European Immigrants and the United States' Rise to the Technological Frontier

Costas Arkolakis
Yale

Sun Kyoung Lee
Yale

Michael Peters
Yale

Abstract

What is the role of immigrants on (American) Growth? To answer this perplex question, we undertake a massive effort of collecting, digitizing, and harmonizing micro and macro economic data from the 19th and early 20th century. The data originate from the historical manufacturing and demographic census of the United States, immigration records datasets and the universe of US patents. To analyze the counterfactual implications of alternative allocations of immigrants, we develop a dynamical trade model where heterogenous firms make innovation and exporting decisions across space and time. The model predicts that the timing and the spatial allocation of immigrant arrivals affect the path of growth outcomes for each location and the aggregate US economy. We use the structural equations arising from the model to interpret empirical findings from difference-in-difference analysis for the importance of the influx of skilled immigrants on the differential growth of US counties. We plan to conduct counterfactual scenarios of alternative allocation of skilled immigrants from different countries across space and time to measure the economic impact of barriers to migration to the United States economy.

1 Introduction

The transformation of the US economy in the last two hundred years has been remarkable. While being primarily rural in the beginning of the 19th century, the US had developed into an essentially industrial nation when the century came to an end. More strikingly, after lagging behind the technological frontier (represented by the UK) for most of the 19th century, the US entered the twentieth century as the global technological leader and the richest nation on the globe (Gordon, 2017). During this time period, which is also referred to as the "The Second Industrial Revolution", the US economy also experienced a massive in-flow of immigrants, mostly from the European continent.

To answer this question, we make a massive effort to collect, digitize and harmonize data that combine longitudinal information on immigrants and their occupations along with measures of economic outcomes such as, output, wages, and innovation at the disaggregated county-industry level. Using this information we plan to investigate to what extent this influx of immigrants was an important contributor to the transformation of the American economic landscape between 1850 and 1940 and provide a comprehensive account of the importance of spatial mobility for the immigrants' life-cycle. To do we develop a dynamical model of innovation where firms in each location innovate taking into account the full stream of future incomes that can be generated there now and in the future. Firms hire production and research workers. Workers, immigrants or natives, can specialize in either of these activities depending on their idiosyncratic abilities that are determined by their comparative advantage. The arrival of immigrants leads to standard scale effects, as in the Romer (1990) model or as we formally argue may even futher boost innovation if immigrants have a comparative advantage on this activity. The model allows the analysis of a range of regional policies, such as one ones that improve labor productivity, improve innovation productivity improvements, or increase population due to arrivals of natives or immigrants.

We use original immigration records and historical passenger lists from the ships heading from Europe to the US.[1] These data sources are real treasure troves for empirical researchers, as they contain direct micro data on immigrants' *pre*-migration occupations.[2] In particular, both in the immigration records and in the passenger lists, all immigrants were required

[1]The immigration database of 13 million immigrants and the passenger lists of around 5 million immigrants leaving for the US via the German port Hamburg, the so-called "Hamburg Passenger Lists" were provided to us for research purposes by the Battery Conservancy and the Archives of the city of Hamburg, respectively. See http://www.castlegarden.org and http://www.germanroots.com/hamburg.html for additional information. To the best of our knowledge, these data sources have yet to be used in empirical research.

[2]We constructed a crosswalk between these published occupational strings and the *Historical International Classification of Occupations (HISCO)*. For more details on the Historical International Classification of Occupations, see http://historyofwork.iisg.nl/index.php.

to give a detailed account of their last occupation in Europe along with other important information such as the time of arrival in the US, the place of residence in Europe, and age. We link these immigration datasets with the restricted-use complete count US Federal demographic decennial censuses from 1850-1940 using modern record-linking techniques. To link these datasets, we exploit the fact that both the immigration records (and passenger lists) and the federal US demographic Census contain time-invariant individual information. In that way we construct a large-scale micro *panel* data set for immigrants with information on their *pre*-migration occupations and their *post*-migration labor market outcomes and spatial mobility patterns over their whole life-time.

Using this integrated dataset, we study one specific mechanism by which immigration could have led to American prosperity: the transmission of *new* knowledge. While anecdotal evidence of this channel, whereby European immigrants brought novel ideas to the American shore, is abundant, there is no systematic evidence whether this mechanism was quantitatively important. To fill that gap, we first show that the information on immigrants' pre-immigration occupations allows us to construct empirical proxies for the flow of novel ideas, or in short, *flows of knowledge*. We then combine this immigrant micro data with novel measures of wages, productivity growth and patent activity. In particular, by digitizing the published results for the historical Manufacturing Census from 1860 to 1939, we construct data on wages and productivity measures at the county-industry level. Secondly, we also complement our data with new county-industry measures of patent activity and patent novelty.

This combination of micro-information on immigrants and macro-measures of productivity and spatial idea creation allows us to relate knowledge flows (as proxied by inflowing immigrants with pre-migration expertise) to data on productivity growth and patent activity for the study period. By doing so, we provide novel evidence on potential mechanisms by which past immigrant settlements could affect economic outcomes (see e.g. Nunn et al. (2017) and Akcigit et al. (2017)). Furthermore, our study period is not only interesting in itself, but it also provides an ideal laboratory to empirically identify the importance of idea flows, which feature prominently in recent theories of economic growth (Kortum, 1997; Lucas Jr and Moll, 2014; Perla and Tonetti, 2014). As communication flows and technology were far less developed than those of today, the importance of embodied knowledge transmission was arguably much more important at that time.

Furthermore, we investigate the role of spatial mobility for the immigrants' earnings life-cycle. We exploit the longitudinal and spatial aspects of our dataset more intensely. In particular, we construct ''spatial-sector''-based earning measures from our newly digitized information on manufacturing wages. In fact, it might be at the heart of understanding earnings differences between natives and immigrants (see eg Abramitzky et al. (2012)) as our

data shows that there a is systematic positive correlation between average wages, urbanization and immigrant shares in the 19th century.[3]

The rest of the paper is structured as follows. Section 2 presents the data and the main facts about immigrants, their skills, and patterns of innovation. Section 3 present the dynamical trade model and its main insights. Section 4 presents reduced form evidence based on a difference-in-difference specification suggested by the model. Section 5 discussed how the counterfactuals are conducted in our model. Section 6 concludes.

[3]This finding is consistent with the findings of a literature in urban economics that finds large city-wage premia in recent data (see e.g. Roca and Puga (2017)).

2 Immigrants and Innovation in the 19th Century

Our analysis relies on four main sources of data. Three datasets are individual-level data: a novel datasets on millions of immigrant entering the US in the 19th and early 20th century, the full count population census for the years 1850 to 1940 and the population of US patents since 1790. Because our theory stresses the role of immigrants for local productivity growth, we combine this data with information on productivity at the county- and city-level, which we measure from the Manufacturing Census. The combination of these datasets allows us to systematically explore the relationship between immigrants' prior expertise, immigrant's location choice, productivity growth and patent activity at the county-industry level. An overview over these datasources is contained in Figure 1.

Historical Immigration Records (1820 - 1914) We construct our immigration database from two primary sources: the Castle Garden Immigration Database and the Hamburg Passenger Lists. The *Castle Garden Immigration Database (1820-1892)* is an educational project of the Battery Conservancy. The database contains the list of all immigrants entering the US via the port of New York between 1820 to 1892. In total, the database comprises approximately 11 million individual micro-records. Through a cooperation with the Battery Conservancy, we have access to the entire Castle Garden Immigration database. We complement this database with original passenger lists of all immigrants leaving from the port of Hamburg to US between 1850 and 1914 called the *Hamburg Passenger List Database (1850-1914).* We have access to the complete records through a cooperation with the Hamburg State Archive. Importantly, these immigration records contain detailed information on the pre-migration occupations of the respective immigrants and a host of demographic information like the name, age or family structure, which we use to match the immigration records to the population census (see below).

The Full Count Population Census (1850-1940) To measure the impact of immigration, we require information on immigrants' characteristics after their arrival in the US. We do so by linking the individual immigration records from the Immigration Database with the Complete Count Federal Demographic Census. We take advantage the complete transcription of Federal Census Records between 1850 to 1940. The US federal demographic census year records exist for all years from 1850 to 1940 every decade except for 1890 (which was lost due to fire) and contains detailed information on individuals including demographic characteristics and employment information at the sectoral and occupational level.

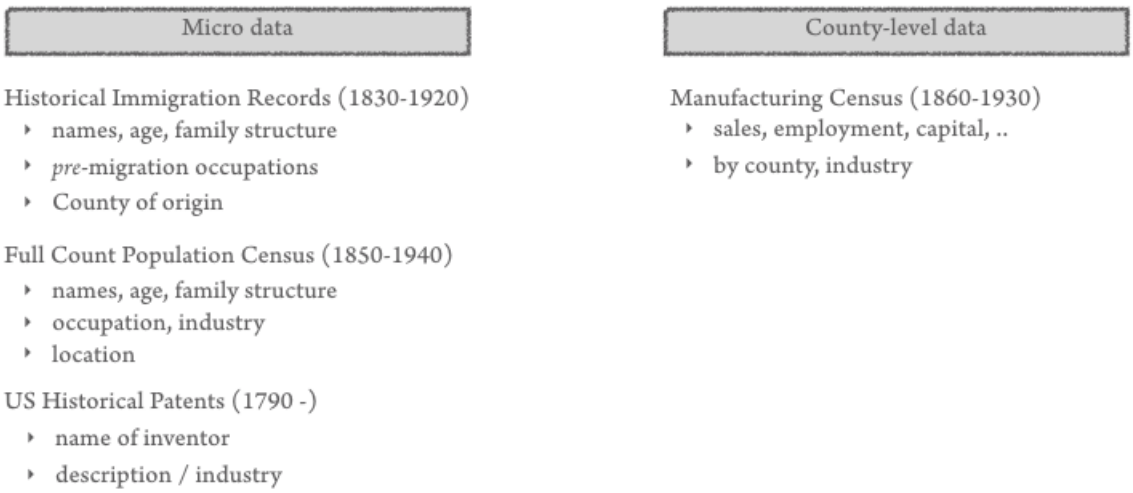


Figure 1: The Data

Historical Data on US Patents To measure innovative activity we exploit information on patenting. The United States Patent and Trademark Office (USPTO) granted millions of patents since 1790 and all patents contain information about the location of the patent and the name of the owner of the patent. We use the information on patenting in two ways. First we study the determinants of patenting at the micro level. This allows us to measure differences innovative activity between natives and immigrants and how the propensity to innovate varies across occupations, sectors and space. Second, we geo-reference the individual patents to connect them to manufacturing census at the county-level. This allows us to study the response of patent activity at the county-level to the inflow of immigrants. We can also directly study whether the arrival of immigrants induced new patent activity in a particular location. To do so, we devise a measure of *spatial idea novelty.* This measure exploits textual analysis to measure the extent to which patents originating in a particular region are similar to the patents that have been invented in that region in the past.[4]

The Manufacturing Census (1870 - 1929) Our empirical strategy heavily relies on spatial variation in productivity growth, innovation activity and the settlement of immigrants. Measuring productivity at a fine spatial resolution in the 19th Century is difficult. First of all, there are no measures of wages at the county-level. While the available individual-level

[4]Intuitively, as in **?**, we measure the similarity between two patents as the correlation in the words, respective patents use. Given this measure of patent-to-patent similarity, we then calculate the similarity of ideas invented in a region as the average patent-to-patent similarity between new patents originated in a particular region and the set of patents stemming from a particular region in the past. See Section B in the Appendix for more details.

data in the decennial Population Census contains county-identifiers, earnings have only been reported starting in 1940. Secondly, information on labor earnings stemming from the National Accounts is available in the 19th Century, but the data does not have a spatial dimension. To overcome this problem, we digitized the published results from the Census of Manufacturers. These tables are published at either the county-industry level or city-industry level and report standard information from firms balance sheets. In particular, they report the number of manufacturing establishments, total number of workers, value of manufactured output, wage-bill and value of the capital stock. As our main measures of spatial productivity, we consider total value added per manufacturing employee, manufacturing value added relative to the wage bill or manufacturing revenue TFP, i.e. $Y_r/\left(K_r^{\alpha}L_r^{1-\alpha}\right)$, where Y_r is total value added and K_r and L_r denote the capital stock and employment. We digitize this data at the *county*-industry level for the years 1860, 1870, 1880 and 1929 and at the *city*-industry level for the years 1880, 1900, 1909, 1919, 1929 and 1939. In Section B in the Appendix, we explain how we harmonize this information and combine it with other sources.

2.1 Record Matching

A key aspect of our empirical work is to match the historical immigration records to the population census and to the microdata on patents. While we relay all details of the matching procedure to Section B in the Appendix, in Table 1 we summarize some aspects for the year 1910. For now we focus on males between 20 and 50 years old, which we suspect to account for the majority of patent holders. In the federal census of 1910 there about 25m. We then try to match individual patents to this population using information on the name and the location of the inventor. In total we observe 0.8m patents, which are issued before 1910.[5] Of these, we can match 350,000 to individuals in the population. Similarly, we turn to the immigration records from the Hamburg Passenger Lists and match individual immigrants to the population census using information on name, the date of arrival, nationality and family characteristics (all of which are available in the population census and the immigration records). We are able to match 100,000 records from the 575,000 potential matches. Finally, the union of matches, i.e. individual migrants, which we can track in the United States and which we can link to a patent, accounts for 20,000 individuals.

This sample forms the basis of our micro-level dataset. In Section 2.3 below we present an analysis of basic empirical regularities. Before turning to this systematic analysis we want to present a particular individual biography, which summarizes the process of this record matching procedure.

[5]In practice we look at the population of patents issued between 1838 and 1910 where the inventors' name is available.

Number of males between 20 and 60 in the population census 1910	
25,272,899	
Number of patents issued before 1910	
Non-matched	Matched
	349,198
Number of immigrants in the Hamburg Passenger Lists	
Non-matched	Matched
50,000	*525,000*

Table 1: Matching Immigrants and Patents to the Population Census in 1910

2.2 A guiding example: The Story of Heinrich Engelhard Steinweg

To illustrate how our novel data can be merged with existing data sources to shed light on the process of technology transfer from old Europe, we start by a particular example: the case of Heinrich Engelhard Steinweg, later known as Henry Engelhard Steinway, the founder of renowned piano manufacturing company *Steinway & Sons*. Heinrich Steinweg left Germany on May 28th, 1850 via the port of Hamburg. This information is declared on the *Hamburg Passenger Lists (1850-1934),* which is available to us through a cooperation with the Archives of the city of Hamburg. His shipment record also indicates his pre-immigration occupation in Germany as *Instrumentenmacher* (instrument maker).[6] As we can see from the same record, shown in Figure 2, his destination was New York and he was accompanied by four family members.

[6]Henry Steinway started working on producing pianos early on with immediate success. But the unstable political climate following the revolutions of 1848 and the limited economic opportunities for a man working outside a guild let him to immigrate to the US. See Claudius Torp, ''Heinrich Engelhard Steinway.'' In Immigrant Entrepreneurship: German-American Business Biographies, 1720 to the Present, vol. 2, edited by William J. Hausman. German Historical Institute.

H Steinweg
in the Hamburg Passenger Lists, 1850-1934

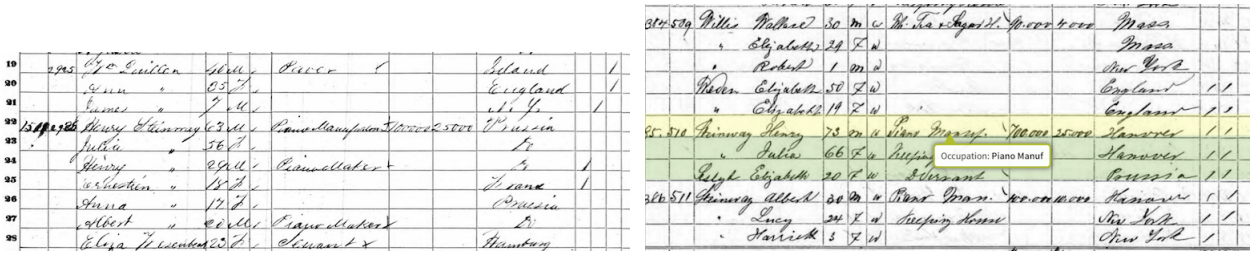
Name:	H Steinweg
Occupation:	Instrumentenmacher
Birth Place:	Seesen, Hannover
Departure Date:	28 Mai 1850 (28 May 1850)
Port of Departure:	Hamburg
Destination:	New York (New York City (All Boroughs))
Port of Arrival:	New York (New York City (All Boroughs))
Ship Name:	Helena Sloman
Captain:	Paulsen, P N
Shipping Clerk:	Rob. M. Sloman
Shipping line:	Rob. M. Sloman
Ship Type:	Dampfschiff
Accommodation:	ohne Angabe
Volume:	373-7 I, VIII A 1 Band 001

Household Members:

Name	Age
H Steinweg	
Steinweg	
Steinweg	7
Steinweg	
Steinweg	23

Figure 2: Heinrich Steinweg in the *Hamburg Passenger Lists, 1850-1934*

We can track Mr. Steinweg, now Mr. Steinway, in the subsequent US population censuses, shown in Figure 3. Both in in 1860 and 1870, Mr. Steinway and his family are recorded to reside in New York. Furthermore, his occupation *piano manufacturer* indicates Steinway's successful transition from a piano maker in Germany to the piano manufacturer of the US.



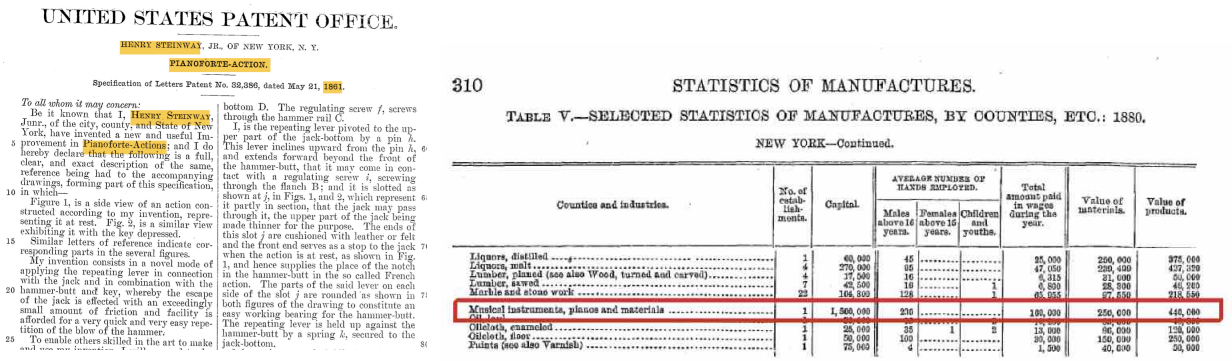
Notes: The figure shows Mr Steinway's US census records in 1860 (left panel) and 1870 (right panel).

