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

Costas Arkolakis
Yale

Sun Kyoung Lee
Yale

Michael Peters
Yale

June 2020

Abstract

We study the role of European Immigration on local and aggregate economic growth in the United States between 1880 and 1920. We employ a big data approach and link, at the individual-level, information from the Population Census, the universe of patents and millions of historical immigration records. We find that immigrants were more prolific innovators than natives, and document large differences in innovation potential across nationalities and regions in the United States. To measure the importance of immigrants for the creation of new ideas and economic growth, we develop a new spatial model of growth through dissemination of knowledge and workers' mobility. The model allows us to use our micro and regional empirical findings to measure immigrants' innovation human capital and the degree of knowledge diffusion which regulates scale effects. We quantitatively analyze the effects of imposing major immigration restrictions on American economic growth in the 19th and early 20th century. We find large, accumulating, losses from these restrictions. Both the scale effects and the exclusion of high-human capital immigrants contribute significantly to these losses.

*We thank Donald Davis, Jose-Antonio Espin-Sanchez, Sam Kortum, Naomi Lamoreaux, Ellen McGrattan, Ferdinando Monte, Dan O'Flaherty, Paul Rhode, and Esteban Rossi-Hansberg for comments and suggestions. We are grateful to Ariel Boyarsky, Shizuka Inoue, Jack Liang, and Roxane Spitznagel for outstanding research assistance. The authors would like to thank the NSF for support under RIDIR grant #1831524 as well as the NBER and Kauffman Foundation for the Entrepreneurship Grant. Michael Peters thanks the Opportunity and Inclusive Growth Institute at the Federal Reserve Bank of Minneapolis for their hospitality. All errors are our own.

1 Introduction

The transformation of the US economy in the last two hundred years has been remarkable. While being primarily rural at 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 technology leader and the richest nation on the globe (Gordon, 2017). During this period, which is also referred to as the "The Second Industrial Revolution", the US economy also experienced a massive inflow of immigrants, mostly from the European continent. This begs the question to what extent such inflows were an important driver of productivity growth in the US.

To answer this question, we undertake a big-data approach and construct a novel micro-data set on immigrants' patent behavior. First, we match the population of US patents to the restricted-use complete count US Federal demographic decennial censuses from 1880-1930 using modern record-linking techniques. Second, we exploit the original immigration records and historical passenger lists from the ships heading from Europe to the US, which contain direct microdata on immigrants' *pre*-migration occupations.[1] Doing so allows us to measure the extent to which skills were indeed geographically portable, i.e. whether immigrants in particular occupations were able to capitalize on such knowledge and generate more patents in the US after their arrival.

These data allow us to precisely measure the main theoretical mechanisms through which such international population inflows could have increased productivity in the US. First, virtually all models of economic growth feature "scale effects", whereby either the level or the growth rate of productivity is increasing in the size of the population. Second, the inflowing immigrants could have played an important role in the process of technological growth through their human capital, whereby Italian textile workers and German craftsmen bring their knowledge of weaving techniques and carpenter skills to the American shore.

We analyze this wealth of data and measure the importance of scale effects and immigrant human capital through the lens of a parsimonious general equilibrium model of innovation. At the heart of the model is the local supply of innovative human capital to generate new varieties. Individuals can either work as production workers or become entrepreneurs by inventing a new product. While goods are tradable, labor markets are local, i.e. workers

[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 haveyet to be used in empirical research.

produce using the technologies brought about by local entrepreneurs. Hence, the size (i.e. scale) and the composition (i.e. human capital) of the local population determine local productivity at each point in time. In that sense, our theory combines the classical idea-based growth model in the spirit of Romer (1990) and a modern model of economic geography, where the spatial distribution of the population determines aggregate economic activity. This combination of our spatial-growth setup with rich micro and regional data allows us to estimate both the scale effect and the human capital of different immigrant groups and separate them from differences in innovation potential between cities and rural areas.

To do so, we proceed in two steps. We first use the variation across individuals within local labor markets to estimate differences in human capital between natives and immigrants of different nationalities. In particular, we show that our theory implies a log linear relationship between patent probabilities and innovation human capital and that we can estimate the human capital of different groups from our matched microdata. We then use the panel dimension of our data across regions to estimate both the strength of scale effects and regional heterogeneity in patent creation. We again exploit the fact that our theory generates a closed-form expression for the law-of-motion of regional patenting, which allows us to estimate these structural parameters directly from the data.

Our empirical results imply that immigrants were an important contributor to US growth between 1880 and 1920. At the micro-level we document that (i) immigrants created - on average - more patents than natives, (ii) that migrants were particularly prolific after being in the US for at least ten years and (iii) that there are stark differences across nationalities with German and British immigrants being far more innovative than Italian immigrants. We also provide direct evidence that these patterns are plausibly due to differences in human capital, the channel, which is at play in our theory. In particular, using the information contained in the historical migration records, we show that immigrants with more innovative pre-migration occupations are more likely to patent after their arrival in the US. We also show that the patents issued by immigrants were qualitatively different from the ones from their native peers in that they show less similarity (as measured from the words in the patent description) to patents issued in the same location in the past. Both of these results are consistent with the idea that immigrants were able to capitalize on their pre-migration knowledge to generate innovations in the US.

At the regional level, the model delivers a structural regression equation to link regional patent growth to the size and composition of the local population, past patent activity, and a county fixed effect.[2] These determinants of regional innovation all have an intuitive

[2]Hence, our theory provides precise structural underpinnings to difference-in-difference approaches suggested in the literature (see Ottaviano et al. (2013), Sequeira et al. (2020); Burchardi et al. (2020)).

structural interpretation. In particular, the effect of past patenting on future patent growth is related to the importance of local knowledge spillovers and coincides with the strength of scale effects and the county fixed effect represents systematic differences in the efficiency of innovation across locations. Our estimates imply that scale effects are sizable and that immigration improves the extent of spatial sorting because immigrants are more likely to live in urban locations, which we estimate to have higher innovation efficiency.

We finally evaluate the importance of immigrants for the process of American Growth between 1880 and 1920 by undertaking a quantitative assessment of the importance of immigration restrictions. We parametrize our model using our empirical findings at the micro and cross-regional level and then study the aggregate effects of immigration restrictions at the federal level. We calibrate the model to the cross-sectional data of the native and immigrant population in 1880 and then solve for the dynamic evolution of the economy under different assumptions on immigration policy. Following Card (2001), we assume that immigrants' initial location choice is related to the existing share of immigrants of their own nationality. This implies that the initial distribution of ancestry determines the spatial exposure to aggregate immigration policy. To isolate the impact of immigrants we abstract from other sources of population growth.

Our model allows us to study the aggregate and local consequences of such a policy. At the aggregate level, we find that immigration restrictions have a significant, accumulating, impact on economic growth. Quantitatively, our model implies that GDP per capita in 1920 would have been 30% lower if all immigration inflows had been completely shut down in 1880. To isolate the importance of general scale effects versus the loss of foreign human capital, we compare immigration restrictions on German and British immigrants relative to Italians. Recall that we estimate the former to have much more innovative human capital than the latter. We indeed find that restricting German and British immigration has more adverse consequences. Without them, GDP pc would have been 10% lower in 1920. In contrast, without Italian immigrants, the aggregate loss in income per capita would have been less than 5%. More than half of this difference can be attributed to the loss of human capital from high skill immigration origins.

Our model also highlights that the consequences of immigration policy at the aggregate level are not homogeneous across space - both because different regions are differentially exposed to immigration inflows and because spatial heterogeneity in innovation efficiency induces spatial sorting by human capital. In our application, we for example find that the geographic distribution of the economic gains attributable to German versus Italian immigrants is vastly different.

Our paper contributes to the literature on growth, innovation human capital, and scale

effects (see for example Jones (2005) and Akcigit (2017) for surveys). The combination of individual-level information on immigrants' patenting behavior and regional data on county-level patent growth is helpful to distinguish regular scale effects from the possibility of immigrants' fostering idea flows, which feature prominently in recent theories of economic growth (Kortum, 1997; Lucas Jr and Moll, 2014; Perla and Tonetti, 2014; Buera and Oberfield, 2020). Empirically, we estimate that scale effects are sizable and consistent with models of semi-endogenous growth as proposed in Jones (1995).

Amidst the rising prominence of immigration flows, especially high-skill immigration, for the world economy (Kerr et al. (2016); Kerr (2018)) there has been a growing interest in estimating the economic effects of immigration. Traditionally, the economics literature was concerned with the short-run consequences on labor outcomes of natives (see e.g. Card (1990), (Burstein et al., 2020), Dustmann et al. (2016) of Peri (2016) for a survey). A set of recent papers focuses (like we do) on the long-run impact. (Sequeira et al., 2020) and Akcigit et al. (2017) document a large positive effect of immigration inflows in the 19th century on economic activity and patent creation today. Hornung (2014) uses data on textile plants to analyze the productivity effects of the Huguenot re-settlement for the 18th century. Burchardi et al. (2020) find a positive effect of immigration inflows on patenting and local growth in the last three decades. Our paper complements these studies both by analyzing patenting behavior at the micro-level and by proposing a structural model, which allows us to interpret and aggregate the empirical findings and perform counterfactual exercises.

We also contribute to a growing literature in spatial economics. Early contributions focus on static models to characterize the spatial allocation of resources taking regional productivity as given - see for example Allen and Arkolakis (2014) or the recent survey by (Redding and Rossi-Hansberg, 2017). A growing set of papers models the evolution of regional productivity explicitly (e.g. Desmet et al. (2018), Nagy (2020), Peters (2019), or Walsh (2019)). To the best of our knowledge, ours is the first paper to connect a micro-founded model of spatial growth with direct evidence on regional patenting at the individual level.

The rest of the paper is structured as follows. Section 2 presents the theory and its main insights. Section 3 presents the data and the main facts about immigrants, their skills, and patterns of innovation. Section 4 presents reduced form evidence based on a difference-in-difference specification suggested by the model. Section 5 contains the counterfactual exercises. Section 7 concludes.

2 Theory

We consider an economy with R regions, denoted by r. Consumers have preferences over a CES bundle of varieties indicated by ω

$$U_r = \left(\int_{\omega=0}^{N} c_r(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right)^{\frac{\sigma}{\sigma-1}} = \left(\sum_{r=1}^{R} \int_{\omega=0}^{N_r} c_r(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right)^{\frac{\sigma}{\sigma-1}},$$

where N_r denotes the number of varieties being produced in region r, $N = \sum_r N_r$ denotes the aggregate number of varieties available in the economy, and $c_r(\omega)$ is the consumption of variety ω in region r. Suppose there are $\{L_r^j\}_{r,j}$ individuals of type $j = 1, 2, 3..., J_r$ in region r that can engage in innovation or work as workers. The type of individuals determines their human capital endowment and ultimately if they will be workers or innovators, as we discuss below.

Varieties are produced with constant returns to scale using only labor and firms compete monopolistically competitive and trade is free. We consider the aggregate final good as a numeraire and denote aggregate income by $\Upsilon = \sum_r \Upsilon_r$ with Υ_r being the income in region r. Standard arguments imply that profits of firm ω in region r are given by

$$\pi_r(\omega) = (p_r(\omega) - w_r) y(\omega) = \frac{1}{\sigma} \left(\frac{\sigma}{\sigma - 1} \right)^{1-\sigma} w_r^{1-\sigma} \Upsilon, \tag{1}$$

where $y(\omega)$ is the quantity of good ω that is produced. Hence, profits are increasing in total income Υ and decreasing in the local wage w_r. Because profits are constant across producers within a location r, we simply denote profits by π_r.

2.1 Creation of Varieties

Varieties are created by individuals engaging in innovation. We therefore also sometimes refer to varieties as ideas and to people who generate varieties as either innovators or entrepreneurs. In our empirical application, we will connect varieties to patents.

Individuals can either start a new firm or work as production workers. Each individual has one unit of time to work as a production worker. In contrast, starting a firm requires a single unit of entrepreneurial human capital. The total return from having h units of human capital and hence h varieties is therefore given by $h\pi_r$, where π_r is given in (1). Individuals differ in their innovative abilities. In particular, suppose region r is inhabited by different types j that draw their entrepreneurial human capital from a distribution $h \sim H_r^j(h)$. Note that the distribution of innovative abilities differs across types and across space. Empirically,

we will relate the different types to immigrants of different nationalities.

An individual, therefore, decides to create a new idea rather than work as a production worker as long as $h\pi_r \geq w_r$. This implies that there is a region-specific human capital cutoff to enter entrepreneurship in region r, h_r^*, which is given by

$$h_r^* = \frac{w_r}{\pi_r} = \frac{1}{\frac{1}{\sigma}\left(\frac{\sigma}{\sigma-1}\right)^{1-\sigma}} \frac{w_r^\sigma}{\Upsilon}, \tag{2}$$

where the second equality uses (1). Hence, high wage locations ''toughen'' the entrepreneurial cutoff both because of higher opportunity costs and because of lower profits conditional on a successful idea. In contrast, a larger market size Υ increases profits and hence induces worse entrepreneurs to engage in idea creation.

2.2 Heterogeneity in Innovative Ability and Aggregate Innovation

We assume that innovative ability, h, is drawn from a Pareto distribution

$$H_r^j(h) = 1 - \left(\frac{\psi_r^j}{h}\right)^\theta, \tag{3}$$

where ψ_r^j determines the average level of human capital of individuals of type j in region r. We put more structure on ψ_r^j below. In anticipation of the introduction of dynamics and growth in Subsection 2.4 we note that it can be specified as a function of time or aggregate variables. At this point, we simply note that some regions might be better suited to generate ideas ($\psi_r^j > \psi_{r'}^j$) and different types could be of different innovative talent within a location ($\psi_r^j > \psi_r^{j'}$). Note also that the share of innovators of type j in region r is given by

$$e_r^j \equiv \left(\frac{\psi_r^j}{h_r^*}\right)^\theta,$$

as everyone with $h \geq h_r^*$ becomes an entrepreneur.

Given (3), the total number of varieties created in region r is given by

$$N_r = \sum_j L_r^j \int_{h_r^*}^\infty h dH_r^j(h) = \frac{\theta}{\theta-1} h_r^* \sum_j L_r^j \left(\frac{\psi_r^j}{h_r^*}\right)^\theta = \frac{\theta}{\theta-1} h_r^* \sum_j L_r^j e_r^j. \tag{4}$$

Hence, the number of active firms in region r is increasing in the size of the population L_r^j, increasing in the supply of human capital ψ_r^j and decreasing in the entrepreneurial cutoff h_r^*.

Furthermore, the total labor payments are given by

$$w_r \sum_j L_r^j P\left(h \leq h_r^*\right) = \left(1 - \frac{1}{\sigma}\right) N_r \left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma} w_r^{1-\sigma} \Upsilon.$$

Using equations (2), (4), letting $\varpi_r^j = L_r^j / L_r$ be the share of people of type j in region r, and simplifying we can solve explicitly for the entrepreneurial human capital cutoff h_r^* as

$$h_r^* = \left(\frac{\theta\sigma - 1}{\theta - 1}\right)^{1/\theta} \left(\sum_j \varpi_r^j \left(\psi_r^j\right)^\theta\right)^{1/\theta}. \tag{5}$$

The solution for h_r^* is akin to a ''CES-weighted average'' of the population-weighted human capital indices ψ_r^j. Notice that if there is no heterogeneity in innovative ability, i.e. $\psi_r^j = \psi_r$, equation (5) implies that $h_r^* = \left(\frac{\theta\sigma-1}{\theta-1}\right)^{1/\theta} \psi_r$, i.e. the entry cutoff is always proportional to the supply of human capital in region r, ψ_r.

This endogenous human capital cutoff h_r^* will be a key statistic to tractably characterize the equilibrium. Conveniently, it allows us to consider arbitrary heterogeneity across groups j and, as we will show below, it will also yield a convenient way to estimate it. Considering the average innovation share in each region r,

$$e_r \equiv \sum_j \varpi_r^j e_r^j = \frac{\theta - 1}{\theta\sigma - 1}, \tag{6}$$

we see that is constant and independent of the supply of human capital ψ_r^j or the size of the population L_r^j. It does, however, depend on the elasticity of substitution σ and the tail of the skill distribution, θ: if products are more substitutable, profit margins decline and less resources are allocated to the creation of ideas. Similarly, the lower θ, i.e. fatter the tail of the talent distribution, the lower the share of entrepreneurs: if superstar entrepreneurs generate many ideas, the entrepreneurial market becomes more competitive and one can economizes on entrepreneurs.

Despite the fact that the share of entrepreneurs is constant across space the number of varieties may differ across locations. Using the variety creation schedule (4) and equation (6) yields

$$N_r = \frac{\theta}{\theta - 1} h_r^* L_r \sum_j \varpi_r^j \left(\frac{\psi_r^j}{h_r^*}\right)^\theta = \frac{\theta}{\theta\sigma - 1} h_r^* L_r. \tag{7}$$

Hence, the number of ideas created in region r is a function both of the size of the population L_r and the entrepreneurial cutoff h_r^*. Given that the cutoff is increasing the level of human capital

ψ_r^j, innovative locations do not have *more* entrepreneurs - they have *better* entrepreneurs and hence generate more ideas in equilibrium. Importantly, given that h^* is fully determined from the composition of the local population ϖ_r^j and that the distribution of human capital ψ_r^j, (7) allows to compute the number of local varieties N_r independently of other equilibrium variables like the regional distribution of wages w_r.

2.3 Equilibrium

To calculate aggregate income in region r, recall that income stems from two sources: labor income and entrepreneurial profits. The Pareto distribution implies that total income of group j in region r is given by

$$w_r L_r^j P\left(h \leq h_r^*\right) + L_r^j \int_{h_r^*}^{\infty} h \pi_r dH_r^j\left(h\right) \quad = \quad L_r^j \left[1 + \frac{1}{\theta - 1}\left(\frac{\psi_r^j}{h_r^*}\right)^{\theta}\right] w_r. \tag{8}$$

Income is naturally increasing in the regional wage w_r and - holding the wage w_r constant - increasing in the share of entrepreneurs $e_r^j = \left(\psi_r^j / h_r^*\right)^{\theta}$, because all infra-marginal entrepreneurs with $h > h_r^*$ earn a rent relative to production workers. Note also that (6) implies that regional income Υ_r is given by

$$\Upsilon_r = \sum_j L_r^j \left[1 + \frac{1}{\theta - 1}\left(\frac{\psi_r^j}{h_r^*}\right)^{\theta}\right] w_r = \frac{\theta \sigma}{\theta \sigma - 1} L_r w_r, \tag{9}$$

i.e. per capita-income is simply proportional to the wage.

To characterize the equilibrium we also need to consider the demand side. Sales per variety are given by

$$x_r\left(\omega\right) \equiv \sigma \pi_r\left(\omega\right) d\omega = \left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma} w_r^{1-\sigma} \Upsilon.$$

Notice that we have considered a simple case where varieties are freely traded across regions. The model can be easily extended to incorporate trade frictions following a voluminous literature in the trade and geography gravity literature (see e.g. Arkolakis et al. (2012); Head and Mayer (2013); Allen and Arkolakis (2014); Donaldson and Hornbeck (2016)) which will simply change the functional form for firm sales per variety. We proceed with this extension in our quantitative application.

Market clearing therefore requires that income equals demand and thus

$$\Upsilon_r = w_r \sum_j L_r^j \left[1 + \frac{1}{\theta - 1}\left(\frac{\psi_r^j}{h_r^*}\right)^\theta\right] = \int_0^{N_r} x_r(\omega)\, d\omega. \tag{10}$$

To summarize, given the solution for the entrepreneurial cutoffs $\{h_r^*\}_r$ from (6), the equation solving for varieties (4), and the market clearing condition, equation (10), we can solve for the distribution of wages $\{w_r\}_r$ across locations.

In the case with no trade frictions equations (6) and (7) imply that regional wages are given by

$$w_r^\sigma = \frac{N_r}{L_r}\left(\frac{\theta\sigma - 1}{\theta\sigma}\right)\left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma}\Upsilon = h_r^*\frac{1}{\sigma}\left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma}\Upsilon. \tag{11}$$

Hence, regional wages are increasing in the number of productive ideas per person, N_r/L_r, which in equilibrium is proportional to the entrepreneurial cutoff h_r^* and hence determined by the supply of human capital $\{\psi_r^j\}_j$. In particular, using (9), (11) determines $\{w_r\}$ as a function of L_r and h_r^*. (11) has the important implication that the cross-sectional variation in wages is *only* determined from the innovation cutoff h_r^*

$$\frac{w_r}{w_m} = \left(\frac{h_r^*}{h_m^*}\right)^{1/\sigma}.$$

Hence, relative wages do not depend on regional scale L_r, but solely on the composition of the population and their human capital endowments ψ_r^j.[3]

2.4 Dynamics and Growth in Space

We now put more structure on the determinants of regional human capital ψ_r^j and incorporate dynamics into our spatial framework. We henceforth use the subindex t to denote the time period of each variable considered. We assume that regional human capital is a function of type-specific and region-specific fixed effects, and following a tradition in the growth literature

[3]Note that our model does feature scale effects in the aggregate. Using (11) and (9) yields

$$w_r^\sigma = h_r^*\left(\frac{\sigma}{\sigma - 1}\right)^{1-\sigma}\frac{\theta}{\theta\sigma - 1}\sum_r L_r w_r.$$

Given that h_r^* is independent of L_r, an increase in the population by 1% increases wages in all locations by $\frac{1}{\sigma-1}$%. These are the usual variety gains of Krugman (1980) and Romer (1990), which operate at the aggregate level, not regional level.

