knowledge required by industries related to $i'$. For instance, if the manufacturing of cars and motorcycles are related in terms of industry specific knowledge, a demand boom in the car sector would cause a shortage of workers with knowledge relevant to the manufacturing of motorcycles in the regions where the car industries are growing. Consequently, we should expect a pioneer firm in the motorcycle industry to hire less workers with industry specific knowledge when the industries related to motorcycle manufacturing are experiencing national level booms. This means the expected correlation, through this mechanism, between the Bartik instrument $B_{ri}$ and the number of workers with industry specific knowledge hired by a pioneer firm $\Phi_f$ should be *negative*.

We define the industry knowledge Bartik shock on industry $i$ in region $r$ as:

$$B_{ri}^{(ind)}(t) = \sum_{i',i'\neq i} g_{i';r}^{(t)} \frac{\phi_{ii'}^{(t)} L_{ri'}^{(t)}}{\sum_{i',i'\neq i} \phi_{ii'}^{(t)} L_{ri'}^{(t)}}, \qquad (9)$$

where $\phi_{ii'}^{(t)}$ is the relatedness between industries $i$ and $i'$, using flows between $t-1$ and $t$, $g_{i;r}^{(t)} = \log(L_{i;r}^{(t)}) - \log(L_{i;r}^{(t-1)})$ is the employment growth of industry $i$ in every region except in region $r$, and $L_{i;r}^{(t)}$ is the number of workers in year $t$ in industry $i$ removing region $r$. $L_{ri'}^{(t)}$ is the number of people working on industry $i'$ in region $r$. Eq. 9 has the same form as the original Bartik shock, since it is an interaction between the national trend ($g_{i';r}^{(t)}$) with the local industrial structure ($L_{ri'}^{(t)}$), but weighted by the similarity with industry $i$ ($\phi_{ii'}^{(t)}$).

Table 4 shows the results of using $B_{ri}^{(ind)}$ as an instrument for industry knowledge $\Phi$ to estimate the effect of industry knowledge in the growth of pioneer firms. Our two-stage least squares estimates confirm the sign of the effect found using OLS.

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Industry knowl. | Three year growth rate | | |
| | *First stage* | *Reduced form* | *Instrumental variable* | *OLS* |
| | (1) | (2) | (3) | (4) |
| Industry knowl. ($\Phi^{(t)}$) | | | $0.502^{**}$ | $0.177^{***}$ |
| | | | (0.256) | (0.032) |
| Bartik shock ($B_{ri}^{(ind)}$) | $-6.899^{***}$ | $-3.465^{**}$ | | |
| | (1.568) | (1.686) | | |
| Growth of industry ($g_{i;r}$) | $0.282^{**}$ | $-0.003$ | $-0.144$ | 0.056 |
| | (0.134) | (0.144) | (0.162) | (0.161) |
| Constant | $-0.634^{***}$ | $0.496^{***}$ | $0.814^{***}$ | $1.898^{***}$ |
| | (0.075) | (0.081) | (0.243) | (0.549) |
| Observations | 1,380 | 1,380 | 1,380 | 1,380 |
| $R^2$ | 0.016 | 0.003 | | 0.234 |
| Adj. $R^2$ | 0.015 | 0.002 | | 0.089 |
| F Statistic | $11.236^{***}$ (df = 2) | 2.129 (df = 2) | | $1.609^{***}$ (df = 220) |
| *Note:* | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 4: Results of using the Bartik shock defined in Eq. 9 as an instrument for the industry specific knowledge brought to a pioneer firm by its first hires ($\Phi$). Our two stage least squares estimates confirm the direction of the effect on growth found using OLS. The F-test for the strength of the instrument yields a statistic of $18.339^{***}$ [28]. Industry knowledge is expressed in standard deviation units.

# 3 Discussion

Here we use the entire work history of Brazil to create measures for the knowledge carried by workers into new activities and study how these different types of knowledge affect the growth and survival of pioneer firms. Pioneer firms–new firms operating in an industry that is new for the region–are of particular interest because their success represents an increase in regional economic diversification. Our work shows that industry specific knowledge is particularly important, since pioneer firms that hire workers with experience in a related industry grow faster and are more likely to survive. Surprisingly, the effect of occupation specific knowledge and general schooling are not significant for pioneer firms, while being important for newly formed non-pioneer firms.

Knowledge diffusion is acknowledged to be a key driver of economic development. In fact, countries and cities have been shown to be more likely to develop new economic activities that are similar to their existing activities [11, 13, 14, 29, 30]. This effect has proven so strong that, at the international level, less than 8 percent of the recorded diversification events between 1970 and 2010 were into unrelated products [31]. Yet, most research on industrial diversification has focused on the macro-level dynamics. Here we contribute to this body of literature by studying the micro-level mechanisms that might lead to this type of observations [32].

The idea that workers carry the knowledge that economies need to grow and diversify is not new. Yet, knowledge and human capital are usually conceptualized as measures of intensity (years of schooling for example). Our evidence suggest that knowledge is better understood in terms of relatedness since workers differ not only in their total knowledge, but also in what this knowledge is about. Here we have shown that general knowledge, measured as average years of schooling, is not a strong determinant of the survival of a pioneer firm, but that the relatedness of knowledge between past and present activities is.

Moreover, we show that for pioneer firms, industry knowledge is a stronger predictor of survival and growth than occupational knowledge. This is an unexpected finding. One explanation for this might be that the first hires of a pioneer company often end up taking some managerial role, while not operating directly as managers. For these roles, industry specific knowledge might be more important than occupation specific knowledge. Another possible explanation could be simply that industry-specific skills take longer to acquire than occupation-specific skills, and hence, firms with more in-house industry experience have an advantage at the outset.

Imagine the case of a salesperson. Salespeople are essential for the growth and survival of firms and have both occupation and industry specific knowledge. The occupation specific knowledge of a salesperson involves knowledge on how to communicate with clients, develop relationships, and close deals. These are skills that can be easily transferred from one firm to the next. The industry specific knowledge required by a salesperson, however, depends strongly on the product or service being sold. A salesperson with experience in selling garments may struggle selling enterprise software, not because she cannot develop a relationship with a client, but because she may lack the knowledge needed to understand the software needs of clients and the engineering capacity of her team. Lacking the experience needed to understand and communicate needs precisely, a salesperson without industry specific knowledge can generate misunderstandings between clients and production teams that could be disastrous for a pioneer company.

Previous work has shown that the founder's experience is a strong predictor of the performance of start-ups [33]. We do not know who the founder of the company is in our data, but we can check whether the observed effect is due to just one employee or if it is a characteristic of the team. We find that an important part of the effect is driven by the most experienced (related) employee, but that there is a significant part that is due to the rest of the team. Even after we remove the most experienced member of the team from the sample and add her as a pioneer specific control, our finding that industry specific knowledge matters remains strong. This suggests that the most experienced employee is not driving all of the observed effect (see SI Appendix).

Another explanation for our results is that workers

from related industries are more likely to have connections to clients, customers, and trustworthy workers, so what they bring is not just their knowledge about the industry, but also their knowledge of the social network where the industry is embedded in [34, 35]. This form of industry specific social capital, can be regarded as a subtype of industry specific knowledge or experience, and also, should be reflected in the locations specific knowledge of a worker, which we find is a significant predictor of the growth and survival of pioneer firms. Unfortunately, there are few data sources that can be used to isolate the effects of skills and location with the pure effects of social capital, so the effects of embeddedness are hard to identify.

These findings add to the literature studying differences between industry and occupation specific knowledge in other contexts [36, 37]. The industry knowledge brought by a firm's manager, for example, has been shown to be very important for the productivity of the firm [22, 38]. In fact, a manager's human capital has been shown to be mainly industry specific [39], in the sense that industry tenure provides a higher wage premium than occupational tenure. For other occupations such as craftsmen, human capital has been shown to be primarily occupation specific. Together with this body of literature, our study suggests that the picture where a job (an occupation for a given industry) is linked to a set of skills only through the occupation might be incomplete.

There is growing evidence of the effects of movement of industry specific human capital on the development of regions. History shows that the migration of skilled workers encourages regional development of new industries. For example, in the sixteenth century, the region around Antwerp was an industrial center for the textile industry, until the anti-Protestant persecution in the late sixteenth century triggered an exodus of Protestant workers. Many of those skilled workers moved to the northern part of the Netherlands and helped develop new textile industries in those cities [40, 41]. Similarly, other studies using pioneer plants have revealed the importance of industry specific human capital [1], but have not compared it with general knowledge or occupational knowledge.

Although our data is specific to Brazil, the great variation in income and industrialization level among Brazilian microregions suggests that our results might generalize. In fact, the richest Brazilian microregion had an average income per capita in 2013 of about USD 28k, which was comparable to that of Spain, Italy, or South Korea; while the poorest microregions had an average income of about USD 5k, which is comparable to that of Paraguay, Jamaica, or Algeria. Moreover, the vast geographic variation of wealth in Brazil makes it an interesting scenario for studying industrial development, since it combines the challenges of middle income countries with the data reporting quality of high income countries. Finally, our results emphasize that in order to fully understand the importance of tacit knowledge for regional industrial diversification it is important to measure knowledge along different dimensions. The work history of individuals may be the key to measure these different types of knowledge.

## Acknowledgements

C.J.F., B.J., and C.A.H. acknowledge support from the MIT Media Lab Consortia, the MIT Skoltech Program, the Masdar Institute of Science and Technology (Masdar Institute), Abu Dhabi, USA—Reference 02/MI/MIT/CP/11/07633/GEN/G/00 and the Center for Complex Engineering Systems (CCES) at King Abdulaziz City for Science and Technology (KACST). We thank comments from Kerstin Enflo, Jian Gao, Tarik Roukny, and Siqi Zheng, as well as data assistance from Manuel Aristarán and Elton Freitas.

## References

[1] R. Hausmann and F. Neffke, "The workforce of pioneer plants," 2016.

[2] P. M. Romer, "Endogenous technological change," *Journal of political Economy*, vol. 98, no. 5, Part 2, pp. S71–S102, 1990.

[3] R. R. Nelson and E. S. Phelps, "Investment in humans, technological diffusion, and economic growth," *The American economic review*, vol. 56, no. 1/2, pp. 69–75, 1966.

[4] R. J. Barro, "Economic growth in a cross section of countries," *The quarterly journal of economics*, vol. 106, no. 2, pp. 407–443, 1991.

[5] E. L. Glaeser, H. D. Kallal, J. A. Scheinkman, and A. Shleifer, "Growth in cities," *Journal of political economy*, vol. 100, no. 6, pp. 1126–1152, 1992.

[6] J. E. Rauch, "Productivity gains from geographic concentration of human capital: evidence from the cities," *Journal of urban economics*, vol. 34, no. 3, pp. 380–400, 1993.

[7] E. L. Glaeser, "Cities, information, and economic growth," *Cityscape*, vol. 1, no. 1, pp. 9–47, 1994.

[8] E. L. Glaeser, "The new economics of urban and regional growth," *The Oxford handbook of economic geography*, pp. 83–98, 2000.

[9] R. J. Barro, "Human capital and growth," *The American Economic Review*, vol. 91, no. 2, pp. 12–17, 2001.

[10] N. Gennaioli, R. La Porta, F. Lopez-de Silanes, and A. Shleifer, "Human capital and regional development," *The Quarterly Journal of Economics*, vol. 128, no. 1, pp. 105–164, 2012.

[11] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann, "The product space conditions the development of nations," *Science*, vol. 317, no. 5837, pp. 482–487, 2007.

[12] R. Hausmann, C. A. Hidalgo, S. Bustos, M. Coscia, A. Simoes, and M. A. Yildirim, "The atlas of economic complexity: Mapping paths to prosperity," 2014.

[13] F. Neffke, M. Henning, and R. Boschma, "How do regions diversify over time? Industry relatedness and the development of new growth paths in regions," *Economic Geography*, vol. 87, no. 3, pp. 237–265, 2011.

[14] F. Neffke and M. Henning, "Skill relatedness and firm diversification," *Strategic Management Journal*, vol. 34, pp. 297–316, Mar. 2013.

[15] D. F. Kogler, D. L. Rigby, and I. Tucker, "Mapping knowledge space and technological relatedness in US cities," *European Planning Studies*, vol. 21, no. 9, pp. 1374–1391, 2013.

[16] R. Boschma, P.-A. Balland, and D. Kogler, "The geography of inter-firm knowledge spillovers in biotech," *The Economics of Knowledge, Innovation and Systemic Technology Policy*, vol. 6, p. 147, 7.

[17] R. Muneepeerakul, J. Lobo, S. T. Shutters, A. Gomézz-Liévano, and M. R. Qubbaj, "Urban economies and occupation space: can they get "there" from "here"?," *PloS one*, vol. 8, no. 9, p. e73676, 2013.

[18] H. A. Simon, "Bounded rationality and organizational learning," *Organization science*, vol. 2, no. 1, pp. 125–134, 1991.

[19] T. Bartik, "Boon or boondoggle? the debate over state and local economic policies," *Who Benefits from State and Local Economic Development Policies*, pp. 1–16, 1991.

[20] A. Cardoso, A. Najar, M. Murat Vasconcellos, J. Levin, S. Rangel, C. Costa Ribeiro, *et al.*, "International microdata scoping studies project: Brazil," *Rio de Janeiro: Economic and Social Research Council (ESRC)*, 2007.

[21] D. d. G. IBGE, "Divisão do brasil em mesorregiões e microrregiões geográficas," tech. rep., IBGE, 1990.

[22] R. P. Castanias and C. E. Helfat, "The managerial rents model: Theory and empirical analysis," *Journal of Management*, vol. 27, no. 6, pp. 661–678, 2001.

[23] M. Delgado, M. E. Porter, and S. Stern, "Defining clusters of related industries," *Journal of Economic Geography*, vol. 16, no. 1, pp. 1–38, 2015.

[24] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate cox proportional hazards models," *Statistics in medicine*, vol. 24, no. 11, pp. 1713–1723, 2005.

[25] D. R. Cox, *Analysis of survival data*. Routledge, 2018.

[26] R. Diamond, "The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000," *The American Economic Review*, vol. 106, no. 3, pp. 479–524, 2016.

[27] O. J. Blanchard, L. F. Katz, R. E. Hall, and B. Eichengreen, "Regional evolutions," *Brookings papers on economic activity*, vol. 1992, no. 1, pp. 1–75, 1992.

[28] J. H. Stock and M. Yogo, "Testing for weak instruments in linear iv regression," 2002.

[29] R. Boschma, "Relatedness as driver of regional diversification: A research agenda," *Regional Studies*, vol. 51, no. 3, pp. 351–364, 2017.

[30] R. Boschma and K. Frenken, "The emerging empirics of evolutionary economic geography," *Journal of economic geography*, vol. 11, no. 2, pp. 295–307, 2011.

[31] F. L. Pinheiro, A. Alshamsi, D. Hartmann, R. Boschma, and C. Hidalgo, "Shooting low or high: Do countries benefit from entering unrelated activities?," *arXiv preprint arXiv:1801.05352*, 2018.

[32] K. Dopfer, J. Foster, and J. Potts, "Micro-meso-macro," *Journal of evolutionary economics*, vol. 14, no. 3, pp. 263–279, 2004.

[33] S. Shane and R. Khurana, "Bringing individuals back in: the effects of career experience on new firm founding," *Industrial and corporate Change*, vol. 12, no. 3, pp. 519–543, 2003.

[34] M. Granovetter, "Economic action and social structure: The problem of embeddedness," *American journal of sociology*, vol. 91, no. 3, pp. 481–510, 1985.

[35] B. Uzzi, "Social structure and competition in interfirm networks: The paradox of embeddedness," *Administrative science quarterly*, pp. 35–67, 1997.

[36] D. Neal, "Industry-specific human capital: Evidence from displaced workers," *Journal of labor Economics*, vol. 13, no. 4, pp. 653–677, 1995.

[37] R. Gibbons and M. Waldman, "Task-specific human capital," *The American Economic Review*, vol. 94, no. 2, pp. 203–207, 2004.

[38] E. E. Bailey and C. E. Helfat, "External management succession, human capital, and firm performance: An integrative analysis," *Managerial and decision economics*, vol. 24, no. 4, pp. 347–369, 2003.