Figure 3: Henry Steinway in the *US Census Schedules 1860 and 1870*

That this successful career trajectory might have been in part due to Mr Steinweg's prior knowledge is consistent with micro data on patenting. Using digitized historical patent data from the United States Patent and Trademark Office, we could extract a number of patents granted to him and his sons' names. For example, Steinway's famous piano-forte patent, dated 1862, is shown in the left panel of Figure 4.

Finally, we can use our newly digitized data from the US Census of Manufactures to learn about the economic magnitude of Mr. Steinway's success. While the US Census of Manufacturers data is not at the plant level (but reported at industry-by-county cells), the information is detailed enough to identify the main manufacturing plant of Steinway & Sons in Queens, NY. As the right panel of Figure 4 shows, the Steinway family had an

enormous impact on manufacturing production in the New York area. The digitized Census of Manufacturers for the year 1880, for example, reveals that this single piano manufacturing plant was one of the most capital intensive sectors in New York City with more than $1.5 millions of capital and sales close to a half a million dollars.



UNITED STATES PATENT OFFICE.

HENRY STEINWAY, JR., OF NEW YORK, N. Y.

PIANOFORTE-ACTION.

Specification of Letters Patent No. 32,386, dated May 21, 1861.

To all whom it may concern:

Be it known that I, HENRY STEINWAY, Junr., of the city, county, and State of New York, have invented a new and useful Improvement in Pianoforte-Actions; and I do hereby declare that the following is a full, clear, and exact description of the same, reference being had to the accompanying drawings, forming part of this specification, in which—

Figure 1, is a side view of an action constructed according to my invention, representing it at rest. Fig. 2, is a similar view exhibiting it with the key depressed.

Similar letters of reference indicate corresponding parts in the several figures.

My invention consists in a novel mode of applying the repeating lever in connection with the jack and in combination with the hammer-butt and key, whereby the escape of the jack is effected with an exceedingly small amount of friction and facility is afforded for a very quick and very easy repetition of the blow of the hammer.

To enable others skilled in the art to make

bottom D. The regulating screw *f*, screws through the hammer rail C.

I, is the repeating lever pivoted to the upper part of the jack-bottom by a pin *h*. This lever inclines upward from the pin *h*, and extends forward beyond the front of the hammer-butt, that it may come in contact with a regulating screw *i*, screwing through the flanch B; and it is slotted as shown at *j*, in Figs. 1, and 2, which represent it partly in section, that the jack may pass through it, the upper part of the jack being made thinner for the purpose. The ends of this slot *j* are cushioned with leather or felt and the front end serves as a stop to the jack when the action is at rest, as shown in Fig. 1, and hence supplies the place of the notch in the hammer-butt in the so called French action. The parts of the said lever on each side of the slot *j* are rounded as shown in both figures of the drawing to constitute an easy working bearing for the hammer-butt. The repeating lever is held up against the hammer-butt by a spring *k*, secured to the jack-bottom.

310 STATISTICS OF MANUFACTURES.

TABLE V.—SELECTED STATISTICS OF MANUFACTURES, BY COUNTIES, ETC.: 1880.

NEW YORK—Continued.

Counties and industries.	No. of establishments.	Capital.	Average number of hands employed. Males above 16 years.	Females above 16 years.	Children and youths.	Total amount paid in wages during the year.	Value of materials.	Value of products.
Liquors, distilled	1	60,000	45			25,000	200,000	375,000
Liquors, malt	4	770,000	95			47,050	220,400	427,329
Lumber, planed (see also Wood, turned and carved)	4	17,500	16			6,315	21,000	50,000
Lumber, sawed	7	42,500	18		1	6,800	28,300	46,260
Marble and stone work	23	164,800	128		1	65,955	97,550	218,590
Musical instruments, pianos and materials	1	1,500,000	230			180,000	250,000	448,000
Oilcloth, enameled	1	25,000	35	1	2	13,000	90,000	128,000
Oilcloth, floor	1	50,000	100			30,000	150,000	250,000
Paints (see also Varnish)	1	75,000	4			1,200	40,000	50,000

Figure 4: Mr Steinway's Pianoforte patent and Steinway & Sons in the *US Census of Manufacturers 1880*

2.3 Patenting & Immigration in the 19th Century: Systematic Evidence

In this section we provide systematic evidence on the patterns of patenting from our matched micro data summarized in Table 1. Using this data we highlight five empirical regularities. These regularities are not only interesting per se but they inform the structure of our theory presented below. We find that:

1. Innovators were rare and showed a clear life-cycle profile
2. Immigrants were equally important for the process of innovation as natives. But there is substantial heterogeneity in patent activity by immigrants' nationality
3. Immigrant and native innovators specialized in different occupations and industries
4. Innovation is highly concentrated in cities
5. Innovation activity is tightly correlated with occupation-based measures of skills

Figure 5 shows two facts about patent activity in the 19th century, which are reminiscent of patent activity today. First of all, patenting was relatively rare - on average about 1.5%

of all individuals held any patents and the vast majority of patentees held a single patent.[7] In the right panel we show that patenting had a clear life-cycle profile. In particular, the probability of patenting peaked at around 45 years.

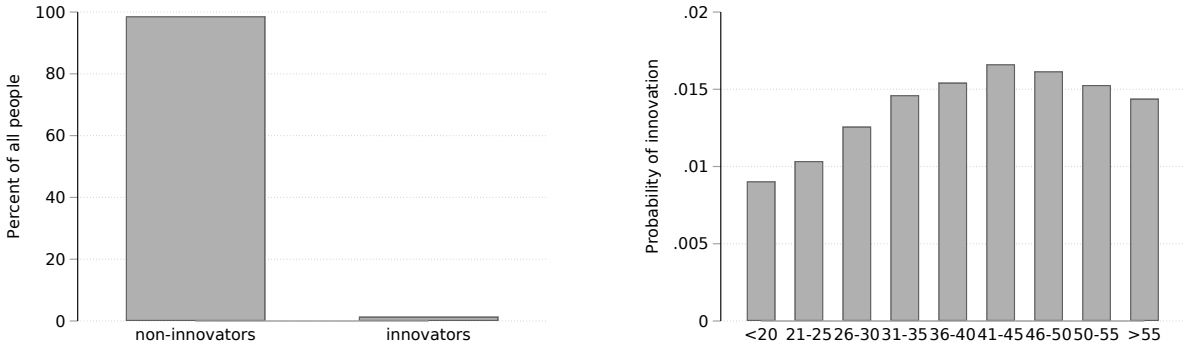


Figure 5: Regularity 1: Patent Propensity & The Patent Lifecycle

Figure 6 documents our second fact, i.e. that immigrants played a similar role for the process of innovation as natives. In the left panel, we show that the share of innovators among immigrants and the number of patents conditional on patenting are very similar between natives and immigrants. In the right panel we show that there is substantial heterogeneity in patent proclivity across nationalities. While the English and the Germans were more likely to patent as natives, the Italians had a lower propensity to patent.

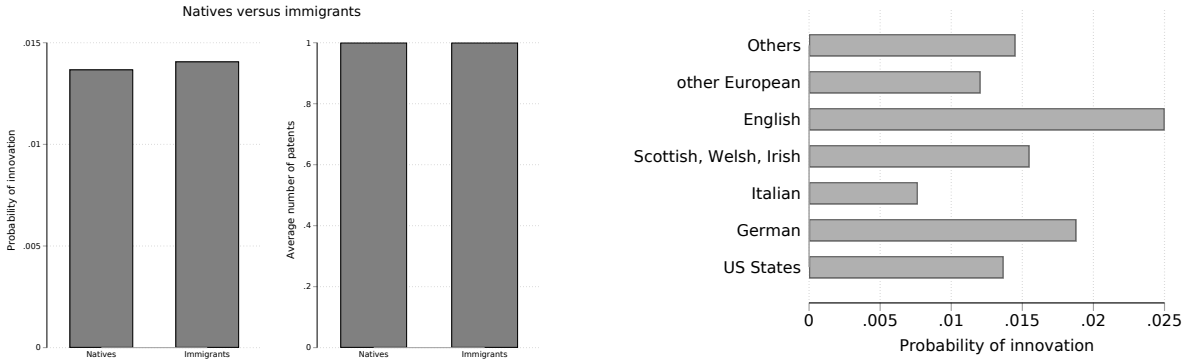


Figure 6: Regularity 2: Patent Activity of Immigrants

Not only was there substantial heterogeneity in patent activity across nationalities, but immigrant and native innovators were also active in different sectors and occupations. This

[7] In the Appendix we document that the distribution of the number of patents conditional on patenting was highly skewed - most patentees had a single patent but some innovators were very productivity owning more than 30 patents.

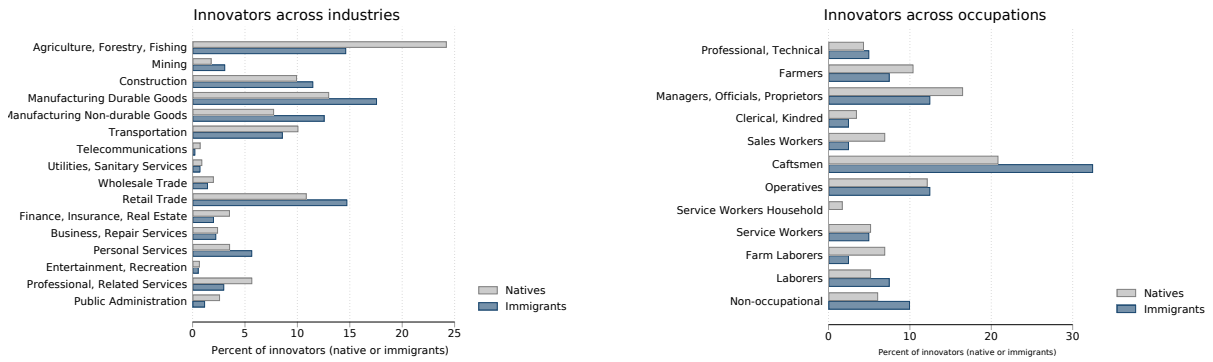


Figure 7: Regularity 3: Heterogeneity in Patent Activity Across Sectors and Occupations

is seen in Figure 7, where we report the distribution of innovators across sectors (left panel) and occupations (right panel) for both immigrants and natives. The figure shows that immigrants' input into the innovation process varies across both industries and occupations. While 25% of all native innovators are in the agricultural sector, this is only true for 15% of all immigrant patentees. Conversely, immigrant inventors are more likely to be in durable manufacturing, non-durable manufacturing and retail trade. A similar pattern emerges at the occupational level. While one third of all immigrant innovators are craftsmen, this it only true for 20% of all natives. Natives are in turn more likely to be farmers and farm laborers and managers, officials and proprietors. This heterogeneity reflects both the different occupational distribution between immigrants and natives and differences in the probability of patenting within sectors and occupations.

Our fourth regularity concerns the spatial distribution of patenting. As expected, most patent holders reside in urban areas. This is seen in Figure 8, which shows the probability of patenting within sectors (left panel) and occupation (right panel) by the urban status of the patent holder. It is clearly seen that living in an urban area increases the probability of patenting by a factor 2 or 3. Reassuringly, this is not the case in the agricultural sector of in the mining industry.

Fifth, we show that - like today - patenting is strongly correlated with skill. Information on individuals' skill in the early 20th century is scarce. However, the US census contains information on educational attainment and earnings potential at the occupational level. This allows us to rank occupations by these measures and connect them to patent activity (aggregated to the occupational level). As we shows in the Appendix, the earnings-based and education-based measures of skills are highly correlated. In Figure 9 we connect the educational ranking of occupations to a measure of innovation output - the number of patents

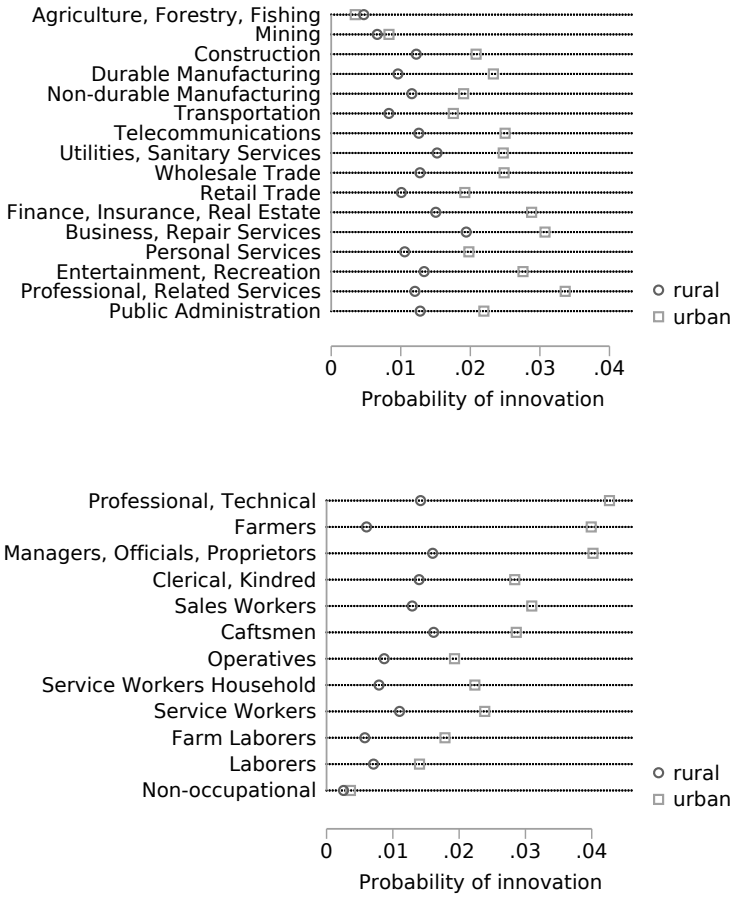


Figure 8: Regularity 4: The Spatial Concentration of Patenting

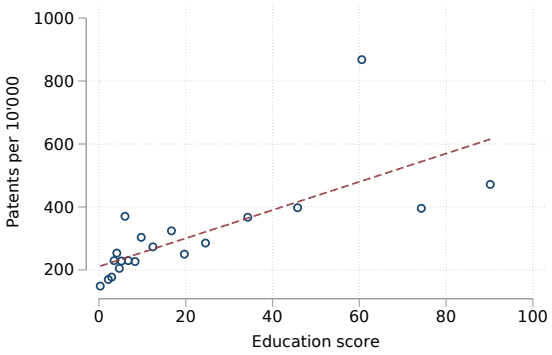


Figure 9: Regularity 5: Patent Activity By Occupation Skill

per 10.000 people working in a particular occupation. The figure shows a clear positive relationship, i.e. occupations with a higher educational score do indeed generate more patent activity.

3 Theory

The country is divided in R regions denoted by r. Time is discrete. We assume the existence of iceberg trade costs between locations. Thus, the final good price of goods from location r in location j is given by

$$p_{rjt} = \tau_{rj} p_{rt},$$

where p_{rt} denotes the price in location r. All workers have identical Constant Elasticity of Substitution (CES) preferences over goods of all locations r arriving in region j, c_{rjt}, at period t given by

$$C_{jt} = \left(\sum_r c_{rjt}^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{1}$$

where ε is the elasticity of substitution across varieties.

Total spending (and income) of the representative agent in region r is denoted by E_{rt}. Workers can work as production and research workers with associated wages in location r, w_r^P and w_r^R. The native population in each region is denoted by L_{rt} and the number of immigrants is denoted by I_{rt}.

Each region produces a final tradable good, which we denote by Y_{jt}. The production of this final good in each location requires a unit continuum of differentiated, non-tradable varieties i, $x_t(i)$, so that

$$Y_{rt} = Z_{rt}^A \left(\int_{i=0}^1 x_t(i)^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

where Z_{rt} is a regional exogenous productivity term. The price of the good produced in in the region is given by

$$p_{rt} = \frac{U_{rt}}{Z_{rt}^A},$$

where U_{rt} denotes the cost of production of the intermediate input bundle and is given by

$$U_{rt} = \left(\int_{i=0}^1 u_{rt}(i)^{1-\sigma} di \right)^{\frac{1}{1-\sigma}},$$

where $u_{rt}(i)$ is the price per unit of variety i in region r at time t.

3.1 Firms and Innovation

Firms are monopolists for their differentiated varieties. Firms differ by location and efficiency. The production function for varieties in region r with efficiency z is given by

$$x_{rt}(z) = z\left(q_{rt}(z)\right)^{\frac{1}{\sigma-1}} h_{rt}(z) \tag{3}$$

where h_{rt} denotes the total amount of efficiency units hired for production by a firm in region r time t, q_{rt} denotes the quality of firm z in region r and z is an exogenous, firm-specific efficiency.[8]

While z is an exogenous firm-characteristic, which is constant, quality q evolves endogenously. In particular, firms can increase their current quality $q_{rt}(z)$ by a factor $i_{rt}(z)$ according to

$$q_{rt+1}(z) = q_{rt}(z)\, i_{rt}(z). \tag{4}$$

Note that we, in principle, allow for depreciation of the firm quality i.e. if firms do not innovate enough, their productivity might decline, $i_t < 1$. As we show below, this will not be the case along a Spatial Balanced Growth Path. To innovate, firms need to hire researchers. More specifically, we assume that to increase their quality by $i_{rt}(z)$, firms pay a cost of

$$c^I_{rt}(q;z) = \frac{1}{\zeta_r Z^I_t} \frac{z^{\sigma-1} q}{Q^{\lambda}_{rt}} \frac{i^{\iota}_t}{\iota} w^R_{rt}, \tag{5}$$

where Q_{rt} is the average productivity in region r at time t

$$Q_{rt} = \mathbb{E}_{rt}\left[z^{\sigma-1} q(z)\right] = \int z^{\sigma-1} q_{rt}(z)\, dF_{rt}(z) \tag{6}$$

and $F_{rt}(z)$ is the cross-sectional distribution of firm efficiencies in region r at time t. Equation (5) stresses that the costs of innovation depend on both firm-level and regional characteristics. On the regional level, they are determined by the prevailing research wage in location r, w^R_{rt}, a fixed region-specific ''innovation efficiency'' ζ_r and a spill-over term Q^{λ}_{rt}. The parameter λ governs the extent to which research cost fall in the existing level of productivity Q_{rt}. As we show below, if $\lambda = 1$ the model becomes an endogenous growth model and if $\lambda < 1$ the model is a semi-endogenous growth model. We also show that this distinction makes precise predictions on the long-run effect of immigrant inflows on regional economic activity in the long-run. As in Atkeson and Burstein (2010) we also assume that innovation costs

[8] The scaling $q^{\frac{1}{\sigma-1}}$ is a normalization that allows to write profits in a linear form, ultimately.

are linear in firm efficiency, $z^{\sigma-1}q$. This scaling implies that the model is consistent with Gibrat's Law where growth is independent of size. Finally, Z_t^I is a time-varying efficiency shifter determining the cost of innovation, which is common across all locations.