(e.g. Romer (1990)) also consider a growth externality,

$$\psi_{rt}^{j} = \underbrace{\zeta_r}_{\text{Location FE}} \quad \underbrace{\varphi^j}_{\text{Type heterogeneity}} \quad \underbrace{N_{rt-1}^{\vartheta}}_{\text{Local accumulation}} . \tag{12}$$

Here, ζ_r is a county fixed effect, which captures regional differences in the creation of human capital across space. Hence, a given unit of human capital might be more productive in for example urban locations relative to the countryside. The parameter φ^j governs differences in human capital endowments across types j. In our application, heterogeneity in innovative ability between natives and immigrants or between immigrants of different nationality will be captured by φ^j. Finally, the term N_{rt-1}^{ϑ} captures the extent to which existing local varieties are an input into the production of new entrepreneurial human capital. This is the usual inter-temporal spillover generating long-term growth in idea-based models of growth (Jones, 2005). The parameter ϑ governs the strength of such spillovers.

The combination of h_{rt}^{*} being akin to a CES average of human capital endowments (see (5)) and the log-additive nature of regional human capital ψ_{rt}^{j} in (12) implies that we can derive a simple dynamic law of motion for the innovation threshold in region r time t, h_{rt}^{*}:

$$h_{rt}^{*} = \left(\frac{\theta\sigma-1}{\theta-1}\right)^{1/\theta}\left(\sum_j \varpi_{rt}^{j}\left(\psi_{rt}^{j}\right)^{\theta}\right)^{1/\theta} = \varrho\left(\sum_j \varpi_{rt}^{j}\left(\varphi^{j}\right)^{\theta}\right)^{1/\theta}\zeta_r\left(h_{rt-1}^{*}\right)^{\vartheta}L_{rt-1}^{\vartheta}, \tag{13}$$

where $\varrho \equiv \left(\frac{\theta\sigma-1}{\theta-1}\right)^{1/\theta}\left(\frac{\theta}{\theta\sigma-1}\right)^{\vartheta}$ is a collection of inconsequential constants. Here we used (7) to substitute for N_{rt-1}.

Equation (13) fully characterizes the dynamic path for $\{h_{rt}\}_t$ given an initial condition h_{r0} and a sequence of population levels $\left\{L_{rt}^{j}\right\}_t$ as these also determine the composition of the local workforce $\left\{\varpi_{rt}^{j}\right\}_t$. In particular, it implies that

$$\ln h_{rt}^{*} = \varsigma + \ln\zeta_r + \vartheta\ln h_{rt-1}^{*} + \vartheta\ln L_{rt-1} + \ln\left[\sum_j \varpi_{rt}^{j}\left(\varphi^{j}\right)^{\theta}\right]. \tag{14}$$

Hence, $\ln h_{rt}^{*}$ follows an AR(1) process with persistence ϑ. In addition there is a regional drift ζ_r brought about through local human capital creation, a positive effect on today's human capital though the size of the past population L_{rt-1}, which positively affects human capital though the creation of non-rival ideas and the composition of human capital.

Equation (12) highlights the three main determinants of the supply of local human capital: regional innovation efficiency $\{\zeta_r\}_r$, individual innovation efficiency $\{\varphi^j\}_j$ and the strength of knowledge spillovers ϑ. Below we show that we can directly estimate all of these objects

using a combination of micro-data and spatial variation. In particular, we show that we can turn (14) into an estimation equation that allows us to recover ϑ and $\{\zeta_r\}$ from regional variation in patent activity. We also show how we can identify human capital differences from our micro data on patenting. The key property that allows us to do so is that the endogenous cutoff h_r^* is akin to a CES aggregator of the human capital of the different types presenting in location r - see equation (5). This CES-structure and the our assumptions on human capital ψ_r^j in (12) imply that the entrepreneurial share of *any* group of individuals $\mathcal{G}$ within a region r is given by

$$e_r^{\mathcal{G}} = \frac{\sum_{j\in\mathcal{G}} L_r^j \left(\frac{\psi_r^j}{h_r^*}\right)^\theta}{\sum_{j\in\mathcal{G}} L_r^j} = \sum_{j\in\mathcal{G}} \varpi_r^{j\mathcal{G}} \left(\frac{\psi_r^j}{h_r^*}\right)^\theta = \frac{\theta-1}{\theta\sigma-1} \frac{\sum_{j\in\mathcal{G}} \varpi_r^{j\mathcal{G}} \left(\psi_r^j\right)^\theta}{\sum_j \varpi_r^j \left(\psi_r^j\right)^\theta} = e_r \frac{\sum_{j\in\mathcal{G}} \varpi_r^{j\mathcal{G}} \varphi_j^\theta}{\sum_j \varpi_r^j \varphi_j^\theta}$$

where $\varpi_r^{j\mathcal{G}} = L_r^j / \sum_{j\in\mathcal{G}} L_r^j$ denotes the share of j types within group $\mathcal{G}$ and $e_r = \frac{\theta-1}{\theta\sigma-1}$ is the innovator share in region r. Hence, the share of innovators within a group *relative* to its regional average, $e_r^{\mathcal{G}}/e_r$, reflects exactly the population-weighted average of human capital within the group, relative to regional average. This relationship allows us to estimate the innovation human capital of different types of immigrants from their patenting behavior.

2.5 The Effects of Immigration Inflows

How does an inflow of immigrants into a location change innovation and economic activity? Equation (14) highlights that immigrants play two distinct roles in the determination of h_{rt}^* and hence wages and income per capita:

1. Immigrants affect local human capital provision $\left[\sum_j \varpi_{rt}^j \left(\varphi^j\right)^\theta\right]^{1/\theta}$. If immigrants are endowed with more (less) innovative human capital than natives, a higher share of immigrants increases (lowers) h_{rt}^* holding h_{rt-1}^* and L_{rt-1} fixed. We refer to this channel as the *local human capital supply channel.*

2. The second is the size effect embedded in L_{rt-1}. Holding h_{rt}^* constant, a larger population increases the number of regional varieties N_{rt}, which in turn leads to higher future innovation cutoff h_{rt+1}. We refer to this channel as the *local scale channel.*

To illustrate these effects more clearly, suppose there are two groups - immigrants and natives - and let their human capital parameters be φ^I (for immigrants) and φ^N (for natives). Suppose that $\varphi^N = 1$ so that φ^I denotes immigrants' innovative potential relative to natives.

If the share of immigrants is small, i.e. $\varpi_{rt}^I \approx 0$, then

$$\ln\left[\sum_j \varpi_{rt}^j \left(\varphi^j\right)^\theta\right] = \ln\left[\left(1-\varpi_{rt}^I\right)\left(\varphi^N\right)^\theta + \varpi_{rt}^I\left(\varphi^I\right)^\theta\right] \approx \left[\left(\varphi^I\right)^\theta - 1\right]\varpi_{rt}^I.$$

Now consider the following experiment: consider a location that contains only a measure of L natives. At t_0, the economy experiences an inflow of immigrants, which increases the share of immigrants to ϖ^I. Hence, the local population also increases by ϖ^I. The immigration shock is permanent. Equation (14) then implies that for $k \geq 0$, this permanent inflow at time t_0 changes the human capital cutoff at time $t_0 + k$ in the following way:

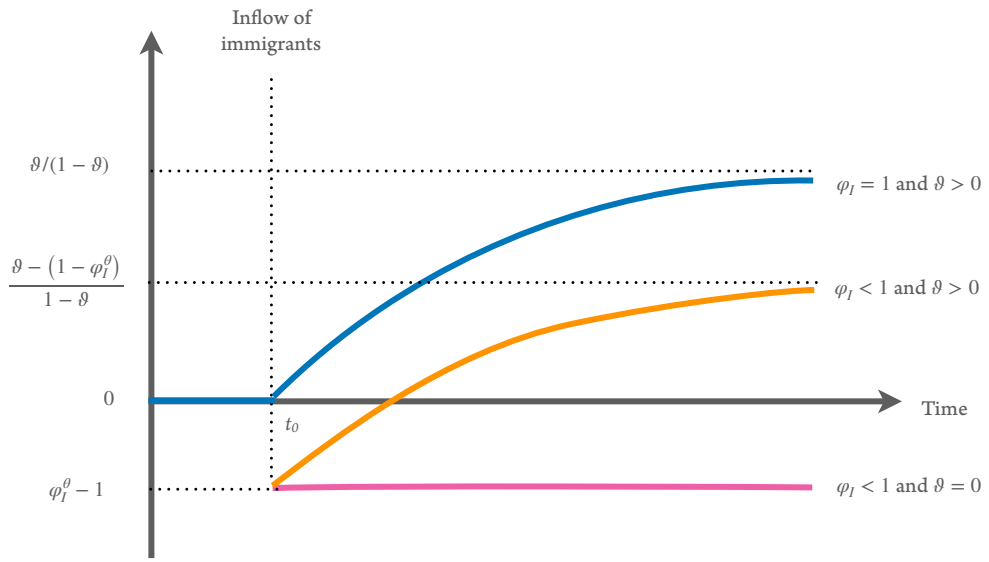
$$\frac{d\ln h_{rt_0+k}^*}{d\varpi^I} = \underbrace{\left(\left(\varphi^I\right)^\theta - 1\right)\sum_{i=0}^{k}\vartheta^i}_{\text{Human capital supply}} + \underbrace{\sum_{i=1}^{k}\vartheta^i}_{\text{Local scale}} = \left(\varphi^I\right)^\theta\sum_{i=0}^{k}\vartheta^i - 1. \tag{15}$$

Recall that regional income per capita is proportional to h_{rt}^* so that (15) also describes the dynamic evolution of income per capita. Equation (15) highlights three aspects: First, through the human capital supply channel, immigrant inflows can be positive or negative. If $\varphi^I < 1$, i.e. immigrants are less innovative than natives, immigrant inflows water down the average human capital pool and local wages decline as per capita idea creation falls. If immigrants are more innovative, $\varphi^I > 1$, the opposite is the case. Second, the aggregate effect of immigrant inflows depend crucially on the parameter ϑ. If $\vartheta > 0$, ideas are non-rival in the creation of new knowledge: *everyone's* human capital in region r benefits from the ideas invented in the past. In that case, the effect of immigration is cumulative. Third, the local scale effect is only active if $\vartheta > 0$ and is always positive: natives and migrants like benefit from the larger population if they can build on their shoulders. Note that the effect of immigration inflows is weakly increasing over time and that

$$\lim_{k\to\infty}\frac{d\ln h_{rt_0+k}^*}{d\varpi^I} = \frac{\left(\varphi^I\right)^\theta}{1-\vartheta} - 1 = \frac{\vartheta - \left(1-\left(\varphi^I\right)^\theta\right)}{1-\vartheta}.$$

Hence, the effect of immigration is a horserace between two forces: the dynamic growth externality captured in ϑ and natives' human capital relative to immigrants $1-\left(\varphi^I\right)^\theta$. If the former is smaller than the latter, immigrant inflows will reduce income per capita both in the short- and the long-run.

In Figure 1 we display the impulse response function in (15) graphically. We consider three cases. In the first case, depicted in solid blue, immigrants are as innovative as natives (i.e. $\varphi^I = 1$) and past ideas aide in creating new ideas ($\vartheta > 0$). The short-run effect of an

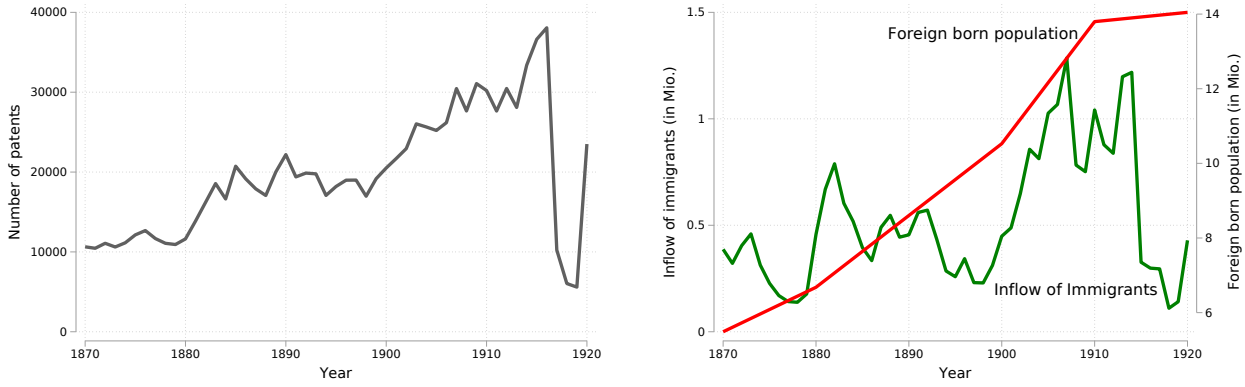


Notes: The figure shows the impulse response function in (15). The immigration shock is permanent and occurs at t_0. We depict three cases, which differ in the innovative human capital of immigrants $\left(\varphi^I\right)$ and the size of the inter-temporal spillover (ϑ).

Figure 1: The Dynamic Impact of Immigrations

immigrant shock is zero as immigrants leave the region's human capital stock unchanged. The long-run effect is positive because both natives and immigrants build on the non-rival ideas created in the past. The second case, depicted in red, is an example of a shock where immigrants are less skilled than natives and there is no inter-temporal spillover, i.e. $\vartheta = 0$. In that case, immigrants reduce regional income per capita permanently, as they reduce average human capital and hence idea creation. The third case, depicted in orange, is akin to the second case, i.e. immigrants are less innovative than natives. However, now ideas are a sufficiently strong factor of idea production, i.e. $\vartheta > 1 - \left(\varphi^I\right)^{\theta}$. As seen in Figure 1, now the effect of immigrants is *ambiguous.* In the short-run the negative effect through their lower human capital dominates and income per capita decline. In the long-run, the dynamic accumulation of ideas where natives' and immigrants' human capital builds on past ideas dominates and the long-run effect is positive.

In the remainder of the paper we explore these implications empirically. In Section 3 we use detailed matched microdata of immigrants and their patent behavior to measure immigrants' relative human capital. In Section 5 we use regional variation to estimate the knowledge spillover elasticity ϑ and the spatial heterogeneity in innovation efficiency. Finally, in Section 6 we use our model and these estimates to quantify the role of immigrants for regional and aggregate growth. To do so, we extend the framework by a number of realistic



Notes: The figure show the number of patents issued each year between 1870 and 1920 (left panel) and the number of immigrants arriving in the US and the stock of the foreign-born population (right panel).

Figure 2: Patenting and Immigration Inflows: 1880 and 1920

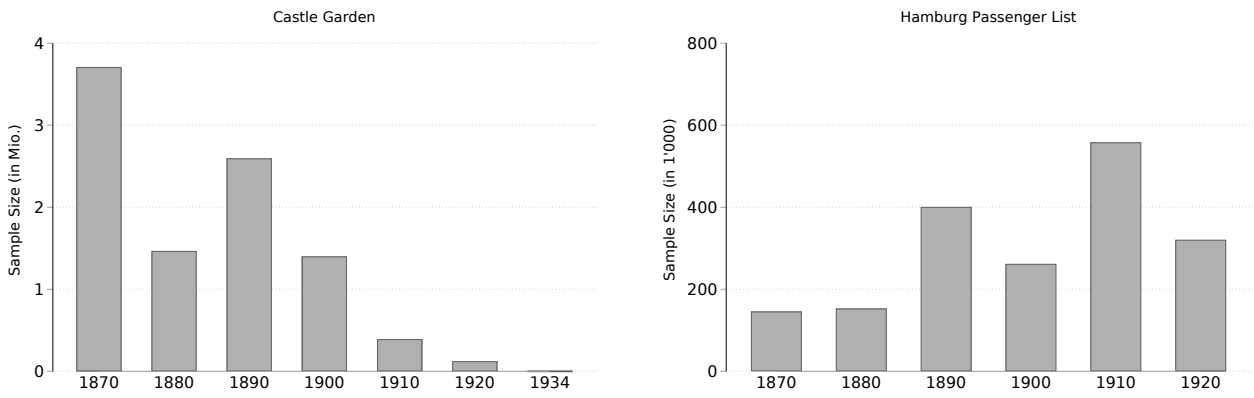
features, in particular trade costs, and internal migration.

3 Empirical Analysis

Guided by our theoretical analysis in Section 2 we now turn to the empirical analysis. We focus on the time period between 1880 and 1920. These four decades, a period often cited as the main part of the era of mass migration (see e.g. Abramitzky and Boustan (2017)), saw a rapid increase in innovative activity and immigrant inflows. Figure 2 shows the aggregate time-series of patent creation (left panel) and immigration inflows (right panel). The number of issued patents rose from roughly 10,000 patents in 1870 to almost 40,000 prior to the First World War. At the same time, the U.S. saw a steady inflow of immigrants. In the period between 1870 and 1900, roughly 400,000 individuals arrived per year. Between 1900 and 1914, this number rose to almost 1 million. Consequently, the number of foreign-born increased from around 6 million people in 1870 to almost 14 million in 1910.

3.1 Data

To conduct our analysis we build a new database on the patent activity at the individual level. We construct our database by linking three datasets: (i) the IPUMS Complete Count U.S. Federal Demographic Census, (ii) the population of patents issued in the U.S., and (iii) novel data on immigrants' pre-migration characteristics from historical immigration records. We describe the data construction in detail in Section A in the Appendix. For the census, we use the restricted-access complete count version of IPUMS (Ruggles et al. (2020)) which



Notes: The figure shows the number of individuals contained in the Caste Garden Data (left panel) and Hamburg Passenger Lists (right panel).

Figure 3: The Historical Immigration Records

has detailed information for the population of individuals residing in the U.S. We specifically use the information on age, occupation, employment status, immigration status and current location of residence. For patent activity, we rely on the publicly available information from the United States Patent and Trademark Office (USPTO), which reports for each patent issued in the U.S. information on the assignee, location, and description of it.

Finally, we rely on two primary sources for our historical immigration records: the Castle Garden Immigration Database and the Hamburg Passenger Lists. The Castle Garden data, compiled by the Battery Conservancy, contains the list of all immigrants entering the US via the port of New York between 1820 to 1914. In total, the database comprises approximately 11 million individual micro-records. The Hamburg Passenger Lists records all passengers leaving from the port of Hamburg to the US between 1850 and 1914 and comprises approximately 6 million records.[4] Importantly, both datasets contain detailed information on immigrants *pre*-migration occupations. We will use this information to measure immigrants' innovative skills (see Section 3.3 below).

In Figure 3 we show the sample size of the Castle Garden Data (left panel) and Hamburg Passenger Lists (right panel) by decade. Figure 3 shows that these two databases complement each other well. The Castle Garden Data tends to oversample immigrants in the early part of our sample, while the Hamburg Passenger Lists cover more immigrants after 1890.

[4]We have access to the complete records through a cooperation with the Hamburg State Archive.

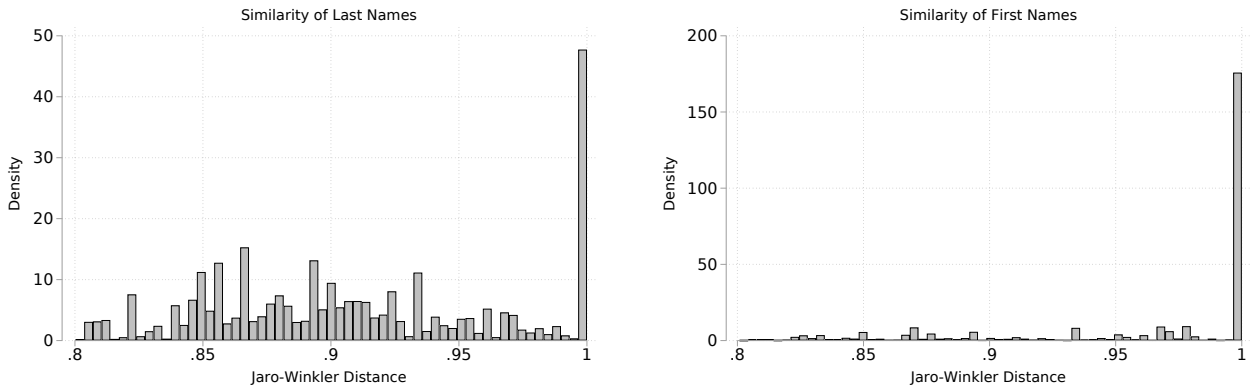
3.2 Constructing the Database: Record Linking

We construct our database by linking patents and historical immigration records to individuals in the population census. Our analysis focuses entirely on males as the majority of patent holders were males between 1880 and 1920. To perform the record linking, we use the fact that the patent records report the name and location (typically, county and state) of the patent assignee and that the immigration records report the name, age and arrival (or departure) year in the U.S. We use this information to match these datasets to the Population Census.