[39] P. Sullivan, "Empirical evidence on occupation and industry specific human capital," *Labour economics*, vol. 17, no. 3, pp. 567–580, 2010.

[40] J. Israel, *Empires and Entrepots: Dutch, the Spanish Monarchy and the Jews, 1585-1713*. Bloomsbury Publishing, 1990.

[41] P. O'Brien, *Urban Achievement in Early Modern Europe: Golden Ages in Antwerp, Amsterdam and London*. Cambridge University Press, 2001.

# The role of industry, occupation, and location specific knowledge in the survival of new firms
# Supplementary Information

C. Jara-Figueroa[a], Bogang Jun[a], Edward Glaeser[b], and Cesar Hidalgo[a,1]

[a]MIT Media Lab, Massachusetts Institute of Technology
[b]Department of Economics, Harvard University
[1]To whom correspondence should be addressed. E-mail: hidalgo@mit.edu

July 17, 2018

## Contents

# 1 Data description: region, industry, and occupation classifications

This section provides descriptive statistics of the spatial aggregation used in the main text, as well as the industry and occupation classifications. More information and descriptive visualizations about Brazilian region, industries, and occuaptions at at every level of aggregation can be found in http://legacy.dataviva.info.

We use Brazil's RAIS (Annual Social Security Information Report) compiled by the Ministry of Labor and Employment (MET) of Brazil between 2002 and 2013. The RAIS dataset uses the National Classification of Economic Activities (CNAE) for industries, and the Brazilian Occupations Classificatyon (CBO) for occupations, both revised by the Brazilian Institute of Geography and Statistics (IBGE).

Information in RAIS is provided on a yearly basis and takes December 31 as the reference date. RAIS was created to administer and control access to unemployment insurance and other pecuniary benefices to workers [1], therefore workers have an incentive to report correctly. Moreover, in the years leading to 2007 the MTE instituted new control and reporting mechanisms (such as fines, and declaration through the Internet) that created strong incentives for firms to comply with the legislation that makes RAIS mandatory. To further improve data quality and ensure compliance, the MTE cross-tabulates registry information from many other official sources, such as the Ministry of Social Security, the Federal Reserve and the Secretary of Federal Revenues (tax authority). As a consequence, MTE estimates that RAIS is annually declared by 98% to 99% of officially existing firms [1].

## 1.1 Microregions

Brazil is divided into 27 states, which are divided into 558 microregions, which are divided into 5570 municipalities. The revision we use here dates back to 1990. We focus on microregions because they span two main indicators where selected to identify microregions: productive structure and spatial interaction [2]. The former implies the analysis of the primary productive structure based on the use of land for agriculture, structure of establishments, productive relationships, technological level and use of capital, and the degree of diversification of farming. The definition relied on data from different sourced, most of them generated between 1980 and 1990.

There are a total of 5,559 municipalities that stay throughout the whole period, and 558 microregions. Figure 1-A shows Brazil divided into microregions, and Figure 1-B shows the state of Sao Paulo divided into municipalities. Table 1 provides basic statistics for properties of both municipalities and microregions, and Figure 2 shows the distribution of the same properties.

## 1.2 Industries

Brazil classifies economic activities according to the National Classification of Economic Activites (CNAE). For example, a retailer of appliances and stereos and video is classified as:

- Section G: Trade
- Division 47: Retail Trade
- Group 475: Retailing of computer and communication equipment; equipment and household goods
- Class 4753-9: Retail Sale Specialized in Appliances and Audio/Video Equipment

For our analysis we use the 3 digit level (group) because at this level the industries are different enough, while still providing a stringent criteria for pioneer firms. Moreover, we can be less concerned about reporting errors coming from the firms not classifying their industry correctly. We identify a major change in the industry classification between 2005 and 2006, so our study is separated into before and after 2006.

Figure 3 shows how workers are distributed over industries at the 2-digit level (division), and at the 1-digit level (sections).

## 1.3 Occupations

The Brazilian Classification of Occupations (CBO) describes and orders occupations according to occupational characteristics that relate to the nature of the workforce (functions, tasks and obligations, etc.) and the content of their work (knowledge, abilities, personal attributes and other requirements required for the occupation).

The Ministry of Labor and Employment is responsible for the management and maintenance of the Brazilian Classification of Occupations. The Brazilian Classification of Occupations underwent an intense revision at the end

of the 1990s. Here we use the resulting new version, CBO-2002, with approximately 10 Major Groups, 47 Major Subgroups, and 596 Basic Groups or Occupational Families. In particular, we use the 4-digit level.

Figure 4 shows the distribution of the workforce according to their occupation at the 2-digit level and 1-digit level (inset).

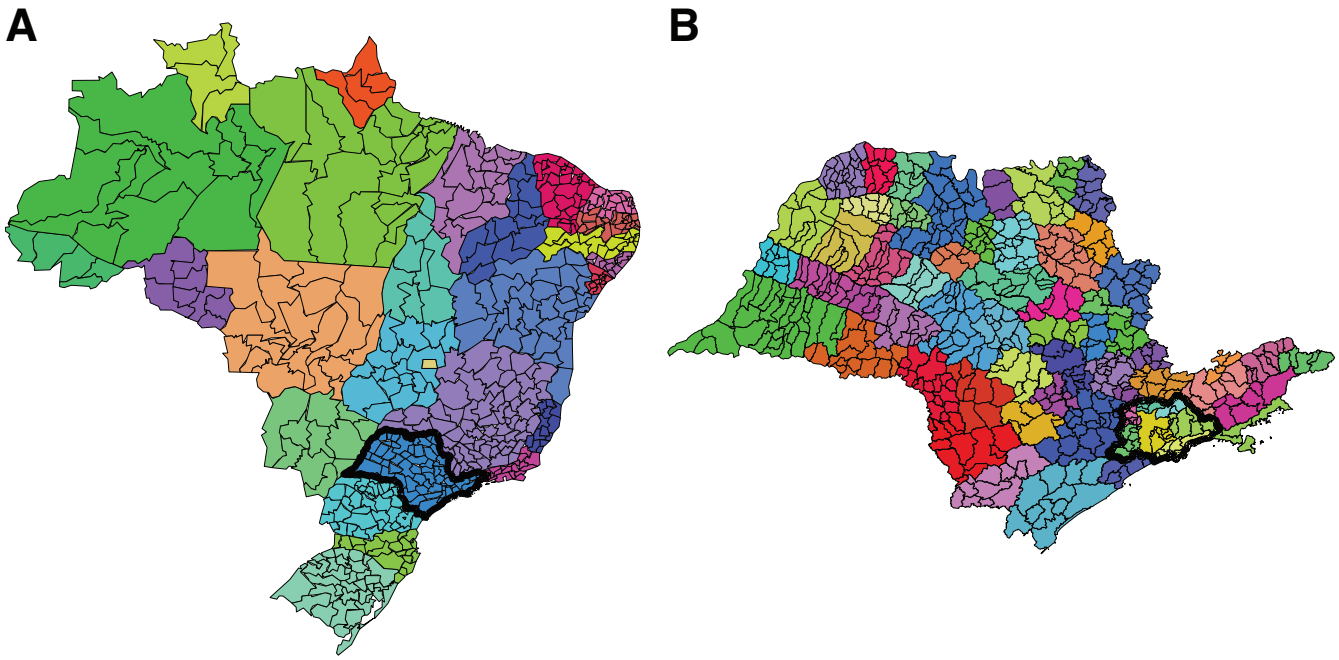


Figure 1: **A** Brazil broken down into states (colors) and microregions (lines); the bold line highlights the state of Sao Paulo. **B** State of Sao Paulo broken down into microregions (colors) and municipalities (lines); the bold line highlights the Sao Paulo mesoregion ("Metropolitana de São Paulo").

| | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| **municipalities:** | | | | | | | |
| Area in km$^2$ | 1.526737e+03 | 5.612542e+03 | 3.565000e+00 | 2.043285e+02 | 4.179250e+02 | 1.026961e+03 | 1.595333e+05 |
| Population | 3.430826e+04 | 2.032201e+05 | 8.050000e+02 | 5.235500e+03 | 1.094300e+04 | 2.356650e+04 | 1.125350e+07 |
| GDP (2010) | 6.781691e+08 | 7.205289e+09 | 7.238000e+06 | 4.452200e+07 | 9.334100e+07 | 2.383680e+08 | 4.436001e+11 |
| GDP per capita (2010) | 1.285257e+04 | 1.485368e+04 | 5.373629e+02 | 5.198712e+03 | 9.819067e+03 | 1.546605e+04 | 3.502279e+05 |
| Average wage (2010) | 1.008558e+03 | 2.773883e+02 | 3.711337e+02 | 8.404957e+02 | 9.540446e+02 | 1.105579e+03 | 5.016095e+03 |
| Education (2010) | 5.963182e+00 | 6.212517e-01 | 2.794101e+00 | 5.588236e+00 | 5.990881e+00 | 6.345485e+00 | 8.988571e+00 |
| number of workers (2010) | 7.927249e+03 | 8.252258e+04 | 1.000000e+00 | 4.500000e+02 | 9.760000e+02 | 2.849500e+03 | 4.873339e+06 |
| **microregions:** | | | | | | | |
| Area in km$^2$ | 1.520991e+04 | 2.948728e+04 | 1.701700e+01 | 2.862540e+03 | 5.562477e+03 | 1.586167e+04 | 3.322373e+05 |
| Population | 3.417915e+05 | 8.778486e+05 | 2.630000e+03 | 9.999300e+04 | 1.738025e+05 | 2.977802e+05 | 1.380483e+07 |
| GDP (2010) | 6.756169e+09 | 2.807352e+10 | 3.363200e+07 | 8.644472e+08 | 1.780288e+09 | 4.249122e+09 | 5.284293e+11 |
| GDP per capita (2010) | 1.395872e+04 | 9.855354e+03 | 3.077026e+03 | 6.215985e+03 | 1.230507e+04 | 1.809657e+04 | 7.029008e+04 |
| Average wage (2010) | 1.116772e+03 | 3.239281e+02 | 6.431218e+02 | 8.938365e+02 | 1.040908e+03 | 1.230896e+03 | 3.703355e+03 |
| Education (2010) | 6.053168e+00 | 4.262041e-01 | 3.936645e+00 | 5.814534e+00 | 6.095327e+00 | 6.328801e+00 | 7.197148e+00 |
| Number of workers (2010) | 7.897415e+04 | 3.120022e+05 | 8.660000e+02 | 9.931500e+03 | 2.071600e+04 | 5.144600e+04 | 5.671684e+06 |

Table 1: Descriptive statistics for municipalities and microregions.

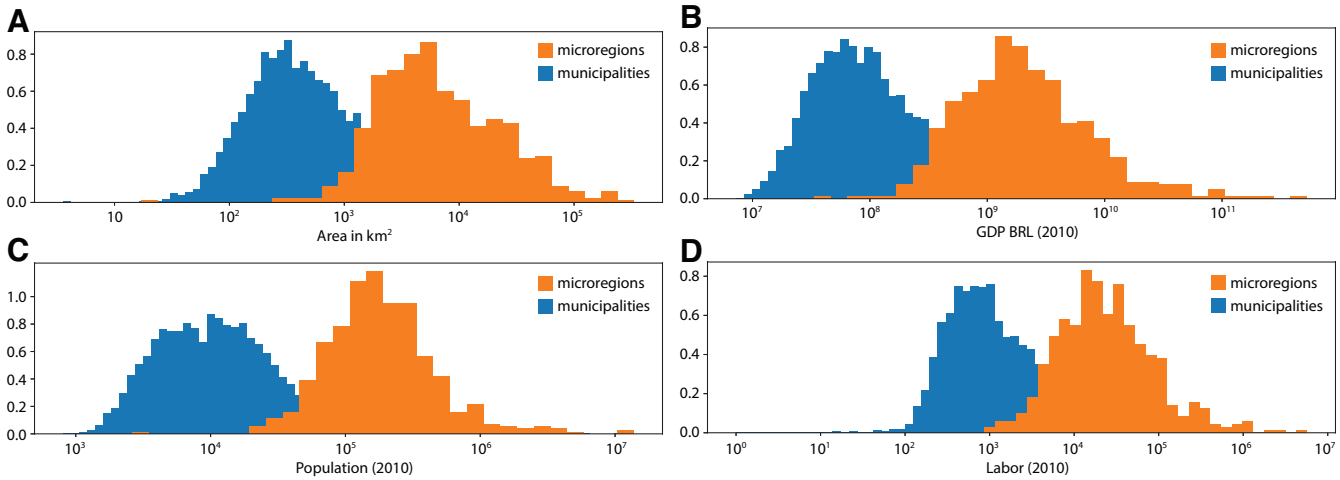


Figure 2: Distribution of characteristics of municipalities and microregions. **A** area, **B** GDP in Brazilian Reais, **C** population, and **D** number of workers.

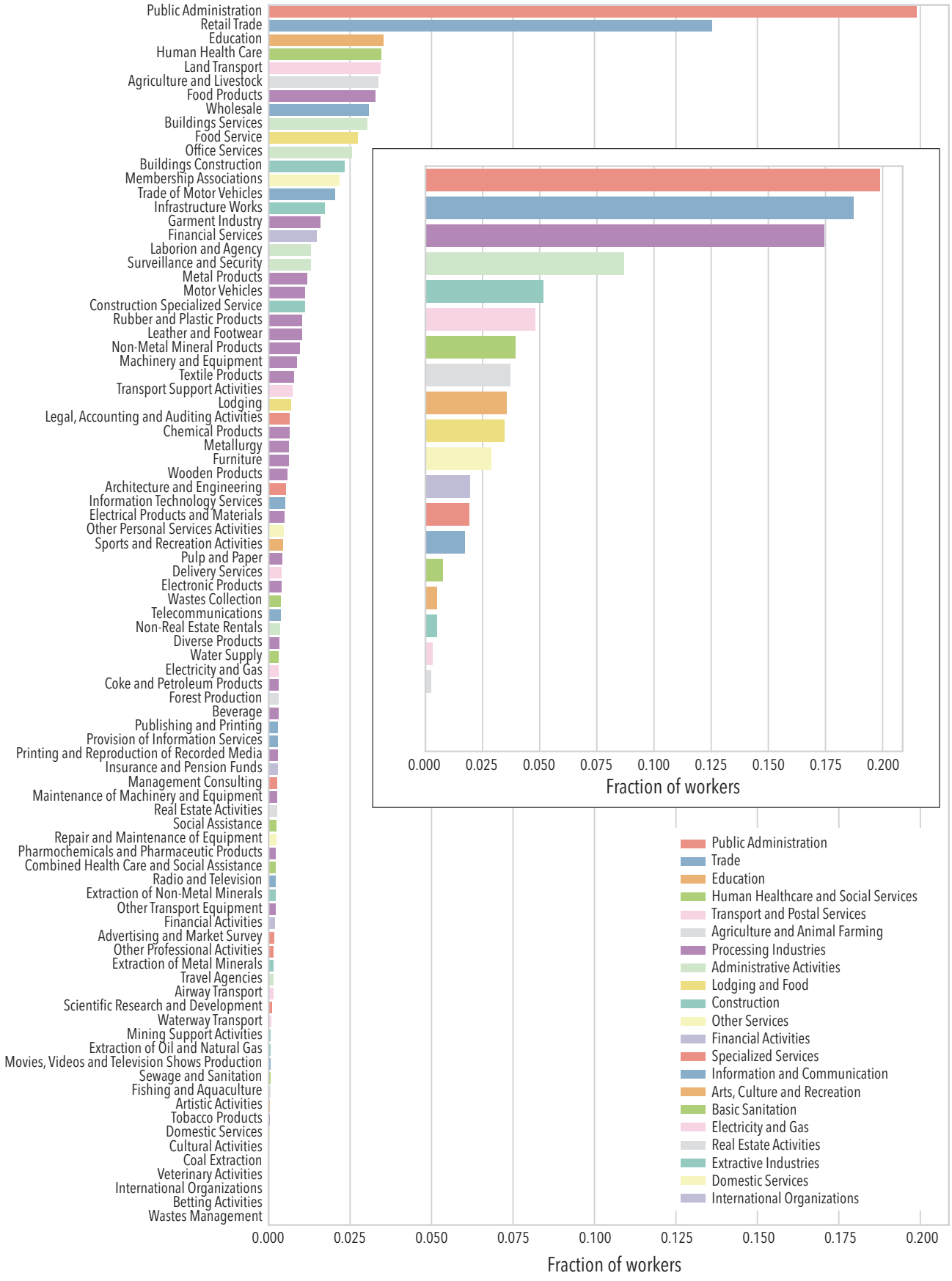


Figure 3: Distribution of workers in industries (average for the 2006-2013 period).