Firm optimization The constant elasticity aggregator across intermediate input producers implies that firms' prices are given by a constant markup over the production cost

$$u_{rt}(z) = \frac{\sigma}{\sigma-1}\frac{w_{rt}^P}{zq_{rt}(z)^{\frac{1}{\sigma-1}}}. \tag{7}$$

Here w_{rt}^P is the wage for production workers in region r at time t. Standard arguments imply that firm z in region r at time t has a variable profit of

$$\pi_{rt}(z) = \frac{1}{\sigma}\left(\frac{u_{rt}(z)}{U_{rt}}\right)^{1-\sigma}E_{rt} = \frac{1}{\sigma}\frac{z^{\sigma-1}q(z)}{Q_{rt}}E_{rt}. \tag{8}$$

Firms' innovation decisions are of course dynamic in nature. Letting the real interest rate be r_t, the value function of a firm in region r at time t with a productivity $z^{\sigma-1}q(z)$ is given

$$V_{rt}\left(z^{\sigma-1}q\right) = \pi_{rt}\left(z^{\sigma-1}q\right) + \max_{i_t}\left[\frac{1}{1+r_t}V_{rt+1}\left(z^{\sigma-1}qi_t\right) - \frac{1}{\zeta_r Z_t^I}\frac{z^{\sigma-1}q}{Q_{rt}^\lambda}\frac{i_t^\iota}{\iota}w_{rt}^R\right]. \tag{9}$$

Hence, $\frac{1}{1+r_t}V_{rt+1}(z^{\sigma-1}qi_t)$ is the expected value of being a firm with productivity qi_t in period $t+1$.

Proposition 1. *Consider the value function $V_{rt}(q)$ in (9). The value function is linear homogeneous in q, i.e. $V_{rt}(z^{\sigma-1}q) = z^{\sigma-1}qv_{rt}$, where*

$$v_{rt} = \frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}} + \frac{\iota-1}{\iota}\frac{w_{rt}^R}{\zeta_r Z_t^I}\frac{i_t^\iota}{Q_{rt}^\lambda}, \tag{10}$$

and the optimal rate of innovation i_{rt} is given by

$$i_t = \left(\frac{v_{rt+1}}{1+r_t}\frac{Q_{rt}^\lambda\zeta_r Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\iota-1}}. \tag{11}$$

Proof. See Section A.1 in the Appendix. □

Proposition 1 shows that innovation incentives i_{rt} are equalized across all firms in location r. This is an implication of the homogeneity of the value function. The policy function for firms' innovation incentives in (11) shows that the optimal innovation rate depends on the discounted future value $\frac{v_{rt+1}}{1+r_t}$ relative to the cost of innovation $\frac{Q_{rt}^\lambda\zeta_r Z_t^I}{w_{rt}^R}$. Note also that by

combining (10) and (11) we can express the value function as a forward looking difference equation

$$v_{rt} = \frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}} + \left(\frac{\iota - 1}{\iota}\right)\left(\frac{Q_{rt}^{\lambda}\zeta_r Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\iota-1}}\left(\frac{v_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}. \tag{12}$$

3.2 Labor Supply

In our model, individuals have two margins for their labor supply decisions: they decide which sector to work in and which location to migrate to. In terms of timing, we assume that individuals first decide on their geographical location r and then on their preferred sector of employment.

Labor supply across sectors We model sectoral labor supply with a simple Roy structure. Individuals are characterized by a single attribute - their immigration status, which we denote by $n \in \{N, I\}$, where $n = N$ denotes ''Natives''and $n = I$ denotes ''Immigrants''. We assume that individual i draws a vector of efficiency

$$\left\{x^P, x^R\right\},$$

where x^P and x^R denotes the efficiency units as a production worker and a research worker. We assume that x^P and x^R are drawn independently from the following Frechet distribution

$$F_{jn}(x) = e^{-h_n^j x^{-\theta}}. \tag{13}$$

Here h_n^j parametrizes the average human capital of an individual with nationality $n$in sector $j \in \{R, P\}$ and θ parametrizes the labor supply elasticity.

Standard arguments imply that the share of people of type n working in sector $j = R, P$ in region r is given by

$$s_{rnt}^j = \frac{h_n^j \left(w_{rt}^j\right)^{\theta}}{h_n^P \left(w_{rt}^P\right)^{\theta} + h_n^R \left(w_{rt}^R\right)^{\theta}}. \tag{14}$$

Similarly, the aggregate level of human capital provided by workers of type n in region r towards sector $j \in \{R, P\}$ is given by

$$H_{rnt}^j = L_{rnt}\Gamma\left(1 - \frac{1}{\theta}\right)\left(h_n^j\right)^{\frac{1}{\theta}}\left(s_{rnt}^j\right)^{\frac{\theta-1}{\theta}}, \tag{15}$$

where L_{rnt} is the number of people of type n in region r. Hence, the aggregate supply of

efficiency units provided to sector j in region r is given by

$$H_{rt}^{j} = L_{rt}\Gamma\left(1-\frac{1}{\theta}\right)\left[(1-\varpi_{rI})\left(h_{N}^{j}\right)^{\frac{1}{\theta}}\left(s_{rNt}^{j}\right)^{\frac{\theta-1}{\theta}} + \varpi_{rI}\left(h_{I}^{j}\right)^{\frac{1}{\theta}}\left(s_{rIt}^{j}\right)^{\frac{\theta-1}{\theta}}\right] \; j \in \{I,N\}, \quad (16)$$

where $\varpi_{rI} = L_{rI}/L_r$ is the population share of immigrants in region r. Note that the respective employment shares s_{rN}^{j} only depend on relative wages (see (14)). Hence, the aggregate supply of human capital towards the research and the production sector also only depends on the relative wage within region r.[9]

For future reference we distinguish two special cases as delineated in Arkolakis et al. (2018). If skills are inelastically provided (which corresponds to the case of $\theta \to 1$), the share of people of group n working in sector j and the total amount of human capital is given by

$$s_{rnt}^{j} = \frac{h_{n}^{j}}{h_{n}^{R}+h_{n}^{P}} \text{ and } H_{rtn}^{j} = L_{rtn}h_{n}^{j}.$$

The second polar case is the case of homogeneity, i.e. $\theta \to \infty$. In that case, workers' occupation choice problem takes a simple cutoff-rule

$$\text{work in sector } R \text{ if and only } h_{rtn}^{R}w_{rt}^{R} \geq h_{rtn}^{P}w_{rt}^{P}.$$

If, for example $h_{rtn}^{R} = h_{rtn}^{P} = 1$, and with an interior solution where all locations produce and innovate we get that wages are equalized across sectors, i.e.

$$w_{rt} = w_{rt}^{R} = w_{rt}^{R}$$

and that the share of people working in the two sectors are fully demand determined.[10]

3.3 Aggregation

Above we have characterized the optimal decisions of firms and workers. To calculate the equilibrium (in particular to solve for equilibrium prices), we need to aggregate these decisions. Exploiting the functional form assumptions of our setup we can characterize the aggregate law of motion for average quality Q_{rt} and express aggregate output directly in terms of production workers' labor supply.

[9]Note that if natives and immigrants are identical, i.e. $h_{N}^{j} = h_{I}^{j}$, (16) reduces to the usual expression $H_{r}^{j} = L_{r}\Gamma\left(1-\frac{1}{\theta}\right)\left(h^{j}\right)^{\frac{1}{\theta}}\left(s_{r}^{j}\right)^{\frac{\theta-1}{\theta}}$, as sectoral employment shares will be equalized given that they face the same prices.

[10]Note that if natives and immigrants differ in h_{rtn}^{j}, generically one group will be fully specialized.

An Aggregate Production Function Our economy aggregates - at the production side - to a standard macro-spatial model. In particular, we can define an aggregate production function of each location that is linear on the total supply of efficiency units of labor in region r and time t and its aggregate productivity is determined by aggregate object of the economy. This result is formalized in the following Lemma

Lemma 1. *Consider the model above. Let H_{rt}^P be the total supply of efficiency units of production workers in region r at time t. Aggregate output of the tradable good in region r is given by*

$$Y_{rt} = A_{rt} \times H_{rt}^P,$$

where the endogenous TFP term A_{rt} is given by

$$A_{rt} = Z_{rt}^A \left(Q_{rt}\right)^{\frac{1}{\sigma-1}}. \tag{17}$$

Proof. See Section A.3 in the Appendix. □

Lemma 1 and (23) show that the evolution of spatial labor supply $\left\{H_{rt}^P, H_{rt}^R\right\}$ fully summarize the evolution of aggregate output $\{Y_{rt}\}$: $\left\{H_{rt}^R\right\}$ determines the evolution of productivity $\{Q_{rt}\}$ from (23) and $\left\{H_{rt}^P\right\}$ determines aggregate output as in a standard model of trade.

Aggregate Trade Between Regions Regional trade flows are statically determined by considering firm prices. Each consumer consumes products from different locations. Given the assumption of CES demand, the market share of location r in the basket of location j at time t is given by

$$\lambda_{rjt} = \frac{\int \left(u_{rt}\left(z\right) dF_{rt}\left(z\right)\right)^{1-\sigma}}{\sum_{r'} \int \left(u_{r't}\left(z\right) dF_{r't}\left(z\right)\right)^{1-\sigma}} = \frac{\left(\frac{w_{rt}^P}{A_{rt}}\right)^{1-\sigma} \int z^{\sigma-1} q\left(z\right) dF_{rt}\left(z\right)}{\sum_{r'} \left(\frac{w_{r't}^P}{A_{r't}}\right)^{1-\sigma} \int z^{\sigma-1} q_{r'}\left(z\right) dF_{r't}\left(z\right)}. \tag{18}$$

Moreover, one can show that production workers receive a constant share of aggregate income. Hence, total spending, E_{rt}, can be written as a function of production worker income

$$E_{rt} \equiv w_{rt}^P H_{rt}^P \frac{\sigma}{\sigma - 1}. \tag{19}$$

We assume that trade is balanced period-by-period, i.e. that product markets clear every period. In other words we have that in equilibrium

$$E_{rt} = \sum_j \lambda_{rjt} E_{jt}. \tag{20}$$

Aggregate labor demand

The market for labor for innovation implies that the total number of efficiency units of workers in innovation must be equal to

$$H_{rt}^R = \frac{1}{\iota} \frac{1}{\zeta_r Z_t^I} \frac{\int z^{\sigma-1} q(z)\, dF_{rt}(z)}{Q_{rt}^\lambda} i_t^\iota = \frac{1}{\iota} \frac{i_t^\iota}{\zeta_r Z_t^I} \frac{1}{Q_{rt}^{\lambda-1}} \tag{21}$$

Combining this equation with the first order condition for innovation equation (10) we obtain the labor demand for innovation

$$H_{rt}^R = \frac{1}{\iota} \left(\frac{v_{rt+1}}{1+r_t} \frac{1}{w_{rt}^R} \right)^{\frac{\iota}{\iota-1}} \left(Q_{rt}^\lambda \zeta_r Z_t^I \right)^{\frac{1}{\iota-1}} Q_{rt}. \tag{22}$$

The Endogenous Law of Motion for Aggregate Quality

Using equation (4) and aggregating among firms given the innovation equation (10) we obtain

$$Q_{rt+1} = Q_{rt} i_{rt}$$

Combining this equation, with (22) we can directly relate the the demand for research workers to the resulting growth rate

$$\frac{Q_{rt+1}}{Q_{rt}} = \left(H_{rt}^R \zeta_r Z_t^I Q_{rt}^{\lambda-1} \iota \right)^{1/\iota}, \tag{23}$$

i.e. the evolution of local productivity Q_{rt} is fully determined from the equilibrium amount of researchers H_{rt}^R.

We exploit these aggregation results to define the equilibrium as a macro system of discrete blocks of equations, statics and dynamic.

3.4 Dynamical Equilibrium

To characterize the equilibrium of the model we need to consider the market clearing of both product and labor markets. 22

Definition 1. *A dynamical equilibrium are sequences of production and research wages* $\left\{ w_{rt}^P, w_{rt}^R \right\}_{rt}$*, per-quality-unit value functions* $\{v_{rt}\}_{rt}$*, innovation choices* $\{i_{rt}\}_{rt}$*, labor alloca-*

tions $\left\{H_{rt}^P, H_{rt}^R\right\}_{rt}$*, regional qualities* $\{Q_{rt}\}_{rt}$*, and consumption demands* $\{c_{jrt}\}_{jrt}$ *such that given an initial level of regional quality* $\{Q_{r0}\}_r$*, labor and good markets clear at each point time,*

1. *firms' innovation choices* $\{i_{rt}\}_{rt}$ *are consistent with* $\{v_{rt}\}_{rt}$*, i.e. solve (11)*
2. *the evolution of qualities* $\{Q_{rt}\}_{rt}$ *is consistent with firms' innovation choices* $\{i_{rt}\}_{rt}$*, i.e. solve (23),*
3. *the per-quality-unit value functions* $\{v_{rt}\}_{rt}$ *solve (10).*
4. Labor markets clear, i.e. labor demand and supply for production and research labor equalize given equations (22), (20), and (16).

Definition 1 makes clear that the general equilibrium consists of a set of dynamic and static equations for an arbitrary number of regions. To solve this daunting dynamic fixed point problem we follow a strategy of modularization, as in Adao et al. (2019), in order to determine sets of equations that can be independently solved taking a subset of the variables of the system at a time. The difference with our approach is that one of the equations of the paper, the firm-innovation module described below, contains dynamic difference equations and not just static. The three modules are as follows:

1. *The Labor Module:* Given production worker wages $\left\{w_{rt}^P\right\}_r$, average product quality $\{Q_{rt}\}_r$ and the value of innovation $\{v_{rt+1}\}_r$, we can use the labor supply equations 16 and the labor demand for research workers (22) to determine $\left\{H_{rt}^P, H_{rt}^R, w_{rt}^R\right\}_r$,
2. *The Trade Module:* Given production worker labor supply $\left\{H_{rt}^P\right\}_r$ and average product quality $\{Q_{rt}\}_r$, production wages $\left\{w_{rt}^P\right\}_r$ are determined from the goods markets. Using (20) we obtain that
$$w_{rt}^P H_{rt}^P = \sum_j \frac{\left(\frac{w_r^P}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}}{\sum_r \left(\frac{w_r^P}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}} w_{jt}^P H_{jt}^P. \tag{24}$$
3. *The Innovation Module: Given wages and labor allocations* $\left\{H_{rt}^P, H_{rt}^R, w_{rt}^P, w_{rt}^R\right\}_r$ and the level of quality $\{Q_{rt}\}_r$, we can solve for i_{rt} from (23). Given i_{rt} we can solve for $\{v_{rt}, v_{rt+1}\}$ from *(10) and (11).*

3.5 Innovation in Space and Time

Starting from the value function, equation 12,

$$v_{rt} = \frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}} + \left(\frac{\iota-1}{\iota}\right)\left(\frac{Q_{rt}^{\lambda}\zeta_r Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\iota-1}}\left(\frac{v_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}.$$

Now use 23, and $E_{rt} = \frac{\sigma}{\sigma-1}w_{rt}^P H_{rt}^P$ to obtain

$$v_{rt} = \frac{1}{\sigma-1}\frac{w_{rt}^P H_{rt}^P}{Q_{rt}} + \left(\frac{\iota-1}{\iota}\right)\iota^{\frac{1}{\iota-1}}\left(\frac{Q_{rt+1}}{Q_{rt}}\right)(Q_{rt+1})^{\frac{1}{\iota-1}}\frac{v_{rt+1}}{1+r_t} \tag{25}$$

We now analyze aspecial case of our model, aspecific factor case, where workers provide their skills inelastically to the two production sectors.

Thecase of no mobility across sectors that corresponds to $\kappa \to 1$ where $H_{rt}^P = \bar{H}_{rt}^P$, $H_{rt}^R = \bar{H}_{rt}^R$ (see Arkolakis et al. (2018)).

Notice that in the second case equation (23) implies that the law of motion of Q_{rt} does not depend on anything else other than $\zeta_r Z_t^I$. In other words, the law of motion of Q_{rt} can be completely determined by this equation and it is country-by-country specific.

Assuming that indeed $\zeta_r Z_t^I$ are fixed and we start from a steady state $Q_{rt} = \bar{Q}_r$. Then using (25) the value function of the firm is given by

$$v_{rt} = \frac{1}{\sigma-1}\frac{w_{rt}^P \bar{H}_{rt}^P}{\bar{Q}_r} + \left(\frac{\iota-1}{\iota}\right)\iota^{\frac{1}{\iota-1}}\frac{v_{rt+1}}{1+r_t}.$$

The solution of the system is then vectors of wages and value function $\{v_{rt}\}, \{w_{rt}^P\}$ that solve the above equation and

$$w_{rt}^P \bar{H}_{rt}^P = \sum_j \frac{\left(\frac{w_{rt}^P}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}}{\sum_r \left(\frac{w_{rt}^P}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}} w_{jt}^P \bar{H}_{jt}^P,$$

with $A_{rt} = Z_{rt}^A \left(\bar{Q}_{rt}\right)^{\frac{1}{\sigma-1}}$.

3.6 The Spatial Balanced Growth Path

We now characterize the balanced growth path (BGP) of this economy. Along the BGP, the distribution of wages and spending across regions is stationary. This requires that productivity grows at the same rate in all regions. The characterization of the BGP is contained in the following proposition.