We conduct our record linking using a machine-learning-based random forest classifier. While we discuss the details of the matching procedure in more detail in the Appendix Section B, intuitively, we search for individuals with similar first and last name and location (to match patents to the Census) and similar first and last name, age and arrival/departure year (to match historical immigration records to the Census). As our measure of ''distance'' between names, we rely on the so-called Jaro-Winkler distance, to measure the distance between two strings (in our context, first and last names). The Jaro-Winkler distance, which is a number between 0 and 1 with 1 denoting a perfect match, is an 'edit distance' — it counts how many operations are required to transform one string into another string. We consider a particular ''patent-census'' or ''immigration record-census'' combination a match, if it has the highest Jaro-Winkler score, provided its Jaro-Winkler score exceeds 0.85, and the observations are in the same geographical location and share the same age and arrival year.

As an example, consider Figure 4 where we display the distribution of the Jaro-Winkler similarity scores for all individuals from the Hamburg Passenger List that we matched to the US Census. In the left panel, we display the distributions for the last names, in the right panel for the first names. Figure 4 shows that for both last and first names, the distribution shows a sizable spike at 1, which refers to a value where the strings align perfectly. Besides, we find relatively higher scores for first names than for last names, reflecting the fact that the latter show more variation than the former and that slight variation in spelling is more likely.

To get a linguistic sense of what these distances mean, consider the following example. In the Hamburger Passenger Lists, we find ''Christian Friedrichsen''. The closest match in the US Census is an individual with the same age whose name is reported as ''Christ Frederickson''. The Jaro-Winkler Distance between the first names is equal to 0.95, the distance between the last names is 0.83, Hence, it does not pass our ''threshold'' of having a Jaro-Winkler distance of at least 0.85 for both names and we will not consider it as a match. In contrast, for the passenger ''Ernest Wodrich'' we find ''Earnest Woodrich'' with the same age and nationality in the sample. The Jaro-Winkler distance is about 0.95 for both the first and the last name and hence we consider it as a match. Finally, ''Morris Brettschneider'' and ''Morris Brettsschneider'' is an example of an almost perfect match — the extra ''s'' in



Notes: The figure shows the distribution of the Jaro-Winkler similarity scores of the last names (left panel) and first names (right panel) for the matches of the Hamburger Passenger Lists and the Population Census.

Figure 4: Distribution of Jaro-Winkler Distances

the Census implies that the Jaro-Winkler distance for the last name is equal to 0.99.

For the remainder, we take this full sample of matches as our baseline sample. However, we also consider various robustness checks where we apply more conservative criteria such as imposing Jaro-Winkler similarity measure of unity (i.e. being a ''perfect'' match).

Matching historical immigration records to the US Census In Table 1 we report a set of summary statistics for the matching rates of the historical immigration records to the US Population Census. In the first two rows, we report the total number of people who are foreign-born as measured in the US Census. In the second row, we report the number of ''recent'' immigrants, which we define as individuals that entered the U.S. within the last 10 years. Note that this information is only contained in the US Census starting in 1900. In rows 3 and 4, we report the number of records we were able to match the historical immigration records from Castle Garden and the Hamburg Passenger Lists respectively. We match roughly between 500,000 and 800,000 individuals per decade. Finally, the last row reports the number of people we can match to both the Hamburg Passenger Lists and Castle Garden. Hence, in total, we arrive at a dataset, which covers the episode of four decades and where we can link more than 3 million immigrants to their immigration records and hence information on their occupations before migrating. This number compares to about 7.5 million census observations of recent immigrants and 25 million records of foreign-born individuals across these years.

In Table 2 we provide some summary statistics of the population of recent immigrants in the census and the subset of such immigrants were able to match.[5] Compared to the native

[5]See also Section E.1 in the Appendix, where we provide a more comprehensive analysis of the quality of

	1880	1900	1910	1920	Total
Number of Obs in Census					
Foreign born	3,464,945	5,742,188	7,876,248	7,738,920	24,822,300
Recent immigrants		1,722,249	3,516,881	2,235,612	7,474,742
Number of people matched from:					
Castle Garden	720,918	580,142	598,789	441,803	2,341,652
Hamburg Passenger Lists	87,860	198,898	318,014	284,414	889,186
HPL & Castle Garden	38,418	52,930	60,401	57,362	209,111

Notes: The table reports the number of foreign born and recent immigrants, i.e. immigrants that arrived within the last ten years as measured from the population census. It also reports the number of matched to the Castle Garden Data, the Hamburg Passenger List and both.

Table 1: Historical Immigration Records: Number of Matches

population, recent immigrants to the U.S. are highly urbanized (two-thirds live in cities) and are more likely to be in the first part of their working life, i.e. between 20 and 40 years old. In the remaining columns, we report the same statistics for our matched data. In terms of the urbanization rate, the age distribution and the sectoral employment shares our matched sample looks reassuringly similar. In terms of the nationality distribution, our matched sample is naturally different. Among individuals matched to the Hamburg Passenger Lists, two-thirds are German and no one is from Italy. This is expected because Italian immigrants overwhelmingly left to the United States from Naples. In contrast, among our matches from the Castle Garden Data, almost 40% are Italians. Note also that by construction both the Castle Garden Data and the Hamburg Passenger List over-sample immigrants from Europe relative to the population of immigrants.

The fact that we have a subset of individuals, that were matched both to the Hamburger Passenger Lists *and* the Castle Garden Database allows us to devise an additional test for the quality of matches. Note that the matching procedures were independent — no information from the Hamburg Passenger Lists was used to match the Castle Garden Data and vice versa. However, both report the port where the respective individual boarded. Given that all individuals appearing in the Hamburg Passenger Lists, by construction, boarded ships in Hamburg, we expect that the distribution of ports of individuals in the Castle Garden Data who are also matched to the Hamburg Passenger List is biased towards ''Hamburg'' *if* our algorithm indeed manages to correctly identify the same individual.

our matching procedure.

	Natives	Recent Immigrants		
		Total	Matched to	
			Castle Garden	HPL
Number of observation	32,857,239	1,717,760	106,459	65,568
Urbanization rate	0.34	0.65	0.69	0.66
Age < 20	0.52	0.23	0.21	0.19
Age ≥ 20 and < 40	0.29	0.62	0.62	0.60
Age ≥ 40 and ≤ 70	0.17	0.14	0.16	0.20
German share		0.14	0.36	0.65
Italian share		0.12	0.41	0.00
British share		0.07	0.00	0.05

Notes: The table compares the characteristics of recent immigrants in the Census (column 1) to the characteristics of immigrants that have been matched to Castle Garden (column 2) and the Hamburg Passenger List (column 3). Recent immigrants are immigrants that arrived within the last decade.

Table 2: Characteristics of Recent Immigrants

This is exactly what we find. In Panel A of Table 3, we report the top five departure ports of individuals in Castle Garden that are also matched to the Hamburg Passenger Lists. Almost 50% of respondents also reported Hamburg (either directly or via transit in Le Havre or Southampton) as their departure port in Castle Garden. This is very different in the sample, where we did *not* manage to match the individual in the Census to the Hamburg Passenger Lists. About 25% of respondents are Italians who left from Naples and Genoa and Hamburg does not even appear as one of the top five departure ports. While we think that Table 3 is encouraging for the match quality of our algorithm, we also want to note that our matched data may include some false positives. In particular, we expect that several individuals in Panel A, which report "Bremen" as their port of departure, are likely incorrectly matched based on a similar German first and last name.

Matching patents to the US Census In Table 4 we summarize the matching rates of patents for each decade. In the first row, we report the total number of patents issued in the respective decade. Note that this number is simply the sum of the annual patent flows displayed in Figure 2 for the respective decade. In the second and third row, we report the number of such patents that we could match individuals in the census. Depending on

Panel A: Castle Garden & Hamburg Passenger Lists			Panel B: Only Castle Garden		
Port	Freq.	Percent	Port	Freq.	Percent
Hamburg	2,657	22.77	Naples	86,905	20.79
Bremen	1,664	14.26	Le Havre	36,262	10.94
Hamburg & Southampton	1,528	13.09	Bremen	35,980	10.85
Hamburg & Havre	1,427	12.33	Bremen & Southampton	25,666	7.74
Bremen & Southampton	1,211	10.38	Genoa	23,642	7.13

Notes: In Panel A we report the top five departure ports reported in Castle Garden for the set of people, which are matched to both the Caste Garden Data and the Hamburg Passenger Lists. Panel B reports the same for the sample in the Castle Garden data which is *not* matched to the Hamburg Passenger Lists.

Table 3: Comparison of Departure Ports

Period	1860-1879	1880-1899	1900-1909	1910-1919
Total number of patents issued	196,307	416,153	304,654	381,131
Number of patents matched to Census	24,597	33,185	183,223	181,223
Share of patents matched	0.13	0.08	0.6	0.48

Table 4: Matching Rate for Patents

the year, we match between 15% and 56% of all patents. In Section C.3 in the Appendix, we discuss the quality of our patent matching procedure in more detail. In particular, we show that our match rates are very similar across industries, suggesting that our matching algorithm is not systematically more successful in matching patents to inventors in specific industries or specific regions.[6]

3.3 Pre-migration skills

As highlighted above, a unique feature of the historical immigration data (both CG and HPL) is the enumeration of immigrants' pre-migration occupations. As both databases were provided to us simply as un-harmonized string variables, we construct a crosswalk to

[6]The matching rates in the later part of the sample are higher because the quality of the name scraping out of the individual patent information improved in the later samples.

the official occupational classification used by the US census. We do so using the lexical database "Wordnet", which is part of the NLTK library in Python.[7] Wordnet contains a large network of synonyms, which are ordered hierarchically. This allows us to classify immigrants' pre-migration occupations to the US census system through shared synonyms. See Section D.3 in the Appendix for details.

4 Immigrants and Patent Creation

Were immigrants an important source of idea creation? Figure 5 takes the first stab at this question by displaying the extensive margin of patenting, i.e. the share of individuals by their immigration status with at least one patent in the period between 1900 and 1909. For immigrants, we distinguish between immigrants that arrived within the last ten years ("recent immigrants") and foreign-born individuals that lived in the US for more than ten years ("assimilated immigrants"). The pattern is striking: assimilated immigrants are by far the most prolific group of innovators.[8] They are almost three times as likely as natives to have a patent. In contrast, recent immigrants are much less likely to file a patent and hardly differ from natives in their patenting behavior.[9]

In Figure 6 we look directly at the relationship between the innovativeness of immigrants and the length of their stay in the US. There is a strong monotone positive relationship: the longer immigrants stayed in the US, the more likely they are to file a patent. Presumably this pattern reflects the gradual accumulation of complementary inputs to create new ideas such as capital, language proficiency, input suppliers, or a customer base.

4.1 Estimating Innovation Human Capital

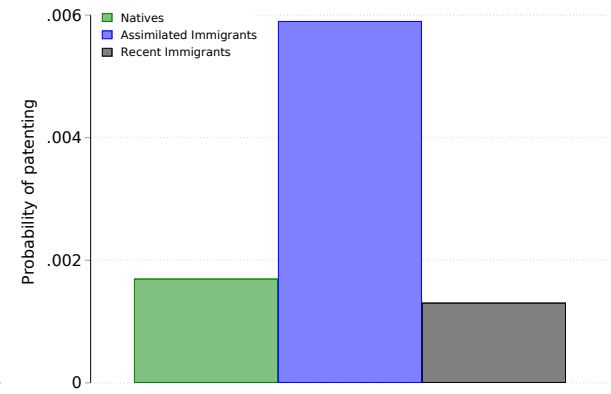
Our theory stresses the importance of innovation human capital, which we summarized by the shifter of the type-specific skill distribution ψ_{rt}^{j}. We now use our micro database on individual patent behavior to estimate the different components of ψ_{rt}^{j}.

To do so we assume that patenting (in the data) is related to the creation of varieties (in

[7] https://wordnet.princeton.edu

[8] In Section E.2 in the Appendix we show that the same pattern also holds true for the intensive margin of patenting.

[9] A potential concern for the patterns depicted in Figure 5 could be that it might simply reflect a particular bias in our matching procedure: maybe immigrants have more distinct names than the native population and hence we simply achieve a higher match rate even though the actual innovation productivity might not differ. In Section C.2 in the Appendix we discuss this in more in detail and provide direct evidence that this is unlikely to be the case. We also think it is unlikely that the recognizability of names changes systematically with the length of the stay in the US as would need to be the case to rationalize the pattern shown in Figure 6.



Notes: The figure shows the extensive margin of patenting for natives and immigrants. We distinguish between assimilated immigrants (that have been in the US for at least 10 years) and recent immigrants (that arrived in the US within the last 10 years). The data refers to the year 1910.

Figure 5: Patent Activity by Natives and Immigrants

the theory). In particular, in terms of measurement, we assume that the observable number of patents of individual i in region r, $\mathcal{P}_{irt}$, is proportional to the number of created varieties conditional on patenting, i.e.

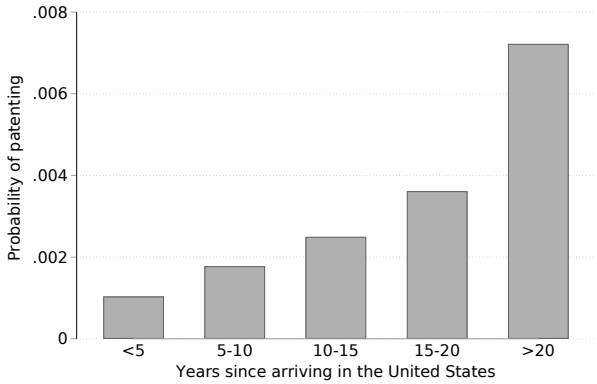
$$\mathcal{P}_{irt} = \begin{cases} 0 & \text{if } h_i < h_{rt}^* \\ \kappa h_i & \text{if } h_i \geq h_{rt}^* \end{cases}, \tag{16}$$

where $\kappa > 0$ is a constant of proportionality. Given the measurement equation in (16), our theory allows us to measure the distribution of innovation human capital φ^j.

Our theory implies that the probability of patenting of individual i in region r, $q_{rti} = P\left[\mathcal{P}_{irt} \geq 0\right]$, is given by

$$\ln q_{irt} = \ln \left(\psi_{rt}^j / h_{rt}^*\right)^\theta \quad = \quad \theta \ln \varphi^j + \ln \left(\zeta_r N_{rt-1}^\vartheta / h_{rt}^*\right)^\theta. \tag{17}$$

Hence, we can use equation (17) to estimate the type specific human capital endowment φ^j (up to scale). To do so, let $\mathcal{I}_{irt}^j$ denote a dummy variable whether individual i is an *immigrant* of type j. We consider an immigrant type as a combination of a nationality (e.g. Germany, Italy) and — in the spirit of Figure 5— whether the immigrant arrived recently (i.e. within the last 10 years) or is already assimilated. Hence, $\mathcal{I}_{irt}^j$ takes the value 0 for natives. Similarly, for an immigrant from Italy who arrived recently, we have $\mathcal{I}_{irt}^{ITA,R} = 1$ (where the superscript "R" refers to "recent") and for an individual born in Germany that lived in the US for more than 10 years we have $\mathcal{I}_{irt}^{GER,A} = 1$ (where the superscript "A" refers



Notes: The figure shows the probability of filing patent by the length of immigrants' stay in the US. The data refers to the year 1910.

Figure 6: Patent Activity and the Length of Stay in the US

to ''assimilated''). Using this notation, we run the regression

$$\ln q_{irt} = \delta_{rt} + \sum_{n} \left(\beta^{nR} \mathcal{I}_{irt}^{n,R} + \beta^{nA} \mathcal{I}_{irt}^{n,A} \right) + x'_{irt}\gamma + u_{irt}, \tag{18}$$

where δ_{rt} is a region-time fixed effect to control for the exogenous region-specific fixed effect ζ_r and the endogenous statistics N_{rt-1} and h_{rt}^* and x'_{irt} is a vector of individual observable characteristics, which could predict innovative human capital φ_j. Our theory, in particular equation (17), implies that the coefficient β^{nk} is given by

$$\beta^{nk} = \theta \ln \left(\varphi^{nk} / \varphi^{N} \right), \tag{19}$$

i.e. β^{nk} reflects exactly the human capital of immigrants with nationality n and status k *relative* to natives. We estimate (18) via logit.[10]

The results of this regression are contained in Table 5. The first column corresponds to the regression version of Figure 5: immigrants have more human capital compared to natives conditional on them being in the country for a sufficiently long time. In column 2, we control for differences in spatial sorting by controlling for urbanization and the size of the population. Doing so reduces both coefficients substantially as immigrants are much more likely to reside in urban, densely populated areas. In terms of our theory, immigrants tend to live in locations that are efficient to generate ideas (i.e. where ζ_r and N_{rt-1} are large relative to the entrepreneurial cutoff h_{rt}^*) that these two observable variables capture a

[10]Strictly speaking, the logit model takes the form $\ln\left(\frac{q_{irt}}{1-q_{irt}}\right) = \ln q_{irt} - \ln\left(1 - q_{irt}\right)$. Because $q_{irt} \approx 0$, estimating (18) via logit will be a good approximation.

	(1)	(2)	(3)	(4)
Recent immigrants	-.112	-.56***	-.603***	-.627***
	(.0811)	(.0545)	(.059)	(.0607)
Assimilated immigrants	1.2***	.832***	.772***	.536***
	(.0548)	(.0352)	(.0388)	(.0357)
Urban share		1.79***		
		(.149)		
(ln) Population		.0191		
		(.0331)		
German				.455***
				(.0485)
Italian				-.837***
				(.107)
English				.633***
				(.0307)
Irish				.109*
				(.06)
N	86,371,820	86,371,820	86,262,764	86,262,764
Year FE	yes	yes	yes	yes
SEA FE	no	no	yes	yes

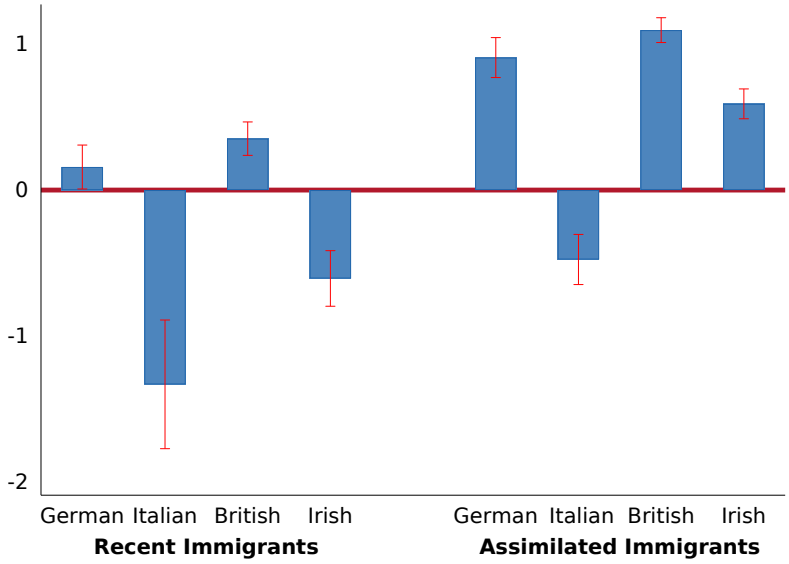
Note: For immigration year 1900-1929. Standard errors are clustered by State Economic Area (SEA). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Estimating Innovation Human Capital

sizable amount of this variation as seen in column 3 where controlling for county-year fixed effects changes the coefficients relatively little.

In columns 4 and 5, we estimate differences in human capital by nationality. For parsimony, we restrict the effect the length of time in the US to be constant across different nationalities - we will come back to this distinction below. Consider first the heterogeneity of human capital across nationalities. There is a clear pattern: English and German immigrants have relatively high entrepreneurial human capital, Italian immigrants have less. Note also that controlling for nationality cuts the coefficient for ''assimilated immigrants'' in half: part of the superior patenting performance of assimilated immigrants is accounted for by differences in the nationality composition because the first wave of immigrants was concentrated among ''innovative'' nationalities like Germany and Britain and the big wave of the Italian immigrants came later.

Quantitatively, equation (19) implies that the human capital of assimilated immigrants exceeds the ones by natives by 1 log point (i.e. 170%) and the recent immigrants' human capital φ^θ is roughly 15% lower (see column 3). Similarly, column 4 implies that German and English immigrants' human capital exceeds that of natives by about 0.55 ($\approx 0.8 - 0.25$) or 1.5 ($\approx 0.8 + 0.7$) log points depending on whether they are assimilated or not. The differences across immigrant nationalities are also sizable. Italian immigrants' human capital



Notes: The figure displays the coefficients of a regression of a patent dummy on region-year fixed effects and the different dummy variables displayed in the window (see (18)).

Figure 7: Heterogeneity in Innovation Human Capital (φ)

is for example 1.5 log points below the ones of German or English immigrants holding the assimilation status constant.

In Figure 7 we display the heterogeneity in innovation human capital across nationalities and immigration types visually. More specifically, we display the coefficients β^{nk} from 18 when we do not assume the effect of assimilation to be necessarily constant across nationalities. Recall that a value of zero corresponds to the human capital of natives.