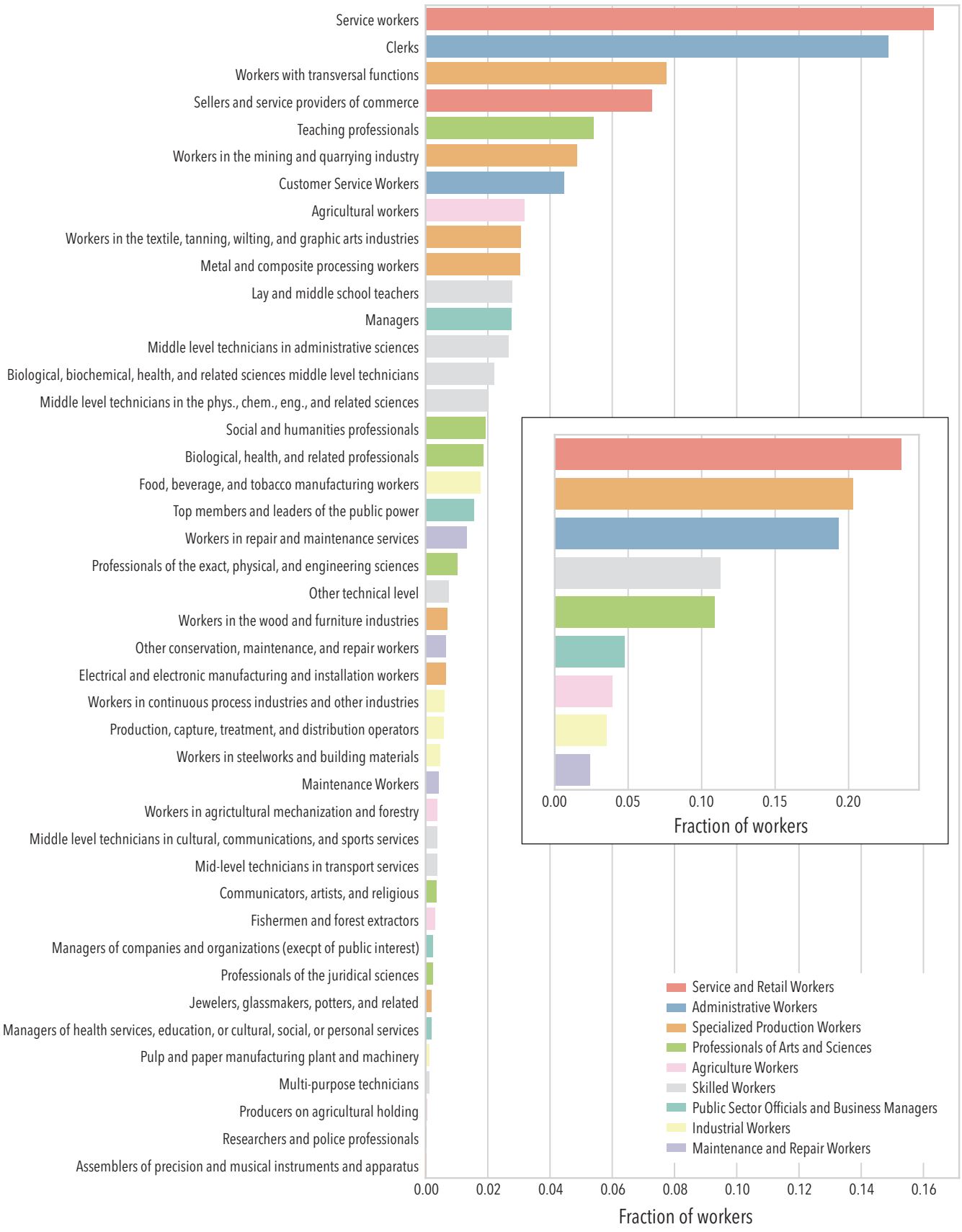


Figure 4: Distribution of workers in occupations (average for the 2006-2013 period).

# 2 Networks of relatedness

## 2.1 Relatedness coefficient

The results presented in the main text rely on a measure of relatedness between industries and occupations built using labor mobility following [3]. The result of the method described in the main text is a weighted directed network for each year starting on 2007 (because the mobility of workers between 2006 and 2007 is used to calculate relatedness for 2007). Figure 5 shows the Pearson correlation for the weights across different years. The networks are relatively stable over time, an important property of a network that aims to capture a property inherent to the similarity between a pair of industries.

Figures 7 and 8 show the full network for 2008 with a label for each individual node.

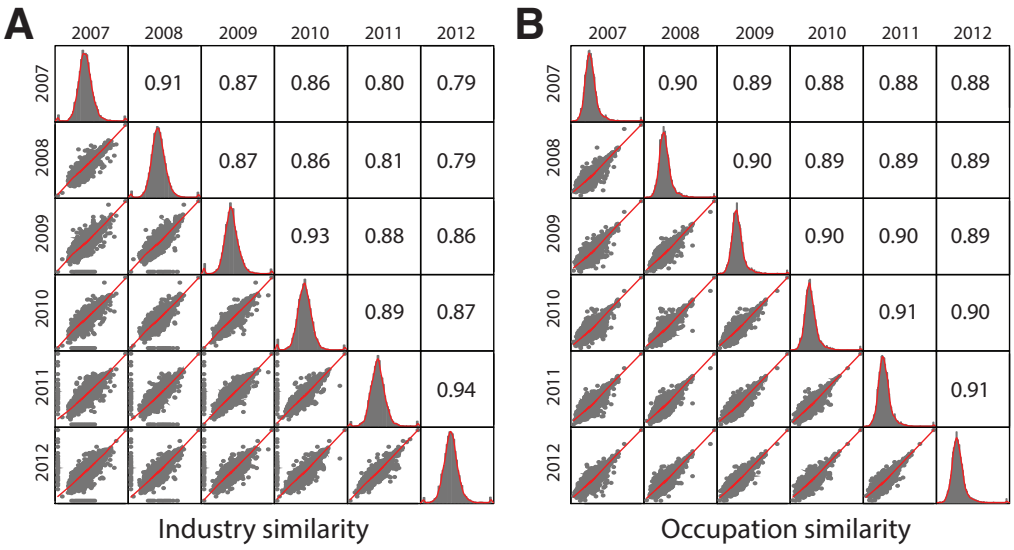


Figure 5: Pearson correlation of the relatedness coefficient across different years.

## 2.2 Alternative relatedness coefficient

Another common way of measuring relatedness between economic activities is to build a bipartite network, and project it into one of its sides [4]. Here we use this method to calculate two alternative measures of relatedness, as robustness checks. First, we calculate relatedness based on co-location of industries and co-location of occupations. Second we calculate relatedness based on co-occurrence of industries in occupations and on co-occurrence of occupations in industries.

The co-location relatedness for industries and for occupations are calculated as:

$$\phi_{ii'}^{L\ (t)} = \frac{N_{ii'}^{(t)}}{\max\{N_i^{(t)}, N_{i'}^{(t)}\}} \tag{1}$$

$$\psi_{oo'}^{L\ (t)} = \frac{N_{oo'}^{(t)}}{\max\{N_o^{(t)}, N_{o'}^{(t)}\}}, \tag{2}$$

where $N_{ii'}^{(t)}$ is the number of regions that have both industry $i$ and industry $i'$ at time $t$, $N_i^{(t)}$ is the number of regions that have industry $i$ at time $t$, $N_{oo'}^{(t)}$ is the number of regions that have workers in both occupations $o$ and $o'$, and $N_o^{(t)}$ is the number of regions that have workers performing occupation $o$. We say that a region $r$ requires industry $i$ when $RCA_{ri}^{(t)} \geq 1$ [5]:

$$RCA_{ri}^{(t)} = \frac{L_{ri}^{(t)}}{\sum_{i'} L_{ri'}^{(t)}} \Bigg/ \frac{\sum_{r'} L_{r'i}^{(t)}}{\sum_{r'i'} L_{r'i'}^{(t)}}\ , \tag{3}$$

where $L_{ri}^{(t)}$ is the number of workers in region $r$ in industry $i$ at time $t$.

In an analogous way we define measures of relatedness based on complementarity:

$$\phi^{C\ (t)}_{ii'} = \frac{\eta^{(t)}_{ii'}}{\max\{\eta^{(t)}_{i}, \eta^{(t)}_{i'}\}} \tag{4}$$

$$\psi^{C\ (t)}_{oo'} = \frac{\eta^{(t)}_{oo'}}{\max\{\eta^{(t)}_{o}, \eta^{(t)}_{o'}\}}, \tag{5}$$

where $\eta^{(t)}_{ii'}$ is the number of occupations required by both industries $i$ and $i'$, $\eta^{(t)}_{i}$ is the number of occupations required by only industry $i$, $\eta^{(t)}_{oo'}$ is the number of industries that required both occupations $o$ and $o'$, and $\eta^{(t)}_{o}$ is the number of industries that require occupation $o$. In the same way we use $RCA^{(t)}_{ri}$ as a criteria for an industry being present in a region, we use $RCA^{(t)}_{io}$ as a criteria for when an industry requires an occupation:

$$RCA^{(t)}_{io} = \frac{L^{(t)}_{io}}{\sum_{i'} L^{(t)}_{i'o}} \Bigg/ \frac{\sum_{o'} L^{(t)}_{io'}}{\sum_{i'o'} L^{(t)}_{i'o'}}\ , \tag{6}$$

where $L^{(t)}_{io}$ is the number of workers in industry $i$ developing occupation $o$ at time $t$.

## 2.3 Pruning for visualization

All of the analysis are conducted using the whole time dependent, weighted, directed network characterized by $\phi^{(t)}_{ii'}$ for industries and $\psi^{(t)}_{oo'}$ for occupations. For visualizing the network, we follow the pruning method described in [4]. We start by building the minimum spanning tree of the weighted network to use as a skeleton, and then fill nodes until a certain threshold. For both networks we use the same threshold of 0.67, which ensures that the average degree of the visualization is between 3 and 4. To layout the nodes we use the Allegro layout implemented in Cytoscape [6].

## 2.4 Louvain clustering

We use Louvain to cluster the nodes of both networks. We run the community detection algorithm on the minimum spanning tree associated with the network, so as to make the communities independent of the chosen threshold. We emphasize the fact that we have not used the communities for any of our main results.

Louvain algorithm [7] is a heuristic method for greedy modularity optimization that returns a multi-level hierarchical scheme. The algorithm consists of two steps that are repeated iteratively. In the first step, modularity is optimized by local changes. We chose a node and calculate the change in modularity if the node joins the community of its immediate neighbors. If the change is positive then the modification remains, otherwise it is rolledback. Step two aggregates the communities obtained in step one, building a new network of communities. On the first one every node is initialized in its own community and then they are grouped only if the modularity increases, this is repeated until the modularity reaches a local optimal. In the second phase all nodes in the same community are grouped into nodes to repeat the first phase again. Here we use the highest level of the dendogram, with the largest communities. Figures 7 and 8 show the result of Louvain for the minimum spanning tree, and Figure 6 shows how the results of the Louvain algorithm converge as we approach the minimum spanning tree threshold.

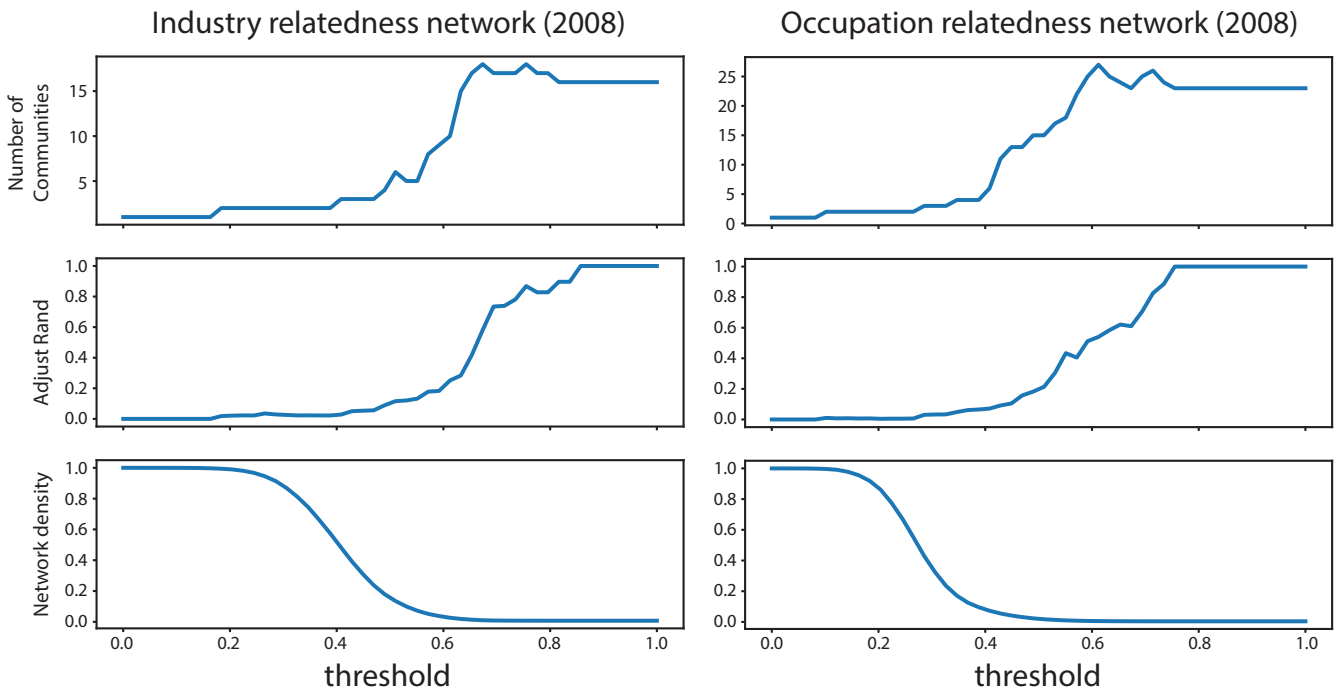


Figure 6: Threshold variation when we run Louvain algorithm on the industry relatedness network for 2008 (left) and on the occupation relatedness network (right) for 2008. The adjusted Rand index is the comparison between the communities from the network built with the given threshold and the minimum spanning tree.