Proposition 2. *Consider the economy above and consider a BGP. Along the BGP, wages* $\left\{w_{rt}^P, w_{rt}^R\right\}_r$ *and spending* $\{E_{rt}\}_r$ *grow at the rate of TFP* A_{rt} *and the population distribution is stationary.*

1. *The growth rate of TFP* A_{rt} *is constant across regions and is given by*

$$1 + g_A = (1 + \overline{g}_Z)\,(1 + \overline{g}_M)^{\frac{1}{1-\lambda}\frac{1}{\sigma-1}},$$

   where $\overline{g}_Z$ *is the exogenous growth rate of* Z_{rt} *and* $\overline{g}_M$ *is the exogenous growth rate of research productivity* M_t*.*

2. *The growth rate of productivity* Q_{rt} *is given by*

$$1 + g_Q = (1 + \overline{g}_M)^{\frac{1}{1-\lambda}}$$

3. *The value function* v_{rt} *is given by*

$$v_{rt} = \frac{1}{1 - \frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}\frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}},$$

   where E_{rt} *is total spending on goods in region* r

4. *The spending share on production workers and researchers is equalized across space, i.e.*

$$\frac{H_{rt}^R w_{rt}^R}{H_{rt}^P w_{rt}^P} = \frac{1}{\sigma - 1}\frac{\frac{1}{\iota}\frac{1+g_A}{1+r}}{1 - \frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}. \tag{26}$$

5. *The distribution of productivity across space satisfies*

$$\frac{Q_{rt}}{Q_{jt}} = \left(\frac{\zeta_r}{\zeta_j} \times \frac{\frac{E_{rt}}{w_{rt}^R}}{\frac{E_{rt}}{w_{rt}^R}}\right)^{\frac{1}{1-\lambda}} \quad \textit{for all } r, j \tag{27}$$

Proof. See Section A.4 in the Appendix. □

The main implication of Proposition 2 is contained in (27): the long run distribution of productivity across space is endogenously determined. Using the fact that $E_{rt} \propto H_{rt}^P w_{rt}^P$ (see (19)) and $H_{rt}^P w_{rt}^P \propto H_{rt}^R w_{rt}^R$ (see (26)), we can express (27) as

$$\ln\left(\frac{Q_r}{Q_j}\right) = \underbrace{\frac{1}{\lambda}\ln\left(\frac{\zeta_r}{\zeta_j}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{\lambda}\ln\left(\frac{H_{rt}^R}{H_{jt}^R}\right)}_{\text{Supply of researchers}}. \tag{28}$$

Hence, region r has high productivity relative to region j if it is relatively efficient to produce new ideas, i.e. $\zeta_r > \zeta_j$, and it is able to attract relatively more researchers, i.e. $H^R_{rt} > H^R_{jt}$. The relative abundance of researchers is of course endogenous and determined from both the trade equilibrium and the system of equations governing labor mobility. Moving costs or the degree of openness across regions as part of the trade module will therefore affect the long-run distribution of productivity across space.

4 Reduced form evidence

In this section, we provide direct evidence for the mechanism in our model. In particular, we show that immigration inflows are positively related to regional productivity growth and to patent activity.

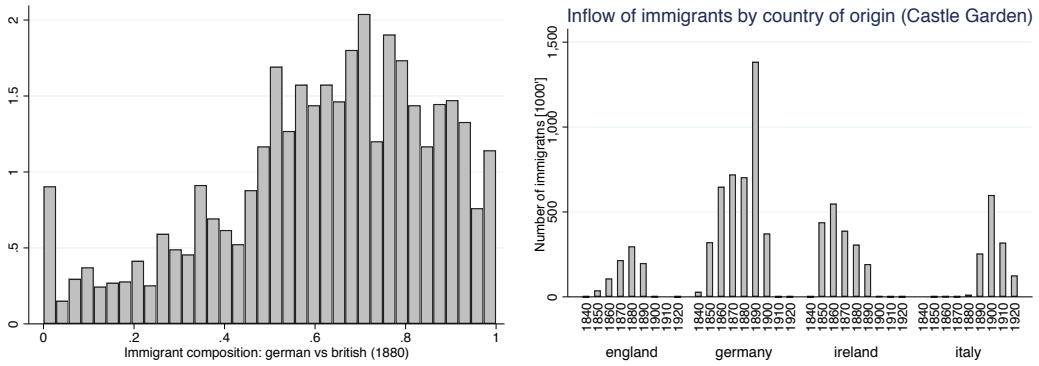
4.1 Constructing an instrument for the allocation of immigrants

In Figure 10 we show the conceptual idea for our instrument, which is based on CARD. We construct predicted immigrant flows from the time-series of the aggregate inflow of immigrants from different countries origin interacted with the existing cross-sectional distribution of immigrants prior to our sample. The time-series variation is shown in the right panel of Figure 10. Two aspects are interesting. First of all, there is a substantial time-series variation *within* immigrant groups. While the flow of Irish immigrants is declining after 1860, the immigration from Germany is increasing and has a peak in the decade between 1880 and 1890. At the same time there is large cross-sectional variation in the composition of the immigrant population across counties in the US. As an example we depict the share of german immigrants relative to immigrants from the UK across US counties in 1880 in the left panel of Figure 10. It is clearly seen that the cross-sectional variation is large. While many counties have a large ''surplus'' in German immigrants relative to British immigrants other counties are much more populated by british immigrants relative to Germans.

Given these two sources of variations, we construct a *predicted* immigrant stock in county r as follows. Let I_{r1860} be the total immigrants in county r in 1860. Also, let I^n_{r1860} denote the total number of immigrants with nationality n in county r in 1860. We then construct the predicted stock of immigrants in county r for all $t > 1860$, I^P_{rt}, as

$$I^P_{rt} = I^P_{rt-1} + \sum_{n \in \{GER,UK,ITA,IRL\}} IF^n_t \times \frac{I^n_{r1860}}{\sum_r I^n_{r1860}}, \tag{29}$$

where IF^n_t is the aggregate inflow of immigrants in year t as depicted in Figure 10. Intu-



Note: The left panel shows the cross-sectional distribution $\frac{I_r^{GER}}{I_r^{GER}+I_r^{UK}}$, where I_r^{GER} (I_r^{UK}) is the number of german (english) immigrants living in region r in the year 1880. In the right panel we display the time-series of immigration inflows by decade for the four main countries of origin.

Figure 10: Constructing the Instrument

itively, we assign the inflowing immigrants from country n according to their cross-sectional distribution in 1860 and then accumulate the regional immigrant stock with these inflow. As we only use (29) as an instrument, we abstract from mortality, which in principle might be specific to particular regions and nationalities.

Because many of our regression utilize the share of immigrants in a particular locality as a regressor, we also construct a predicted immigrant share as

$$s_{Irt}^P = \frac{I_{rt}^P}{I_{rt}^P + L_{Nrt}},$$

where L_{Nrt} is the native population in region r at time t.

4.2 Immigrants and Regional Productivity Growth

Recall that aggregate productivity in region r is given by $A_{rt} = Z_{rt}^A \left(Q_{rt}\right)^{\frac{1}{\sigma-1}}$ (see (17)). Hence, regional productivity growth is given by

$$\ln A_{rt} - \ln A_{rt-1} = \frac{1}{\sigma - 1}\left(\ln Q_{rt} - \ln Q_{rt-1}\right) + \ln Z_{rt}^A - \ln Z_{rt-1}^A$$

Using (23) we get that

$$\ln Q_{rt+1} = \frac{\iota + \lambda - 1}{\iota} \ln Q_{rt} + \ln H_{rt}^R + \ln \zeta_r + \ln Z_t^I.$$

Hence,

$$\ln A_{rt} = \frac{\lambda - 1 + \iota}{\iota} \ln A_{rt-1} + \frac{1}{\sigma - 1} \ln H^R_{rt-1} + \frac{1}{\sigma - 1} \ln \zeta_r + \frac{1}{\sigma - 1} \ln Z^I_{t-1} + \ln Z^A_{rt} - \frac{\lambda - 1 + \iota}{\iota} \ln Z^A_{rt-1}.$$

This suggests a specification of

$$\ln A_{rts} = \delta_s + \delta_t + \rho \ln A_{rt-1s} + \beta \ln H^R_{rt-1} + X'_{rt}\gamma + u_{rst},$$

where δ_s denotes a set of sector fixed effects to control for sector heterogeneity (which is not present in our theory), $X'_{rt}\gamma$ denotes a set of sectoral controls, which controls for regional differences in research technology ζ_r and the time fixed effect δ_t controls for the aggregate state of research efficiency $\ln Z^I_{t-1}$.

Changes in the supply of immigrants affect the provision of research human capital in region r, H^R_{rt}. Our model implies that (see (16))

$$\begin{aligned} \ln H^R_{rt} &= \varsigma + \ln L_r + \ln \left[(1 - \varpi_{rIt}) \left(h^R_N \right)^{\frac{1}{\theta}} \left(s^R_{rNt} \right)^{\frac{\theta-1}{\theta}} + \varpi_{rIt} \left(h^R_I \right)^{\frac{1}{\theta}} \left(s^R_{rIt} \right)^{\frac{\theta-1}{\theta}} \right] \\ &= \varsigma + \ln L_r + \frac{\theta - 1}{\theta} \ln s_{rNt} + \frac{1}{\theta} \ln h^R_N + \ln \left[(1 - \varpi_{rIt}) + \varpi_{rIt} \left(\frac{h^R_I}{h^R_N} \right)^{\frac{1}{\theta}} \left(\frac{s^R_{rIt}}{s^R_{rNt}} \right)^{\frac{\theta-1}{\theta}} \right], \end{aligned} \tag{30}$$

where ς is an inconsequential constant.[11] Log-linearizing the last term around the case of "no immigrants", i.e. $\varpi_{rIt} = 0$, we get that

$$\ln A_{rts} = \delta_s + \delta_t + \rho \ln A_{rt-1s} + \beta \ln L_{rt-1} + \beta \frac{\theta - 1}{\theta} \ln s^R_{rNt-1} + \frac{1}{\theta} \ln h^R_N + \alpha_{rt-1} \varpi_{rIt-1} + X'_{rt}\gamma + u_{rst}, \tag{31}$$

where

$$\alpha_{rt-1} = \beta \left(\left(\frac{h^R_I}{h^R_N} \right)^{\frac{1}{\theta}} \left(\frac{s^R_{rIt-1}}{s^R_{rNt-1}} \right)^{\frac{\theta-1}{\theta}} - 1 \right). \tag{32}$$

Equations (31) and (32) show that the relationship between productivity growth $\ln A_{rts} - \ln A_{rt-1s}$ and the share of immigrants ϖ_{rI}, depends crucially on the extent of comparative advantage. If immigrants have a comparative advantage in the research sector (i.e. $h^R_I > h^R_N$) and have a high employment shares in research $s^R_{rIt} > s^R_{rNt}$, our theory predicts a positive relationship between the share of immigrants and productivity growth, i.e. $\alpha_{rt-1} > 0$. If immigrants and natives are identical in terms of their skills, productivity growth and the share of immigrants should not be related once the size of workforce L_{rt-1} and the research employment share of natives s_{rNt} (which under homogeneity will coincide with the

[11] In fact, $\varsigma = \ln \left(\Gamma \left(1 - \frac{1}{\theta} \right) \right)$.

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV
Immigrant share	0.566***	0.429***	0.450***	0.490***	0.742***	0.683***
	(0.068)	(0.105)	(0.110)	(0.162)	(0.129)	(0.186)
ln Native research share		0.156***	0.168***	0.169***		0.133***
		(0.036)	(0.038)	(0.038)		(0.044)
ln Pop		-0.118***	-0.116***	-0.115***		-0.114***
		(0.017)	(0.017)	(0.017)		(0.017)
High pre-immig skills				-0.012		
				(0.034)		
Immigrant share x High pre-immig skills				-0.074		
				(0.186)		
ln Productivity	0.324***	0.388***	0.388***	0.389***	0.399***	0.385***
	(0.019)	(0.025)	(0.025)	(0.025)	(0.023)	(0.026)
Urban share			-0.121	-0.115		-0.130
			(0.139)	(0.139)		(0.140)
Manufacturing share			-0.120	-0.108		-0.154
			(0.127)	(0.128)		(0.136)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4625	2162	2162	2162	2540	2115
R^2	0.445	0.539	0.539	0.539	0.520	0.537

Notes: Robust standard errors in parentheses.

Table 2: Immigrants and Productivity Growth (Equation (31))

employment share of immigrants) is controlled for.

In Table 2 we report the results of estimating (31), when we restrict α_{rt-1} to be constant across time and space. In column 1, we report the relationship between future productivity and the share of immigrants, when we only control for the current level of productivity. There is a strong positive relationship. Quantitatively, a change in the immigrant share by 0.1 percentage points, i.e. from say 0.1 to 0.2, increases regional productivity ten years later by 5%. In column 2, we estimate the specification in (31) without any regional controls. This specification is consistent with the model if there is no systematic heterogeneity in spatial research productivity ζ_r. The coefficient on the research employment share of natives, $\ln s_{rNt-1}$, is positive and significant. In terms of structural parameters of the model, the coefficient should be equal to $\frac{1}{\sigma-1}\frac{\theta-1}{\theta}$. Empirically, total population size has a negative effect, even though the theory predicts a positive relationship. Below, when we consider patent activity instead of local productivity, we indeed find a positive relationship with local population. The relationship between future productivity and the immigrant share is slightly smaller compared. In column 3 we control for additional local characteristics, which we think

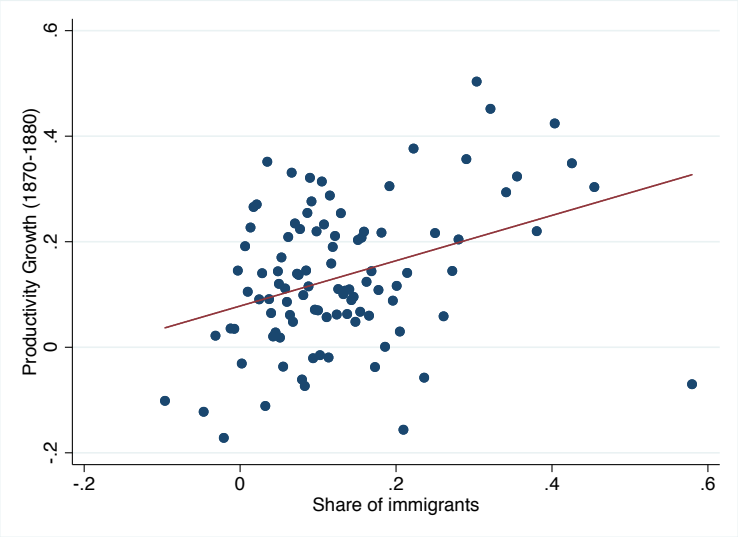


Figure 11: Immigrants and Productivity Growth

of proxies for local research productivity. We utilize the share of the urban population and the manufacturing employment share. Both of these seems to be unrelated to future productivity growth and hence leave the coefficients on the other explanatory variables unchanged.

In the last two columns, we report the IV specification, where we instrument the actual immigrant share with the predicted share of immigrants as explained in Section 4.1 above. Using the instrument increases the coefficients slightly. As in the OLS specification, the coefficient on the immigrant share decreases slightly once we control for the share of researchers in the native population. Finally, in Figure 11, we depict the specification in column 3 of Table 2 graphically. We report a binscatter plot for 100 quantiles of the distribution of immigrant shares after all explanatory variables in columns 3 of Table 2 are controlled for. For ease of readability, we report productivity growth as the dependent variable. Figure 11 shows that the correlation between immigrants and productivity growth is robustly positive and not driven by particular outliers.

4.3 Immigrants and Regional Patent Activity

Our theory also makes predictions on the relationship between regional immigration inflows and the creation of patents. We adopt the following measurement approach. Let $\mathcal{P}_{rt}$ be the *stock* of patents filed in region r up to year t. We assume that Q_{rt} is proportional to the

number of parents, i.e.[12]

$$Q_{rt} = \kappa \mathcal{P}_{rt}.$$

Hence, the model implies that (see (23))

$$\frac{\mathcal{P}_{rt}}{\mathcal{P}_{rt-1}} = \frac{Q_{rt}}{Q_{rt-1}} = \frac{Q_{rt-1}^{\iota+\lambda-1} H_{rt-1}^{R} \zeta_r Z_{t-1}^{I}}{Q_{rt-1}} = Q_{rt-1}^{\iota+\lambda-2} H_{rt-1}^{R} \zeta_r Z_{t-1}^{I}.$$

This implies that

$$\ln\left(\frac{\mathcal{P}_{rt}}{\mathcal{P}_{rt-1}}\right) = (\iota + \lambda - 2) \ln Q_{rt-1} + \ln H_{rt-1}^{R} + \ln \zeta_r + \ln Z_{t-1}^{I}. \tag{33}$$

Using again the approximation for $\ln H_{rt-1}^{R}$ (see (30)), equation (33) suggests the regression

$$\ln \mathcal{P}_{rst} = \delta_s + \delta_t + \rho \ln \mathcal{P}_{rst-1} + \beta \ln L_r + \eta \ln s_{rNt} + \mu \varpi_{rIt-1} + X_{rt}' \gamma + u_{rst},$$

where the theory implies that $\rho = \iota + \lambda - 1$, $\beta = 1$, $\eta = \frac{\theta-1}{\theta}$ and $\mu > 0$ if and only if $\left(\frac{h_I^R}{h_N^R}\right)^{\frac{1}{\theta}} \left(\frac{s_{rIt-1}^R}{s_{rNt-1}^R}\right)^{\frac{\theta-1}{\theta}} > 1$. The time fixed effects control for the state of the research technology, $\ln Z_{t-1}^{I}$, and $X_{rt}'\gamma$ contains a set of observable regional characteristics which control for the systematic variation in research efficiency across space, ζ_r.

We report the empirical results in Table 3, which has the exact same structure as Table 2 above. Columns 1 - 3 report the OLS specification, where we only control for the level of the patent stock $\mathcal{P}_{rst-1}$ (column 1), additional determinants of the supply of human capital (column 2) and additional regional characteristics (column 3). The difference between Table 3 and 2 is informative about the economic mechanism, in particular the extent of knowledge diffusion. While immigrants seems to have a robust positive effect of productivity growth, there is little evidence on a direct effect on patenting once the research share of the native population is controlled for. Hence, immigrants might have been less important for the *direct* increase in patent activity but affected productivity growth *indirectly* through knowledge diffusion to the native population who subsequently sorted into research-intensive occupations. The last two columns contain the IV specification, which are again qualitatively similar to the results using OLS. In Figure 12, we depict the specification in column 3 of Table 3 graphically.