Figure 7 shows two patterns. First, within nationalities, assimilated migrants tend to have higher patenting rates, consistent with our results above. Note also that the increase in innovation human capital due to the assimilation is roughly constant across nationality. Secondly, the life-cycle variation in patenting is small relative to the cross-nationality variation. In particular, while German and British immigrants have higher patent rates than the native population, immigrants from Italy add relatively less to the nation's knowledge stock.

4.2 Is Knowledge Portable? Direct Evidence on the Role of Human Capital

So far we have shown that immigrants played an important role for patent creation and that they differed systematically by their nationality and the length of their stay in the United States. Our theory interprets these differences as reflecting differences in human capital. In this section we provide direct evidence for this human capital interpretation.

Innovation and Pre-migration Knowledge If human capital was an important input into the production of knowledge and human capital is at least partially portable across space, we would expect that migrants who were innovative in their home countries in Europe were more likely to keep innovating after their arrival in the US. In the absence of data on immigrants' patenting behavior prior to migrating, we cannot directly test this implication. As an alternative, we rely on the information on migrants' pre-migration occupations.

According to our model, human capital is reflected in patent creation. Hence, if individuals within a particular occupation are more likely to generate patents than others, such occupations apparently attract individuals that are abundant in innovation human capital. Leveraging this intuition, we rank immigrants' *pre*-migration occupations by their innovativeness in the US and then ask whether immigrants who used to work in innovative occupations in Europe are indeed more likely file a patent in the US.

To implement this procedure, we construct an occupation-based ranking of innovative potential from the regression

$$P_{irt} = \delta_{rt} + \sum_{o} \phi_{ot} O_{iort} + u_{irt}, \tag{20}$$

where - as before - P_{irt} is an indicator variable for individual i's patent activity and δ_{rt} is a set of county-year fixed effects. O_{iort} is an indicator variable, whether individual i works in occupation o. Hence, the coefficients of interest, ϕ_o, measure the average probability of individuals in occupation o to engage in patenting. Note that by controlling for county-year fixed effects, ϕ_o is not identified from the variation of occupations across space, but only compare different occupations within locations. We estimate specification (20) only for natives to limit concerns of reverse-causality.

Because we measure occupations at the three digit level, (20) yields more than 200 estimated fixed effects. In Table 6 we report the top ten occupations based on our estimates of $\hat{\phi}_o$.[11] It is reassuring that occupations with a focus on engineering and other natural sciences are the most innovative occupations. Moreover, there is a striking degree of correlation across years.

With these estimates at hand, we then calculate individual i's *pre*-migration innovation score as $PMI_i = \hat{\phi}_o$, if individual i worked in occupation o prior to migrating. We then consider estimate the regression

$$P_{irt} = \delta_{rt} + \chi PMI_i + \sum_{j} \beta^j \mathcal{I}^j_{irt} + x'_{irt}\gamma + u_{irt}, \tag{21}$$

[11] The estimates for all occupations are available upon request.

Rank	1900	1910	1920
1	Social sciences (n.e.c.)	Professional workers	Physics
2	Engineers, metallurgical	Engineers, mechanical	Engineers, chemical
3	Engineers, chemical	Designers	Engineers, mechanical
4	Professional workers	Engineers, mining	Professional workers
5	Chemistry	Engineers, electrical	Engineers (n.e.c.)
6	Engineers, mechanical	Chemists	Technicians (n.e.c.)
7	Engineers, electrical	College presidents and deans	Biological scientists
8	Engineers (n.e.c.)	Draftsmen	Engineers, electrical
9	Airplane pilots and navigators	Engineers (n.e.c.)	Designers
10	Miscellaneous natural scientists	Managers and superintendents, building	Geologists and geophysicists

Table 6: Innovativeness Ranking of Occupations

on the population of immigrants in the US.

In Table 7 we report the results of estimating (21). Columns 1 and 2 show that our measure of the pre-migration innovation score is highly correlated with patent activity, both unconditionally and within regions. In column 3 we control for the nationality of immigrants. This reduces the effect of pre-migration knowledge substantially, reflecting the fact that the skill-composition of immigrants (as measured by their pre-migration skills) is heterogeneous and accounts for the superior innovation performance of for example British and German immigrants. Finally, in the last column, we control for a full set of current occupation fixed effects. Hence, immigrants that worked in occupations with a high innovation score have - on average - more patents and this effect is largely explained by them working in more innovative occupations after their arrival. This is consistent with human capital being portable and an important input in the production of knowledge.

What Do Immigrants Invent? Measuring Patent Novelty As a second source of evidence for the importance of human capital, we also looked at the type of innovations generated by immigrants. We are particularly interested whether immigrants' innovations were qualitatively different from the ''typical'' innovation created in a particular region.

While we describe the details of this measure in Section D.4 in the Appendix, the main idea is simple. The patent data contains a detailed description of the idea being patented. Using textual analysis we can therefore compare whether two patents are similar, i.e. whether they rely on similar words. We use such textual analysis to devise a novelty score for each patent by comparing it to the set of patents, which have been issued in the same location in the past. Hence, if a patent uses very different words than typically used in patents stemming from this region, we think of this patent as novel.

Our procedure yields a novelty score for each patent, which we express as a z-score, i.e. with a unit variance and a mean of zero. Given this score n_{it}, we then consider a regression

	(1)	(2)	(3)	(4)
Pre-migration occupation	15.4***	13.8***	3.95	2.18
	(1.83)	(1.97)	(2.95)	(6.7)
German			.674***	.717***
			(.106)	(.104)
Italian			-1.25***	-1.18***
			(.182)	(.177)
English			1.23***	1.36***
			(.0733)	(.0811)
Irish			.731***	.881***
			(.0854)	(.095)
N	3,548,129	3,495,868	3,495,868	3,493,100
Year FE	yes	yes	yes	yes
SEA FE	no	yes	yes	yes
Occupation FE	no	no	no	yes

Year 1860-1929. Standard errors are clustered by State Economic Area (SEA).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Pre-migration Knowledge And Patent Activity

of the form

$$n_{irt} = \delta_{rt} + \delta_{IND} + \sum_j \beta^j \mathcal{I}^j_{irt} + x'_{irt}\gamma + u_{irt}, \tag{22}$$

where n_{irt} is the novelty score of patent i and δ_{IND} are industry fixed effects. The remaining variables are defined as above. Table 8 reports the estimation results based on 22.

In column 1 we show the simple bivariate correlation between patent novelty and the immigration status of the inventor. We find that this correlation is negative. In column 2 we control for region fixed effects. We find that now assimilated immigrants have more novel patents than natives. This difference is due to the fact that immigrants systematically settle in more innovative locations, where more patents are generated. This makes the average patent more similar to the typical patent - see column 3 where we directly control for the number of patents and which are negatively correlated with novelty score. Finally, in column 4 we control for industry fixed effects. This further raises the correlation between the immigration status and the novelty of the patent.

As with our results regarding the importance of inmigrants' pre-migration occupations, these findings are consistent with the presence of portable human capital. If immigrants were indeed able to bring ideas from the old continent to the US, we would expect that such knowledge would provide immigrants with a comparative advantage in patenting, that this would be particular pronounced for immigrants with innovative occupations in their home

	Standardized Novelty Score			
	(1)	(2)	(3)	(4)
Recent immigrants	-.222***	-.00468	-.0331	.0183
	(.0521)	(.0146)	(.0215)	(.0134)
Assimilated immigrants	-.141***	.0145*	.00242	.0181***
	(.0316)	(.0074)	(.0116)	(.00657)
(ln) Number of patents			-.134***	
			(.0156)	
R^2	.0053	.158	.0895	.246
N	364,446	364,356	364,446	358,510
Decade FE	yes	yes	yes	yes
County FE	no	yes	no	yes
Industry FE	no	no	no	yes

Year 1900-1919. Standard errors are clustered by State Economic Area (SEA).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Immigrants and Innovation Novelty

country and that their ideas would be relatively novel. Tables 5., 7 and 8 are qualitatively consistent with this narrative.

5 Immigrant Inflows and Local Patent Creation

In Section 3 we used our microdata to estimate immigrants' and natives' innovation human capital φ^j. To do so we exploited the within-region variation in patent activity. Hence, this variation is not suited to discipline two aspects of our theory, which operate at the regional level: differences in "innate" innovation potential across localities (ζ_r) and the strength of inter-temporal spillovers in the idea production function (ϑ). In this section we use variation in patent creation across regions to estimate these objects. We will use these estimates in our quantitative analysis in Section 6.

5.1 Immigration Inflows and Regional Patent Activity

Before turning to the structural estimation of ϑ and $\{\zeta_r\}_r$, we study the relationship between immigrant inflows and subsequent patent activity in a regression framework and show that our cross-regional results are qualitatively consistent with our findings from the micro data. We know that the number of patents generated in region r, $\mathcal{P}_{rt}$, is proportional to $h^*_{rt}L_{rt}$. To link these objects to the data, we assume that the contemporaneous flow of patents $\mathcal{P}_{rt}$ is measured as the number of patents issued between t and $t+1$ and that L_{rt} and h_{rt} are

determined from the stock of the population and its composition at t.

Using that h^*_{rt} follows the endogenous AR(1) process given in (14), our theory implies that the number of patents issued in region r is given by

$$\begin{aligned}
\ln \mathcal{P}_{rt} &= \text{const} + \ln h^*_{rt} + \ln L_{rt} \\
&= \text{const} + \frac{1}{\theta} \ln \left(\sum_j \varpi^j_{rt} \left(\varphi^j \right)^\theta \right) + \ln \zeta_r + \vartheta \ln \left(h^*_{rt-1} L_{rt-1} \right) + \ln L_{rt} \\
&= \text{const} + \frac{1}{\theta} \ln \left(\sum_j \varpi^j_{rt} \left(\varphi^j \right)^\theta \right) + \ln \zeta_r + \vartheta \ln \mathcal{P}_{rt-1} + \ln L_{rt}. \qquad (23)
\end{aligned}$$

For simplicity consider (as in the Section 2.5 above) the case of two types: immigrants and natives. Hence, (23) suggests the regression

$$\ln \mathcal{P}_{rst} = \delta_r + \delta_t + \delta_s + \beta \ln \mathcal{P}_{rst-1} + \chi \varpi^I_{rt} + \phi \ln L_{rt} + u_{rt}, \qquad (24)$$

where $\mathcal{P}_{rst}$ is the number of patents in industry s in region r and δ_s is a set of sector fixed effect. The theory also implies that $\beta = \vartheta < 1$, that $\phi = 1$ and that $\chi = \frac{1}{\theta} \left((\zeta_I / \zeta_N)^\theta - 1 \right)$ is related to the average human capital of immigrants relative to natives.

In Table 9 we report the results from estimating (24) at the county-industry level for the years 1900, 1910 and 1920 when we measure ϖ^I_{rt} as the share of recent immigrants, i.e. the share of foreign born that arrived within the last 10 years.

In column 1 we only include year and industry fixed effect and find a large and positive effect of recent immigration on patent creation. Columns 2 and 3 highlights that this positive effect is to a large extent due to spatial sorting: once county fixed effects are controlled for, the coefficient drops substantially and is no longer significant.

Our empirical analysis in Section 3 highlighted the difference between recent and assimilated immigrants because migrants' innovation human capital is sharply increasing in the length of their stay in the US. This implies that if we were to measure the share of immigrants not solely as ''recent arrivals'' but rather as the share of foreign born, we would expect larger effects in the cross-section. Column 4 shows that this is exactly what we find. In terms of our theory, the coefficient χ in (24) now reflects the average human capital of all foreign born and not only recent immigrants.

Endogeneity The first four columns of Table 9 estimate (24) using OLS. This is problematic if immigrant inflows are correlated with shocks to patent creation conditional on county fixed effects. In terms of our theory, this would for example be the case if the regional innovation

	OLS				IV	
	(1)	(2)	(3)	(4)	(5)	(6)
Share of immigrants	1.902***	0.337	0.389		1.151**	3.229**
	(0.227)	(0.279)	(0.276)		(0.584)	(1.356)
Share of foreign born				0.588**		
				(0.237)		
(ln) $\text{Population}_{\text{rt}}$	0.312***	0.202***	0.215***	0.193***	0.195***	0.186***
	(0.0118)	(0.0481)	(0.0499)	(0.0483)	(0.0480)	(0.0490)
(ln) $\text{Patents}_{\text{rit-1}}$	0.599***	0.503***	0.503***	0.503***	0.503***	0.504***
	(0.0106)	(0.0109)	(0.0109)	(0.0109)	(0.0109)	(0.0110)
$\text{Urban share}_{\text{rt}}$			-0.140			
			(0.0947)			
R^2	0.666	0.700	0.700	0.700	0.243	0.239
N	72373	72157	72157	72555	72555	72555
First stage F					67.25	35.90
County FE	no	yes	yes	yes	yes	yes
Year FE	yes	yes	yes	yes	yes	yes
Industry FE	yes	yes	yes	yes	yes	yes

Standard errors in parentheses

Note: Year 1900-1930. Standard errors clustered by county. Immigrants are people who immigrated within the last ten years for a given census year. Foreign born are all of the immigrants

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regional patenting and immigration inflows

efficiency ζ_r was time-varying and immigrants would sort towards regions that experienced an increase in ζ_{rt}. In the last two columns of Table 9 we therefore estimate (24) with an instrumental variable strategy.

In column 5 we use a traditional 'shift-share' instrument for immigrant inflows, exploiting the fact that immigrants are likely to settle in places where their ancestors were living in the past (Altonji and Card (1991); Card (2001)). Intuitively, we predict immigrant inflows from the time series of actual immigrant inflows from different origin countries, interacted with the distribution of ancestors across space in the previous period. More formally, let I_o^t be the aggregate immigrant inflows of origin country o at time t, we construct our first instrument for regional immigrant inflows as

$$IV_{rt}^1 = \sum_o \frac{A_{ort-1}}{A_{ot-1}} I_{ot}, \tag{25}$$

where $\frac{A_{or}^{t-1}}{A_o^{t-1}}$ is the share of ancestors in region r of an origin country o in the previous period - see Section E.3 in the Appendix for more details of the instrument construction. Using this shift-share instrument for the share of recent immigrants yields an estimate for χ, which is larger then the OLS estimate (see column 5).

In column 6 we use an alternative instrument following the approach of Burchardi et al. (2020). Burchardi et al. (2020) argue that the traditional shift-share instrument in (32) will be invalid if the distribution of past ancestry is correlated with future patent growth. To address this issue they propose a modified version of (32), where not the entire ancestry distribution is used to predict future immigration, but only an exogenous component. Again, see Section E.3 in the Appendix for more details. Column 6 again shows a positive relationship between immigration inflows and patent activity. The points estimate is larger than with the ''traditional'' shift-share instrument, but is also more imprecisely estimated.

Our preferred specifications are the OLS estimates in columns 3 and 4. They are consistent with our theory, which implies that the inclusion of the county fixed effects and past patent activity appropriately controls for any endogeneity of immigrant inflows. In addition, our theory provides an explanation for why one might expect the IV estimates to be upward biased. Our theory, in particular the discussion in Section 2.5, highlights that the productivity effects of immigrant inflows are *cumulative*. Hence, past immigration will drive both future immigration in the spirit of the shift-share instrument and the local creation of varieties, which reduces the cost of innovation. While our theory implies that this latter effect is indeed fully controlled for by including $\ln \mathcal{P}_{rst-1}$ as a dependent variable, it is important to highlight that the validity of the instruments relies on this particular functional form.

In any case, because we will not rely on these estimates for our quantitative exercise in

Section 6, none of our quantitative results hinges on the numbers reported in Table 9.

5.2 Estimation of Structural Parameters

We can also use (23) to directly estimate the inter-temporal spillover elasticity ϑ and the regional heterogeneity in innovation efficiency ζ_r. To first-order, (23) implies that[12]

$$\ln \mathcal{P}_{rt} \quad = \quad \text{const} + \sum_{j \neq N} \varpi_{rt}^{j} \frac{\left(\varphi^{j}\right)^{\theta} - 1}{\theta} + \ln \zeta_r + \vartheta \ln \mathcal{P}_{rt-1} + \ln L_{rt}. \tag{26}$$

Hence, we can estimate the scale elasticity ϑ and the spatial innovation heterogeneity ζ_r from the regression

$$\ln \mathcal{P}_{rst} \quad = \quad \delta_r + \vartheta \ln \mathcal{P}_{rst-1} + \delta_s + \delta_t + \sum_{j \neq N} \beta^{j} \varpi_{rt}^{j} + \gamma \ln L_{rt} + u_{rst}. \tag{27}$$

Note that we again include industry fixed effects δ_s to allow for heterogeneity across industries, which is not modeled in our single-industry model. This implies that we only identify ζ_r up to a common scale. This, however, is inconsequential as the number of varieties N_{rt} does not have a natural scale.

In Table 10 we report the results for ϑ based on (27). In the first column we directly estimate (27) without imposing the restrictions from the theory that $\gamma = 1$ and that $\beta^{j} = \frac{1}{\theta}\left[\left(\varphi^{j}\right)^{\theta} - 1\right]$. For brevity we do not report the ten different coefficients β^{j} (note that we have five nationalities and both recent and assimilated immigrants). We estimate that ϑ =0.563.[13]

Note that the theory restricts the population elasticity γ to be equal to unity. As seen in Table 10, this condition is not satisfied. Similarly, the theory also implies that the coefficients

[12]Note that

$$\ln\left(\sum_{j} \varpi_{rt}^{j} \left(\varphi^{j}\right)^{\theta}\right) = \ln\left(\left(\varphi^{N}\right)^{\theta} + \sum_{j \neq N} \varpi_{rt}^{j}\left[\left(\varphi^{j}\right)^{\theta} - \left(\varphi^{N}\right)^{\theta}\right]\right) \approx \sum_{j \neq N} \varpi_{rt}^{j} \frac{\left(\varphi^{j}\right)^{\theta} - \left(\varphi^{N}\right)^{\theta}}{\left(\varphi^{N}\right)^{\theta}} = \sum_{j \neq N} \varpi_{rt}^{j}\left[\left(\varphi^{j}\right)^{\theta} - 1\right],$$

where the last equality used our normalization that $\varphi^{N} = 1$. Hence,

$$\ln \mathcal{P}_{rt} = \text{const} + \sum_{j \neq N} \varpi_{rt}^{j} \frac{\left(\varphi^{j}\right)^{\theta} - 1}{\theta} + \ln \zeta_r + \vartheta \ln \mathcal{P}_{rt-1} + \ln L_{rt}.$$

[13]Note that this is qualitatively consistent with our cross-county results reported in Table 9, where we found that $\varphi \approx 0.5$

on the shares ϖ_{rt}^j should coincide with the estimates of human capital. Again this is not borne out empirically. In column two of Table 10 we therefore estimate equation (27) when imposing these restrictions. Hence, we use our estimates of $(\varphi^j)^\theta$ to directly construct $\sum_{j\neq N}\beta^j\varpi_{rt}^j$ and then consider as the dependent variable $\ln \mathcal{P}_{rst} - \ln L_{rt} - \sum_{j\neq N}\beta^j\varpi_{rt}^j$. Doing so yields an elasticity of $\vartheta = 0.548$, which is very similar to the non-restricted estimate.

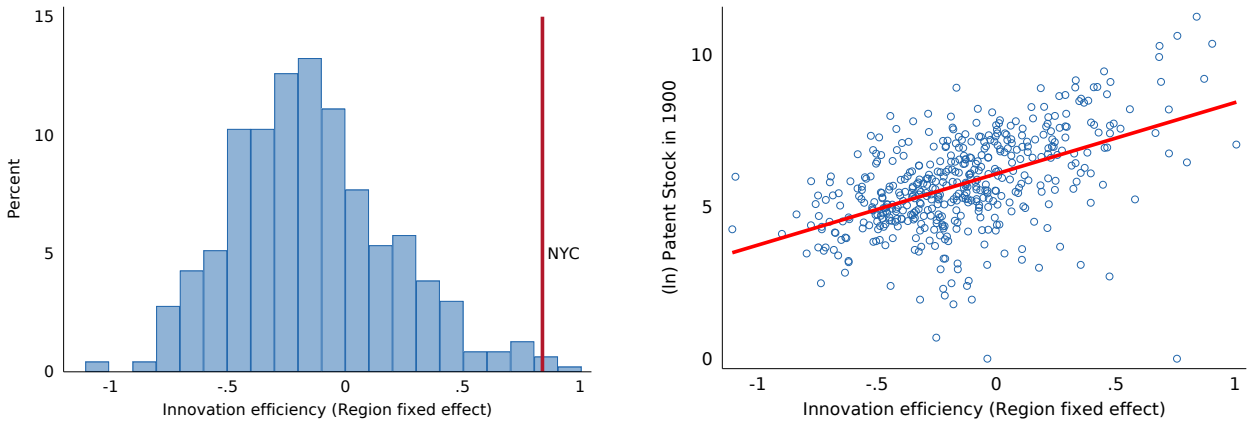
	(1)	(2)
(ln) $\text{Patents}_{\text{rst-1}}$	0.563***	0.548***
	(0.0141)	(0.0144)
(ln) $\text{Population}_{\text{rt}}$	0.303***	
	(0.0550)	
R^2	0.751	0.700
N	33814	33814
Decade FE	yes	yes
SEA FE	yes	yes
Industry FE	yes	yes
Population composition $\{\varpi_{rt}^j\}$	yes	no
Restriction from theory	no	yes

Table 10: Estimation of scale elasticity ϑ

In addition to the scale elasticity ϑ, we also rely on equation (27) to estimate the distribution of spatial innovation efficiency ζ_r. This distribution is displayed in Figure 8. In the left panel we show the histogram of the estimated fixed effects δ_r from 27. For comparison we also display our estimate for New York City (NYC), which we estimate to be very productive in generating new varieties. In Figure 8 we show the cross-sectional correlation between these fixed effects estimated and the observed stock of patents as of 1900, i.e. prior to our sample used to estimate (27). Our theory suggests that this relationship is positive. In fact, we show explicitly below that in a stationary equilibrium, the relationship between ζ_r and patent activity is monotone and log-linear. Figure 8 shows that reassuringly this is also the case in the data. There is strong positive relationship, which is approximately log-linear.