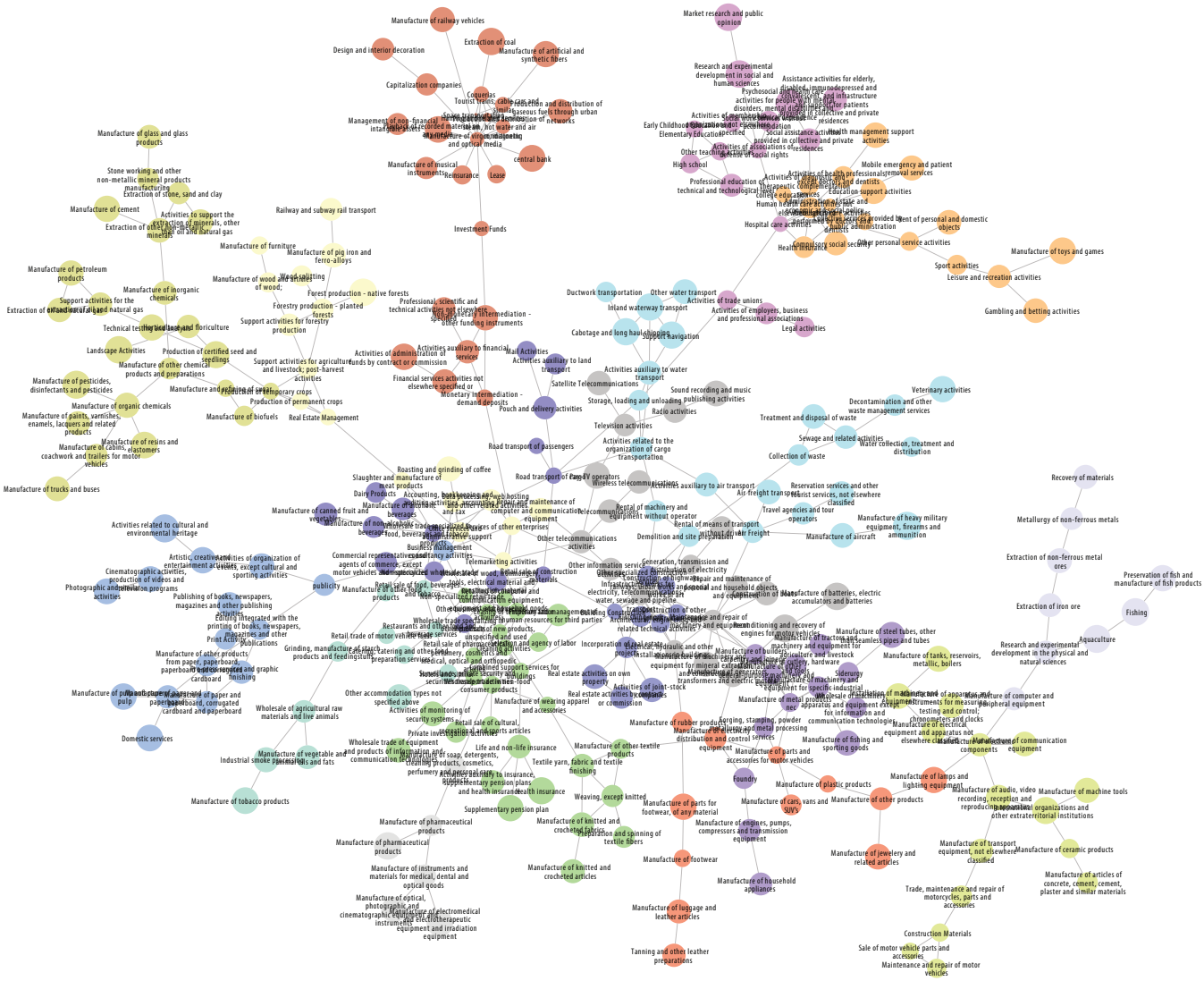


Figure 7: Network of relatedness between industries for 2008.

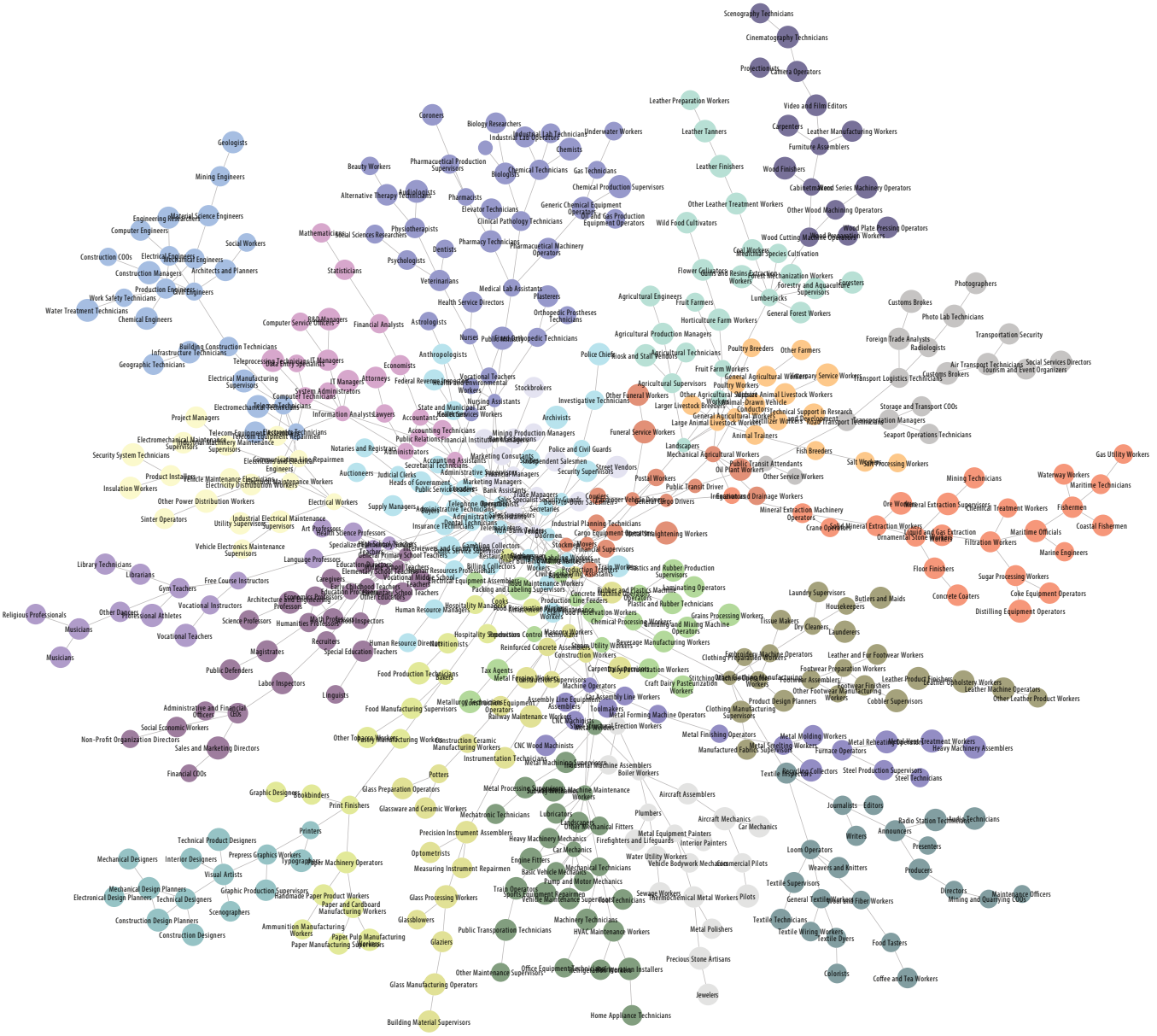


Figure 8: Network of relatedness between occupations for 2008.

# 3 The knowledge of new firms

## 3.1 Identifying pioneer firms

Pioneer firms are firms that are i) new firms, and ii) bring a new industry to their region. We operationalize these two characteristics in the following way: i) there is no record of the company for the six years before what we consider to be the starting year, and ii) there is no record of the industry being in the region for at least two years before the pioneer's starting year.

The fact that we need to track the work histories for a pioneer's first hires for at least two years before the company started, imposes some constraints on the pioneers we can study. For example, we cannot study pioneers that started on 2006 because, due to the change in industry classification between 2005 and 2006, we cannot compare the industry experience of the first hires with the industry of the pioneer. By the same token we cannot study any pioneers from 2007. The need for tracking the work history of first hires implies that we can only study pioneers that started either on 2005, 2008, 2009, 2010, 2011, 2012, or 2013.

Together with the aforementioned restriction, we add the extra constrain that the company stays alive for at least one year after its foundation. This condition is not necessary for the analysis, but if provides a more stringent criteria for the entry of companies, and is there as a way to address some of the concerns we might have regarding reporting issues. Moreover, adding this condition means that we can now not only focus on the workforce on the first year, but on the workforce during the first two years. The survival at the second year restriction means that we have to rule out pioneers that started on 2013. In the main text we show the results for companies that started after 2006, and in this SM we also show the results for 2005 pioneers without the second year restriction.

In the main text we have presented the results for the three year survival rate, which means three years **after** their starting year. Since our data is right-censored, we do not know whether companies that started in 2012 survived on their third year. This forces us to study only companies that started either on 2008, 2009, or 2010. Moreover, because in some cases companies will skip a year without that meaning that they are dead, we only consider companies to be dead if they fail to report for two consecutive years. Since we need to check for this last condition, this forces us to drop companies that started in 2010.

## 3.2 Measures of knowledge content; industry and occupation

The main text describes the two measures of knowledge content of a firm in terms of industry knowledge $\Phi$ and occupation knowledge $\Psi$ of its first hires. Here we will dive into the details of how these measures were constructed.

We define a company's first hires as all the employees that worked in the company during the first two years of its existence (or during only the first year for companies that started in 2005). Because the first hires can be unemployed on the previous year, we track their work histories going back two years before the company has started. We then compare the previous work experience with their particular position in the new company; in terms of occupation and industry. Since a worker can have performed more than one job, in different industries or occupations, we take the highest similarity between the work experience and the job she is performing in the new company. For example, say that one of the first employees of a web consulting firm founded in 2008 is a salesperson, who worked as a salesperson in an insurance company in 2007 and as a developer in a database consulting company in 2006. This employee's industry knowledge is given by the similarity between database consulting and web consulting (higher), not by the similarity between insurance and web consulting (lower). Likewise, this employee's occupation knowledge is equal to one, since she has prior experience as salesperson.

Because we are using the work experience of first hires to estimate the industry and occupation knowledge of new companies, we drop every new company that has over 75% of their workforce coming from outside the labor market. In other words, we drop companies for which we cannot certainly asses the work experience of their first hires.

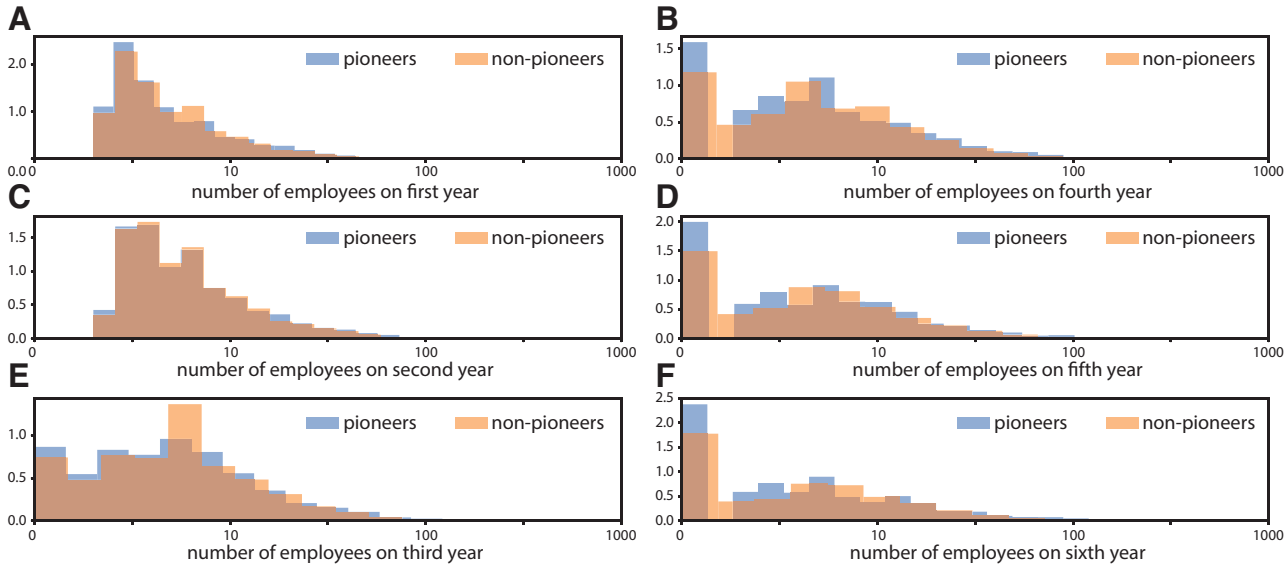


Figure 9: Distribution of number of employees for pioneer and non-pioneer new firms.

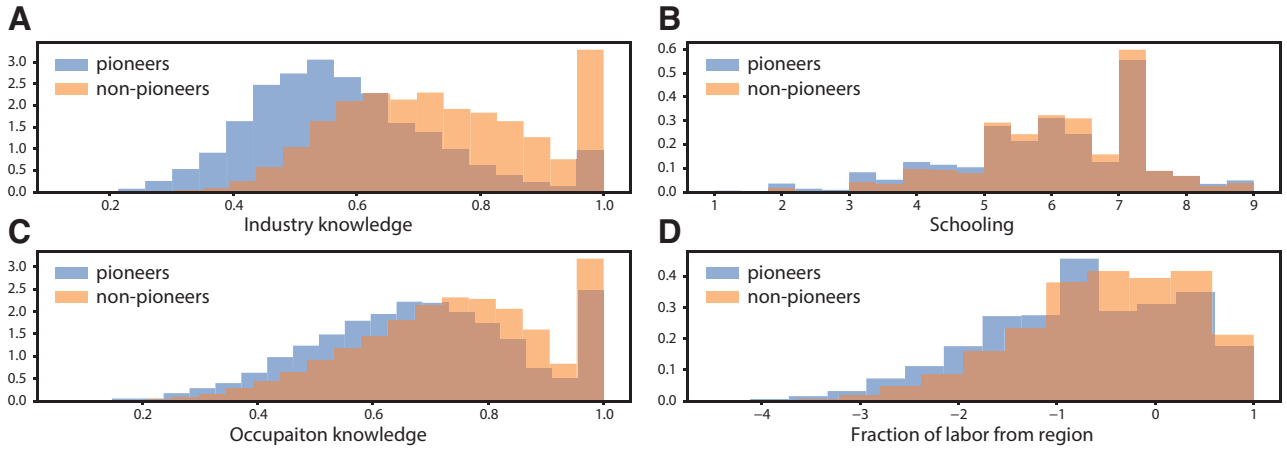


Figure 10: Distribution of knowledge measures of employees for pioneer and non-pioneer new firms.

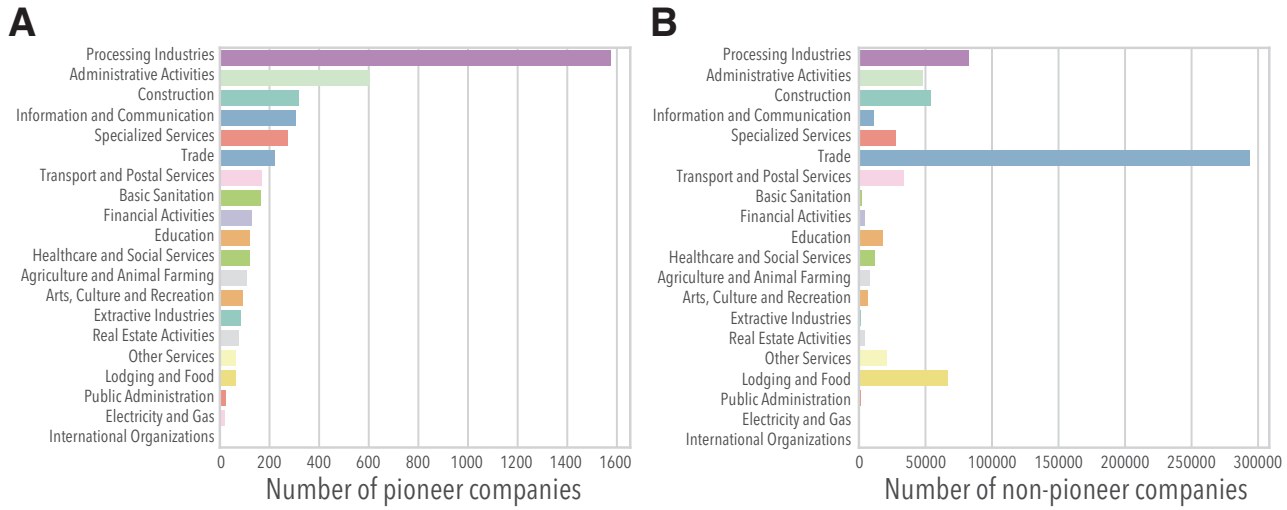


Figure 11: Distribution of pioneer and non-pioneer companies into industries.

# 4 Regressions for survival and growth

Here we expand on the analysis conducted in the main text by exploring other specifications of the logistic regression that explains the three year survival rate.