[12]The *flow* of new patents in year t, i.e. the number of patents filed at year t, N_{rt}^{Pat}, is therefore given by $N_{rt}^{Pat} = \mathcal{P}_{rt} - \mathcal{P}_{rt-1}$.

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV
Immigrant share	0.813***	0.144	0.182	-0.019	0.465**	-0.212
	(0.103)	(0.111)	(0.113)	(0.143)	(0.215)	(0.195)
ln Native research share		0.577***	0.600***	0.599***		0.641***
		(0.067)	(0.069)	(0.068)		(0.070)
ln Pop		0.188***	0.189***	0.185***		0.195***
		(0.057)	(0.057)	(0.056)		(0.058)
High pre-immig skills				-0.016		
				(0.036)		
Immigrant share x High pre-immig skills				0.368*		
				(0.197)		
ln Stock of patents	0.914***	0.755***	0.755***	0.754***	0.909***	0.756***
	(0.008)	(0.035)	(0.035)	(0.035)	(0.012)	(0.036)
Urban share			0.048	0.019		0.059
			(0.084)	(0.087)		(0.084)
Manufacturing share			-0.262*	-0.290**		-0.166
			(0.140)	(0.140)		(0.144)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2641	1518	1518	1518	1732	1490
R^2	0.880	0.927	0.928	0.928	0.912	0.929

Notes: Robust standard errors in parentheses.

Table 3: Immigrants and Patent Activity

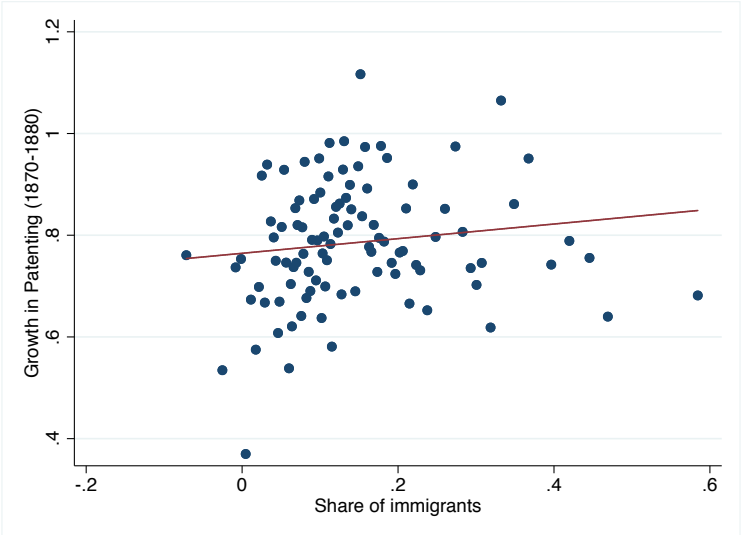


Figure 12: Immigrants and Patent Activity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS						IV	
Immigrant share	0.012***	-0.001	0.009***	0.010***	0.005***	0.003*	0.027***	0.006**
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)
ln Native research share		0.007***						0.005***
		(0.001)						(0.001)
ln Pop		0.003***	0.003***	0.003***	0.003***	0.003***		0.003***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)
High pre-immig skills						-0.001*		
						(0.000)		
Immigrant share x High skills						0.004**		
						(0.002)		
Patent novelty	0.144***	0.111***	0.127***	0.214***	0.208***	0.207***	0.131***	0.113***
	(0.034)	(0.043)	(0.033)	(0.023)	(0.022)	(0.022)	(0.034)	(0.042)
Urban share					-0.001	-0.001		-0.002
					(0.000)	(0.000)		(0.002)
Manufacturing share					0.012***	0.012***		0.009***
					(0.001)	(0.001)		(0.003)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3282	1635	3282	6321	6321	6321	2113	1606
R^2	0.729	0.209	0.744	0.733	0.737	0.737	0.046	0.208

Notes: Robust standard errors in parentheses.

Table 4: Immigrants and Patent Novelty

4.4 Immigrants and Patent Novelty

We finally report the same regression but using spatial patent novelty as our left-hand-side variable. Our textual analysis provides a measure of *spatial idea novelty* for each region that we relate it systematically to the spatial inflow of immigrants. Columns 1 - 6 report the OLS specification, where we only control for the level of the patent novelty (column 1), additional controls (2-4) and the supply of human capital (column 6) and the IV specification with the level of patent novelty and the full set of controls and supply of human capital proxies. Almost all the specificationz yield a statistically significant role for the immigrants for local patent novelty that go beyond the scale effects and remain significant even after controlling for various proxies of human capital.

5 Counterfactuals (Preliminary)

To conduct the counterfactuals we build on the procedure of Dekle et al. (2007) on conditioning the unobservables on actual data. We define $\hat{x} = x'/x$ and apply this definition to all the

equilibrium equations of the model. For the trade module, equation 24,

$$\hat{w}_{rt}^P \hat{H}_{rt}^P w_{rt}^P H_{rt}^P = \sum_j \hat{x}_{ijt} x_{ijt} \hat{w}_{jt}^P \hat{H}_{jt}^P w_{jt}^P H_{jt}^P, \tag{34}$$

where

$$\hat{x}_{ijt} = \frac{\left(\frac{\hat{w}_r^P}{\hat{A}_{rt}} \hat{\tau}_{rj}\right)^{1-\varepsilon}}{\sum_r x_{ijt} \left(\frac{\hat{w}_r^P}{\hat{A}_{rt}} \hat{\tau}_{rj}\right)^{1-\varepsilon}}$$

and

$$\hat{A}_{rt} = \hat{Z}_{rt}^A \left(\hat{Q}_{rt}\right)^{\frac{1}{\sigma-1}}.$$

The innovation module uses equations *(10) and (11)*

$$\hat{v}_{rt} v_{rt} = \frac{1}{\sigma} \frac{E_{rt}}{Q_{rt}} \frac{\hat{E}_{rt}}{\hat{Q}_{rt}} + \frac{\iota - 1}{\iota} \left(\frac{\hat{v}_{rt+1}}{1 + r_t} v_{rt+1}\right)^{\frac{\iota}{\iota-1}} \left(\frac{\hat{Q}_{rt}^\lambda \zeta_r \hat{Z}_t^I}{\hat{w}_{rt}^R} \frac{Q_{rt}^\lambda Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\iota-1}} \iff$$

$$\hat{v}_{rt} v_{rt} = \frac{1}{\sigma - 1} \frac{w_{rt}^P H_{rt}^P}{Q_{rt}} \frac{\hat{w}_{rt}^P \hat{H}_{rt}^P}{\hat{Q}_{rt}} + \frac{\iota - 1}{\iota} \left(\frac{\hat{v}_{rt+1}}{1 + r_t}\right)^{\frac{\iota}{\iota-1}} \left(\frac{\hat{Q}_{rt}^\lambda \hat{Z}_t^I}{\hat{w}_{rt}^R}\right)^{\frac{1}{\iota-1}} (v_{rt+1})^{\frac{\iota}{\iota-1}} \left(\frac{Q_{rt}^\lambda \zeta_r Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\iota-1}}$$

Notice that, using equation 22 we have

$$H_{rt}^R = \frac{1}{\iota} \left(\frac{v_{rt+1}}{1 + r_t} \frac{1}{w_{rt}^R}\right)^{\frac{\iota}{\iota-1}} \left(Q_{rt}^\lambda \zeta_r Z_t^I\right)^{\frac{1}{\iota-1}} Q_{rt}$$

so that

$$\hat{v}_{rt} v_{rt} = \frac{1}{\sigma - 1} \frac{w_{rt}^P H_{rt}^P}{Q_{rt}} \frac{\hat{w}_{rt}^P \hat{H}_{rt}^P}{\hat{Q}_{rt}} + (\iota - 1) (\hat{v}_{rt+1})^{\frac{\iota}{\iota-1}} \left(\frac{\hat{Q}_{rt}^\lambda \hat{Z}_t^I}{\hat{w}_{rt}^R}\right)^{\frac{1}{\iota-1}} \frac{H_{rt}^R w_{rt}^R}{Q_{rt}} \iff$$

$$\hat{v}_{rt} v_{rt} Q_{rt} = \frac{1}{\sigma - 1} w_{rt}^P H_{rt}^P \frac{\hat{w}_{rt}^P \hat{H}_{rt}^P}{\hat{Q}_{rt}} + (\iota - 1) (\hat{v}_{rt+1})^{\frac{\iota}{\iota-1}} \left(\frac{\hat{Q}_{rt}^\lambda \hat{Z}_t^I}{\hat{w}_{rt}^R}\right)^{\frac{1}{\iota-1}} H_{rt}^R w_{rt}^R \tag{35}$$

Same equations imply that

$$v_{rt} Q_{rt} = \frac{1}{\sigma - 1} w_{rt}^P H_{rt}^P + \frac{\iota - 1}{\iota} H_{rt}^R w_{rt}^R$$

With this equation we can calibrate the initial

$$v_{r0} Q_{r0},$$

with knowledge of $w_{r0}^P H_{r0}^P, H_{r0}^R w_{r0}^R$.

Also notice that equation 22 in changes is

$$\hat{H}_{rt}^{R}=\left(\frac{\hat{v}_{rt+1}}{1+\hat{r}_{t}}\frac{1}{\hat{w}_{rt}^{R}}\right)^{\frac{\iota}{\iota-1}}\left(\hat{Q}_{rt}^{\lambda}\hat{Z}_{t}^{I}\right)^{\frac{1}{\iota-1}}\hat{Q}_{rt} \tag{36}$$

The second equation for the innovation module (23) can be directly written in changes

$$\frac{\hat{Q}_{rt+1}}{\hat{Q}_{rt}}=\hat{i}_{t}=\left(\hat{H}_{rt}^{R}\hat{Z}_{t}^{I}\hat{Q}_{rt}^{-(1-\lambda)}\right)^{1/\iota} \tag{37}$$

Thus, given knowledge of $\left\{w_{rt}^{j}H_{rt}^{j}\right\}$ for $j=P,R$and changes in $\left\{\hat{Z}_{rt}^{A},\hat{Z}_{t}^{I}\right\}$ we can solve for $\hat{Q}_{rt+1},\hat{v}_{rt+1},\hat{w}_{rt}^{R},\hat{w}_{rt}^{P},\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}$ as we summarize in the next proposition

Proposition 3. *Conditional on the initial levels of* $\left\{w_{rt}^{j}H_{rt}^{j}\right\}$ *for* $j=P,R$*and* $t=0$ *and changes in* $\left\{\hat{Z}_{rt}^{A},\hat{Z}_{t}^{I}\right\}$ *the changes in* $\left\{\hat{Q}_{rt},\hat{v}_{rt},\hat{w}_{rt}^{R},\hat{w}_{rt}^{P},\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}\right\}$ *can be determined with the solution of equations 34, 35, 37, and 36, and the labor supply equations for* $\left\{\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}\right\}$.

We plan to implement simulation to measure the impact of immigration on local and aggregate growth.

6 Conclusions

We have developed an empirical and theoretical framework to analyze the role of human capital and spatial policies on economic growth. Our big data historical approach allows us to analyze decades of data on the American economy and the associated effects of the influx of immigrants. We couple these data with a model of forward looking innovating firms that allows us to evaluate the empirical data using structural relationships that arise from the theory. In future work we plan to exploit the micro aspect of the data to fully understand the process of knowledge creation by immigrants and to provide more definitive conclusions on the impact of immigration on American growth.

References

Abramitzky, R., Boustan, L. P., Eriksson, K., August 2012. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. American Economic Review 102 (5), 1832--56.
URL https://ideas.repec.org/a/aea/aecrev/v102y2012i5p1832-56.html

Adao, R., Arkolakis, C., Esposito, F., 2019. Spatial linkages, global shocks, and local labor markets: Theory and evidence.

Akcigit, U., Grigsby, J., Nicholas, T., 2017. Immigration and the rise of american ingenuity. Tech. rep., National Bureau of Economic Research.

Arkolakis, C., Ramondo, N., Rodríguez-Clare, A., Yeaple, S., 2018. Innovation and production in the global economy. American Economic Review 108 (8), 2128--73.

Atkeson, A., Burstein, A., 2010. Innovation, firm dynamics, and international trade. Journal of Political Economy 118 (3), 433--489.

Dekle, R., Eaton, J., Kortum, S., 2007. Unbalanced trade. Tech. rep., National Bureau of Economic Research.

Feigenbaum, J. J., 2016. Automated census record linking: A machine learning approach.

Gordon, R. J., 2017. The rise and fall of American growth: The US standard of living since the civil war. Princeton University Press.

Kortum, S., 1997. Research, patenting, and technological change. Econometrica 65 (6), 1389--1419.

Lucas Jr, R. E., Moll, B., 2014. Knowledge growth and the allocation of time. Journal of Political Economy 122 (1), 1--51.

Nunn, N., Qian, N., Sequeira, S., 2017. Migrants and the making of america: The short and long run effects of immigration during the age of mass migration.
URL https://ideas.repec.org/a/ces/ifodic/v15y2017i3p50000000000049.html

Perla, J., Tonetti, C., 2014. Equilibrium imitation and growth. Journal of Political Economy 122 (1), 52--76.

Roca, J. D. L., Puga, D., 2017. Learning by working in big cities. The Review of Economic Studies 84 (1), 106--142.

Romer, P. M., 1990. Endogenous technological change. The Journal of Political Economy 98 (5), S71--S102.

Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., Skiena, S., 2017. Nationality classification using name embeddings. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17. ACM, New York, NY, USA, pp. 1897--1906. URL http://doi.acm.org/10.1145/3132847.3133008

A Theoretical Appendix

A.1 Proof of Proposition 1

Consider the value function in (9), given by

$$V_{rt}\left(z^{\sigma-1}q\right)=\pi_{rt}\left(z^{\sigma-1}q\right)+\max_{i_t}\left[\frac{1}{1+r_t}V_{rt+1}\left(z^{\sigma-1}qi_t\right)-w_{rt}^R\frac{z^{\sigma-1}}{\zeta_r^I Z_t^I}\frac{q}{Q_{rt}^\lambda}\frac{i_t^\iota}{\iota}\right].$$

We conjecture that $V_{rt}(q)$ is linear in q, i.e. takes the form

$$V_{rt}(q)=v_{rt}q, \tag{38}$$

and we will determine v_{rt}. Using, (38) and (8) we get that

$$v_{rt}z^{\sigma-1}q=\frac{1}{\sigma}\frac{z^{\sigma-1}q}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\max_{i_t}\left[\frac{1}{1+r_t}v_{rt+1}iz^{\sigma-1}q-w_{rt}^R\frac{1}{\zeta_r^I Z_t^I}\frac{z^{\sigma-1}q}{Q_{rt}^\lambda}\frac{i_t^\iota}{\iota}\right],$$

so that indeed

$$v_{rt}=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\max_{i_t}\left[\frac{1}{1+r_t}v_{rt+1}i-w_{rt}^R\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^\lambda}\frac{i_t^\zeta}{\zeta}\right].$$

The optimality condition for i_t reads

$$\frac{1}{1+r_t}v_{rt+1}=w_{rt}^R\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^\lambda}i_t^{\iota-1}.$$

This implies that the optimal innovation rate is given by

$$i_t=\left(\frac{1}{1+r_t}\frac{v_{rt+1}}{w_{rt}^R}Q_{rt}^\lambda\zeta_r^I Z_t^I\right)^{\frac{1}{\iota-1}}, \tag{39}$$

and that the value function is given by

$$v_{rt}=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)w_{rt}^R\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^\lambda}i_t^\iota,$$

where i_t is given in (39). Substituting for i_t in (39) we can also express the value function as a forward looking difference equation

$$\begin{aligned}v_{rt}&=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)w_{rt}^R\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^\lambda}\left(\frac{1}{1+r_t}\frac{v_{rt+1}}{w_{rt}^R}Q_{rt}^\lambda\zeta_r^I Z_t^I\right)^{\frac{\iota}{\iota-1}}\\&=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)\left(\frac{Q_{rt}^\lambda\zeta_r^I Z_t^I}{w_{rt}^R}\right)^{\frac{1}{\zeta-1}}\left(\frac{v_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}.\end{aligned}$$

A.2 Deriving the labor supply relationships

To derive the results in Section 3.2, we rely heavily on the max stability of the Frechet distribution. In particular, if the S dimensional vector $[x_s]_s$ is iid Frechet distributed across

s, with

$$P\left(x_s \leq z\right) = F_s\left(z\right) = e^{-h_s z^{-\theta}},$$

the variable

$$a = \max_s \left[\lambda_s x_s\right] \tag{40}$$

is distributed according to

$$P\left(a \leq \alpha\right) = e^{-\Lambda^\theta \alpha^{-\theta}}$$

where

$$\Lambda = \left(\sum_s \lambda_s^\theta h_s\right)^{1/\theta}$$

Using this result we can derive the expressions for average income, average human capital and the relative employment shares. In particular,

$$E\left[a\right] = \Gamma\left(1 - \frac{1}{\theta}\right) \times \Lambda$$

Deriving average income Total income of individual i is given by

$$y^i = \left\{w^P x^P, w^R x^R\right\}.$$

Hence,

$$P\left(y^i \leq \overline{y}\right) = e^{-W^\theta \overline{y}^{-\theta}}$$

where

$$W^i = \left(\left(w^P\right)^\theta h_i^P + \left(w^R\right)^\theta h_i^R\right)^{1/\theta}.$$

Hence, if for example individual i if skill type k and nationality n in region r, W^i is given by

$$W_{rn}^k = \left(\left(w_r^P\right)^\theta h_n^{Pk} + \left(w_r^R\right)^\theta h_n^{Rk}\right)^{1/\theta}.$$

Deriving the total supply of human capital by occupation (equation (15)) To derive the aggregate supply of human capital of individuals of type (n, k) in occupation j, note that the *average* number of efficiency units provided to occupation o is

$$E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] = E\left[z_n^{ok} | z_n^{ok} = \max_j \left\{\frac{w^j}{w^o} z_n^{jk}\right\}\right].$$

Using (40) with $\lambda_j = \frac{w^j}{w^o}$ this object is given by

$$\begin{aligned} E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] &= \Gamma\left(1-\frac{1}{\theta}\right)\left(\sum_j \left(\frac{w^j}{w^o}\right)^\theta h_n^{jk}\right)^{1/\theta} \\ &= \Gamma\left(1-\frac{1}{\theta}\right)\frac{\left(\sum_j \left(w^j\right)^\theta h_n^{jk}\right)^{1/\theta}}{w^o} \end{aligned}$$

Also note that the share of people of type (n,k) working in occupation j in region r is given by

$$s_{rn}^{jk} = P\left(w_r^j z_n^{jk} = \max_o \left\{w_r^o z_n^{ok}\right\}\right) = \frac{h_n^{jk}\left(w_r^j\right)^\theta}{\sum_j h_n^{jk}\left(w_r^j\right)^\theta} = h_n^{jk}\frac{\left(w_r^j\right)^\theta}{\left(W_{rn}^k\right)^\theta}.$$

Hence, letting the mass of workers of type (n,k) in region r working in occupation j be

$$L_{rn}^{jk} = L_{rn}^k s_{rn}^{jk},$$

we get that

$$\begin{aligned} H_{rn}^{jk} &= L_{rn}^{jk} E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] \\ &= L_{rn}^k s_{rn}^{jk}\Gamma\left(1-\frac{1}{\theta}\right)\frac{W_{rn}^k}{w_r^j} \\ &= L_{rn}^k\Gamma\left(1-\frac{1}{\theta}\right)s_{rn}^{jk}\left(\frac{h_n^{jk}}{s_{rn}^{jk}}\right)^{1/\theta} = L_{rn}^k\Gamma\left(1-\frac{1}{\theta}\right)\left(h_n^{jk}\right)^{1/\theta}\left(s_{rn}^{jk}\right)^{\frac{\theta-1}{\theta}}. \end{aligned}$$

Alternatively we can also express this object as

$$H_{rn}^{jk} = L_{rn}^k\Gamma\left(1-\frac{1}{\theta}\right)s_{rn}^{jk}\frac{W_{rn}^k}{w_r^j} = L_{rn}^k\Gamma\left(1-\frac{1}{\theta}\right)h_n^{jk}\left(\frac{w_r^j}{W_{rn}^k}\right)^{\theta-1}.$$

These expressions are exactly (15).