6 The Aggregate and Local Effects of Immigration Restrictions: A Quantitative Illustration

We next use the theory to perform a quantitative exploration of the role of migration for American growth. We implement our exercise at the level of the 452 SEAs in the US. We



Notes: The right panel shows a histogram of the estimated fixed effects from (27). For comparison we also indicate the fixed effects of NYC. The left panel shows a scatter plot of these fixed effects against the stock of patent in 1900.

Figure 8: The Distribution of Innovation Efficiency ζ_r

take 1880 as our initial period and consider a period a 5-year interval. We study the episode of 1880-1920. We furthermore enrich the theoretical framework in two aspects.

We first proceed to specify spatial friction of moving goods following a voluminous literature in international trade. In particular, we assume that regions in the continental United States trade their differentiated products subject to an iceberg cost $\tau_{rd} \geq 1$, where we normalize $\tau_{rr} = 1$. The price of the product from region i in region j is thus $p_{rd} = p_r \tau_{rd}$ and the price index in region j can be now written as

$$P_d^{1-\sigma} = \sum_r N_r p_r^{1-\sigma} \tau_{rd}^{1-\sigma}. \tag{28}$$

The market clearing condition in the presence of trade costs can be now written as

$$\sigma w_r L_r = h_r^* L_r \left(\frac{\sigma}{\sigma - 1} \right)^{1-\sigma} w_r^{1-\sigma} \sum_d \tau_{rd}^{1-\sigma} \frac{Y_d}{P_d^{1-\sigma}}. \tag{29}$$

Notice that in this specification, the elasticity of trade flows to trade costs is $1-\sigma$. We measure trade costs following Donaldson and Hornbeck (2016). These trade costs are symmetric and are constructed by using waterway and railroad distances in 1870 and 1890 between counties. For our purposes we use just the 1870 trade costs.[14]

[14]We use a SEA to county crosswalk available on the IPUMS website to convert the county level codes that they provide to SEA codes. This provides us with only 430 SEA codes of the 452 SEAs available in our 1880 U.S. census data. In order to address this further, for every missing SEA we use the NBER county to county great-circle distance data to impute and recover data for the remaining SEAs.

Second, we need to further specify the spatial labor supply function as this will determine the persistence of local immigration inflows. To do so, we borrow from a large literature on economic geography and assume that workers can move with probability ψ and that conditional on being allowed to move, they draw a multiplicative location-specific preference shock from a Fréchet distribution with shape ε. This implies that labor supply of workers of each type j evolves according to the law of motion

$$L_{rt}^{j} = (1-\psi)\, L_{rt-1}^{j} + \psi \frac{\left(\frac{\overline{w}_{rt}^{j}}{P_{rt}}\right)^{\varepsilon}}{\sum_{d=1}^{R}\left(\frac{\overline{w}_{dt}^{j}}{P_{dt}}\right)^{\varepsilon}} L_{t-1}^{j} + I_{rt}^{j} \quad \text{for} \quad j = 1, ..., J_r \tag{30}$$

where $\overline{w}_{rt}^{j}$ is given by equation (8) and I_{rt}^{j} denotes the number of immigrants of type j arriving in region r at time t from overseas. These immigrant inflows, which we take as exogenous, are our main ''policy variable'' in that we quantify the aggregate and spatial losses of restricting the inflow of immigrations of particular types.

Finally, we model immigrants' assimilation process and hence the supply of human capital in a parsimonious way. First of all, as in the data, we assume that the inflowing immigrants I_{rt}^{j} are all ''recent'' immigrants and assimilate within one period. Secondly, in the absence of a fully specified demographic model, we assume that a fraction α of assimilated immigrants turn into natives as a reduced form way to capture the fact that children of immigrants are natives.

In Figure 9 we display the timing of our quantitative model. Given the state variables for the stock of workers of different type and the measure of available varieties in each region, $\left[L_{rt-1}^{j}, N_{rt-1}\right]$, the equilibrium then determines the allocation of labor, new variety creation, the wages and the innovation threshold, $\left[L_{rt}^{j}, N_{rt}, w_{rt}, h_{rt}^{*}\right]$ from the evolution of human capital (12), the spatial labor supply function (30), the equilibrium cutoff h_{rt}^{*}, equation (2), and the labor market clearing condition, equation (29). Given the initial conditions and the structural parameters of the model, we can calculate a sequence of dynamic equilibria. In particular, we can calculate the counterfactual effects of restricting immigrant inflow to the US.

6.1 Calibration

To calculate a dynamic equilibrium, we need to take a stand on two objects: the structural parameters and the initial conditions of the system. Consider first the structural parameters. Our model is parsimoniously parameterized and consists only of seven parameters. We need to discipline the elasticity of substitution σ, the spatial labor supply function (ψ, ε),

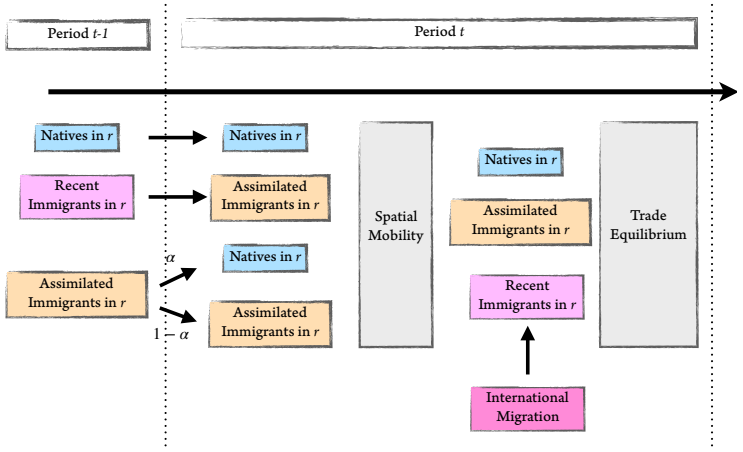


Figure 9: Timing of Quantitative Model

the distribution of human capital $\left(\theta, \{\varphi^j\}_j\right)$, the strength of scale effects ϑ and the spatial distribution of innovation efficiency $\{\zeta_r\}_r$.

For local idea production summarized by $\{\varphi^j\}_j$, ϑ and $\{\zeta_r\}_r$, we rely on our estimates of Sections 3 and 5. Our estimates for the distribution of human capital $\{\varphi^j\}_j$ are displayed in Figure 7 and the distribution of innovation efficiency $\{\zeta_r\}_r$ is shows in Figure 8. For the scale elasticity ϑ, we rely on our estimate in Table 10, which yielded $\vartheta = 0.55$.

The remaining parameters we adopt the following strategy. For the elasticity of substitution σ we assume a value of five, which is in between the median and average estimates reported by Broda and Weinstein (2006), within the range of estimates of Oberfield and Raval (2014) across different sectors (three to seven), and consistent with values of trade elasticity estimates reported in the trade literature (see e.g, Anderson (2010)). Our model implies that the labor share is given by $\frac{\theta\sigma-1}{\theta\sigma}$ with the remainder going to entrepreneurial payments. Targeting an entrepreneurial share of 10% yields a value of θ of two. For the two parameters of the labor supply function, (ε, ψ) we rely on estimates from the literature. We follow Peters (2019) that sets ε of about 2. For the moving shock ψ, we rely on the estimates of Eckert and Peters (2018), who use matched Census distributed by IPUMS, to measure that 55% of people remain in their commuting zone between 1880 and 1900. Since each period in our model is 5 years and SEAs are typically bigger than commuting zones we set $(1-\psi)^4 = 0.55$ so that $\psi \approx 0.14$ as an upper bound of the mobility probability. Our final set of parameters is summarized in Table 11.

In terms of our initial conditions, we take the empirically observed population distribution $\left[L^j_{r1880}\right]_r$ for natives and our five immigrant groups (Italians, Germans, British, Irish, Other) as given. Consistent with our timing assumption in the model (see Figure 9), all such

Structural Parameter		Value
θ	Shape parameter of pareto	2
σ	Elasticity of substitution	5
ε	Spatial elasticity of labor supply	2.5
ψ	Moving probability	0.14
ϑ	Scale elasticity of existing ideas	0.55
φ^j	Relative human capital	See Figure 7
ζ_r	Regional efficiency shifter	See Figure 8

Notes: The table reports the structural parameters used for our quantitative analysis.

Table 11: Structural Parameters

immigrants are assimilated. For the initial distribution varieties N_{r1880}, we use the model to predict the distribution of varieties under the assumption that is it a steady-state taking the population distribution in 1880 as given.[15] Because we calculate our main counterfactual results relative to a baseline with the empirically observed immigration inflows, this choice is not essential for our results.

6.2 The Economic Effects of Immigration Restrictions

To quantify the economic role of European immigrants, we consider the following experiment. Consider the US in 1880. In the following four decades, millions of immigrants from different countries are going to enter the United States. In Table 12 we report these aggregate inflows from the different countries of origin. To allocate these aggregate inflows across space in the US, we assume that immigrants were, in fact, avid believers of the traditional shift-share instrument and dispersed across the US in proportion of the prevailing distribution of immigrants of their nationality (or ancestry). Formally, we assume that the inflow of recent immigrants of nationality n to region r at time t, L_{rt}^{Rn}, is given by

$$L_{rt}^{Rn} = \frac{L_{rt-1}^{An}}{\sum_d L_{dt-1}^{An}} \times I_t^n, \tag{31}$$

[15] It is easy to show that the stationary distribution of varieties for a given population size is given by

$$N_{rt} = \left(\frac{1}{1+n}\right)^{\frac{1}{1-\vartheta}\frac{\vartheta}{1-\vartheta}} \left(\frac{\theta}{\theta\sigma-1}\left(\frac{\theta\sigma-1}{\theta-1}\right)^{1/\theta}\left(\sum_j \varpi_r^j \left(\varphi^j\right)^\theta\right)^{1/\theta} \zeta_r L_{rt}\right)^{\frac{1}{1-\vartheta}},$$

where n denotes the growth rate of the population.

Year	British	Irish	German	Italian	Others
1885	400,196	365,207	920,215	108,216	1,243,318
1890	410,704	308,854	524,966	159,444	806,658
1895	236,259	233,922	457,894	304,811	1,073,864
1900	92,500	171,788	121,178	298,950	689,016
1905	129,477	166,881	154,928	838,424	1,965,207
1910	340,041	178,059	173,794	1,092,051	3,130,033
1915	308,530	137,410	161,195	1,104,833	3,462,477
1920	63,348	29,035	13,032	125,083	941,949

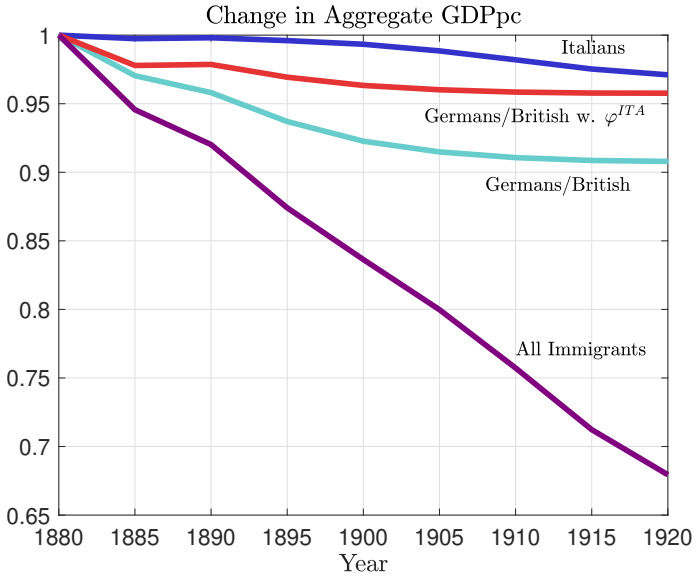
Notes: The table reports the number of immigrants entering the United States between 1880 and 1920.

Table 12: Aggregate Inflows of Immigrants: 1880 - 1920

where L_{rt-1}^{An} is the number of assimilated immigrants of nationality $n$in region r. Given the allocation rule (31), we can calculate the dynamic equilibrium of our economy for any matrix of immigration inflows $\{I_t^n\}_{tn}$.

To quantify the role of immigrants for US growth, we consider various forms of immigration restrictions, which could have been imposed in 1880. To measure both the aggregate and local effects of such restrictions, we proceed in the following way. We first calculate the equilibrium time-path for the actually observed immigration inflows, i.e. the matrix $\{I_t^n\}_{tn}$ reported in Table 12, while keeping the native population fixed in the 1880 to focus on their role of immigrants. We then calculate various policies like (i) restricting all immigration (i.e. setting $I_t^n = 0$ for all n and t), (ii) restricting immigration from Germany and Britain and (iii) restricting immigration from Italy. We are particularly interested in the comparison between the German/British restriction and the Italian restriction for three reasons. First, and most importantly: we estimated in Section 3 that German and British immigrants had higher innovation human capital than Italians. Hence, this comparison can shed light on the relative importance of human capital and pure scale effect. Second, the spatial exposure of different regions to these aggregate shocks is very different because the initial distribution of immigrants in 1880 differed by nationality. Finally, as seen in Table 12, the time path of aggregate inflows is very different. The big wave of Italian immigrants entered the US in the early 20th century, while they hey-days of German immigration were the late 19th century.

Aggregate Impact In Figure 10 we display the aggregate impact of immigration restrictions as implied by our model. Consider for example the purple line, which depicts the time path of aggregate GDP per capita (GDPpc) of an economy without any international immigration relative the baseline where immigration is given in Table 12. Our model suggests



Notes: The figure shows the change in aggregate income per capita for different counterfactuals relative to the baseline economy, where international immigration is given in Table 12, and native populations are given in their 1880 level. We report the counterfactual of shutting down all immigration (purple line), immigration from Germany and Britain (turquoise line) and immigration from Italy (blue line). In the red line we depict the aggregate impact shutting down immigration from Germany and Britain if German and British immigrants had the same human capital as Italian Immigrants, i.e. $\varphi^{Rn} = \varphi^{RI}$ and $\varphi^{An} = \varphi^{AI}$ for $n = GER$ and UK.

Figure 10: The Aggregate Impact of Immigration Restrictions

that the US economy could have lost 15% of GDPpc by 1900 and almost 1/3 of GDPpc by 1920, if it had closed its international borders by 1880.

To put this number into perspective, we consider two "partial" immigration restrictions. In turquoise we show the impact of restricting entry to all immigrants from Germany and Britain. Doing so would have reduced income per capita by roughly 10% by 1920. The time path of such losses is interesting: the losses accumulate quite quickly until roughly 1905, when British and German immigrants were important. In the early 20th century, the economic losses are much smaller as immigration from these countries became less important in the aggregate. In blue we depict the losses, which are attributable to an exclusion of Italian immigrants starting in 1880 (i.e. 40 years before the Quota-Act, which essentially restricted immigration from Italy). The economic consequences of such a policy are far less severe. First of all, given the small number of Italian immigrants prior to 1900, there is hardly an impact prior to 1900. Secondly, given the relatively low innovation human capital of Italian immigrants, the US economy would lose out on relatively fewer novel ideas by prohibiting Italians from entering the country.

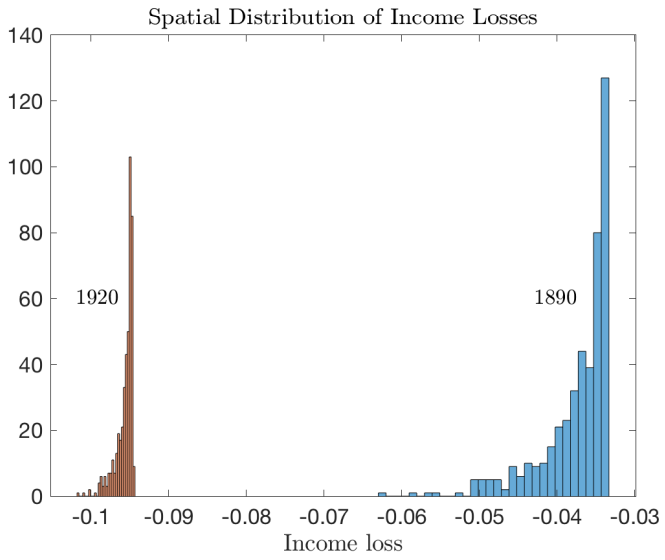
The difference between the German/British and the Italian counterfactual is suggestive

for the importance of human capital. However, a precise decomposition into the human capital channel and the scale effect channel it is still difficult because German and British immigrants differed both in their human capital and their aggregate importance. To see that differences in human capital as estimated from our micro data are quantitatively important, consider the red line, where we depict a hypothetical situation of restricting German British immigration if they had the same human capital as Italians. If most of our effect is due to scale effects, the red and the turquoise line should be close to each other as the number of people affected by the policy is identical. If in contrast, human capital was the sole driver of the impact of immigration, the red and the blue line should be equal as the average human capital endowment of immigrants is the same. Figure 10 shows that both channels are important. Differences in human capital account for about 2/3 of the differences, while differences in the size of the affected population account for about 1/3.

Spatial Impact The effects of shutting down international migration are spatially unbalanced. First, given the distribution of immigrants in 1880, different regions are going to be differentially exposed to changes in aggregate immigration inflows, if - as assumed in (31) - immigrants allocate spatially in proportion to the existing distribution. Second, the spatial heterogeneity in innate innovation efficiency ζ_r implies that high-ζ are particularly attractive for individuals with ample innovation human capital, while low-ζ regions attract groups, that have a comparative advantage in production rather the innovation.

In Figure 11 we display the distribution of regional income losses induced by restricting immigration inflows from Germany and Britain. In the blue histogram we depict the losses in the short-term, i.e. in 1890, and in the red histogram we depict the long-term consequences in 1920. The first thing to note is that the losses from immigration are cumulative: by 1920, the losses are two to three times larger as in 1890. This is both due to the fact that the ''treatment'' is larger as immigration restrictions would further reduce population inflows between 1890 and 1920 and that according to our theory, immigration restrictions are cumulative in the presence of positive scale effect, i.e. $\vartheta > 0$ - recall the discussion in Section 2.5, in particular Figure 1. Secondly, the dispersion of losses is larger in the short-run than in the long-run. This is the consequence of the existence of spatial linkages through spatial mobility and trade.

To see the aforementioned spatial heterogeneity in exposure directly, we depict the geography of losses in local income per capita for each SEA in the two maps in Figure 12. In the left map, colored in blue, we depict the distribution of losses of immigration restrictions for Germany and Britain. In the right map, colored in green, we depict such losses for immigration restrictions from Italy. We color-code regions in five quantiles, with darkest



Notes: The figure shows the distribution of income losses at the SEA level of shutting down immigration from Germany and Britain in 1880. The blue histogram reports the distribution of income losses in 1890. The red histogram reports the distribution of income losses in 1920.

Figure 11: The Spatial Impact of Immigration Restrictions

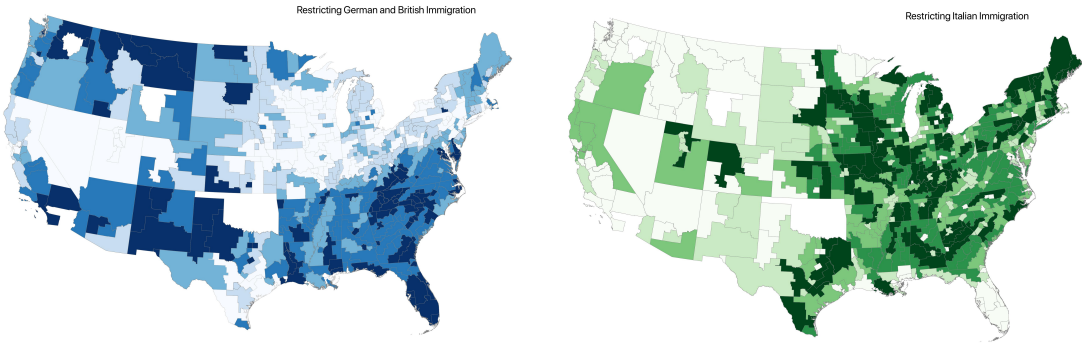
color indicating the largest losses, i.e. the highest exposure. The two maps in Figure 12 show that the economic impact of immigration restrictions differs strongly across space.