## 4.1 Interaction terms

First, we add the interaction term between industry and occupation knowledge to check whether there is a substitution effect. Table 2 shows the results including the interaction term for the three year survival rate, and Table 3 for the three year growth rate. The coefficient of the interaction term is negative and significant, implying that there exists diminishing returns. In other words, when occupation specific knowledge is sufficient enough, industry specific knowledge does not contribute as much. Figure 12 shows the average marginal effect for models (6) and (7) from Table 2. In both cases, industry knowledge remains positive and significant, providing an average 5% increase in change of survival for each additional unit of standard deviation, even after controlling for year, industry, and region characteristics.

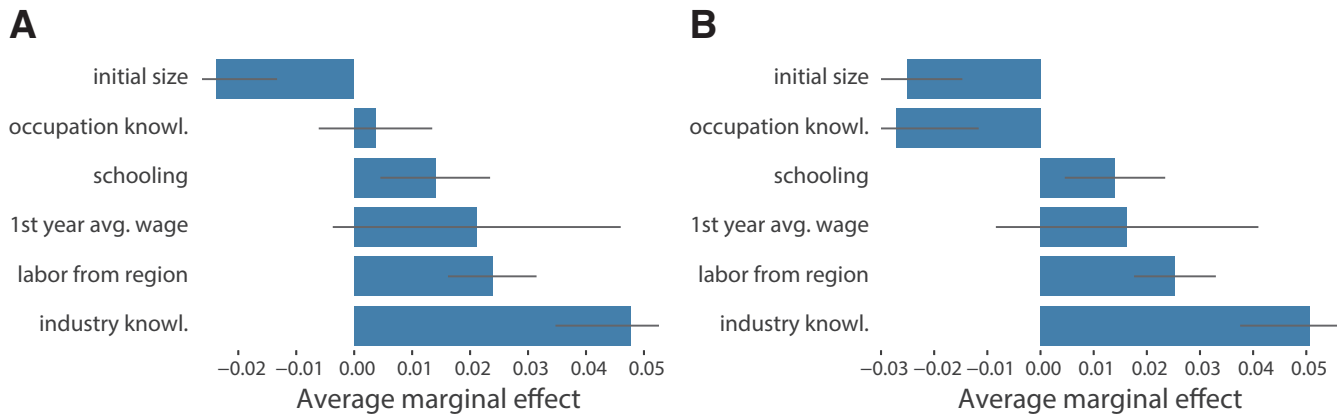


Figure 12: Average marginal effect for the survival at the third year for **A** model (6) and **B** model (7) from Table 2. Knowledge variables are normalized. The normalization is done over all new companies, not only pioneers.

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | survival at third year | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Industry knowl. | | $0.466^{***}$ | | | | $0.457^{***}$ | $0.488^{***}$ |
| ($\Phi$) | | (0.114) | | | | (0.123) | (0.124) |
| Occupation knowl. | | | $0.184^{**}$ | | | 0.035 | $-0.263$ |
| ($\Psi$) | | | (0.085) | | | (0.092) | (0.163) |
| Industry:occupation knowl. | | | | | | | $-0.245^{**}$ |
| ($\Phi \cdot \Psi$) | | | | | | | (0.098) |
| Years of schooling | | | | $0.163^{*}$ | | 0.134 | 0.135 |
| ($edu$) | | | | (0.086) | | (0.091) | (0.091) |
| Local knowledge | | | | | $0.239^{***}$ | $0.228^{***}$ | $0.244^{***}$ |
| ($\rho$) | | | | | (0.071) | (0.072) | (0.073) |
| Initial size | $-0.246^{***}$ | $-0.251^{***}$ | $-0.261^{***}$ | $-0.226^{**}$ | $-0.235^{**}$ | $-0.227^{**}$ | $-0.242^{**}$ |
| ($\log(n_0)$) | (0.093) | (0.095) | (0.094) | (0.092) | (0.093) | (0.096) | (0.097) |
| Average wage | 0.208 | 0.136 | 0.188 | 0.137 | 0.342 | 0.202 | 0.157 |
| ($\log(w)$) | (0.220) | (0.233) | (0.221) | (0.224) | (0.235) | (0.257) | (0.259) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Obs. | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 |
| McFadden | 0.2128 | 0.2265 | 0.2161 | 0.2153 | 0.22125 | 0.23674 | 0.24157 |
| AICc | 1635.926 | 1619.117 | 1633.894 | 1634.998 | 1626.58 | 1612.714 | 1608.572 |
| Log Likelihood | $-558.142$ | $-548.392$ | $-555.780$ | $-556.332$ | $-552.123$ | $-541.142$ | $-537.718$ |
| *Note:* | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 2: The effects of industry, occupation, and general knowledge on the survival rate of pioneers at the third year with the interaction term. The interaction term shows that industry and occupation knowledge are substitutes. All knowledge variables are expressed in standard deviation units.

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | growth at third year | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Industry knowl. | | $0.174^{***}$ | | | | $0.169^{***}$ | $0.175^{***}$ |
| ($\Phi$) | | (0.029) | | | | (0.029) | (0.029) |
| Occupation knowl. | | | 0.033 | | | $-0.041^{*}$ | $-0.089^{**}$ |
| ($\Psi$) | | | (0.022) | | | (0.023) | (0.037) |
| Years of schooling | | | | 0.023 | | | |
| ($edu$) | | | | (0.025) | | | |
| Local knowledge | | | | | 0.014 | | |
| ($\rho$) | | | | | (0.019) | | |
| Industry:occupation knowl. | | | | | | | $-0.046^{*}$ |
| ($\Phi \cdot \Psi$) | | | | | | | (0.024) |
| Initial size | $-0.393^{***}$ | $-0.394^{***}$ | $-0.395^{***}$ | $-0.391^{***}$ | $-0.393^{***}$ | $-0.399^{***}$ | $-0.405^{***}$ |
| ($\log(n_0)$) | (0.031) | (0.030) | (0.031) | (0.031) | (0.031) | (0.030) | (0.030) |
| Average wage | $0.231^{***}$ | $0.209^{***}$ | $0.228^{***}$ | $0.221^{***}$ | $0.238^{***}$ | $0.214^{***}$ | $0.212^{***}$ |
| ($\log(w)$) | (0.071) | (0.069) | (0.071) | (0.072) | (0.072) | (0.066) | (0.066) |
| Constant | 0.408 | 0.837 | 0.426 | 0.448 | 0.351 | 0.485 | 0.546 |
| | (0.800) | (0.768) | (0.800) | (0.799) | (0.802) | (0.602) | (0.602) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 |
| $R^2$ | 0.324 | 0.343 | 0.325 | 0.324 | 0.324 | 0.279 | 0.281 |
| Adjusted $R^2$ | 0.194 | 0.216 | 0.194 | 0.194 | 0.194 | 0.198 | 0.199 |
| F Statistic | $2.490^{***}$ (df = 222) | $2.699^{***}$ (df = 223) | $2.487^{***}$ (df = 223) | $2.481^{***}$ (df = 223) | $2.480^{***}$ (df = 223) | $3.422^{***}$ (df = 140) | $3.425^{***}$ (df = 141) |
| *Note:* | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3: The effects of industry, occupation, and general knowledge on the growth rate of pioneers at the third year, including interaction term between industry and occupation knowledge. All knowledge variables are expressed in standard deviation units.

## 4.2 Region controls

Table 4 complements our results by replacing the region fixed effects with explicit controls for region characteristics such as population, gdp per capita, average years of schooling in the region, industry knowledge available in the region, and non-pioneer survival rate. Non-pioneer survival rate is the survival rate of non-pioneer new firms in the region, and it is meant to capture how competitive the region is. Industry knowledge available in the region is calculated as the average industry relatedness of the workers in the region.

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | survival at third year | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Industry knowl. | | $0.405^{***}$ | | | | $0.351^{***}$ | $0.370^{***}$ |
| ($\Phi$) | | (0.105) | | | | (0.111) | (0.110) |
| Occupation knowl. | | | $0.213^{***}$ | | | 0.094 | $-0.144$ |
| ($\Psi$) | | | (0.076) | | | (0.082) | (0.139) |
| Industry:occupation knowl. | | | | | | | $-0.197^{**}$ |
| ($\Phi \cdot \Psi$) | | | | | | | (0.088) |
| Years of schooling | | | | $0.166^{**}$ | | $0.141^{*}$ | $0.137^{*}$ |
| ($edu$) | | | | (0.079) | | (0.082) | (0.081) |
| Local knowledge | | | | | $0.174^{***}$ | $0.151^{**}$ | $0.161^{**}$ |
| ($\rho$) | | | | | (0.064) | (0.064) | (0.064) |
| Initial size | $-0.215^{***}$ | $-0.217^{***}$ | $-0.228^{***}$ | $-0.192^{**}$ | $-0.204^{**}$ | $-0.194^{**}$ | $-0.204^{**}$ |
| ($\log(n_0)$) | (0.081) | (0.083) | (0.083) | (0.081) | (0.081) | (0.083) | (0.083) |
| Average wage | 0.103 | 0.035 | 0.069 | 0.015 | 0.193 | 0.030 | $-0.003$ |
| ($\log(w)$) | (0.183) | (0.189) | (0.185) | (0.188) | (0.190) | (0.202) | (0.201) |
| Population of region | $-0.066$ | $-0.061$ | $-0.054$ | $-0.072$ | $-0.104$ | $-0.092$ | $-0.089$ |
| | (0.101) | (0.102) | (0.100) | (0.102) | (0.102) | (0.102) | (0.103) |
| GDP per capita of region | 0.005 | 0.014 | 0.019 | $-0.012$ | $-0.065$ | $-0.058$ | $-0.052$ |
| | (0.125) | (0.127) | (0.126) | (0.125) | (0.127) | (0.129) | (0.130) |
| Average years of schooling of region | $-0.323^{*}$ | $-0.327^{*}$ | $-0.335^{*}$ | $-0.369^{*}$ | $-0.285$ | $-0.337^{*}$ | $-0.339^{*}$ |
| | (0.192) | (0.192) | (0.192) | (0.190) | (0.189) | (0.189) | (0.190) |
| Industry knowl. available in region | 0.864 | $-0.558$ | 0.986 | 1.019 | 1.171 | 0.060 | $-0.449$ |
| | (1.389) | (1.458) | (1.388) | (1.390) | (1.396) | (1.479) | (1.499) |
| Non-pioneer survival rate | 0.761 | 0.203 | 0.589 | 0.741 | 0.784 | 0.260 | 0.383 |
| | (1.666) | (1.708) | (1.695) | (1.676) | (1.671) | (1.705) | (1.714) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Obs. | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 |
| McFadden | 0.0893 | 0.1016 | 0.0950 | 0.0925 | 0.0951 | 0.1101 | 0.1138 |
| AICc | 1491.009 | 1475.824 | 1485.172 | 1488.639 | 1485.05 | 1470.591 | 1467.529 |
| Log Likelihood | $-645.694$ | $-636.975$ | $-641.648$ | $-643.382$ | $-641.588$ | $-630.967$ | $-628.302$ |
| *Note:* | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 4: The effects of industry, occupation, and general knowledge on the survival rate of pioneers at the third year, using region controls instead of region fixed effects. All knowledge variables are expressed in standard deviation units.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | growth rate at third year | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Industry knowl. | | $0.176^{***}$ | | | | $0.182^{***}$ |
| ($\Phi$) | | (0.030) | | | | (0.032) |
| Occupation knowl. | | | $0.047^{**}$ | | | $-0.017$ |
| ($\Psi$) | | | (0.022) | | | (0.023) |
| Years of schooling | | | | 0.010 | | $-0.0003$ |
| ($edu$) | | | | (0.025) | | (0.025) |
| Local knowledge | | | | | 0.023 | 0.015 |
| ($\rho$) | | | | | (0.019) | (0.019) |
| Initial size | $-0.410^{***}$ | $-0.411^{***}$ | $-0.412^{***}$ | $-0.409^{***}$ | $-0.410^{***}$ | $-0.410^{***}$ |
| ($\log(n_0)$) | (0.032) | (0.031) | (0.031) | (0.031) | (0.032) | (0.031) |
| Average wage | $0.259^{***}$ | $0.236^{***}$ | $0.254^{***}$ | $0.253^{***}$ | $0.270^{***}$ | $0.244^{***}$ |
| ($\log(w)$) | (0.070) | (0.068) | (0.070) | (0.070) | (0.070) | (0.069) |
| Population of region | 0.023 | 0.021 | 0.024 | 0.023 | 0.019 | 0.018 |
| | (0.035) | (0.035) | (0.035) | (0.035) | (0.035) | (0.035) |
| GDP per capita of region | $-0.091^{**}$ | $-0.080^{*}$ | $-0.087^{**}$ | $-0.093^{**}$ | $-0.102^{**}$ | $-0.088^{**}$ |
| | (0.042) | (0.041) | (0.042) | (0.042) | (0.043) | (0.042) |
| Average years of schooling of region | 0.072 | 0.079 | 0.070 | 0.068 | 0.074 | 0.082 |
| | (0.056) | (0.055) | (0.056) | (0.058) | (0.056) | (0.057) |
| Industry knowl. available in region | 0.541 | 0.055 | 0.569 | 0.545 | 0.563 | 0.040 |
| | (0.473) | (0.474) | (0.472) | (0.473) | (0.474) | (0.476) |
| Non-pioneer survival rate | 0.744 | 0.523 | 0.688 | 0.743 | 0.732 | 0.528 |
| | (0.577) | (0.572) | (0.578) | (0.578) | (0.578) | (0.572) |
| Constant | $-1.573$ | $-0.975$ | $-1.540$ | $-1.520$ | $-1.601$ | $-0.984$ |
| | (1.022) | (1.008) | (1.022) | (1.038) | (1.021) | (1.021) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 |
| $R^2$ | 0.222 | 0.243 | 0.224 | 0.222 | 0.223 | 0.243 |
| Adjusted $R^2$ | 0.165 | 0.187 | 0.167 | 0.165 | 0.166 | 0.186 |
| F Statistic | $3.929^{***}$ (df = 93) | $4.370^{***}$ (df = 94) | $3.932^{***}$ (df = 94) | $3.887^{***}$ (df = 94) | $3.901^{***}$ (df = 94) | $4.236^{***}$ (df = 97) |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 5: The effects of industry, occupation, and general knowledge on the growth rate of pioneers at the third year using region controls instead of region fixed effects. All knowledge variables are expressed in standard deviation units.