A.3 Proof of Lemma 1

The production function for the final good in (2) is given by

$$Y_{rt} = Z_{rt}X_{rt} = Z_{rt}\left(\int_{i=0}^1 x_{rt}\left(z\right)^{\frac{\sigma-1}{\sigma}} dF\left(z\right)\right)^{\frac{\sigma}{\sigma-1}}.$$

Letting U_{rt} be the price index for the bundle X_{rt}, we get that

$$\frac{u_{rt}\left(z\right)x_{rt}\left(z\right)}{U_{rt}X_{rt}} = \left(\frac{u_{rt}\left(z\right)}{U_{rt}}\right)^{1-\sigma} \Longrightarrow x_{rt}\left(z\right) = \left(\frac{u_{rt}\left(z\right)}{U_{rt}}\right)^{-\sigma} X_{rt}.$$

We have that $u_{rt}(z) = \frac{\sigma}{\sigma-1} w_{rt}^p / q(z)^{1/(\sigma-1)} z$ and $U_{rt} = \left(\int_0^1 u_{rt}(z)^{1-\sigma} F(z) dz \right)^{1/(1-\sigma)}$. Substituting the production function (3) and $\frac{u_{rt}(z)}{U_{rt}} = \left(\frac{z^{\sigma-1} q(z)}{Q} \right)^{\frac{1}{1-\sigma}}$, we get that total employment of production workers at firm i is given by

$$h_{rt}(z) = z^{-1} q_{rt}(z)^{-\frac{1}{\sigma-1}} x_{rt}(z) = z^{-1} q_{rt}(z)^{-\frac{1}{\sigma-1}} \left(\frac{z^{\sigma-1} q_{rt}(z)}{Q_r} \right)^{\frac{\sigma}{\sigma-1}} X_{rt} = z^{\sigma-1} q_{rt}(z) \left(\frac{1}{Q_r} \right)^{\frac{\sigma}{\sigma-1}} X_{rt}$$

Hence, total labor demand is

$$\int h_{rt}(z) dF(z) = \left(\frac{1}{Q_{rt}} \right)^{\frac{\sigma}{\sigma-1}} X_{rt} \int z^{\sigma-1} q(z) dF(z) = \left(\frac{1}{Q_{rt}} \right)^{\frac{\sigma}{\sigma-1}} X_{rt} Q_{rt} = (Q_{rt})^{\frac{-1}{\sigma-1}} X_{rt}.$$

Labor market clearing implies that $\int h_{rt}(z) dF(z) = H_{rt}^P$. Hence,

$$X_{rt} = (Q_{rt})^{\frac{1}{\sigma-1}} H_{rt}^P.$$

Substituting into the production function yields

$$Y_{rt} = Z_{rt} (Q_{rt})^{\frac{1}{\sigma-1}} H_{rt}^P.$$

A.4 Characterization of the Balanced Growth Path (BGP)

In this section, we characterize the details of the balanced growth path (BGP). Along the BGP the allocation of people is constant across space and all aggregate variables grow at some constant rate, g^i where i is the relevant variable and i could be potentially different for different i. Along the balanced growth path interest rates are also constant, $r_t = r$. We assume that $\frac{Z_{rt+1}^A}{Z_{rt}^A} = 1 + \bar{g}_Z$, i.e. the exogenous component of productivity grows at rate g (which is the same for all regions). We also assume the aggregate research productivity M_t grows at a constant rate, i.e. $\frac{M_{t+1}}{M_t} = 1 + \overline{g}_M$

To have a balanced growth path we need that aggregate productivity A_{rt} grows at the same rate in all regions. Hence, see Lemma 1, we need that

$$1 + g_A = \frac{A_{rt+1}}{A_{rt}} = \frac{Z_{rt+1}^A}{Z_{rt}^A} \left(\frac{Q_{rt+1}}{Q_{rt}} \right)^{\frac{1}{\sigma-1}} = (1 + \overline{g}_Z)(1 + g_Q)^{\frac{1}{\sigma-1}}.$$

The BGP growth rate of productivity Q is therefore given by

$$1 + g_Q = \left(\frac{1 + g_A}{1 + g_Z} \right)^{\sigma-1}. \tag{41}$$

Using (23), this implies that the innovation rate in region r, i_{rt}, has to be constant across

locations and time. Using (11) this implies that

$$1 + g_Q = i = \left(\frac{1}{1+r} \frac{\zeta_r^I Z_t^I}{Q_{rt}^{1-\lambda}} \frac{v_{rt+1} Q_{rt}}{w_{rt}^R} \right)^{\frac{1}{\iota - 1}}. \tag{42}$$

Moreover, the value function v_{rt} given by equation (10) can be written as

$$\frac{v_{rt} Q_{rt}}{w_{rt}^R} = \frac{1}{\sigma} \frac{E_{rt}}{w_{rt}^R} + \frac{\iota - 1}{\iota} \frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I} i^{\iota}. \tag{43}$$

First note that w_{rt}^R and E_{rt} grow at the same rate in a stationary equilibrium. Hence, conjecture that along a BGP both $\frac{v_{rt} Q_{rt}}{w_{rt}^R}$ and $\frac{Q_{rt}^{1-\lambda}}{Z_t^I}$ are constant. This implies that

$$1 + g_Q = (1 + \overline{g}_M)^{\frac{1}{1-\lambda}}. \tag{44}$$

(41) therefore implies that the growth rate of TFP is given by

$$1 + g_A = (1 + \overline{g}_Z) (1 + \overline{g}_M)^{\frac{1}{1-\lambda} \frac{1}{\sigma - 1}}. \tag{45}$$

Under our price normalization, wages w_{rt}^R and total spending E_{rt} are growing at the rate of aggregate TFP g_A. Hence, for $\frac{v_{rt} Q_{rt}}{w_{rt}^R}$ to be constant, the growth rate of the value function given by

$$\frac{v_{rt+1}}{v_{rt}} = (1 + g_v) = \frac{w_{rt+1}^R / w_{rt}^R}{Q_{rt+1} / Q_{rt}} = \frac{1 + g_A}{1 + g_Q} = \frac{(1 + \overline{g}_Z) (1 + \overline{g}_M)^{\frac{1}{1-\lambda} \frac{1}{\sigma - 1}}}{(1 + \overline{g}_M)^{\frac{1}{1-\lambda}}} = (1 + \overline{g}_Z) (1 + \overline{g}_M)^{\frac{1}{1-\lambda} \left(\frac{2-\sigma}{\sigma - 1} \right)} \tag{46}$$

In fact, the value function can be solved explicitly along the BGP.

Proposition 4. *Let $r > g_A$. Along the BGP the value function is given by*

$$\frac{v_{rt} Q_{rt}}{w_{rt}^R} = \frac{1}{1 - \frac{\zeta - 1}{\zeta} \frac{1 + g_A}{1 + r}} \frac{1}{\sigma} \frac{E_{rt}}{w_{rt}^R}.$$

Proof. From (42) and the fact that v_{rt} grows at at rate $1 + g_v$ given in (46), we get that

$$1 + g_Q = i = \left(\frac{1}{1+r} \frac{\zeta_r^I Z_t^I}{Q_{rt}^{1-\lambda}} \frac{v_{rt+1} Q_{rt}}{w_{rt}^R} \right)^{\frac{1}{\iota - 1}} = \left(\frac{1}{1+r} \frac{\zeta_r^I Z_t^I}{Q_{rt}^{1-\lambda}} \frac{(1 + g_v) v_{rt} Q_{rt}}{w_{rt}^R} \right)^{\frac{1}{\zeta - 1}}.$$

Hence, we can solve for $\frac{Q_{rt}^{1-\lambda}}{\varphi_r^I M_t}$ as

$$\frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I} = \frac{1}{1+r} \frac{1}{(1 + g_Q)^{\zeta - 1}} \frac{(1 + g_v) v_{rt} Q_{rt}}{w_{rt}^R}.$$

Substituting into (43) yields

$$\begin{aligned}\frac{v_{rt}Q_{rt}}{w_{rt}^R} &= \frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R} + \frac{\iota-1}{\zeta}\frac{1}{1+r}\frac{1}{(1+g_Q)^{\zeta-1}}\frac{(1+g_v)\,v_{rt}Q_{rt}}{w_{rt}^R}i^{\zeta} \\ &= \frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R} + \frac{\iota-1}{\iota}\frac{1+g_Q}{1+r}\frac{(1+g_v)\,v_{rt}Q_{rt}}{w_{rt}^R}\end{aligned}$$

so that

$$\frac{v_{rt}Q_{rt}}{w_{rt}^R} = \frac{1}{1-\frac{\iota-1}{\iota}\frac{1+g_Q}{1+r}(1+g_v)}\frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R} = \frac{1}{1-\frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}\frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R}.$$

□

The spatial productivity distribution Along the BGP, the spatial distribution of productivity Q_{rt} is stationary as all regions grow at the same rate. The *level* of productivity is, however, determined endogenously. In particular, (42) implies that

$$\frac{\varphi_r^I M_t}{Q_{rt}^{1-\lambda}}\frac{v_{rt+1}Q_{rt}}{w_{rt}^R} = \frac{\varphi_j^I M_t}{Q_{jt}^{1-\lambda}}\frac{v_{jt+1}Q_{jt}}{w_{jt}^R} \text{ for all } r, j. \tag{47}$$

This implies that relative productivities Q_{rt} and Q_{jt} are given by

$$\frac{Q_{rt}}{Q_{jt}} = \left(\frac{\varphi_r^I}{\varphi_j^I} \times \frac{\frac{v_{rt}Q_{rt}}{w_{rt}^R}}{\frac{v_{jt}Q_{jt}}{w_{jt}^R}}\right)^{\frac{1}{1-\lambda}}.$$

Hence, long-run differences in productivity across space Q_{rt}/Q_{jt} are governed by

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\frac{v_{rt}Q_{rt}}{w_{rt}^R}}{\frac{v_{jt}Q_{jt}}{w_{jt}^R}}\right)}_{\text{Endogenous value of innovation}}. \tag{48}$$

Using Proposition 4, (48) can be written as

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\frac{E_{rt}}{N_{rt}w_{rt}^R}}{\frac{E_{jt}}{N_{jt}w_{jt}^R}}\right)}_{\text{Endogenous Market Size}}. \tag{49}$$

Finally, note that we can express (49) also in terms the labor supply. To do so note that along the BGP payments to researchers and production workers are equalized.

Proposition 5. *Consider a BGP. Then*

$$\frac{w_{rt}^R H_{rt}^R}{w_{rt}^P H_{rt}^P} = \frac{1}{\sigma - 1} \frac{\frac{1}{\iota}\frac{1+g_A}{1+r}}{1 - \frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}.$$

Proof. The total demand for efficiency units in the research sector is given by

$$\frac{1}{\zeta_r^I Z_t^I} \frac{q}{Q_{rt}^\lambda} \frac{i_t^\iota}{\iota} w_{rt}^R$$

$$H_{rt}^R = \int_q \frac{1}{\zeta_r^I Z_t^I} \frac{q}{Q_{rt}^\lambda} \frac{i^\iota}{\iota} dF_{rt}(q) = \frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I} \frac{i^\iota}{\iota}.$$

Using that $i = 1 + g_Q$ and $\frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I} = \frac{1}{1+r} \frac{1}{\left(1+g_Q\right)^{\zeta-1}} \frac{(1+g_v) v_{rt} Q_{rt}}{w_{rt}^R}$ (see proof of Proposition 4), we get that

$$H_{rt}^R = N_{rt} \frac{1+g_Q}{1+r} \frac{(1+g_v) v_{rt} Q_{rt}}{w_{rt}^R} \frac{1}{\iota} = \frac{1+g_A}{1+r} \frac{v_{rt} Q_{rt}}{w_{rt}^R} \frac{1}{\iota}.$$

Using the result in Proposition 4, this implies that

$$H_{rt}^R w_{rt}^R = \frac{1+g_A}{1+r} \frac{1}{1 - \frac{\iota-1}{\iota}\frac{1+g_A}{1+r}} \frac{1}{\sigma} \frac{1}{\iota} E_{rt}.$$

Hence, the payments to researchers are a constant fraction of revenue along the BGP. And because production workers also receive a constant fraction of revenue (see (19)), we get

$$\frac{H_{rt}^R w_{rt}^R}{w_{rt}^P H_{rt}^P} = \frac{1}{\sigma - 1} \frac{\frac{1}{\iota}\frac{1+g_A}{1+r}}{1 - \frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}.$$

□

Using Proposition 5, the long run distribution of productivity in (49) can also be expressed in terms of (endogenous) amount of resources in research

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda} \ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda} \ln\left(\frac{H_{rt}^R}{H_{jt}^R}\right)}_{\text{Resources employed in research}}, \tag{50}$$

or the number of production workers and the relative cost of research

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda} \ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda} \ln\left(\frac{H_{rt}^P}{H_{jt}^P}\right)}_{\text{Market size by production workers}} + \underbrace{\frac{1}{1-\lambda} \ln\left(\frac{w_{rt}^P / w_{rt}^R}{w_{jt}^P / w_{jt}^R}\right)}_{\text{Relative cost of research}} \tag{51}$$

B Empirical Appendix

C Record Linking

In this section we provide details on the procedure and quality of the linking microdata. We link four types of data: (1) we match individuals in the US census over time, (2) & (3) we match immigrants to the US Census (both from the Hamburg Passenger Lists and Castle Garden), and (4) we match the information on patenting to the US Census. We discuss the quality of these matches in turn in the sections below.

C.1 Census-to-Census Matching of individuals over time

We have access to the complete individual-level complete-count US demographic federal census records from 1850-1940.[13] By linking our novel immigration records to the US census records, we can measure the entire life-cycle of immigrants since they entered the US.

Our record linking procedure has the following characteristics: (1) we rely on all *complete-count* US Federal Census records with *occupation* and *industry information* (a newly transcribed variable from the original census records), (2) we link people at *more than two points in time,* and (3) we use both *individual and household level information* to improve the matching of individuals.

These three elements are important for this study. The availability of the occupation information both before and after entering the US enables us to measure individuals' skills and to investigate novel economic problems, such as occupational transitions along immigrants' life-cycle. Also, while many record matching methods match individuals only at two points in time, this horizon may not be long enough to systematically analyze spatial mobility along the life-cycle. We match individuals multiple points in time (up to three or four times) so that we follow individuals for multiple decades.

Finally, most existing historical record matching practices drop non-unique potential matches. This may introduce a systematic bias of matched records, as for example records with relatively common names would be systematically excluded. We therefore develop new methods to find link records across sources and time.

C.2 Matching Immigrants from the Hamburg Passenger Lists to the US Census

Our goal is to extract as many unique matches as possible by imposing rules such that we can eliminate false positives. From both datasets (i.e. Hamburg Passenger Lists and US Demographic Census), we take advantage of the overlap in information sets to extract unique matches. Table 5 shows the final number of matches after completing the pruning procedure. The variable ''Post First Stage Matches'' shows the number of possible matches as outputted

[13]We use restricted complete-count US demographic census from 1850 to 1940 to link individual-level records by implementing machine learning approach (random forest classification). This record linking methodology is similar in spirit to Minnesota Population Center (MPC) Record record linkage project of the 1850-1930 sample census records to the 1880 complete count census records. However, MPC implemented a support vector machines (SVM) to automate the record linking. Feigenbaum (2016) discusses a machine learning approach to census record linking and compares different possible matching algorithms.

from the first stage linking procedure where age, first and last name similarities were primary linking criteria.

Table 5: Hamburg to US Demographic Census Statistics

Record Type	Germany	Italy	UK	Austria	Poland	Hungary	Total
Immigration Record	49,390	2,804	684	129,332	96,578	156,861	437,649
1910 Census Record	1,343,333	887,044	676,429	706,524	20,029	287,229	3,920,588
Post First Stage Matches	227,604	4,983	4,194	259,734	3,741	142,558	642,814
Final Matches	26,136	1,237	251	39,615	2,091	28,711	98,041

The pruning procedure itself uses three rules applied iteratively to the data. In particular, since applying any one rule may delete potential matches that could have been found via another rule we apply rules in every possible order and extract unique matches after each application. We also extract any unique matches that were discovered after the application of the first stage. In the case that the application of a rule introduces a duplicate observation in either the census or Hamburg data we keep the earlier iteration of that observation. This permutation-based approach extracts about 25% of the possible unique matches as a pose to a linear approach which would yield only about 3-5% of the possible matches.