7 Conclusions

Between 1880 and 1920, more than 20 million immigrants made the United States their home. In this paper, we evaluate the role that immigration wave played to increase US productivity growth.

At the heart of our analysis is a novel micro-data set on individual patent behavior, which we construct by matching the population of US patents to the restricted-use complete count US Federal demographic decennial censuses from 1880-1920. To empirically measure immigrants' human capital, we augment this data by merging it to millions of original immigration records and historical passenger lists, both of which contain detailed information on immigrants' occupations before migrating. Our final dataset covers more than 400,000 patents and contains linked pre- and post-migration occupational information for more than 3 million immigrants over the four decades between 1880 and 1920.

Using this data we document three results. First, we show that immigrants were more prolific innovators than natives, in particular after having lived in the United States for more than a decade. Second, we find stark differences in patenting across nationalities with



Notes: The two maps show the distribution of income losses at the SEA level of shutting down immigration from Germany and Britain (left map in blue) and Italy (right map in green). In the left map we show income losses in 1910, in the right map we show income losses in 1920.

Figure 12: The Spatial Heterogeneity of Immigration Restrictions

German and British immigrants being much more innovative than, for example, immigrants from Italy. The fact that the composition of sending countries changed between 1880 and 1920 and that the regional distribution of immigrants in the US was very heterogeneous implies that the inflow of innovation human capital varied both across time and across space. Third, we estimate large spatial differences in innovation efficiency and show that immigrants were predominantly living in such locations.

We interpret this evidence through the lens of a novel model of spatial growth. The model stresses the local supply of innovative human capital as the main determinant of local productivity. How such local productivity gains affect other locations and the aggregate economy depends on trade linkages and the process of spatial mobility. We exploit the ability of the theory to tightly connect to the empirical evidence and aggregate the empirical estimates of human capital for different immigrant groups to quantify the importance of immigrants in a realistic setting featuring many locations, trade costs, and frictional population mobility.

As an application, we study the regional and aggregate economic consequences of immigration restrictions. At the aggregate level, we find income per capita in 1920 would have been 30% lower if the US has closed its borders to international migrants in 1880. We also find immigrants' human capital was an important contributor to US growth: while restricting German and British immigration after 1880 would have reduced income per capita by 10%, curtailing immigration from Italy would have had much smaller effects.

References

Abramitzky, Ran and Leah Boustan, ''Immigration in American Economic History,'' *Journal of Economic Literature*, 2017, *55* (4), 1311--45.

_ , Leah Platt Boustan, Katherine Eriksson, James J Feigenbaum, and Santiago Pérez, ''Automated Linking of Historical Data,'' Technical Report 2020.

Akcigit, Ufuk, ''Economic Growth: The Past, the Present, and the Future,'' *Journal of Political Economy*, 2017, *125* (6), 1736--1747.

_ , John Grigsby, and Tom Nicholas, ''Immigration and the Rise of American Ingenuity,'' *American Economic Review*, May 2017, *107* (5), 327--31.

Allen, Treb and Costas Arkolakis, ''Trade and the topography of the spatial economy,'' *Quarterly Journal of Economics*, 2014, *129* (6), 1085--1140.

Altonji, Joseph G and David Card, ''The effects of immigration on the labor market outcomes of less-skilled natives,'' in ''Immigration, trade, and the labor market,'' University of Chicago Press, 1991, pp. 201--234.

Anderson, James E., ''The Gravity Model,'' *forth. Annual Review of Economics*, 2010.

Arkolakis, Costas, Arnaud Costinot, and Andres Rodríguez-Clare, ''New Trade Models, Same Old Gains?,'' *American Economic Review*, 2012, *102* (1), 94--130.

Blevins, Cameron and Lincoln A. Mullen, ''Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction,'' *Digital Humanities Quarterly*, 2015, *9* (3).

Broda, Cristian and David Weinstein, ''Globalization and the Gains from Variety,'' *Quarterly Journal of Economics*, 2006, *121* (2), 541--585.

Buera, Francisco J. and Ezra Oberfield, ''The Global Diffusion of Ideas,'' *Econometrica*, January 2020, *88* (1), 83--114.

Burchardi, Konrad B, Thomas Chaney, Tarek Alexander Hassan, Lisa Tarquinio, and Stephen J Terry, ''Immigration, Innovation, and Growth,'' Working Paper 27075, National Bureau of Economic Research May 2020.

Burstein, Ariel, Gordon Hanson, Lin Tian, and Jonathan Vogel, ''Tradability and the Labor-Market Impact of Immigration: Theory and Evidence From the United States,'' *Econometrica*, 2020, *88* (3), 1071--1112.

Card, David, ''The Impact of the Marial Boatlift on the Miami Labor Market,'' *Industrial and Labor Relations Review*, 1990.

_ , ''Immigrant inflows, native outflows, and the local labor market impacts of higher immigration,'' *Journal of Labor Economics*, 2001, *19* (1), 22--64.

Desmet, Klaus, Dávid Krisztián Nagy, and Esteban Rossi-Hansberg, ''The Geography of Delevopment,'' *Journal of Political Economy*, 2018, *126* (3), 903--983.

Diamond, Rebecca, Tim McQuade, and Franklin Qian, ''The Effects of Rent Control Expansion on Tenants, Landlords, and Inequality: Evidence from San Francisco,'' *American Economic Review*, September 2019, *109* (9), 3365--94.

Donaldson, Dave and Richard Hornbeck, ''Railroads and American Economic Growth: A ''Market Access'' Approach,'' *The Quarterly Journal of Economics*, 02 2016, *131* (2), 799--858.

Dustmann, Christian, Uta Schönberg, and Jan Stuhler, ''Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment,'' *The Quarterly Journal of Economics*, 2016.

Eckert, Fabian and Michael Peters, ''Spatial Structural Change,'' 2018. Working Paper.

Gordon, Robert J, *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War*, Princeton University Press, 2017.

Head, Keith and Thierry Mayer, *Gravity equations: Workhorse, toolkit, and cookbook*, Centre for Economic Policy Research, 2013.

Hornung, Erik, ''Immigration and the diffusion of technology: The Huguenot diaspora in Prussia,'' *The American Economic Review*, 2014, *104* (1), 84--122.

Jones, Charles, ''Growth and Ideas,'' 2005, *1, Part B*, 1063--1111.

Jones, Charles I, ''R&D-based Models of Economic Growth,'' *Journal of Political Economy*, 1995, *103* (4), 759--784.

Jr, Robert E Lucas and Benjamin Moll, ''Knowledge Growth and the Allocation of Time,'' *Journal of Political Economy*, 2014, *122* (1), 1--51.

Kelly, Bryan T., Dimitris Papanikolaou, Amit Seru, and Matt Taddy, ''Measuring Technological Innovation Over the Long Run,'' January 2020. Working Paper.

Kerr, Sari Pekkala, William Kerr, Çağlar Özden, and Christopher Parsons, "Global Talent Flows," *Journal of Economic Perspectives*, Fall 2016, *30* (4), 83--106.

Kerr, William, *The Gift of Global Talent: How Migration Shapes Business, Economy & Society*, Stanford University Press, 2018.

Kortum, Samuel, "Research, Patenting, and Technological Change," *Econometrica*, 1997, *65* (6), 1389--1419.

Krugman, Paul, "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, 1980, *70* (5), 950--959.

Lee, Sun Kyoung, "Essays in History and Spatial Economics with Big Data," *Doctoral Dissertation*, 2019. Working Paper.

Manson, Steven, Jonathan Schroeder, David Van Riper, and Steven Ruggles, "IPUMS National Historical Geographic Information System: Version 14.0 [Database]," 2019.

Nagy, Dávid Krisztián, "Hinterlands, City Formation and Growth: Evidence from the US Westward Expansion," 2020. Working Paper.

Oberfield, Ezra and Devesh Raval, "Micro data and macro technology," Technical Report, National Bureau of Economic Research 2014.

Ottaviano, Gianmarco IP, Giovanni Peri, and Greg C Wright, "Immigration, offshoring, and American jobs," *American Economic Review*, 2013, *103* (5), 1925--59.

Peri, Giovanni, "Immigrants, Productivity, and Labor Markets," *The Journal of Economic Perspectives*, 2016, *30* (4), 3--29.

Perla, Jesse and Christopher Tonetti, "Equilibrium imitation and growth," *Journal of Political Economy*, 2014, *122* (1), 52--76.

Peters, Michael, "Market Size and Spatial Growth - Evidence from Germany's Post-War Population Expulsions," *Unpublished manuscript*, 2019.

Petralia, Sergio, Pierre-Alexandre Balland, and David Rigby, "HistPat Dataset," 2016.

Redding, Stephen J. and Esteban Rossi-Hansberg, "Quantitative Spatial Economics," *Annual Review of Economics*, 2017, *9* (1), 21--58.

Romer, Paul M., ''Endogenous Technological Change,'' *The Journal of Political Economy*, 1990, *98* (5), S71--S102.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek, ''Integrated Public Use Microdata Series: Version 10.0 [Machine-readable database],'' 2020.

Sequeira, Sandra, Nathan Nunn, and Nancy Qian, ''Immigrants and the Making of America,'' *The Review of Economic Studies*, 2020, *87* (1), 382--419.

Walsh, Conor, ''Firm Creation and Local Growth,'' 2019. Working Paper.

A Empirical Appendix

Our analysis relies on four main data sources: the complete count population census for the years 1850 to 1940, the population of US patents since 1790, the Hamburg Passenger Lists and the Immigration Records from Castle Garden. In this section we describe these datasets.

A.1 Population Census Data: *IPUMS Complete Count U.S. Census (1850-1940)*

Our main source of information is the restricted-access version of the complete count US Population Census for the years 1850-1940 (Ruggles et al., 2020). These individual-level micro data exist for all decades except for 1890 (which was lost due to fire) and contain detailed demographic information for each individual residing in the US.

We heavily rely on the fact that the restricted-access version of the data reports the name of each individual, which we use to link individuals to patents and immigration records (see Section B, where we describe the linking procedure in more detail). We also use the information on individuals' locations, their age, their immigration status, their birthplace and their employment structure, in particular their occupation.

A.2 Historical Immigration Records *(1820 - 1914)*

We construct our immigration database from two primary sources: the Castle Garden Immigration Database, which captures *inflows* into the US, and the Hamburg Passenger Lists, which captures *outflow* from Europe to the US.

A.2.1 The *Castle Garden Immigration Database (1820-1914)*

The Castle Garden database contains the list of all immigrants entering the US via the port of New York between 1820 to 1914. In total, the database comprises approximately 11 million individual micro-records. Castle Garden was America's first official immigration center, and the captain of each arriving ship prepared a Customs Passenger List and filed with it at the port of arrival. This enabled a systematic collection of data on immigration to the United States.

For our analysis we exploit the fact that the Castle Garden Data contains details information on immigrants' pre-migration occupations. In order to link this information to the Population Census, we rely on a host of demographic information such as name, age, family structure and arrival year (see Section B.1 for details of the record linking process.).

A.2.2 The Hamburg Passenger Lists (1850-1914)

The The Hamburg Passenger Lists (HPL) contain passenger lists of ships that departed from the port of Hamburg, Germany from 1850-1934 . We got access to this database through a cooperation with the State Archive of Hamburg. Due to privacy concerns we only have access to the data until 1914.

We have access to approximately 4.6 million individual records. Among them, roughly 77% were headed to the United States. Like the immigration records from Castle Garden, the Hamburg Passenger Lists report the pre-migration occupation of each immigrant. In addition, they usually contain information such as the name, gender, age, nationality, departure date, ship name, route and final destination.

Because the HPL records were originally written in German and some key information such as nationality was not reported in certain time periods, we process the HPL data using translation, classification, imputation and harmonization techniques of key variables. Details of these "pre-linking" data preparation steps of the HPL data are available in Section D.1.

A.3 *Historical Data on US Patents*

To measure innovative activity, we exploit information on patenting. We build our patent-dataset from two sources: the HistPat (Petralia et al. (2016)) and Google's Patent Search Engine (https://patents.google.com/). The United States Patent and Trademark Office (USPTO) granted millions of patents since 1790, and digitized version of these data since 1836 became publicly available through USPTO Patent database. However, pre-1975 patents are mostly in a format that cannot be directly used for research (i.e. not in a machine-readable format for quantitative analyses).

The HistPat and Google Patent search engines tackles these challenges by cleaning and processing this original data. Google (and Reed Tech) processed original patent documents through the Optical Character Recognition (OCR) process the HistPat takes these OCR-processed data and constructs a systematic database through a standard text0mining algorithm. Therefore, the HistPat releases Patent Number, Patent Grant Year, Geographic Information such as city, county, and state for all patents between 1836 and 1975.

Given these digital sources, first, we take the HistPat data and extracts the patent year and geographic inforamtion, and harmonize the geography at the SEA level using county-boundary files by decade from NHGIS (Manson et al. (2019), see Section D.2 for details). Then, we use Google's patent search engine, and collect 1. Patent assignee names, 2. entire patent description. 3. Patent Industry classification which follows the Cooperative Patent Classification (CPC) for all patents.[16] We harmonize the CPC-based patent industry classification into IND1950 which is a standard 1950 Industry classification of US Census Bureau (Ruggles et al. (2020) also recodes information about industry into 1950 Census Bureau industrial classification system to enhance comparability of industry data across years included in the IPUMS).

After we construct this patent-level data that fits our research need, we link patent records to individuals in the population census, using the name and location of patent assignees (see Section B.2 for details of this procedure). This patent-census linked data enables us to study the determinants of patenting at the micro-level. We then take the patent description of each patent, and devise a measure of novelty by performing textual analysis. In doing so, we closely follow Kelly et al. (2020) (See Section D.4 describes for details). This textual

[16]The Cooperative Patent Classification (CPC) has been jointly developed by the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). Google Patent Database lists CPC-based Classification for every patent.

analysis-based patent novelty measure allows us to study how a particular patent is more novel (or ''innovative'') than others beyond classification.

B Record Linking

In this section, we provide details on the procedure to link (i) immigrants from the Hamburg Passenger Lists and Castle Garden Database to the US Census and (ii) patent assignees from the US Patent records to the US Census.

We implement a supervised discriminative machine learning approach to link historical records without time-invariant individual identifier(s). The essence of this approach is that we use training data (as ''teaching-material'') to train the algorithm on how to identify the potential matches based on certain discrepancies in the data.[17] We create a training dataset which contain both ''true'' and ''false'' matches and their characteristics. By taking this training data, we build a prediction model, or *learner*, which will enable us to predict the outcome for new, unseen objects. A well-designed learner armed with a solid training dataset should accurately predict outcomes for new, unseen objects.

Our end-goal is to use the training data as inputs to predict the output values (i.e. predicting the pair of observations are potentially a match or not). Throughout this process, we tune a matching algorithm that matches individual records from various sources (discussed in Section A) while minimizing both false positives and false negatives and reflecting inherent noises in historical records.

We explored various models for model selection, and we ultimately chose the machine-learning approach of *random forest classification.* This method is more conservative in matching records, the number of unique matches are significantly higher than for other machine-learning based algorithms such as the standard Support Vector Machine model (used by IPUMS in linking complete count 1880 and 1% sample of census records), and it is more flexible in dealing with missing values than many other competing algorithms such as XGBoost (used by Ancestry.com for providing users the potential search outcomes). Additionally, it is also computationally efficient (Lee (2019)).

Abramitzky et al. (2020) compares how various automated algorithms including machine-learning approach to record linking perform relative to each other and relative to the manually linked data. Abramitzky et al. (2020) concludes that automated methods generate a very low (less than 5%) false positive rates, and that coeffiicient estimates and parameters of interest are very similar when using linked samples based on each of the different automated methods.

Overall, finding of such supports that our record linking algorithm (as a class of automated record linking algorithm) would be reliable by generating very low false positives, while computationally efficient.

[17] For example, Heinrich Engelhard Steinweg, the founder of prominent piano manufacturing company, *Steinway & Sons*, anglicized his names into ''Henry E. Steinway.'' Therefore, in linking his records across censuses, string comparison measures called Jaro-Winkler distance of his first (Heinrich vs. Henry), middle (Engelhard vs E.) and last name (Steinweg vs Steinway) would show name discrepancies) even if his birth year and birthplace may be the same across different records

B.1 Linking Immigrants to the US Census

B.1.1 Linking Immigrants from the Hamburg Passenger Lists to the US Census

To link individuals appearing in the Hamburg Passenger Lists to the US Census, we use information on name, race, age, departure year and nationality. We proceed in two steps. We first identify a set of individuals with the same nationality and race, whose age and names are similar. Specifically, we consider all individuals whose age in the Passenger Lists and the Census differ by at most two years and whose Jaro-Winkler distance of first name and last name is at least 0.8 (1 being the maximum Jaro-Winkler string similarity measure and 0 being the minimum).

If multiple individuals from the US Census satisfy these criteria (i.e. are potential matches), we iteratively apply three rules to identify a unique match. First, we restrict the Jaro-Winkler distance for the last name to be greater than or equal to 0.85. Second, for individuals over the age of 25 at the time of immigration, we impose that their marital status must be stable across datasets. Finally, we impose the departure year in the Hamburg Passenger List and the year of immigration in the census data to not be more than three years apart. See Table 13 where we summarize this procedure.

Data Linked	Rules for Searching Potential Matches	Rules for Pinning down Unique Matches
Hamburg Passenger Lists to Census	- Search within the same birthplace and race: 1. Jaro-Winkler Distance of first and last name (minimum of 0.8) 2. Age (Census vs. HPL Birth Year, differences up to 2 years)	- Given the potential set of matches: 1. Last name Jaro-Winkler distance ≥ 0.85 2. If Age ≥ 25, Census Marital Status = Marital Status 3. $\lvert$Census Immigration Year - Departure Year$\rvert \leq 3$
Castle Garden Census to Census	- Search within the same birthplace and race: 1. Jaro-Winkler Distance of first and last name (minimum of 0.8) 2. Age (Census vs. CG Birth Year, differences up to 2 years)	- Given the potential set of matches: 1. Last name Jaro-Winkler distance ≥ 0.85 2. If Age ≥ 21, Census Marital Status = Marital Status 3. $\lvert$Census Immigration Year - Arrival Year$\rvert \leq 3$
Patent to Census	- Search within the same State Economic Area: 1. Jaro-Winkler Distance of first and last name (minimum of 0.8)	- Given the potential set of matches: 1. Last name Jaro-Winkler distance ≥ 0.85 2. Patent Year - Birth Year ≤ 80

Table 13: Record Linking Rules

In Table 14 we report the number of matches and match rate between HPL and Census for different nationalities over time. Not surprisingly, among the matched records from the Hamburg Passenger Lists, German immigrants in account for the largest share of immigrants. In contrast, Italian immigrants are not the primary immigrant group as Italian immigrant did not typically depart to America via Hamburg. However, Russian immigrants rapidly rose in the beginning of the twentieth century, making up one of the largest immigrant group in linked HPL-Census records, their match rate is systematically lower than those of German immigrants. It is also noteworthy that match rate given each nationality tends

to decrease over time as the number of immigrants increased. In Section C.1 we discuss potential explanations for such differences in match rates over time and by nationality in greater detail.

	1850	1860	1870	1880	1900	1910	1920
Number of Natives	8,979,637	11,881,405	16,510,374	14,702,972	33,017,727	39,735,657	46,122,485
Number of Foreign-borns	1,263,283	2,210,386	2,963,059	3,464,945	5,742,188	7,876,248	7,738,920
Number of Matches: CG and Census	175,170	385,503	552,971	720,918	580,142	598,789	441,803
Number of Matches: Germany	56,262	154,282	246,804	343,975	273,702	256,698	175,946
Match Rate: Germany	(0.28)	(0.27)	(0.25)	(0.25)	(0.11)	(0.11)	(0.08)
Number of Matches: Ireland	89,297	161,003	193,091	208,857	89,464	72,920	48,316
Match Rate: Ireland	(0.39)	(0.32)	(0.27)	(0.24)	(0.09)	(0.08)	(0.05)
Number of Matches: Italy		38	84	542	55,676	101,264	85,874
Match Rate: Italy		(0.02)	(0.03)	(0.05)	(0.10)	(0.14)	(0.12)
Number of Matches: UK	26,838	61,315	100,971	143,078	94,911	84,103	61,989
Match Rate: UK	(0.26)	(0.30)	(0.28)	(0.25)	(0.13)	(0.12)	(0.09)
Number of Matches: HPL and Census		12,438	47,947	87,860	198,898	318,014	284,414
Number of Matches: Germany		11,182	41,466	74,116	154,233	196,156	143,619
Match Rate: Germany		(0.41)	(0.33)	(0.29)	(0.23)	(0.25)	(0.17)
Number of Matches: UK		903	4,091	8,328	19,467	17,710	13,799
Match Rate: UK		(0.27)	(0.27)	(0.22)	(0.14)	(0.13)	(0.10)
Number of Matches: Italy		2	10	43	604	2,177	2,027
Match Rate: Italy		(0.02)	(0.04)	(0.06)	(0.15)	(0.36)	(0.30)
Number of Matches: Russia					4,904	49,705	45,468
Match Rate: Russia					(0.19)	(0.24)	(0.16)
Number of Matches: Patent and Census	3,243	6,884	10,958	22,405	96,526	97,807	75,752
Number of Natives	2,501	4,715	6,827	133,543	67,140	67,480	57,233
Number of Foreign borns	742	2,169	4,131	9,051	29,386	30,327	18,519

Table 14: Number of Matches and Match Rate

Notes: This table shows the number of observations that were uniquely matched between dataset of interests (HPL/CG/Patent) and census, and corresponding match rate by nationality over study period. As HPL and CG observations finishes around 1914, we only link immigration records against census records up to 1920 and skip linking 1930 and 1940 census records at all. However, for patent and census linking, as patent records are available for the 1930s and 1940s records also, we still match patent records against 1930 and 1940 demographic census records. Table for other years or nationality are available upon request.