## 4.3 Access to capital

Here we resort to five variables that capture how easy it is for companies to access loans in the year and microregion they started. We use the number of banks at the microregion level obtained from RAIS, the Balance of Credit operations at the state level provided by Brazil's central bank [8], the average interest rate at the microregion level of all the "Indirectly contracted operations" (Operações contratadas na forma indireta automática) from BNDES [9] (the largest development bank in Brazil), and the total number of operation and the average value of each operation funded by BNDES. Table 6 shows the results for different specifications. We note that even though the coefficients are not significant, the three BNDES variables add to the predictive power of the model (higher McFadden and lower AICc). In all of the specifications, the coefficient of industry knowledge remains consistent.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | survival at third year | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Industry knowl. | $0.349^{***}$ | $0.352^{***}$ | $0.354^{***}$ | $0.337^{***}$ | $0.337^{***}$ | $0.347^{***}$ |
| ($\Phi$) | (0.111) | (0.111) | (0.112) | (0.112) | (0.113) | (0.114) |
| Occupation knowl. | 0.093 | 0.089 | 0.089 | 0.096 | 0.098 | 0.091 |
| ($\Psi$) | (0.082) | (0.082) | (0.082) | (0.084) | (0.084) | (0.084) |
| Years of schooling | $0.140^{*}$ | $0.141^{*}$ | $0.145^{*}$ | $0.139^{*}$ | $0.135^{*}$ | $0.144^{*}$ |
| ($edu$) | (0.082) | (0.082) | (0.082) | (0.082) | (0.082) | (0.082) |
| Local knowledge | $0.151^{**}$ | $0.156^{**}$ | $0.155^{**}$ | $0.165^{**}$ | $0.164^{**}$ | $0.167^{***}$ |
| ($\rho$) | (0.064) | (0.064) | (0.065) | (0.064) | (0.064) | (0.065) |
| Initial size | $-0.194^{**}$ | $-0.192^{**}$ | $-0.195^{**}$ | $-0.184^{**}$ | $-0.183^{**}$ | $-0.185^{**}$ |
| ($\log(n_0)$) | (0.083) | (0.083) | (0.083) | (0.083) | (0.084) | (0.084) |
| Average wage | 0.035 | 0.002 | 0.059 | −0.004 | −0.016 | 0.012 |
| ($\log(w)$) | (0.202) | (0.207) | (0.207) | (0.208) | (0.208) | (0.215) |
| population of region | −0.091 | −0.034 | −0.081 | −0.049 | −0.098 | −0.080 |
| | (0.102) | (0.173) | (0.103) | (0.106) | (0.140) | (0.195) |
| GDP per capita of region | −0.058 | −0.009 | −0.001 | −0.013 | −0.107 | −0.057 |
| | (0.129) | (0.176) | (0.138) | (0.140) | (0.188) | (0.202) |
| Average years of schooling of region | $-0.336^{*}$ | $-0.348^{*}$ | $-0.351^{*}$ | $-0.366^{*}$ | $-0.364^{*}$ | $-0.381^{*}$ |
| | (0.189) | (0.190) | (0.192) | (0.193) | (0.191) | (0.196) |
| Industry knowl. available in region | 0.086 | 0.048 | 0.113 | 0.355 | 0.363 | 0.398 |
| | (1.479) | (1.476) | (1.481) | (1.478) | (1.481) | (1.477) |
| Non-pioneer survival rate | 0.285 | 1.054 | 0.520 | 1.959 | 1.784 | 2.354 |
| | (1.706) | (1.890) | (1.736) | (2.343) | (2.318) | (2.370) |
| N. of banks in region | | −0.052 | | | | −0.097 |
| | | (0.182) | | | | (0.244) |
| State Balance of Credit Operations | | | −0.061 | | | −0.058 |
| | | | (0.059) | | | (0.069) |
| BNDES average interest rate | | | | 0.049 | | 0.066 |
| | | | | (0.099) | | (0.113) |
| BNDES number of operations | | | | | 0.039 | 0.110 |
| | | | | | (0.102) | (0.125) |
| BNDES average operation value | | | | | 0.069 | 0.093 |
| | | | | | (0.191) | (0.196) |
| Constant | $5.115^{*}$ | 4.084 | $5.122^{*}$ | 3.115 | 3.398 | 2.425 |
| | (2.761) | (3.315) | (2.793) | (3.169) | (3.885) | (4.340) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| McFadden | 0.1098 | 0.1112 | 0.1106 | 0.1308 | 0.1309 | 0.1324 |
| AICc | 1470.067 | 1470.435 | 1471.248 | 1442.792 | 1444.991 | 1449.733 |
| Observations | 1,629 | 1,625 | 1,629 | 1,599 | 1,599 | 1,599 |
| Log Likelihood | −630.692 | −629.726 | −630.149 | −615.792 | −615.753 | −614.701 |
| *Note:* | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 6: Survival at the third year for pioneer companies, controlling for access to loan variables: number of banks in the microregion, the Balance of Credit operations at the state level provided by Brazil's central bank, the average interest rate at the microregion level of all the "Indirectly contracted operations" from BNDES, Brazil's largest development bank, and the total number of operation and the average value of each operation funded by BNDES.

## 4.4 Separating the knowledge of the top, and the knowledge of the rest

The results presented in the main text use a simple average to aggregate over the work experience of all the employees inside a pioneer firm. Yet, this simple average has the drawback that it might be capturing the effect of most knowledgeable individual in the company. Here we explore the question of whether it is just one employee (the most experienced one) who is driving the effect, or if there is also a team level effect. Table 7 separates industry knowledge into the knowledge of individual with experience in the most related industry, and the average knowledge of all the other workers. The coefficient of the average relatedness decreases, but it remains significant, suggesting that the observed effect is not because of the most skilled employee, but rather a combination of the knowledge possessed by the team. Table 8 shows the separation between the knowledge of the most experienced individual and the rest of the team for the growth of firms.

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | survival rate at third year | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Industry knowl.: average | | $0.524^{***}$ | | | | | $0.510^{***}$ | |
| ($\Phi_{average}$) | | (0.155) | | | | | (0.165) | |
| Industry knowl.: max | | | $0.216^{**}$ | | | | | $0.181^{*}$ |
| ($\Phi_{max}$) | | | (0.091) | | | | | (0.097) |
| Industry knowl.: average removing max | | | $0.311^{***}$ | | | | | $0.352^{***}$ |
| ($\Phi_{withoutMax}$) | | | (0.119) | | | | | (0.123) |
| Occupation knowl. | | | | $0.217^{*}$ | | | 0.012 | 0.007 |
| ($\Psi$) | | | | (0.117) | | | (0.128) | (0.128) |
| Years of schooling | | | | | 0.173 | | 0.138 | 0.172 |
| ($edu$) | | | | | (0.112) | | (0.116) | (0.117) |
| Local knowledge | | | | | | $0.944^{***}$ | $0.906^{***}$ | $0.927^{***}$ |
| ($\rho$) | | | | | | (0.318) | (0.329) | (0.332) |
| Initial size | $-0.206^{**}$ | $-0.183^{*}$ | $-0.263^{**}$ | $-0.215^{**}$ | $-0.179^{*}$ | $-0.189^{*}$ | $-0.144$ | $-0.214^{**}$ |
| ($\log(n_0)$) | (0.100) | (0.100) | (0.105) | (0.101) | (0.099) | (0.099) | (0.100) | (0.103) |
| Average wage | $-0.168$ | $-0.239$ | $-0.285$ | $-0.178$ | $-0.249$ | $-0.064$ | $-0.202$ | $-0.266$ |
| ($\log(w)$) | (0.235) | (0.243) | (0.245) | (0.235) | (0.247) | (0.247) | (0.271) | (0.274) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,350 | 1,350 | 1,350 | 1,350 | 1,350 | 1,350 | 1,350 | 1,350 |
| McFadden | 0.2502 | 0.2624 | 0.2685 | 0.2533 | 0.2523 | 0.2582 | 0.2714 | 0.2784 |
| AICc | 1,363 | 1,352 | 1,348 | 1,362 | 1,363 | 1,356 | 1,350 | 1,346 |
| Log Likelihood | $-415.326$ | $-408.570$ | $-405.198$ | $-413.608$ | $-414.154$ | $-410.870$ | $-403.562$ | $-399.723$ |
| *Note:* | | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 7: Effects of industry, occupation, and general knowledge on the survival rate of pioneers at the third year, separating the most knowledgeable individual from the rest. Sample sizes are smaller than in previous regressions because we have dropped all pioneers for which we cannot distinguish an individual with more knowledge than everyone else (they all come from the same industry, for example). All knowledge variables are expressed in standard deviation units.

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | growth rate at the third year | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Industry knowl.: average | | $0.247^{***}$ | | | | | $0.239^{***}$ | |
| ($\Phi_{average}$) | | (0.038) | | | | | (0.040) | |
| Industry knowl.: max | | | $0.128^{***}$ | | | | | $0.122^{***}$ |
| ($\Phi_{max}$) | | | (0.025) | | | | | (0.027) |
| Industry knowl.: average removing max | | | $0.094^{***}$ | | | | | $0.094^{***}$ |
| ($\Phi_{withoutMax}$) | | | (0.032) | | | | | (0.032) |
| Occupation knowl. | | | | $0.103^{***}$ | | | 0.013 | 0.017 |
| ($\Psi$) | | | | (0.031) | | | (0.031) | (0.031) |
| Years of schooling | | | | | 0.002 | | $-0.012$ | $-0.0002$ |
| ($edu$) | | | | | (0.032) | | (0.031) | (0.031) |
| Local knowledge | | | | | | $0.184^{**}$ | 0.146 | 0.148 |
| ($\rho$) | | | | | | (0.093) | (0.091) | (0.091) |
| Initial size | $-0.506^{***}$ | $-0.498^{***}$ | $-0.534^{***}$ | $-0.509^{***}$ | $-0.505^{***}$ | $-0.504^{***}$ | $-0.499^{***}$ | $-0.532^{***}$ |
| ($\log(n_0)$) | (0.033) | (0.032) | (0.032) | (0.032) | (0.033) | (0.032) | (0.032) | (0.032) |
| Average wage | $0.287^{***}$ | $0.251^{***}$ | $0.237^{***}$ | $0.279^{***}$ | $0.286^{***}$ | $0.307^{***}$ | $0.273^{***}$ | $0.254^{***}$ |
| ($\log(w)$) | (0.076) | (0.072) | (0.073) | (0.075) | (0.077) | (0.076) | (0.074) | (0.074) |
| Constant | 1.421 | $1.913^{**}$ | $2.089^{**}$ | 1.472 | 1.425 | 1.114 | $1.637^{*}$ | $1.832^{**}$ |
| | (1.000) | (0.892) | (0.921) | (0.961) | (1.002) | (0.999) | (0.892) | (0.918) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,157 | 1,157 | 1,157 | 1,157 | 1,157 | 1,157 | 1,157 | 1,157 |
| $R^2$ | 0.404 | 0.427 | 0.429 | 0.409 | 0.404 | 0.406 | 0.429 | 0.431 |
| Adjusted $R^2$ | 0.263 | 0.291 | 0.293 | 0.268 | 0.262 | 0.264 | 0.291 | 0.292 |
| F Statistic | $2.862^{***}$ (df = 221) | $3.139^{***}$ (df = 222) | $3.146^{***}$ (df = 223) | $2.909^{***}$ (df = 222) | $2.847^{***}$ (df = 222) | $2.870^{***}$ (df = 222) | $3.106^{***}$ (df = 225) | $3.113^{***}$ (df = 226) |
| *Note:* | | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 8: Effects of industry, occupation, and general knowledge on the growth rate of pioneers at the third year, separating the most knowledgeable individual from the rest for growth. Sample sizes are smaller than in previous regressions because we have dropped all pioneers for which we cannot distinguish an individual with more knowledge than everyone else (they all come from the same industry, for example). All knowledge variables are expressed in standard deviation units.

## 4.5 Alternative operational definition of pioneers

Next, we present the results over a much more stringent definition of pioneer firms. Here, we take all firms that are new in a region where their industry of operation was not present in the two years before, and that there were less than 30 workers in that industry in all the neighboring regions combined. This criteria reduces significantly the number of pioneer firms, so adding fixed effects leads to over-parametrization of the model. Table 9 shows the result of the three year survival rate, and Table 10 for the three year growth rate for this very stringent definition of pioneer firms. The effect of industry knowledge is still observable even with this very reduced number of pioneer firms.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | survival rate at the third year | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Industry knowl. | | 0.395** | | | | 0.390** |
| ($\Phi$) | | (0.171) | | | | (0.181) |
| Occupation knowl. | | | 0.122 | | | 0.011 |
| ($\Psi$) | | | (0.137) | | | (0.147) |
| Years of schooling | | | | 0.064 | | 0.055 |
| ($edu$) | | | | (0.145) | | (0.148) |
| Local knowledge | | | | | 0.031 | 0.002 |
| ($\rho$) | | | | | (0.115) | (0.117) |
| Initial size | −0.136 | −0.120 | −0.139 | −0.122 | −0.134 | −0.109 |
| ($\log(n_0)$) | (0.151) | (0.154) | (0.151) | (0.154) | (0.151) | (0.157) |
| Average wage | −0.394 | −0.457 | −0.403 | −0.420 | −0.376 | −0.480 |
| ($\log(w)$) | (0.300) | (0.301) | (0.299) | (0.306) | (0.307) | (0.316) |
| Population of region | −0.287 | −0.268 | −0.279 | −0.287 | −0.294 | −0.270 |
| | (0.190) | (0.190) | (0.190) | (0.190) | (0.192) | (0.193) |
| GDP per capita of region | −0.021 | 0.014 | −0.012 | −0.026 | −0.034 | 0.009 |
| | (0.240) | (0.242) | (0.240) | (0.240) | (0.244) | (0.248) |
| Average years of schooling of region | −0.483 | −0.510 | −0.484 | −0.501 | −0.479 | −0.523 |
| | (0.365) | (0.371) | (0.363) | (0.366) | (0.365) | (0.372) |
| Industry knowl. available in region | 0.644 | −0.591 | 0.533 | 0.712 | 0.644 | −0.533 |
| | (2.568) | (2.684) | (2.587) | (2.579) | (2.571) | (2.695) |
| Non-pioneer survival rate | −5.615 | −6.017 | −5.545 | −5.512 | −5.631 | −5.926 |
| | (4.275) | (4.299) | (4.311) | (4.296) | (4.267) | (4.321) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 604 | 604 | 604 | 604 | 604 | 604 |
| McFadden | 0.1890 | 0.2000 | 0.1905 | 0.1893 | 0.1891 | 0.2003 |
| AICc | 619.109 | 616.027 | 621.028 | 621.624 | 621.745 | 624.077 |
| Log Likelihood | −211.797 | −208.902 | −211.402 | −211.700 | −211.761 | −208.831 |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 9: Effects of industry, occupation, and general knowledge on the survival rate of pioneers at the third year, using a definition of pioneers that relies on both microregion and neighboring microregions. All knowledge variables are expressed in standard deviation units.

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | growth rate at the third year | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Industry knowl. | | 0.177*** | | | | 0.187*** |
| ($\Phi$) | | (0.054) | | | | (0.057) |
| Occupation knowl. | | | 0.054 | | | −0.024 |
| ($\Psi$) | | | (0.040) | | | (0.041) |
| Years of schooling | | | | −0.063 | | −0.079* |
| ($edu$) | | | | (0.044) | | (0.044) |
| Local knowledge | | | | | 0.074** | 0.073** |
| ($\rho$) | | | | | (0.035) | (0.034) |
| Initial size | −0.379*** | −0.382*** | −0.378*** | −0.388*** | −0.377*** | −0.393*** |
| ($\log(n_0)$) | (0.058) | (0.058) | (0.059) | (0.059) | (0.058) | (0.059) |
| Average wage | 0.218** | 0.191* | 0.214** | 0.256** | 0.263** | 0.283*** |
| ($\log(w)$) | (0.108) | (0.105) | (0.108) | (0.109) | (0.107) | (0.105) |
| Population of region | −0.015 | −0.012 | −0.013 | −0.012 | −0.031 | −0.025 |
| | (0.057) | (0.057) | (0.057) | (0.056) | (0.057) | (0.057) |
| GDPp per capita of region | −0.048 | −0.016 | −0.038 | −0.042 | −0.073 | −0.037 |
| | (0.074) | (0.074) | (0.074) | (0.075) | (0.075) | (0.075) |
| Average years of schooling of region | 0.091 | 0.095 | 0.089 | 0.108 | 0.103 | 0.129 |
| | (0.111) | (0.108) | (0.112) | (0.112) | (0.115) | (0.110) |
| Industry knowl. available in region | −0.440 | −0.691 | −0.464 | −0.511 | −0.350 | −0.693 |
| | (0.939) | (0.919) | (0.941) | (0.942) | (0.937) | (0.913) |
| Non-pioneer survival rate | 1.180 | 0.911 | 1.174 | 1.161 | 1.022 | 0.721 |
| | (0.979) | (0.962) | (0.975) | (0.980) | (0.986) | (0.982) |
| Constant | −1.771 | −1.197 | −1.803 | −2.121 | −1.771 | −1.591 |
| | (1.574) | (1.576) | (1.581) | (1.611) | (1.574) | (1.625) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 510 | 510 | 510 | 510 | 510 | 510 |
| $R^2$ | 0.260 | 0.282 | 0.263 | 0.263 | 0.266 | 0.291 |
| Adjusted $R^2$ | 0.118 | 0.142 | 0.119 | 0.119 | 0.124 | 0.147 |
| F Statistic | 1.833*** (df = 82) | 2.012*** (df = 83) | 1.831*** (df = 83) | 1.832*** (df = 83) | 1.864*** (df = 83) | 2.019*** (df = 86) |
| *Note:* | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 10: Effects of industry, occupation, and general knowledge on the growth rate of pioneers at the third year, using a definition of pioneers that relies on both microregion and neighboring microregions. All knowledge variables are expressed in standard deviation units.