The three rules we apply are as follows. First, we restrict the Jaro-Winkler distance for last name to be greater than or equal to 0.85 (1 being the maximum and 0 being the minimum of Jaro-Winkler distance). For individuals over the age of 25 when they immigrated, we impose that their marital status must be stable across datasets. Finally, we impose the year of immigration in the Census data and the departure year in the Hamburg data may not be more than three years apart.

C.3 Matching Immigrants from the Castle Garden Immigration Records to the US Census

Similar to Section C.2, we link immigrants from the Castle Garden Database to US Census. Primarily, we use individual-level given and last name, age and year of immigration to link immigrants from the Castle Garden database to the US demographic census. We match four major immigrants groups by sending country (Germany, Italy, Ireland, and the UK) from 1850 and 1930. This procedure yields millions of matched immigrant records which we can track their lives for multiple decades.

C.4 Matching Patents to Individuals

We link the universe of US Patent published pre-1910. Each of these patents has (via HistPat) an associated county and state, and (most, but not all, have) an associated inventor name. The inventor name is first parsed into a first and last name, where the last name is the last word in the inventor name (once words like "Jr", "Sr" are removed), and the first name is simply all proceeding words (and therefore it may include middle name). Then, given this first and last name and county and state, consider the universe of individuals who, as of the 1910 census, were residing in that county and state. This assumes that individuals do not

move after patenting. We consider individuals of age over 20 to be a potential inventor at the time of inventing one's first patent.

Measure the Jaro-Winkler distance between the first name of patenters and the listed first name of individuals (males) in the census, and similarly for the last name. For each patent, consider the matched census individual as simply that who has the highest match value. This produced a dataset of 349,198 inventors with information both from the US Patents and US Demographic Census (Inventor Name, County and State of Patents, Census Individual) linked triples.

D Analysis of Hamburg Passenger Lists (HPL)

In this section we provide a comprehensive analysis of the information contained in the HPL. The HPL database contains passenger lists of ships that departed from the port of Hamburg, Germany from 1850-1914 (the data basically stops before the World War I). The database was obtained from the Hamburg State Archive, and we have approximately 4.6 million records of individuals, with approximately 77% of whom were headed to the United States. The database usually contains information such as name, gender, age/birthyear, occupation, nationality, departure date and ship name and we discuss some key information of our interests in the following.

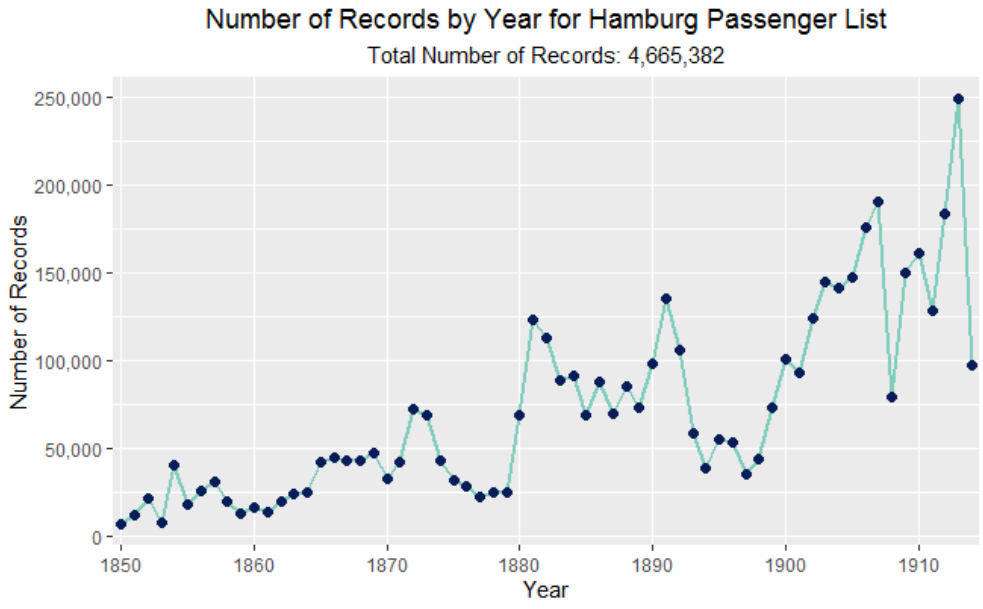


Figure 13: Number of Passengers over Time

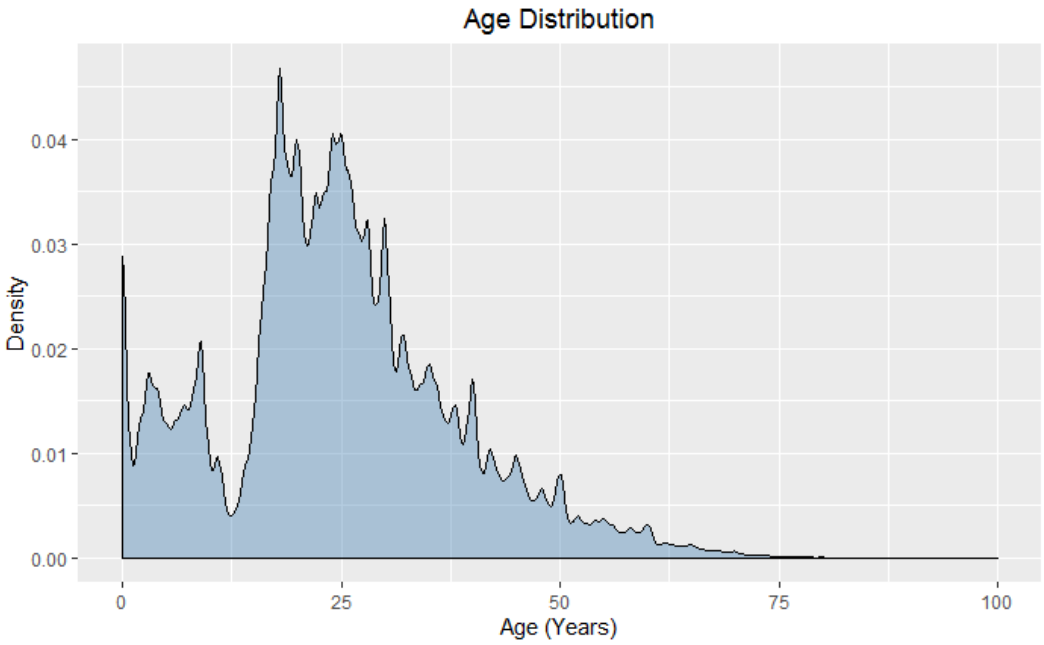


Figure 14: Age Distribution of Passengers

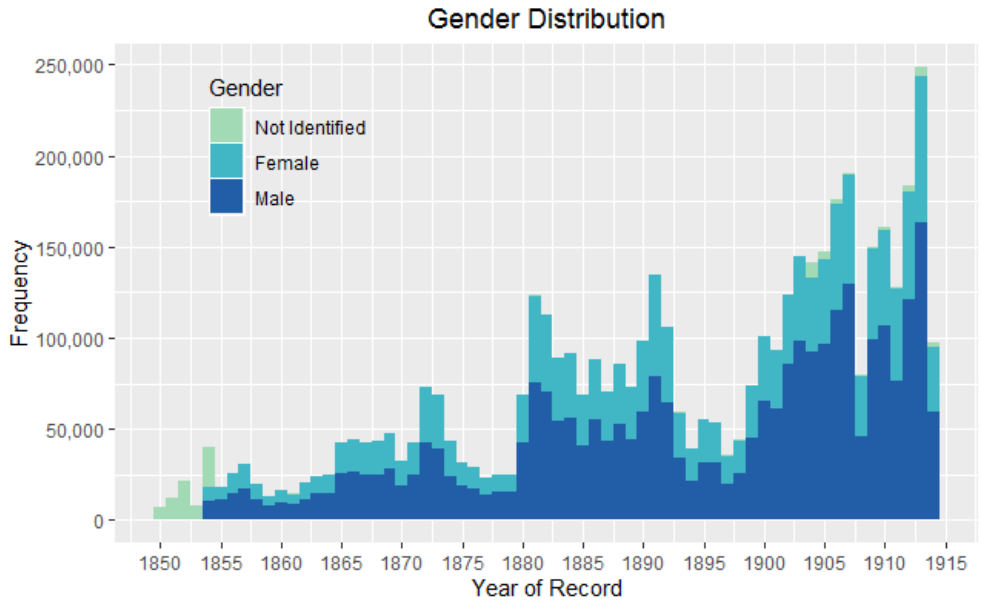


Figure 15: By Gender

Occupation (Beruf) Transcribed information on people's occupations from the Hamburg Passenger Lists (HPL) were cleaned, translated, standardized and coded following 1950

Census Bureau occupation classification system and the Historical Internattional Standard classification of occupations (HISCO) to enhance comparability across years. 53% of HPL records have occupational responses and as in Figure 17, occupation was almost always available for working age males. The most frequently appearing occupations are farm laborers, general laborers, managers, followed by manufacturing workers such as tailors, shoemakers and cabinetmakers.

Non-Empty Records	As Percentage of Total Records	Top 10 Occupations Recorded	Top 10 Occupations Percentage
2,544,925	52.55%	1,923,315	75.58%

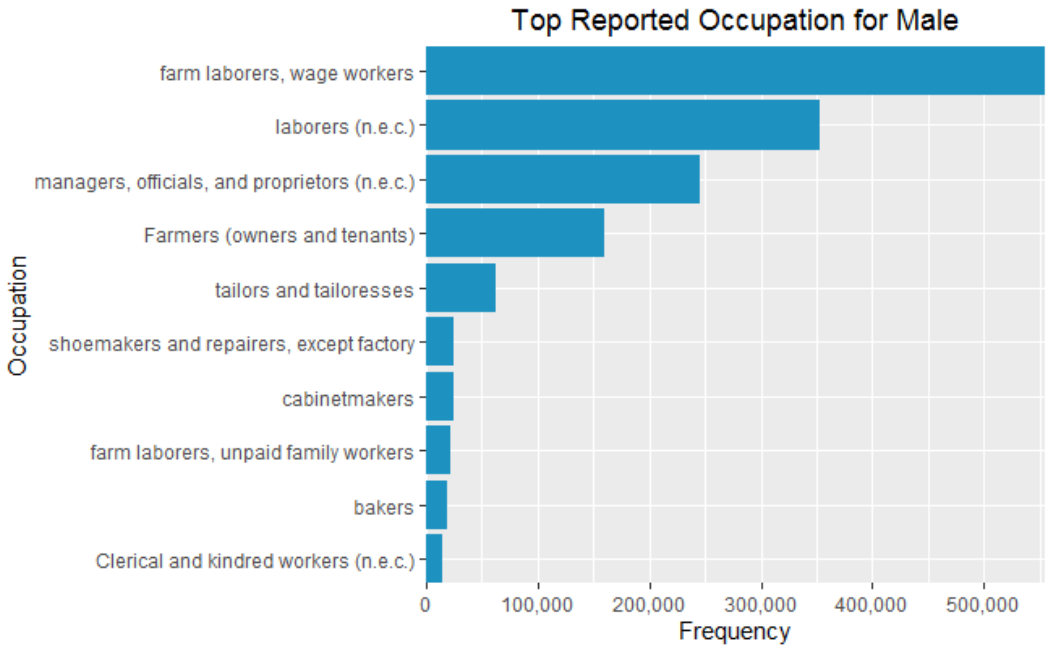


Figure 16: Top Reported Occupations

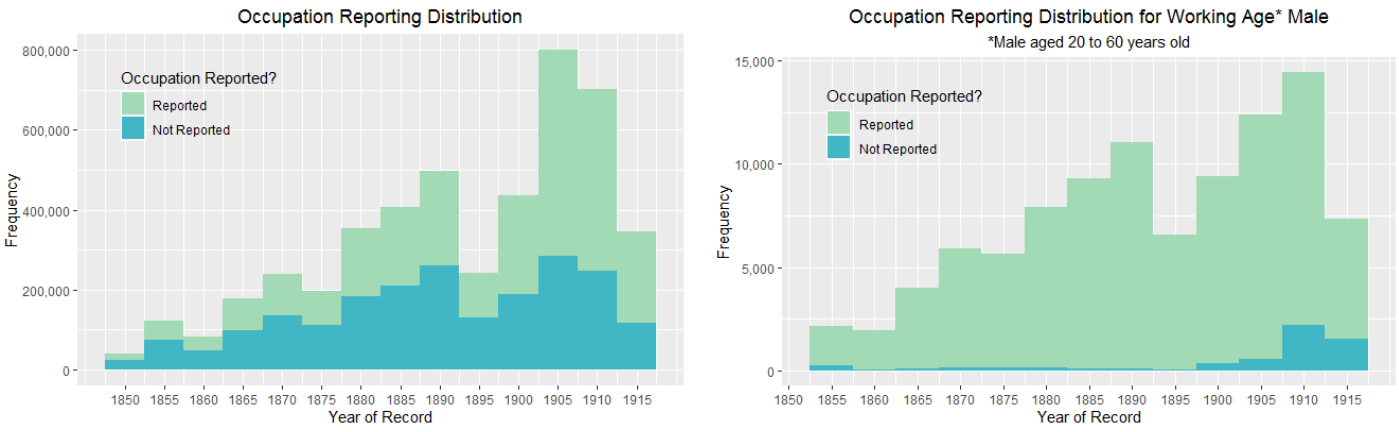


Figure 17: Occupation Reporting Pattern

Destinations (Zielort) HPL records passengers' destinations, and this information is available for approximatley 25% of records and we infer passengers' destinations based on ship route and departure information which are universally available. As in Figure 18, destinations from the port of Hamburg vary significantly and majority of passengers headed to North America (such as New York, Chicago, and Quebec in Canada) whereas some headed to South America (such as Buenos Aires and Rio de Janeiro) and Europe (such as London in the UK). As in Figure 19, passengers' destination were almost never reported between 1860 and 1870 and the destination became more availalbe since then.

# Destination Available Records	% of Destination-available Records	Top 1 Place (New York)	% US Destination
1,168,347	24.36%	533,329	77.31%

Table 6: Destination Summary

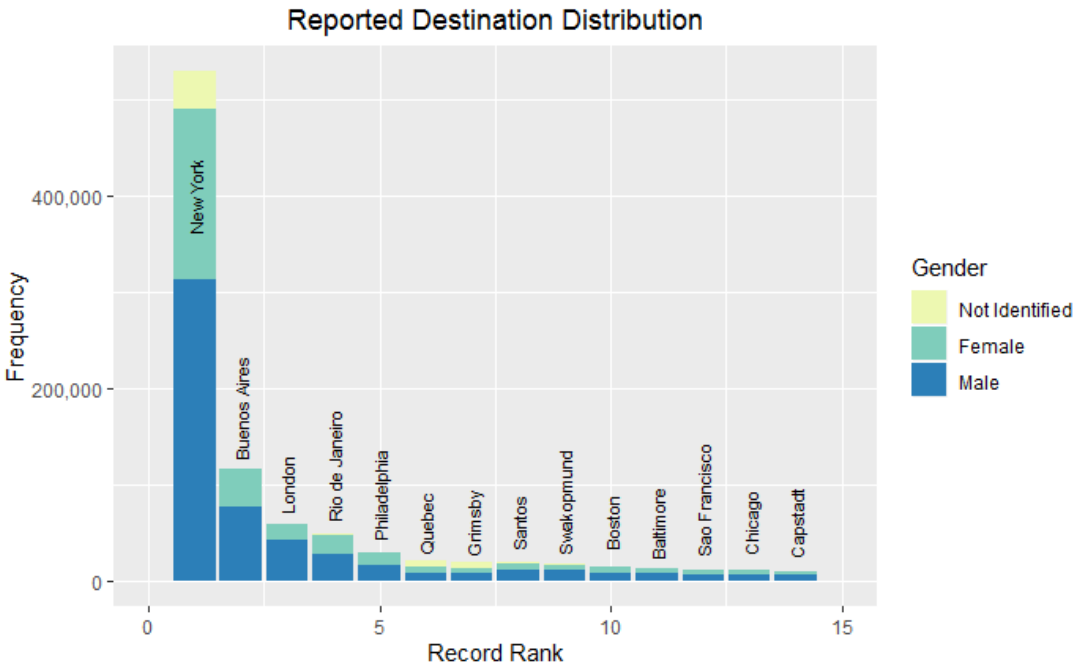


Figure 18: Top Destinations

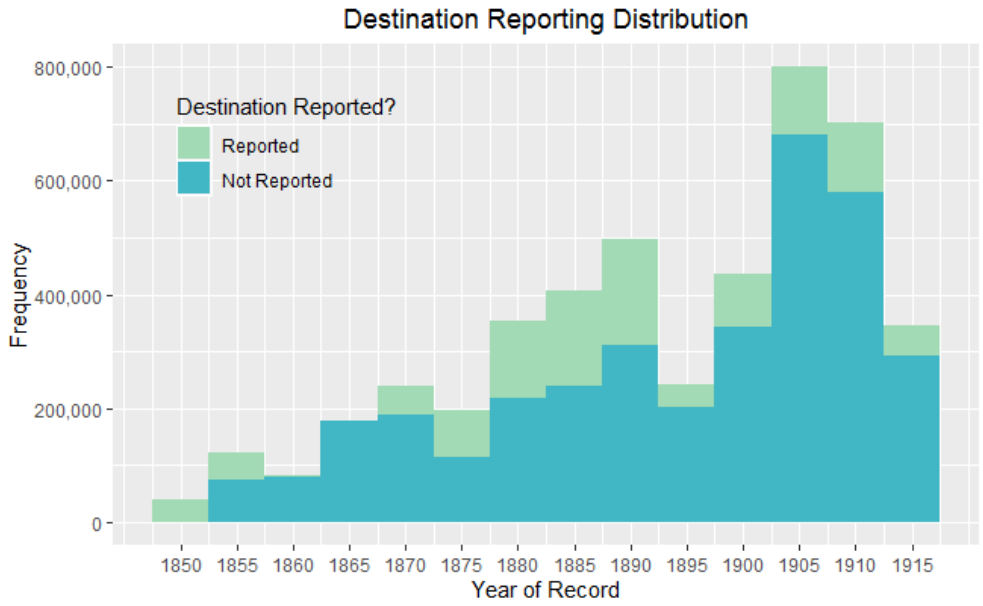


Figure 19: Destination Reporting Over Time

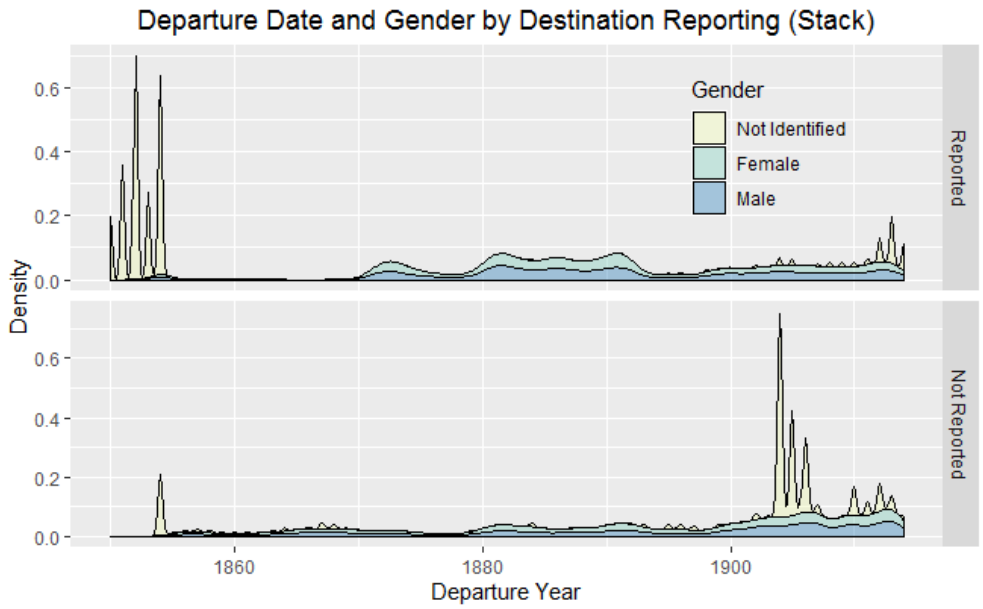


Figure 20:

Origins

- Nationality

Nationality is reported for 45.85% of the records, and this information became mostly available since 1898. The major nationality of HPL records were Russia, Austria, Hungary, the United States, followed by Poland (Table 7, Figure 22 and Figure 21). We do not directly observe nationality in our pre-1898 data, so we use an imputation procedure based on passengers' names and last residences.[14]

	% Available	Russia	Austria	Hungary	United States
Records	2,138,970	816,412	386,902	334,945	168,893
Percentage (%)	45.85*	38.17	18.09	15.66	7.90

Table 7: Nationality Summary

[14]We impute nationality by combining first and last name, age and marital status of individuals (Ye et al. (2017)).

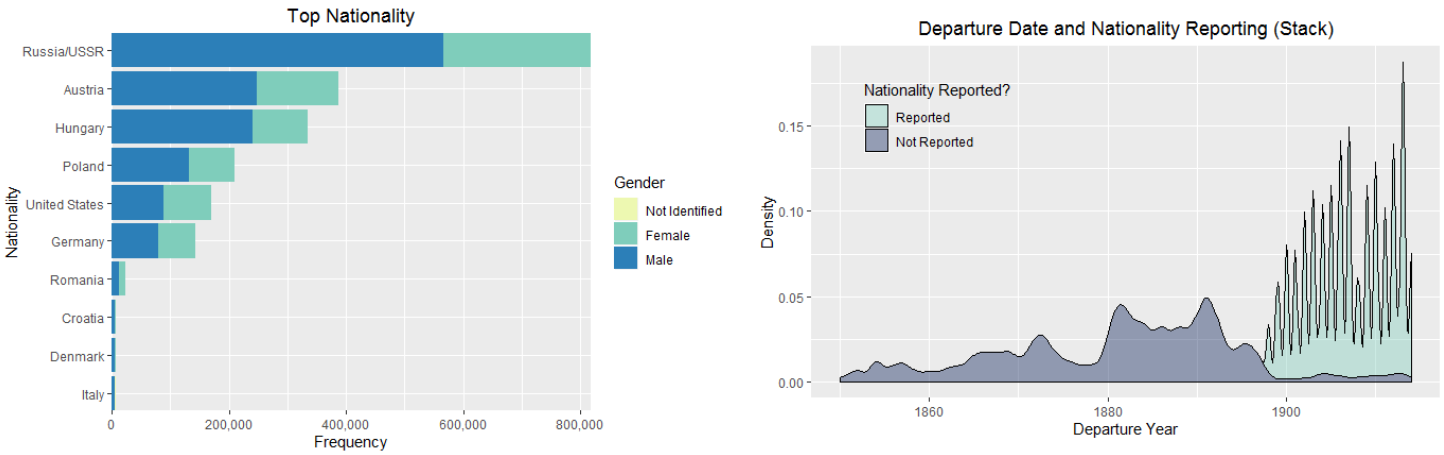


Figure 21: Reported nationalities under current administrative division

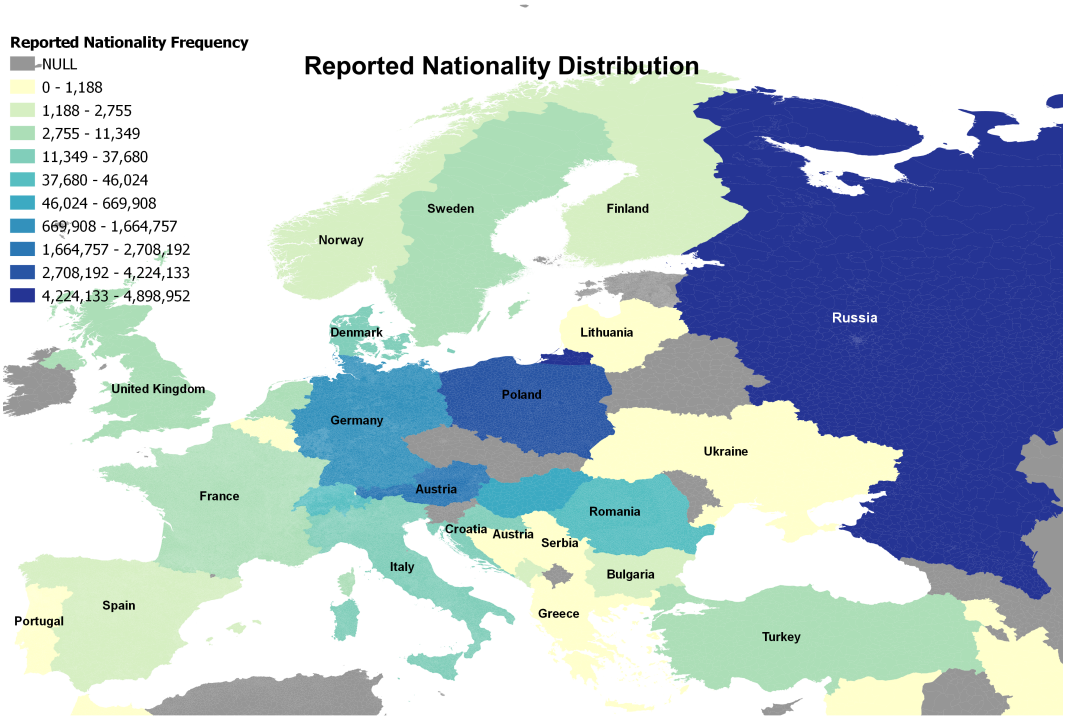


Figure 22: Reported Nationality Distribution (European Area)

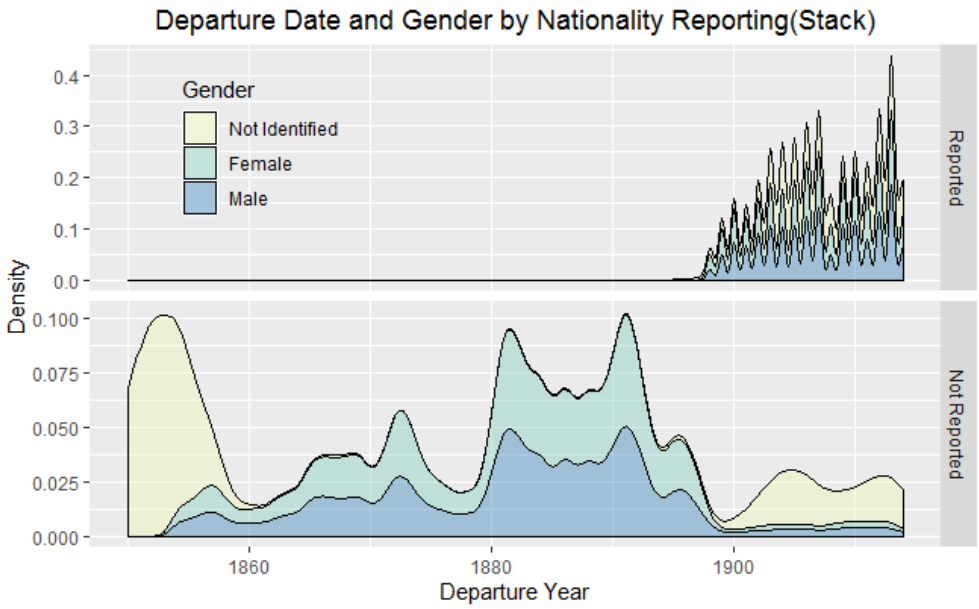


Figure 23: Nationality is reported evenly among gender groups

- Birthplace (Geburtsort)

Birthplace information was available for only 1.89% of the HPL records and it is only available for pre-1860 passengers. Due to its limited coverage of information, we do not use birthplace information as one of linking criteria.

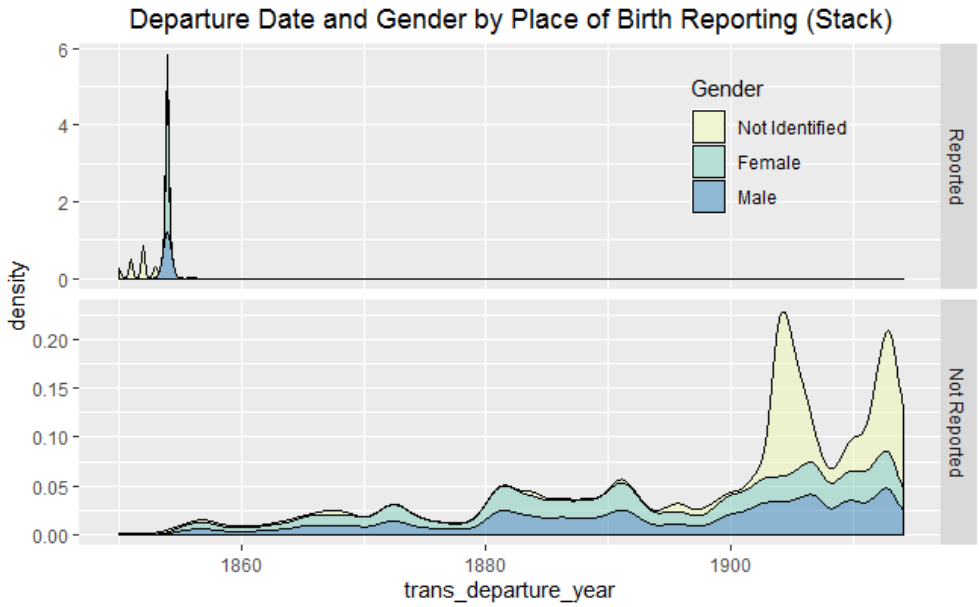


Figure 24: Almost all places of birth are reported before 1860

- Place of Residence(Wohnort)

HPL also records passengers' last residential locations (Table 8). We map transcribed texts of last residence information, harmonize and classify them at the country level using IPUMS geographic classification system. The most frequent last residences are Germany, Poland, Austria and Hungary (Figure 25).

#Obs Last Residence Aavailable	% of Available Records	Reported Missing	Top 10 Places Reported	Top 10 Places Percentage
2,242,447	48.07%	127,787	1,474,037	65.73%

Table 8: Top places of Residence and frequency

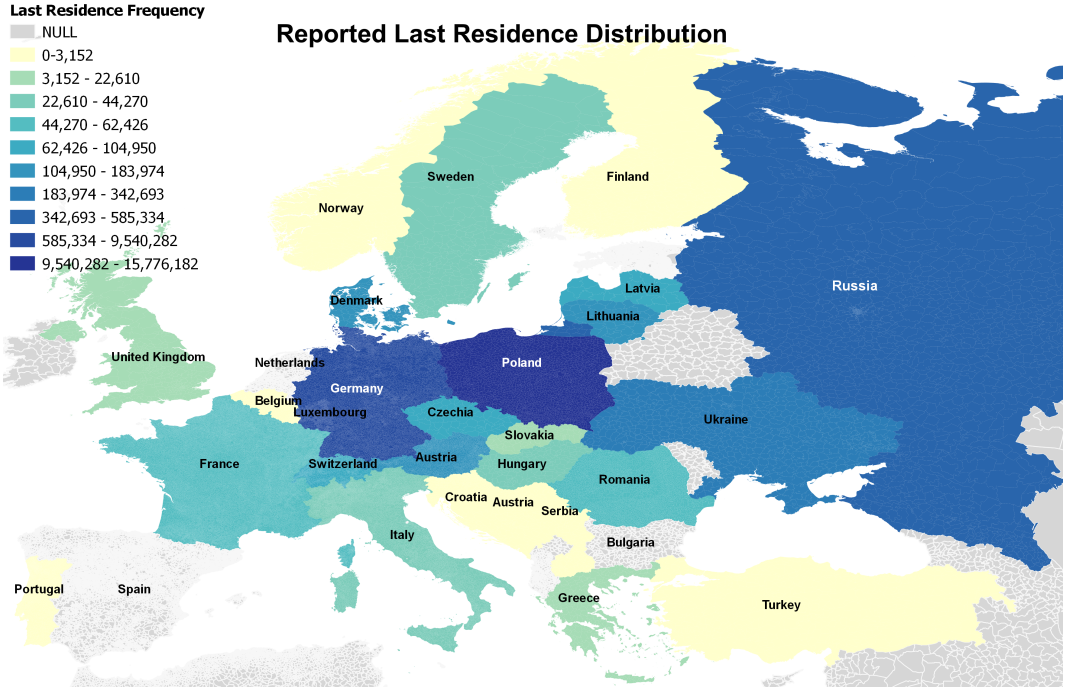


Figure 25: Reported Residence Distribution

E Occupational Classifications

E.1 Hamburg Passenger Lists (HPL) Data Harmonization

As the occupation information in the *Hamburg Passenger Lists* in only available in German and they are neither integrated nor coded with any standardized classification system (recall Mr Steinweg's occupation "Instrumentenmacher" in Hamburg Passenger Lists record), we translate the reported German occupational string and codify them in the occupational system of the US Federal Census. Similarly, the occupational classification available in the *Castle*

Garden Database, while available in English, is also not classified. Therefore, we construct occupational crosswalk between Castle Garden occupation strings and occupational measures which are consistently available in US demographic census (i.e. variaables ''OCCHISCO'' and ''OCC1950'').

E.2 Industry Classification and Harmonization

We also match the reported sectoral groups in the Census of Manufacturers to standard classification of industries that follows the 1950 Census Bureau industrial classification system.[15] The digitized *Census of Manufactures* is more detailed than 1950 Census Bureau industrial classification system. To enhance comparability, we create industry crosswalk from reported industry strings in the Census of Manufacturers to 1950 Census Bureau industrial classification. The same applies to the historical patent data, where we use a combination of the detailed patent descriptions and the patent classifications to assign individuals patents to particular sectors of production.

Finally, we merge all aforementioned datasets at a unified spatial level. While the Population Census and the Patent Data is already geo-referenced, we perform geo-referencing using the original geographical information in the Manufacturing Census. We integrate all county aggregates from 1860 to 1940 at consistent levels of aggregation. As counties and state boundaries have changed over time, we take the county shapefiles from National Historical Geographic Information System, and create the geographic location-based crosswalks of counties across time.

References

Abramitzky, R., Boustan, L. P., Eriksson, K., August 2012. Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. American Economic Review 102 (5), 1832--56.
URL https://ideas.repec.org/a/aea/aecrev/v102y2012i5p1832-56.html

Adao, R., Arkolakis, C., Esposito, F., 2019. Spatial linkages, global shocks, and local labor markets: Theory and evidence.

Akcigit, U., Grigsby, J., Nicholas, T., 2017. Immigration and the rise of american ingenuity. Tech. rep., National Bureau of Economic Research.

Arkolakis, C., Ramondo, N., Rodríguez-Clare, A., Yeaple, S., 2018. Innovation and production in the global economy. American Economic Review 108 (8), 2128--73.

[15] Details are available here: https://usa.ipums.org/usa-action/variables/IND1950#description_section

Atkeson, A., Burstein, A., 2010. Innovation, firm dynamics, and international trade. Journal of Political Economy 118 (3), 433--489.

Dekle, R., Eaton, J., Kortum, S., 2007. Unbalanced trade. Tech. rep., National Bureau of Economic Research.

Feigenbaum, J. J., 2016. Automated census record linking: A machine learning approach.

Gordon, R. J., 2017. The rise and fall of American growth: The US standard of living since the civil war. Princeton University Press.

Kortum, S., 1997. Research, patenting, and technological change. Econometrica 65 (6), 1389--1419.

Lucas Jr, R. E., Moll, B., 2014. Knowledge growth and the allocation of time. Journal of Political Economy 122 (1), 1--51.

Nunn, N., Qian, N., Sequeira, S., 2017. Migrants and the making of america: The short and long run effects of immigration during the age of mass migration.
URL https://ideas.repec.org/a/ces/ifodic/v15y2017i3p50000000000049.html

Perla, J., Tonetti, C., 2014. Equilibrium imitation and growth. Journal of Political Economy 122 (1), 52--76.

Roca, J. D. L., Puga, D., 2017. Learning by working in big cities. The Review of Economic Studies 84 (1), 106--142.

Romer, P. M., 1990. Endogenous technological change. The Journal of Political Economy 98 (5), S71--S102.

Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., Skiena, S., 2017. Nationality classification using name embeddings. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17. ACM, New York, NY, USA, pp. 1897--1906.
URL http://doi.acm.org/10.1145/3132847.3133008