B.1.2 Linking Immigrants from the Castle Garden Immigration Records to the US Census

Our procedure to link immigrants appearing in the Castle Garden data to the Census uses a procedure akin to the one described in Section B.1.1 above. First, we identify a set of potential matches by insisting on similar first and last names (i.e. a Jaro-Winkler distance of minimum 0.8), a maximum age difference of two years and a common nationality and race. Then we identify unique matches by imposing the same three rules as above: stricter last name similarity, stable marital status, and a gap between the departure year and the immigration year gap of less than 3 years This procedure is again summarized in Table 13.

Table 14 shows the number of matches and the match rate between the Castle Garden data and Population Census for different nationalities . German immigrants make up roughly 30-40% of the linked data between Castle Garden and the Census. Note that Germans account for 20% of all immigrants in Castle Garden (i.e. their match rate is relatively high).

In contrast, Table 14 shows that the match rate Irish and British immigrants is systematically lower. We also see the rapid rise of Italian immigration at the turn of the century.

B.2 Linking Patent-Holding Inventors to the US Census

To link patents to individuals in the US Census, we rely on the information contained in HistPat: the name of the patentee and the county where the patent was issued. To account for the fact that the place where the patent was issued does not necessarily coincide with the living location of the patentee, we harmonize the patent and Census data at the level State Economic Areas (SEA).

We first take the entire patent data and block it first by the decade that a patent was filed and within each decade by the SEA of each filing. We then link this to the appropriate decade of the census for each SEA using the Jaro-Winkler distance of first and last name between census records and patent inventor records. Again we identify the set of potential matches by insisting on a Jaro-Winkler distance of at least 0.8. We also restrict attention to individuals that were at least 16 years old at the time the patent was issued.

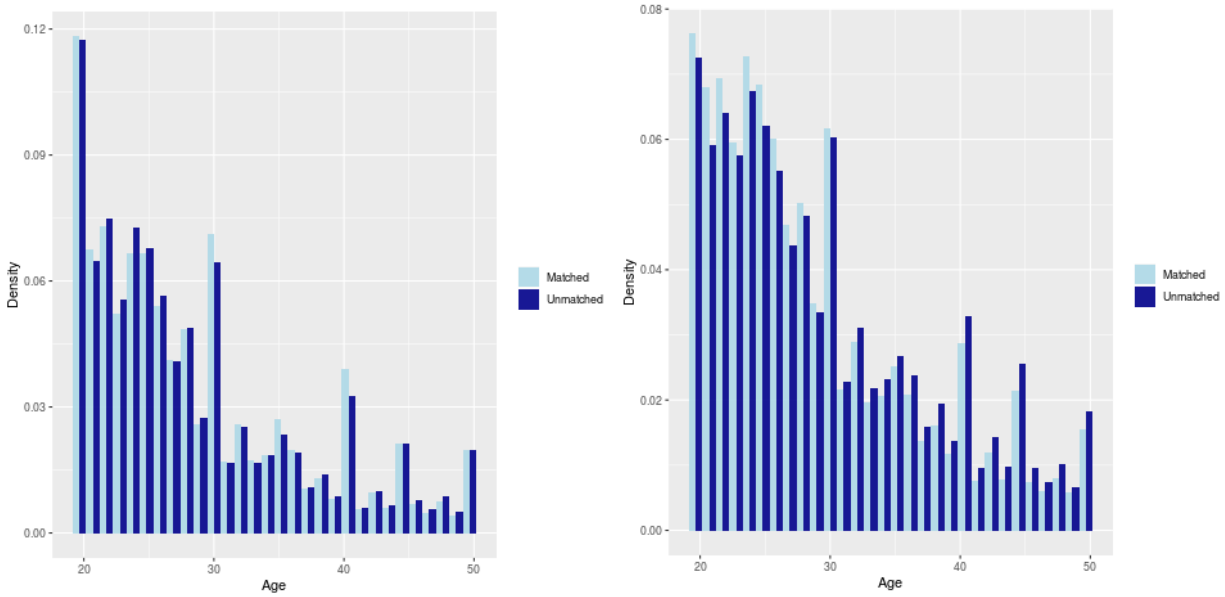
To then determine a unique match, we again use the same iterative procedure as above. Specifically, we impose the first rule of a Jaro-Winkler distance to be at least 0.85. We also restrict the patentee to be at most 80 years old at the time of patenting. Again, see Table 13 for a summary of this procedure.

C Evaluating the Quality of the Linked Data

In this section, we discuss the quality of our record linking algorithm and linked data. We show 1) there are no systematic differences between matched and unmatched observations, 2) our record linking algorithm does not create a bias by over-representing some groups while under-representing others.

C.1 Representativeness

We first compare the matched and unmatched individuals from Immigration records. Figures 13 and 14 show the relative age distribution of matched and unmatched observations from the immigration records. For Figure 13 we take all males aged between 20 and 50 years at the time of census who arrived in the last 10 years (based on the census year) from Castle Garden Immigration Arrival records, and plot the relative age distribution of CG observations that were uniquely linked to the census records in Light Blue, and plot the corresponding of CG observations that were not linked to the census in Dark Blue. If a certain age group from the CG records was more likely to be linked to the census than other groups, the relative age distributions of the matched and unmatched would look different. However, respective age distributions are not systematically different between the matched and unmatched.

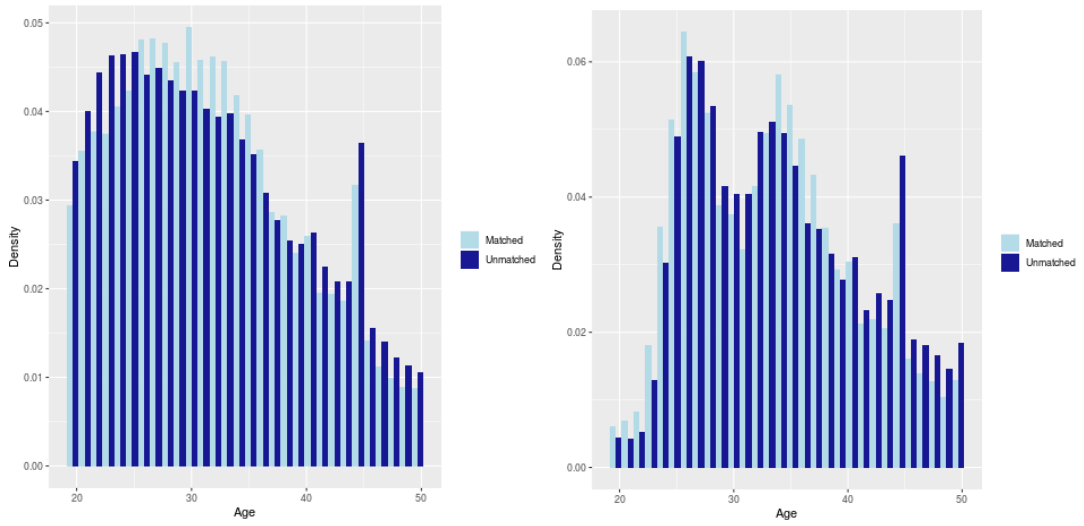


(a) 1860 (b) 1880

Notes: The above figures show the age distribution of Castle Garden Data that is uniquely linked to the Census (Light Blue) against Castle Garden Data that is unmatched to the Census (Dark Blue). The left figure plots for the 1860 data, and the right figure for the 1880 data. Figures for other years are available upon request.

Figure 13: Age Distribution of the Matched and Unmatched (Castle Garden)

Figure 14 shows the same distributions for the Hamburg Passenger Lists. Again, we find that the age distribution of matched individuals (shown in light blue) is very similar to the one of the individuals that we are unable to match (dark blue).



(a) 1910 (b) 1920

Notes: The above figures show the age distribution of Hamburg Passenger Lists Data that is uniquely linked to the Census (Light Blue) against Hamburg Passenger Lists Data that is unmatched to the Census (Dark Blue). The left figure plots for the 1910 data, and the right figure for the 1920 data. Figures for other years are available upon request.

Figure 14: Age Distribution of the Matched and Unmatched (Hamburg Passenger Lists)

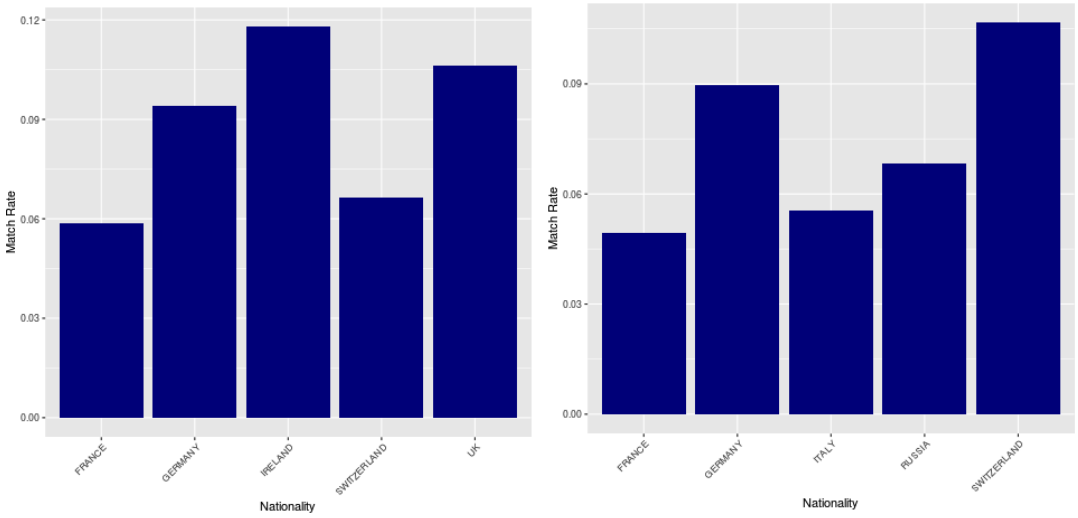
While the previous section looks at the age structure between the matched and unmatched, this section analyzes if the ratio between the matched and unmatched systematically differ by nationality.

Figure 15 shows that the match ratio (i.e. $\frac{\#\ matched}{\#\ unmatched}$) from CG records does not vary substantially across nationalities. To make this ratio comparable, as we did in previous section, we again restrict observations to recent arrivals who arrived in the last 10 years at the time of census, and to be 20-50 years old at the time of census. The left panel of Figure 15 shows the top five nationalities from CG records based on 1860 census. It shows that France, Germany, Ireland, Switzerland and the UK were the top five immigrant-sending countries between 1850 and 1860, and the match ratio differs across countries is approximately 0.1. As the composition of immigrants change over time, the top immigrant-sending countries also change. By 1900, immigration from UK and Ireland decreased in relative magnitude, whereas Russia and Italy rose as one of the biggest immigrant-sending countries.

Figure 16 shows our match ratio from HPL records does not vary too much by nationality. Again, we strict the observations to recent arrivals, aged between 20 and 50. It is noteworthy that the match ratio varies over time for the same country. The most dramatic case is Croatia/Serbia where the match ratio increased from 0.1 in 1910 to well over 0.3 in 1920. For Russia the match ratio decreased from 0.2 to 0.15.

There are many reasons why the match ratio could differ by nationality — to enumerate some, Anglicization of first and last names; popularity of common names; and composition of immigrants (for example, if there are many Italian immigrants such as "23 years old Giovanni Rossi from Italy" with similar age groups, common names and common characteristics such as marital status); the number of immigrants who returned to their home countries (for

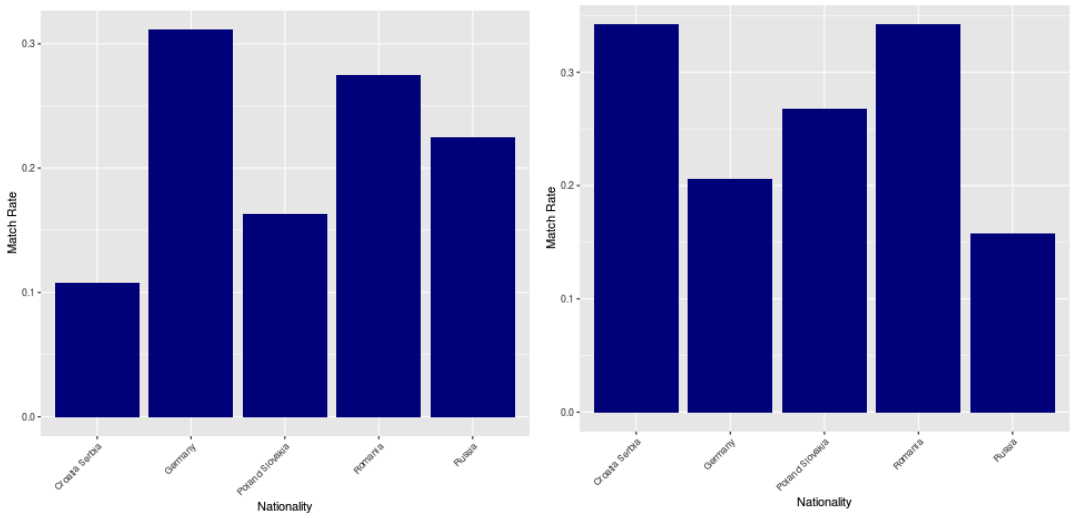
example, if a certain nationality group was much more likely to return to their home countries and therefore, we only see their Immigration Arrival records but do not see their presence in the census records, the match ratio can be lower for such group).



(a) 1860 (b) 1900

Notes: The above figures show the ratio of (the number of observations matched to census/the number of unmatched observations in Castle Garden). The top five nationalities are based on the count of immigrants' birth place. The left figure plots for the 1860 data, and the right figure for the 1900 data. Figures for other years are available upon request.

Figure 15: Ratio of the Matched and Unmatched (Castle Garden) by Nationality



(a) 1910 (b) 1920

Notes: The above figures show the ratio of (the number of observations matched to census/the number of unmatched observations in Hamburg Passenger Lists). The top five nationalities are based on the count of immigrants' birth place. The left figure plots for the 1910 data, and the right figure for the 1920 data. Figures for other years are available upon request.

Figure 16: Ratio of the Matched and Unmatched (Hamburg Passenger Lists) by Nationality

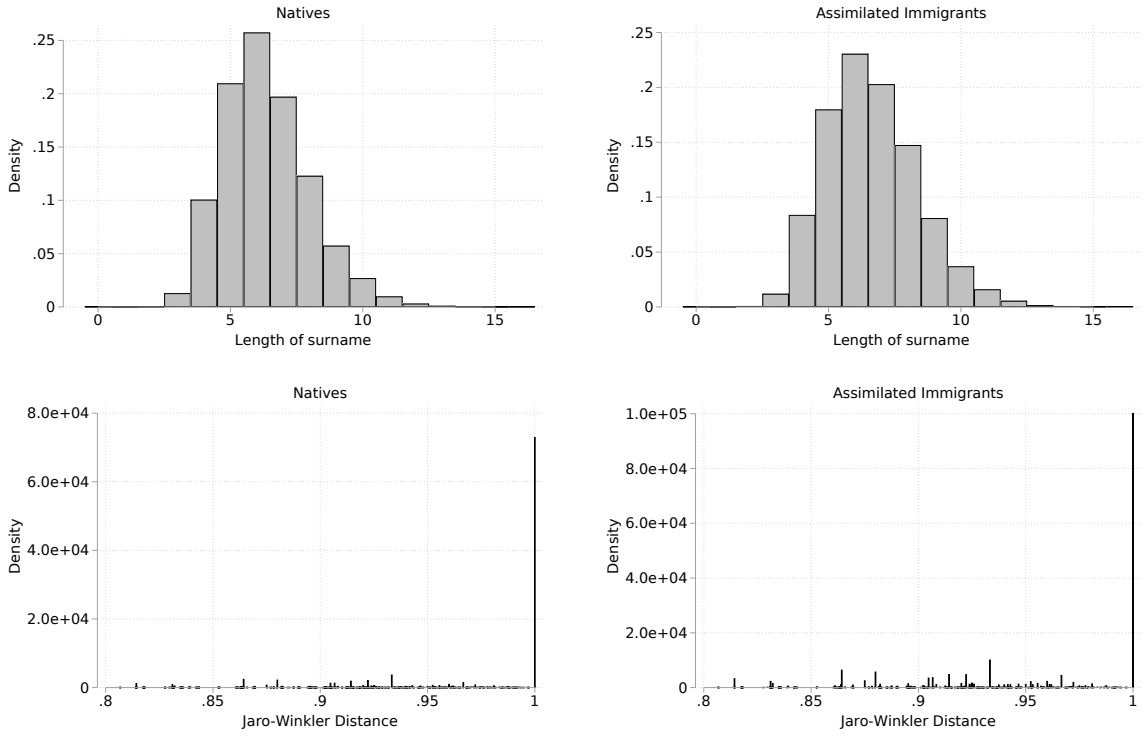
C.2 Matching Patents to Immigrants and Natives

In Section 4 we showed that an important aspect of systematic differences in immigrants' entrepreneurial human capital is the length of their stay in the US. A potential concern could be that it might simply reflect a particular bias in our matching procedure: maybe immigrants have more distinct names than the native population and hence we simply achieve a higher match rate even though the actual innovation productivity might not differ.

To see that this concern is unlikely to be important, consider Figure 17. In the top row we report the distribution of the length of last names for natives (left panel) and assimilated immigrants (right panel). These distributions are quite similar, i.e. it is not the case the immigrants have substantially longer names, which might be easier to find on the patent records. In the bottom panel, we report the distribution of the Jaro-Winkler distances for natives and immigrants. Again, if migrants and natives differed in the ease with which their names could be matched, we would expect these distributions to be different. Figure 17, however, shows that they are very similar.

C.3 Patent Matching Rates by Industry

In Table 15 we report the marginal distribution of patents across industries and our match rate by industry for all decades. More than 80% of all patents between 1860 and 1920 are for durable and non-durable goods. The match rate of our algorithm is very similar across years but improves markedly over time. This is due to the fact the information on inventor names



Notes: In the top row we depict a histogram of the length of the last names of natives (left panel) and assimilated immigrants (right panel). In the bottom row we depict the distribution of the Jaro-Winkler distances of the matches of natives (left panel) and assimilated immigrants (right panel).

Figure 17: Quality of Patent Matches: Natives vs. Immigrants

(i.e. Optical Character Recognition-based transcription quality) is much more complete in the 20th century than it is in the 19th century.

D Data Processing

D.1 Data Harmonization: Hamburg Passenger Lists

The original information of the Hamburg Passenger Lists is in German and in a non-structured format. To make this information operation we therefore had to translate, harmonize and classify it. The three pieces of information we used are the the pre-migration occupation, the final destination and the nationality.

Occupation (*"Beruf"*) We cleaned, translated, standardized and coded occupation information from the Hamburg Passenger Listsfollowing the 1950 Census Bureau occupation information classification system ("OCC1950") and the Historical International Standard classification of occupations ("OCCHISCO") to enhance comparability across years. 53% of HPL records have occupational responses and this information is almost always available for working-age males. The most frequently appearing occupations are farm laborers, general

Table 15: Matching Rates By Industry

Decade	1860-1879		1880-1899		1900-1909		1910-1919	
Industry	Total	Matched rate	Total	Matched rate	Total	Matched rate	Total	Matched rate
Agriculture, Forestry, and Fishing	0.09	0.12	0.05	0.07	0.04	0.60	0.04	0.45
Mining and Construction	0.02	0.13	0.02	0.09	0.02	0.62	0.02	0.51
Durable Goods	0.66	0.13	0.69	0.08	0.71	0.61	0.72	0.48
Nondurable Goods	0.16	0.12	0.15	0.08	0.13	0.57	0.13	0.46
Transportation	0.02	0.13	0.02	0.08	0.02	0.61	0.02	0.50
Telecommunication	0.00	0.10	0.01	0.06	0.01	0.64	0.01	0.57
Utilities and Sanitary Services	0.04	0.13	0.05	0.08	0.05	0.62	0.05	0.49
Entertainment and Recreation	0.00	0.09	0.01	0.06	0.01	0.61	0.01	0.51
Other			0.00		0.00	0.69	0.00	0.72

laborers, managers, followed by manufacturing workers such as tailors, shoemakers, and cabinet makers.

Destinations (*"Zielort"*) We translated the ship route and final destination information to identify passengers who were headed to the United States. The majority of passengers departing from the port of Hamburg were headed to North America, however, still some passengers were headed to South America (such as Buenos Aires and Rio de Janeiro) and other parts of Europe.