## 4.6 Alternative definitions of relatedness to predict survival and growth

We use the definitions of relatedness based on co-location and complementarity, defined in this SM, as a way to check that our results presented in the main text are not specific to the relatedness measure based on labor mobility. The industry knowledge and occupation knowledge of each firm are calculated following Eqs. [5] and [6] from the main text. Thus, we define the co-location industry knowledge ($\Phi^L$), the co-location occupation knowledge ($\Psi^L$), the complementarity industry knowledge ($\Phi^C$), and the complementarity occupation knowledge ($\Psi^C$). Table 11 compares the results using all measures of relatedness for survival, and Table 12 compares the results for growth.

| | *Dependent variable:* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | survival at the third year | | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Industry knowl. | | 0.466*** | | | | | | 0.457*** | | |
| ($\Phi$) | | (0.114) | | | | | | (0.123) | | |
| Occupation knowl. | | | 0.184** | | | | | 0.035 | | |
| ($\Psi$) | | | (0.085) | | | | | (0.092) | | |
| Co-location industry knowl. | | | | 0.317** | | | | | 0.234* | |
| ($\Phi^L$) | | | | (0.123) | | | | | (0.137) | |
| Co-location occupation knowl. | | | | | 0.237** | | | | 0.154 | |
| ($\Psi^L$) | | | | | (0.093) | | | | (0.102) | |
| Complementarity industry knowl. | | | | | | 0.453*** | | | | 0.414*** |
| ($\Phi^C$) | | | | | | (0.122) | | | | (0.135) |
| Complementarity occupation knowl. | | | | | | | 0.196** | | | 0.026 |
| ($\Psi^C$) | | | | | | | (0.092) | | | (0.103) |
| Years of schooling | | | | | | | | 0.134 | 0.138 | 0.113 |
| ($edu$) | | | | | | | | (0.091) | (0.090) | (0.090) |
| Local knowledge | | | | | | | | 0.228*** | 0.214*** | 0.220*** |
| ($\rho$) | | | | | | | | (0.072) | (0.071) | (0.071) |
| Initial size | −0.246*** | −0.251*** | −0.261*** | −0.246*** | −0.265*** | −0.252*** | −0.260*** | −0.227** | −0.231** | −0.228** |
| ($\log(n_0)$) | (0.093) | (0.095) | (0.094) | (0.094) | (0.094) | (0.094) | (0.094) | (0.096) | (0.094) | (0.094) |
| Average wage | 0.208 | 0.136 | 0.188 | 0.176 | 0.168 | 0.136 | 0.148 | 0.202 | 0.212 | 0.205 |
| ($\log(w)$) | (0.220) | (0.233) | (0.221) | (0.227) | (0.220) | (0.230) | (0.223) | (0.257) | (0.246) | (0.252) |
| Constant | 2.216 | 3.566* | 2.437 | 2.669 | 2.568 | 3.128 | 2.562 | 2.890 | 2.222 | 2.423 |
| | (1.843) | (2.154) | (1.833) | (1.979) | (1.868) | (2.024) | (1.862) | (2.120) | (1.970) | (2.014) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 | 1,632 |
| McFadden | 0.2128 | 0.2265 | 0.2161 | 0.2177 | 0.2174 | 0.2229 | 0.2160 | 0.2367 | 0.2291 | 0.2319 |
| AICc | 1,635.926 | 1,619.117 | 1,633.894 | 1,631.6 | 1,632.042 | 1,624.252 | 1,634.014 | 1,612.714 | 1,623.582 | 1,619.588 |
| Log Likelihood | −558.142 | −548.392 | −555.780 | −554.633 | −554.855 | −550.960 | −555.840 | −541.142 | −546.576 | −544.579 |
| *Note:* | | | | | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 11: Effects of industry, occupation, and general knowledge on the survival rate of pioneers at the third year, using different measures of industry and occupation similarity. All knowledge variables are expressed in standard deviation units.

| | *Dependent variable:* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | growth rate at third year | | | | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Industry knowl. | | 0.174*** | | | | | | 0.185*** | | |
| ($\Phi$) | | (0.029) | | | | | | (0.031) | | |
| Occupation knowl. | | | 0.033 | | | | | −0.029 | | |
| ($\Psi$) | | | (0.022) | | | | | (0.023) | | |
| Co-location industry knowl. | | | | 0.187*** | | | | | 0.193*** | |
| ($\Phi^L$) | | | | (0.033) | | | | | (0.035) | |
| Co-location occupation knowl. | | | | | 0.047* | | | | −0.014 | |
| ($\Psi^L$) | | | | | (0.024) | | | | (0.025) | |
| Complementarity industry knowl. | | | | | | 0.181 | | | | 0.195*** |
| ($\Phi^C$) | | | | | | | | | | (0.036) |
| Complementarity occupation knowl. | | | | | | | 0.040* | | | −0.029 |
| ($\Psi^C$) | | | | | | | (0.024) | | | (0.026) |
| Years of schooling | | | | | | | | 0.012 | 0.004 | 0.006 |
| ($edu$) | | | | | | | | (0.025) | (0.026) | (0.026) |
| Local knowledge | | | | | | | | 0.007 | 0.005 | 0.005 |
| ($\rho$) | | | | | | | (0.019) | (0.019) | (0.019) | |
| Initial size | −0.393*** | −0.394*** | −0.395*** | −0.391*** | −0.395*** | −0.390*** | −0.394*** | −0.391*** | −0.390*** | −0.389*** |
| ($\log(n_0)$) | (0.031) | (0.030) | (0.031) | (0.030) | (0.031) | (0.030) | (0.031) | (0.030) | (0.030) | (0.030) |
| Average wage | 0.231*** | 0.209*** | 0.228*** | 0.214*** | 0.226*** | 0.202*** | 0.219*** | 0.208*** | 0.216*** | 0.208*** |
| ($\log(w)$) | (0.071) | (0.069) | (0.071) | (0.070) | (0.071) | (0.070) | (0.072) | (0.071) | (0.072) | (0.073) |
| Constant | 0.408 | 0.837 | 0.426 | 0.664 | 0.446 | 0.718 | 0.456 | 0.843 | 0.648 | 0.699 |
| | (0.800) | (0.768) | (0.800) | (0.761) | (0.801) | (0.761) | (0.803) | (0.770) | (0.765) | (0.769) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 | 1,376 |
| $R^2$ | 0.324 | 0.343 | 0.325 | 0.341 | 0.326 | 0.340 | 0.325 | 0.344 | 0.341 | 0.340 |
| Adjusted $R^2$ | 0.194 | 0.216 | 0.194 | 0.214 | 0.195 | 0.212 | 0.195 | 0.215 | 0.212 | 0.210 |
| F Statistic | 2.490*** (df = 222) | 2.699*** (df = 223) | 2.487*** (df = 223) | 2.677*** (df = 223) | 2.495*** (df = 223) | 2.656*** (df = 223) | 2.491*** (df = 223) | 2.665*** (df = 226) | 2.636*** (df = 226) | 2.620*** (df = 226) |
| *Note:* | | | | | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 12: Effects of industry, occupation, and general knowledge on the growth rate of pioneers at the third year, using different measures of industry and occupation relatedness. All knowledge variables are expressed in standard deviation units.

# 5 Cox Proportional Hazards Model

Here we will use the Cox Proportional Hazards model as a way to address right censoring in our data, and to extend our analysis beyong the three year survival rate.

## 5.1 Firms from 2005

First, we start with the set of pioneer firms for which we have the longest survival data for; the pioneer firms that started operating in 2005. Figure 13 shows the Kaplan-Meier curve for pioneer and non-pioneer firms that started operating in 2005. Table 13 shows the estimates of the Cox proportional hazards model for this set of pioneer firms. These results confirm what we found using a logistic regression. Figure 14 shows the marginal effects at the mean for model (6) from Table 13, for varying degrees of industry knowledge and occupation knowledge.

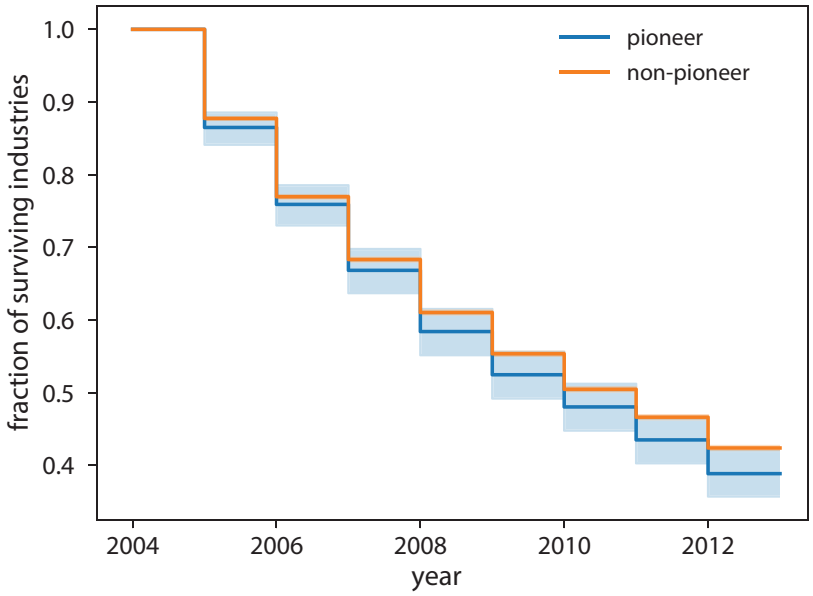


Figure 13: Kaplan-Meier Curves for pioneers and non-pioneers that started in 2005. Shaded area correspond to the 95% confidence interval of the Kaplan-Meier estimator.

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | death probability | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Industry knowl. | | $-0.214^{**}$ | | | | $-0.181^{**}$ | $-0.186^{**}$ |
| ($\Phi$) | | (0.089) | | | | (0.092) | (0.093) |
| Occupation knowl. | | | $-0.107^{*}$ | | | $-0.038$ | $-0.006$ |
| ($\Psi$) | | | (0.059) | | | (0.063) | (0.100) |
| Industry:occupation knowl. | | | | | | | 0.031 |
| ($\Phi \cdot \Psi$) | | | | | | | (0.072) |
| Years of schooling | | | | $-0.129^{**}$ | | $-0.105^{*}$ | $-0.106^{*}$ |
| ($edu$) | | | | (0.057) | | (0.058) | (0.058) |
| Local knowledge | | | | | $-0.145^{***}$ | $-0.144^{***}$ | $-0.143^{***}$ |
| ($\rho$) | | | | | (0.047) | (0.048) | (0.048) |
| Initial size | $-0.102$ | $-0.109$ | $-0.100$ | $-0.120$ | $-0.131$ | $-0.152^{*}$ | $-0.150^{*}$ |
| ($\log(n_0)$) | (0.082) | (0.083) | (0.082) | (0.083) | (0.083) | (0.084) | (0.084) |
| Average wage | $-0.087$ | $-0.026$ | $-0.068$ | $-0.038$ | $-0.153$ | $-0.046$ | $-0.039$ |
| ($\log(w)$) | (0.140) | (0.142) | (0.139) | (0.145) | (0.139) | (0.145) | (0.146) |
| Population of region | 0.006 | 0.004 | 0.008 | 0.015 | 0.027 | 0.035 | 0.038 |
| | (0.087) | (0.087) | (0.087) | (0.087) | (0.087) | (0.086) | (0.087) |
| GDP per capita of region | 0.058 | 0.020 | 0.048 | 0.051 | 0.105 | 0.062 | 0.059 |
| | (0.113) | (0.114) | (0.113) | (0.113) | (0.113) | (0.115) | (0.116) |
| Average years of schooling of region | $-0.197$ | $-0.167$ | $-0.184$ | $-0.153$ | $-0.170$ | $-0.104$ | $-0.110$ |
| | (0.141) | (0.141) | (0.141) | (0.141) | (0.140) | (0.141) | (0.142) |
| Industry knowl. available in region | $-0.848$ | 0.161 | $-0.785$ | $-0.846$ | $-1.024$ | $-0.203$ | $-0.145$ |
| | (0.932) | (1.031) | (0.928) | (0.937) | (0.925) | (1.014) | (1.027) |
| Non-pioneer survival rate | $-0.098$ | $-0.051$ | 0.025 | $-0.129$ | $-0.252$ | $-0.156$ | $-0.164$ |
| | (0.821) | (0.822) | (0.824) | (0.826) | (0.822) | (0.831) | (0.830) |
| Obs. | 462 | 462 | 462 | 462 | 462 | 462 | 462 |
| $R^2$ | 0.012 | 0.026 | 0.019 | 0.023 | 0.032 | 0.054 | 0.054 |
| Max. Possible $R^2$ | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| Log Likelihood | $-1,537$ | $-1,534$ | $-1,535$ | $-1,534$ | $-1,532$ | $-1,527$ | $-1,527$ |
| Wald Test | 5.650 | 11.580 | 9.070 | 10.790 | $15.660^{**}$ | $25.840^{***}$ | $26.320^{***}$ |
| | (df = 7) | (df = 8) | (df = 8) | (df = 8) | (df = 8) | (df = 11) | (df = 12) |
| LR Test | 5.709 | 12.150 | 9.004 | 10.685 | $15.060^{*}$ | $25.630^{***}$ | $25.813^{**}$ |
| | (df = 7) | (df = 8) | (df = 8) | (df = 8) | (df = 8) | (df = 11) | (df = 12) |
| *Note:* | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 13: Effects of industry, occupation, and general knowledge on the death rate of pioneers that started on 2005, using Cox proportional hazards model. Knowledge variables are expressed in standard deviation units.

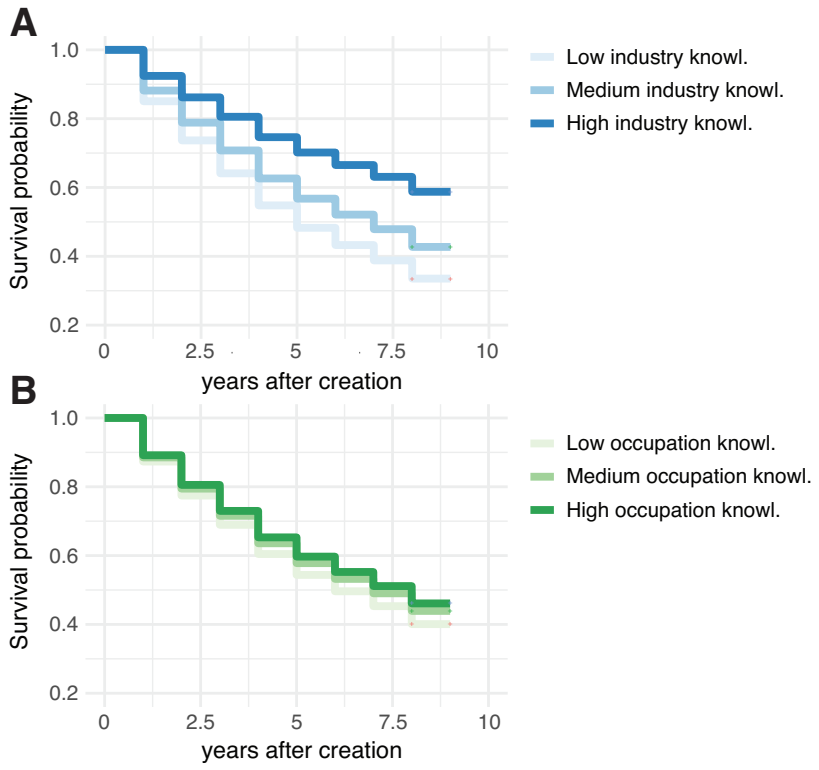


Figure 14: Cox model (6) from Table 13 for pioneer firms that started on 2005. The curves correspond to the marginal effects at the mean for the model. *low* means the smallest observed value of knowledge among pioneers, *medium* means the median of the observed values among pioneers, and *high* means the maximum observed value among pioneers.

## 5.2 Firms after 2006

Next, we run the Cox proportional hazards model for firms that started after 2008. Figure 15 shows the Kaplan Meier curves for low and high values of industry knowledge, occupation knowledge, average years of schooling, and local knowledge (fraction of workers hired from the region). Table 14 shows the result for the Cox regression. The overall results presented in the main text and in this supplementary information (Table 2) are confirmed. Figure 16 shows the marginal effect at the mean of industry and occupation knowledge for models (6) and (7) from Table 14.

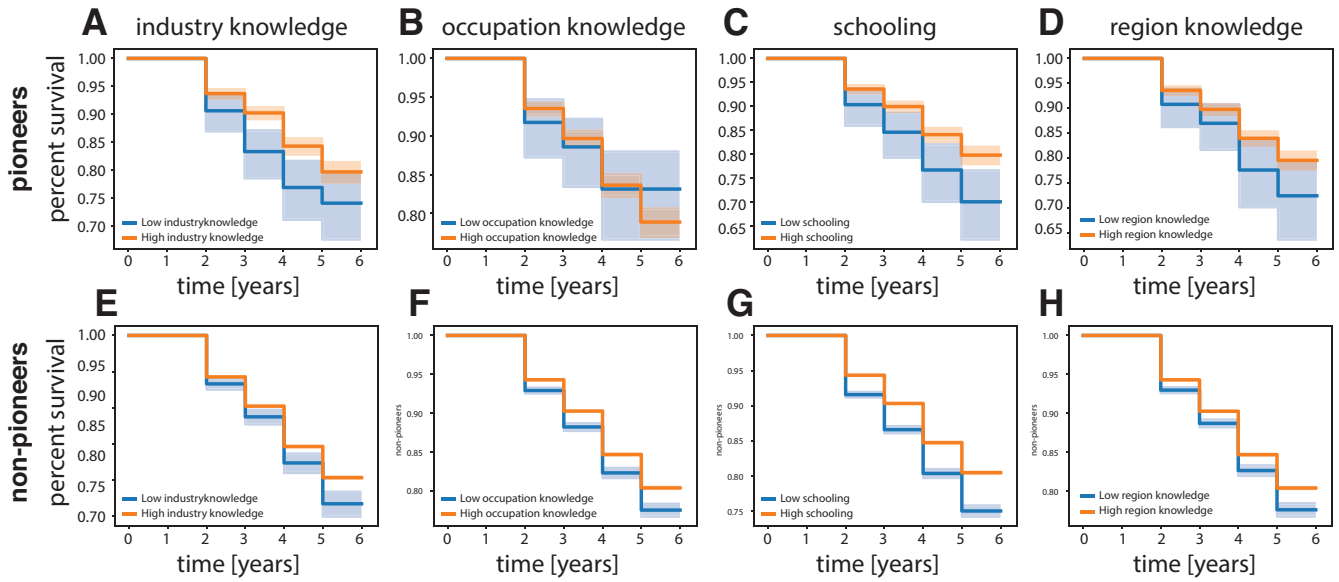


Figure 15: Kaplan-Meier Curves for pioneers and non-pioneers grouped along different knowledge dimensions. For all plots, *low* means below the 25 percentile and *high* means above the 25 percentile. Shaded area corresponds to the 95% confidence interval of the Kamplan-Meier estimator.

| | *Dependent variable:* | | | | | | |
|---|---|---|---|---|---|---|---|
| | death probability | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Industry knowl. ($\Phi$) | | $-0.244^{***}$ (0.075) | | | | $-0.182^{**}$ (0.082) | $-0.194^{**}$ (0.082) |
| Occupation knowl. ($\Psi$) | | | $-0.188^{***}$ (0.060) | | | $-0.124^{*}$ (0.065) | 0.101 (0.099) |
| Industry:occupation knowl. ($\Phi \cdot \Psi$) | | | | | | | $0.199^{***}$ (0.065) |
| Years of schooling ($edu$) | | | | $-0.143^{**}$ (0.061) | | $-0.115^{*}$ (0.062) | $-0.114^{*}$ (0.062) |
| Local knowledge ($\rho$) | | | | | $-0.178^{***}$ (0.048) | $-0.159^{***}$ (0.048) | $-0.179^{***}$ (0.048) |
| Initial size ($\log(n_0)$) | 0.075 (0.066) | 0.077 (0.066) | 0.092 (0.067) | 0.057 (0.066) | 0.066 (0.066) | 0.067 (0.067) | 0.083 (0.067) |
| Average wage ($\log(w)$) | $-0.142$ (0.158) | $-0.098$ (0.159) | $-0.116$ (0.158) | $-0.078$ (0.160) | $-0.246$ (0.162) | $-0.131$ (0.166) | $-0.114$ (0.165) |
| Year f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Industry f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Region f.e. | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Obs. | 2,684 | 2,684 | 2,684 | 2,684 | 2,684 | 2,684 | 2,684 |
| $R^2$ | 0.183 | 0.187 | 0.186 | 0.185 | 0.187 | 0.193 | 0.195 |
| Max. Possible $R^2$ | 0.879 | 0.879 | 0.879 | 0.879 | 0.879 | 0.879 | 0.879 |
| Log Likelihood | $-2{,}567.403$ | $-2{,}561.952$ | $-2{,}562.462$ | $-2{,}564.637$ | $-2{,}560.664$ | $-2{,}551.952$ | $-2{,}547.381$ |
| Wald Test | 243.400 (df = 226) | 253.270 (df = 227) | 254.170 (df = 227) | 248.610 (df = 227) | 248.220 (df = 227) | 253.690 (df = 230) | 256.460 (df = 231) |
| LR Test | $543.731^{***}$ (df = 226) | $554.632^{***}$ (df = 227) | $553.612^{***}$ (df = 227) | $549.262^{***}$ (df = 227) | $557.208^{***}$ (df = 227) | $574.633^{***}$ (df = 230) | $583.775^{***}$ (df = 231) |
| *Note:* | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 14: Effects of industry, occupation, and general knowledge on the death rate, using Cox proportional hazards model, for firms that started after 2008. All knowledge variables are expressed in standard deviation units.

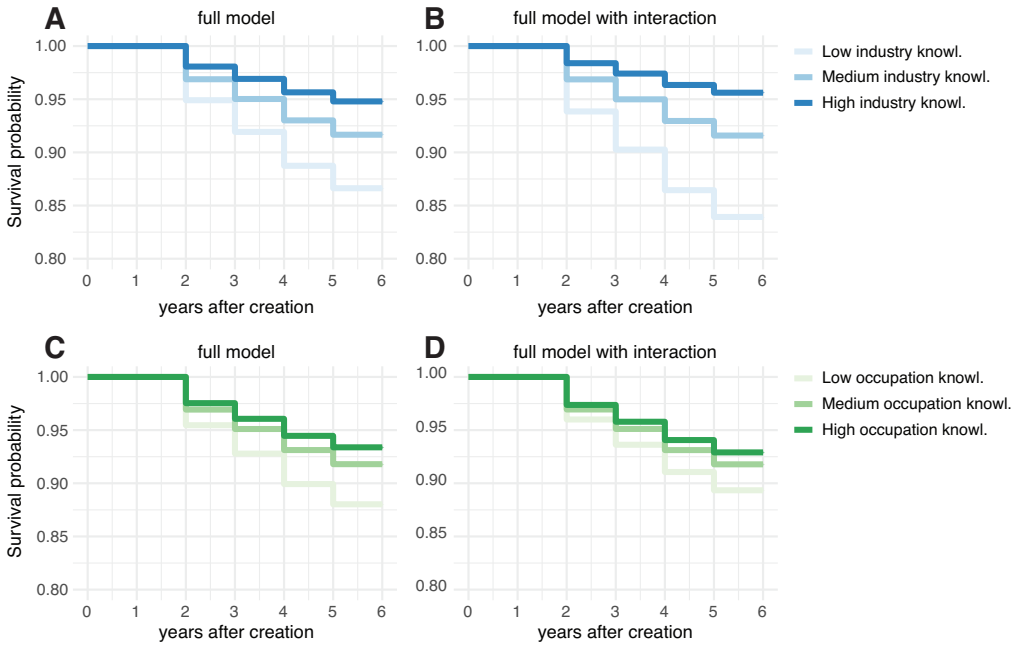


Figure 16: Cox model. Figures **A** and **C** correspond to model (6) from Table 14 and Figures **C** and **E** to model (7) from Table 14. In all cases, *low* means the smallest observed value of knowledge among pioneers, *medium* means the median of the observed values among pioneers, and *high* means the maximum observed value among pioneers.

# 6 Bartik instruments for supply of workers

The definition of the Bartik instrument that we use in the main text relies on the measure of relatedness based on labor flow $\phi_{ii'}^{(t)}$. A high positive value of $B_{ri}^{(ind)}(t)$ means that a lot of the industries related to industry $i$ that are present in region $r$ have grown at the national level. A negative value of $B_{ri}^{(ind)}(t)$ means that a lot of the industries related to industry $i$ that are present in region $r$ have declined at the national level. Small values of $B_{ri}^{(ind)}(t)$ mean that either the industries related to $i$ have not experienced significant shocks, or that these industries are not present in region $r$.

There is a legitimate concern regarding potential forward and backward linkages between industry $i$ and its related industries. In other words, the industries related to industry $i$ that are present in region $r$ might lead to growth of industry $i$ because of an increase in demand for $i$ or an increase in supply for $i$. This mechanism, however, would lead to a positive correlation between the shock and the industry knowledge in the first stage, which is something we do not observe. Yet, this will technically lead to a potential violation of the exclusion restriction.

To solve this potential violation of the exclusion restriction we use the 2010 Input-Output Matrix published by the Brazilian Institute of Geography and Statistics (IBGE) [10]. Input-Output data for Brazil is only available for 67 economic activities, a higher level of aggregation than the one we use in the main text. The lack of granularity has the drawback that if we use this classification of 67 economic activities to define pioneer firms, we significantly reduce the number of pioneers.

We use the input-output information to correct the measure of relatedness we use in the Bartik instrument, such that relatedness is now defined by labor flows after controlling for input-output. The new Bartik instrument is defined as:

$$\tilde{B}_{ri}^{(ind)}(t) = \sum_{i',i'\neq i} g_{i';r}^{(t)} \frac{\tilde{\phi}_{ii'}^{(t)} L_{ri'}^{(t)}}{\sum_{i',i'\neq i} \tilde{\phi}_{ii'}^{(t)} L_{ri'}^{(t)}}, \tag{7}$$

$$\tilde{\phi}_{ii'}^{(t)} = \begin{cases} \frac{\hat{\delta}_{ii'}^{(t)} - \min_{ii'}\{\hat{\delta}_{ii'}^{(t)}\}}{\max_{ii'}\{\hat{\delta}_{ii'}^{(t)}\} - \min_{ii'}\{\hat{\delta}_{ii'}^{(t)}\}} & , \quad i \neq i' \\ 1 & , \quad i = i' \end{cases} \tag{8}$$

where $\hat{\delta}_{ii'}^{(t)}$ is now the residual of a regression explaining labor flows as a function of the size of industries, their growth rates, and their input-output linkages [3]:

$$F_{i\leftrightarrow i'}^{(t)} = \beta_0 + \beta_1 g_{ii'}^{(t)} + \beta_2 \tilde{L}_{ii'}^{(t)} + \beta_3 X_{ii'}^{(t)} + \delta_{ii'}^{(t)}, \tag{9}$$

where $F_{i\leftrightarrow i'}^{(t)}$ is the total flow of workers in log-scale going from $i$ to $i'$ and from $i'$ to $i$ between year $t-1$ and $t$. $g_{ii'}^{(t)} = \max\{g_i^{(t)}, g_{i'}^{(t)}\}$ is the maximum growth rate in the number of employees $g_i^{(t)} = \ln L_i^{(t)} - \ln L_i^{(t-1)}$ between both industries, $\tilde{L}_{ii'}^{(t)} = \max\{L_i^{(t)}, L_{i'}^{(t)}\}$ is the maximum number of employees between both industries, in log-scale, and $L_i^{(t)}$ is the number of employees of industry $i$ in year $t$, also in log-scale. In this supplementary material we have added $X_{ii'}^{(t)}$, which is the logarithm of the maximum between $\frac{x_{i\to i'}}{x_{i\to}}$, $\frac{x_{i\to i'}}{x_{\to i'}}$, $\frac{x_{i'\to i}}{x_{i'\to}}$, and $\frac{x_{i'\to i}}{x_{\to i}}$, as done in [11], where $x_{i\to i'}$ is the output of $i$ going to $i'$, $x_{\to i}$ is the input of $i$, and $x_{i\to}$ is the output of $i$.

Table 15 shows the result using the Bartik shock adjusted using input-output data for pioneers defined as in the main text, and for pioneers defined using only the 67 sectors from the input-output data. The instrument is weak in both cases, and the standard errors are high. Yet, the point estimate confirms the direction of the relation we report in the main text.

| | *Dependent variable:* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Industry knowl. | Three year growth rate | | | Industry knowl. | Three year growth rate | | |
| | *First stage* | *Reduced form* | *Instrumental variable* | *OLS* | *First stage* | *Reduced form* | *Instrumental variable* | *OLS* |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Industry knowl. ($\Phi$) | | | 2.325 (2.029) | $0.468^{***}$ (0.157) | | | 0.513 (2.474) | $0.692^{**}$ (0.331) |
| Bartik shock adjusted ($\tilde{B}_{ri}$) | $-0.912^{***}$ (0.301) | $-2.120$ (1.768) | | | $-1.449^{**}$ (0.603) | $-0.743$ (3.607) | | |
| Growth of industry ($g_{i;r}$) | 0.097 (0.075) | $-0.904^{**}$ (0.441) | $-1.128^{**}$ (0.467) | $-1.027^{**}$ (0.432) | 0.112 (0.113) | $-0.027$ (0.673) | $-0.084$ (0.665) | $-0.094$ (0.652) |
| Constant | $0.683^{***}$ (0.014) | $0.492^{***}$ (0.084) | $-1.096$ (1.308) | 0.101 (0.109) | $0.645^{***}$ (0.028) | $0.420^{**}$ (0.165) | 0.090 (1.450) | $-0.015$ (0.206) |
| N. of industries for pioneers | 284 | 284 | 284 | 284 | 67 | 67 | 67 | 67 |
| Observations | 1,377 | 1,377 | 1,377 | 1,377 | 320 | 320 | 320 | 320 |
| $R^2$ | 0.007 | 0.005 | | 0.010 | 0.019 | 0.0002 | | 0.014 |
| Adjusted $R^2$ | 0.006 | 0.003 | | 0.009 | 0.012 | $-0.006$ | | 0.007 |
| F Statistic | $4.854^{***}$ | $3.388^{**}$ | | $7.115^{***}$ | $2.996^{*}$ | 0.025 | | 2.194 |
| F Statistic for instrument | $9.159^{***}$ | | | | $5.766^{**}$ | | | |
| *Note:* | | | | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 15: The effects of industry specific knowledge on growth rate using Bartik shock adjusted by input-output linkages. Industry knowledge is expressed in standard deviation units.

# References


[1] A. Cardoso, A. Najar, M. Murat Vasconcellos, J. Levin, S. Rangel, C. Costa Ribeiro, *et al.*, "International microdata scoping studies project: Brazil," *Rio de Janeiro: Economic and Social Research Council (ESRC)*, 2007.

[2] D. d. G. IBGE, "Divisão do brasil em mesorregiões e microrregiões geográficas," tech. rep., IBGE, 1990.

[3] F. Neffke and M. Henning, "Skill relatedness and firm diversification," *Strategic Management Journal*, vol. 34, pp. 297–316, Mar. 2013.

[4] C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann, "The product space conditions the development of nations," *Science*, vol. 317, no. 5837, pp. 482–487, 2007.

[5] B. Balassa, "Trade liberalisation and "revealed" comparative advantage," *The manchester school*, vol. 33, no. 2, pp. 99–123, 1965.

[6] I. Lütkebohle, "Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization." `http://www.cytoscape.org/`, 2017. [Online; accessed 01-April-2018].

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[8] Banco Central Do Brasil, "Time Series Management System - v2.1." `https://www3.bcb.gov.br/sgspub`, 2018. [Online; accessed 30-May-2018].

[9] Banco Nacional Do Desenvolvimento, "Central de Downloads." `https://www.bndes.gov.br/wps/portal/site/home/transparencia/centraldedownloads`, 2018. [Online; accessed 30-May-2018].

[10] I. C. de Contas Nacionais, "Matriz de insumo-produto : Brasil : 2010," tech. rep., Instituto Brasileiro de Geografia e Estatística, 2016. Accessed: 2018-01-29.

[11] D. Diodato, F. Neffke, and N. O'Clery, "Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time," tech. rep., Utrecht University, Department of Human Geography and Spatial Planning, Group Economic Geography, 2018.