Nationality (*"Nationalitaet"*) We rely on the information on immigrants' nationality to match immigration records to the US Census. This information is only reported in the Hamburg Passenger List starting in 1898. For al records prior to 1898,

we impute this information. Specifically, we use a name-based nationality and ethnicity classification tool called "NamePrism" (which is also used by Diamond et al. (2019)), which predicts an individual's nationality and ethnicity based on their first and last name.

We also use another algorithm to impute gender based on first name (Blevins and Mullen (2015)). As these tools also provide corresponding "reliability" measures in the form of probabilities, we restrict our analysis to observations where such "success probabilities" are at least 50%.

D.2 Geography Harmonization

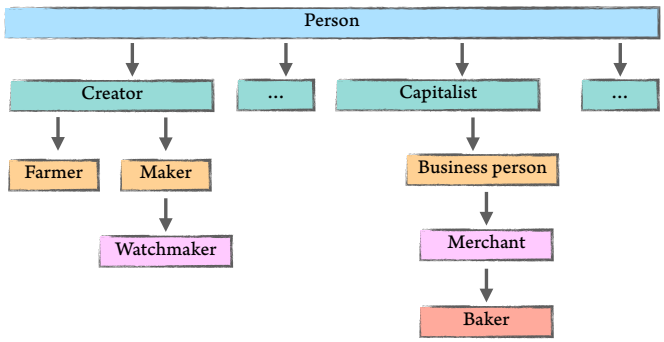
We merge all the aforementioned datasets at a unified spatial level. As counties and state boundaries have changed over time, we take the county shape files from National Historical Geographic Information System (Manson et al. (2019)) and create a time-consistent crosswalk of counties in constant borders.

D.3 Pre-Migration Occupations

Both the Castle Garden data and Hamburg Passenger Lists contain information on the occupation the respective immigrant had before entering the United States. We refer to such occupations as pre-migration occupation. To make this information, which in the raw data is simple a string variable, workable we aimed to harmonize these pre-immigration occupations to the census occupation codes.

To do so we exploit the lexical database ‘‘Wordnet’’, which is part of the NLTK library in Python.[18] Wordnet groups words into a set of synonyms called synsets. It is basically a network of related words based on synonyms. Wordnet also offers different measures of how these words are related, one of which is a 'word similarity' measure. It is based on the depth of the Wordnet taxonomy of the two words and their lowest common subsumer. We use this structure to harmonize the different pre-migration occupations of the Castle Garden Data into a set of roughly 300 groups.[19] We perform the same step for the census occupations. Given these two lists, the list of Castle Garden occupations and occupations from the Census , we then iterate over the the first list to find the census occupation with the highest similarity score.

In Figure 18 and Table 16 we provide two examples of our procedure. Figure 18 shows the hierarchical structure of four occupations - farmer, watchmaker, baker and merchant - in Wordnet. Consider for example a watchmaker. A watchmaker is a part of all ‘‘makers’’ as they share a large set of synonyms. ‘‘Makers’’ in turn are a subset of ‘‘creators’’, a property which they share with farmers. Similarly, bakers and the broader class of merchants are a subset of business persons and capitalists.



Notes: The figure shows an example of four occupations (farmer, watchmaker, baker and merchant) and their relation Wordnet.

Figure 18: The Hierarchical Structure in Wordnet

In Table 16 we report the mapping from the reported occupations in the Castle Garden Data (left column) to the eventual occupations in the census (right column). Via Wordnet we are able to deduce the class ‘‘repairmen’’ for the ‘‘watch repairer’’ or the general class ‘‘housekeeper’’ to the kitchen maid. Table 16 also shows what could go wrong: both the piano maker and the pasta maker are classified as ‘‘makers’’, which find their closest match in the census category ‘‘tool maker’’. While this might be reasonable for the artisan of musical

[18] https://wordnet.princeton.edu

[19] If we cannot find a word in Wordnet, we split the occupation into sub-occupations and search for these, eg ‘‘taxi driver’’ becomes ‘‘taxi’’ and ‘‘driver’’. If still neither exists, we use ‘‘wordninja’’ (another python library) that splits words without white spaces, eg ‘‘breadbaker’’ becomes ‘‘bread’’ and ‘‘baker’’. While iterating over the occupations, we always make sure to only use words that mean what we are looking for, eg in the above example of ‘‘taxi’’ and ‘‘driver’’, we are obviously interested in the word ‘‘driver’’ and not ‘‘taxi’’. We do so by requiring that each word has a person as a hypernym.

instruments, the pasta maker (most likely an Italian immigrant) should arguably be assigned the occupational code of a baker.

Castle Garden Occupation	Class in Wordnet	Census occupation
MIDWIFE	midwife	Midwives
PIANO MAKER/MANUFACTURER	maker	Tool makers
PASTA MAKER	maker	Tool makers
SCULPTOR	artist	Artists and art teachers
WATCH REPAIRER	repairman	Mechanics and repairmen
KITCHEN MAID	housekeeper	Housekeepers, private household

Table 16: Classification of Occupations: Example

D.4 Measuring the Novelty of Patents

To provide direct evidence for the importance of human capital, we aim to measure whether the patents by immigrants are systematically different. To do so, we exploit the patent descriptions using the data that construct (See Section A.3 for patent-data construction procedure and details). As in Kelly et al. (2020) we measure the similarity of patents relative to previous patents using textual analysis.

More specifically, we identify terms that are most diagnostic of a document's topical content through the "Term Frequency Backward Inverse Document Frequency" ($TFBDIF$) transformation of word counts. Consider a word w and patent document d. We then calculate:

$$TFIDF_{dw} = TF_{dw} \times BIDF_{w}.$$

Here TF_{dw} is the relative frequency of word w within the patent document d

$$TF_{dw} = c_{dw} / \sum_{k} c_{dk}$$

and the Backward Inverse Document Frequency $BIDF_w$ measures the log frequency of patent documents containing the term w in any patent granted *prior* to patent d, i.e.

$$BIDF_{w} = \ln\left(\frac{\text{Number of documents before } d}{1 + \text{Number of documents before } d \text{ containing } w}\right).$$

Hence, $BIDF_w$ captures how novel ("ground-breaking") the word w in patent d was relative to the prior stock of patents that existed.

Hence, $TFIDF_{dw}$, the product of these two, measures the importance of the word w in the given document d. Words that appear infrequently (low TF) or common terms that appear in many document (high $BIDF$) tends to have lower $TFBIDF$. On the other hand, a term that appear intensively in a particular patent document, but do not in most other previous documents could be an indication that the term w was the novel element of the patent.

Finally, we construct the similarity between the patent pair (i, j) in the following way. For both patents, we construct $TFBIDF$ for each term w in patent i

$$TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t},\ where\, t \equiv min(i, j)$$

and put in in a vector V_d of dimension which is the number of unique words in all documents (of which many of them are 0s). After each $TFBIDF$ vector is normalized to have unit length, we calculate the similarity between two vectors (''patents'') as

$$\rho_{i,j} = V_i \cdot V_j.$$

If the two patents $i$and j use the identical set of words with same proportion, $\rho_{i,j}$ will take the similarity value of 1. In contrast, if the two patents do not have any overlapping words, $\rho_{i,j}$ will take the value of 0. We therefore use the measure

$$n_{ij} = 1 - \rho_{ij}$$

as our measure of patent novelty.

E Additional Empirical Results

In this section we report some additional empirical results to complement our empirical analysis in the main text.

E.1 Natives and Immigrants in the Census

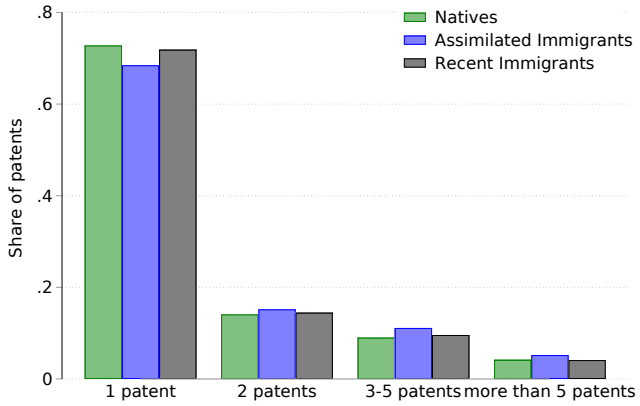
In Table 17 we provide additional evidence on the different demographic make-up between natives and immigrants. The differences are stark. The foreign born population is much more urbanized and hence much less likely to work in agriculture. The age distribution is also strikingly different as children are less likely to immigrate. Finally, the last rows show the changing nationality decomposition of the foreign-born population. While almost a third of the immigrant population was of German descent in 1880, by 1920, Italians have emerged as the dominant nationality of the foreign born.

	1880		1900		1910		1920	
	Natives	Foreign-Borns	Natives	Foreign-Borns	Natives	Foreign-Borns	Natives	Foreign-Borns
Observations	21'877'294	3'629'755	32'857'239	5'729'369	39'735'657	7'876'248	46'122'485	7'738'920
Share urban	0.22	0.47	0.34	0.60	0.40	0.68	0.46	0.73
Sh. <=20	0.56	0.13	0.52	0.11	0.49	0.12	0.47	0.08
Sh. >20, <=40	0.28	0.46	0.29	0.46	0.31	0.48	0.30	0.45
Sh. >40, <=70	0.15	0.39	0.17	0.39	0.18	0.36	0.21	0.42
Share agric.	0.29	0.24	0.24	0.18	0.23	0.13	0.18	0.11
Share manuf.	0.05	0.17	0.05	0.14	0.06	0.18	0.07	0.21
Sh. German		0.30		0.25		0.17		0.11
Sh. Italian		0.01		0.06		0.11		0.12
Sh. British		0.14		0.11		0.09		0.08
Sh. Irish		0.24		0.13		0.08		0.06
Sh. Russian		0.01		0.04		0.11		0.11
Sh. Austrian		0.01		0.03		0.09		0.05

Table 17: Comparison Natives and Foreign-Borns in the Census

E.2 The Intensive Margin of Patenting for Natives and Immigrants

In Section 4 we documented that immigrants, in particular assimilated immigrants, are more likely to engage in patenting than natives. In Figure 19 we show the same pattern for the intensive margin by focusing on the distribution of the number of patents conditional on patenting. Like for the extensive margin, foreign born individuals that lived in the US for more than 10 years are particularly prolific as their distribution dominates the one of natives and recent immigrants in the first-order-stochastic dominance sense.



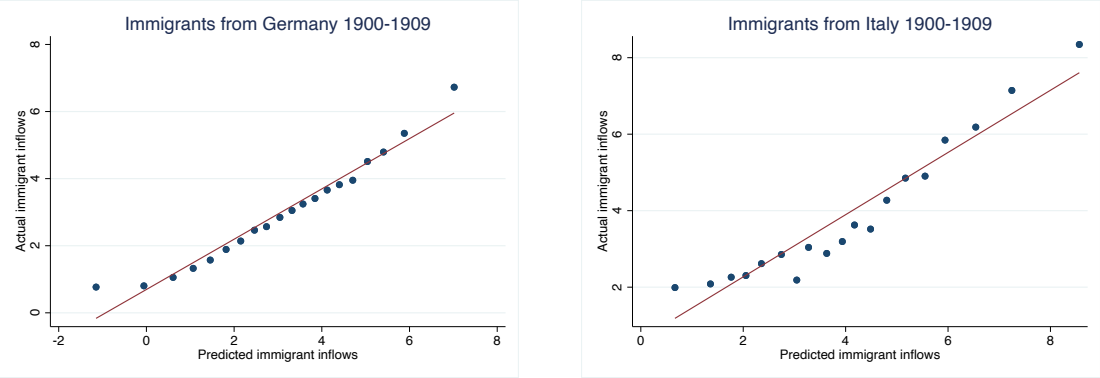
Notes: The figure shows the intensive margin of patenting for natives, assimilated immigrants, that have been in the US for more than 10 years, and recent immigrants, i.e. immigrants that arrived within the last 10 years. The data refers to the year 1910.

Figure 19: Patent Activity by Natives and Immigrants

E.3 Construction of the Instrument for Specification (24)

In this section we describe the construction of our instrumental variable used in the last two columns of Table 9 in more detail. We follow the 'shift-share' literature and construct an instrument for immigrant inflows (Altonji and Card (1991); Card (2001)). The motivation is that immigrants are likely to settle in places where their ancestors were living in the past. We construct predicted immigrant inflows from the time series of actual immigrant inflows from different origin countries, interacted with the distribution of ancestors across space in the previous period. More formally, let I_o^t be the immigrant inflows of origin country o at time t. We can construct our instrument in region r at time t, $Card_r^t$ as

$$IV_{rt} = \sum_o \frac{A_{ort-1}}{A_{ot-1}} I_{ot}, \tag{32}$$



Notes: The figure displays a binscatter plot between the predicted (log) number of immigrants and the actual (log) immigrants from Germany (left panel) and Italy (right panel). We calculate the predicted number of immigrants according to (32).

Figure 20: Predicted versus actual immigrant inflows

where $\frac{A_{or}^{t-1}}{A_{o}^{t-1}}$ is the share of ancestors in region r of an origin country o in the previous period.[20] Intuitively, immigrant inflows from country o at time t are allocated according to the distribution of their ancestors across space at time *t-1*.[21] Figure 20 shows predicted versus actual inflows in year 1910 for Germany and Italy, two of the largest immigrant countries at that time. Actual and predicted inflows are closely aligned along the 45 degree line with the slope for Germany being slightly more precisely estimated than for Italy. We find similar positive and strong relationships for the other origin countries in our analysis.

The first stage of our regression is

$$\frac{I_r^t}{Pop_r^t} = \gamma_1 \cdot \frac{IV_{rt}}{\widehat{Pop_r}^t} + \gamma_2 \cdot Ctrls_r^t + \varepsilon_r^t, \tag{33}$$

where Pop_r^t is the population in region r at time t, $Card_r^t$ is defined as in (32) and $Ctrls_r^t$ are the control variables in region r at time t. Importantly, $\widehat{Pop_r}^t$ is predicted population and defined as

$$\widehat{Pop_r}^t = Pop_r^{t-1} + IV_{rt} \tag{34}$$

We calculate predicted population $\hat{Pop}_r^t$ for the same reason we implement an IV identification strategy in the first place. In particular, an endogenous choice by the immigrants of where to settle inevitably implies that population itself is endogenous. Therefore, an analysis based on the first stage in (33) without accounting for this endogeneity problem would not be valid. Moreover, since in (34) only past population is incorporated, any contemporaneous population related forces will not impact predicted population $\hat{Pop}_r^t$.

[20] Our panel data is unbalanced. To create these ancestry shares, we have to rely on the shares from previous periods, if they don't exist in *t-1*.

[21] Our construction of the instrument abstracts from mortality, which in general could be specific to origin country o or region r.

We estimate the same regression specifications that we used for the OLS regressions in Table 9. Table 18 shows the first stage of this regression as in (33). For all specifications the coefficients remain significantly positive and precisely estimated. The t-statistics (in brackets) suggest that there is no weak IV problem. Moreover, specification (II) and (IV) are estimated at the county-year level, instead of the county-industry-year level. Figure 21 plots the predicted values from column VI of table 18 and shows that our instrument successfully predicts actual census inflow shares.

	Share of immigrants					
	I	II	III	IV	V	VI
Pred. Share of Immigrants$_{rt}$	0.509***	0.435***	0.220***	0.224***	0.220***	0.246***
	(18.27)	(22.37)	(9.19)	(13.13)	(9.13)	(8.02)
(*ln*) *Population*$_{rt}$			0.00398*	0.00184	0.00415**	0.00278
			(1.90)	(1.21)	(1.97)	(1.20)
Urban share$_{rt}$			0.0194***	0.00987**	0.0192***	0.0210***
			(3.44)	(2.23)	(3.41)	(3.20)
(*ln*) *Patents*$_{rit-1}$						-0.000184
						(-1.45)
R^2	0.482	0.511	0.857	0.818	0.858	0.862
N	102093	9594	102010	9389	98728	72555
County FE	no	no	yes	yes	yes	yes
Year FE	no	no	yes	yes	yes	yes
Industry FE	no	no	no	no	yes	yes

Notes: t-statistic in parentheses. The predicted share of immigrants are people who are predicted to have immigrated within the last ten years for a given census year over predicted population. Year 1900-1930. Standard errors clustered by county.

Table 18: First Stage in shares and exog. population

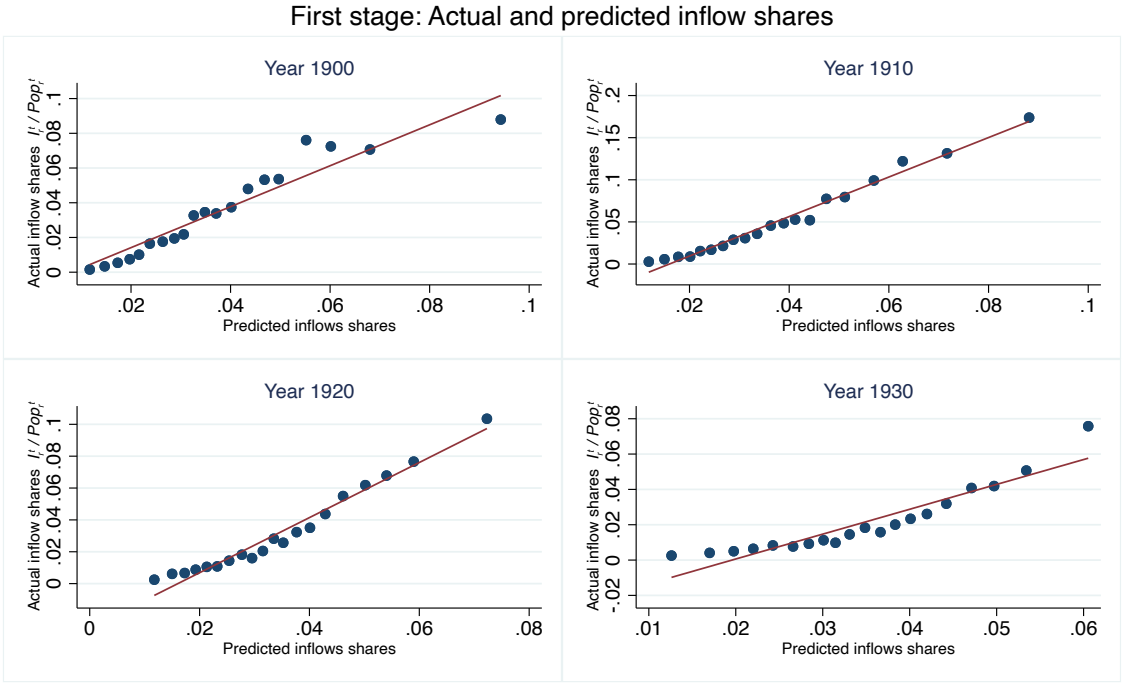
E.3.1 An Alternative IV strategy

The strategy based on (32) guards against the possible critique that migrants might settle in prosperous and innovative locations. However, the main issue with this approach is that not only where migrants settle now, but also where migrants have settled in the distant past, could potentially be correlated with unobserved factors that also affect innovation. This would lead to endogenous ancestry shares, $\frac{A_{or}^{t-1}}{A_{ot-1}}$, and hence an invalid IV.

In recent paper Burchardi et al. (2020) propose a modified version of (32), where not the entire ancestry distribution is used to predict future immigration, but only an exogenous component. More precisely, these propose to use

$$IV_{rt}^{Ex} = \sum_{o} \frac{\hat{A}_{ort-1}}{\hat{A}_{ot-1}} I_{o}^{t}, \tag{35}$$

where the inflow of immigrants from origin country o, I_{ot}, is "distributed" according to the distribution of the exogenous pre-existing ancestry $\hat{A}_{ort-1}$. Such exogenous ancestors, $\hat{A}_{ort-1}$,



Note: Actual immigrant inflow shares are from the census and defined as I_r^t / Pop_r^t. Predicted immigrant inflow shares are defined as the predicted values from the first stage regression in table 14 (column VI).

Figure 21: First stage in (exogenous) shares

evolve according to

$$\widehat{A}_{or}^{t} = \sum_{\tau=1900}^{t} \hat{\beta}^{\tau} \frac{I_r^{\tau}}{I^{\tau}} I_o^{\tau} + \widehat{A}_{or}^{t-1} \tag{36}$$

where $\hat{\beta}^{\tau}$ is the coefficient from the regression

$$A_{or}^{t} = \sum_{\tau=1900}^{t} \beta^{\tau} \frac{I_r^{\tau}}{I^{\tau}} I_o^{\tau} + b_2 \cdot Ctrls_{or}^{t} + \delta_o^t + \delta_r^t + \delta_t + \varepsilon_{or}^t. \tag{37}$$

Hence, we first run the regression in (37) and identify β^{τ} from the relationship between the stock ancestors of origin o in region r and the *aggregate* inflow of immigrants of origin o, the the share of *all* immigrants heading towards region r. Hence, the stock A_{or}^t is the combination of an origin-specific push factor ($I_{o\tau}$) and a region specific pull factor (I_r^{τ}/I^{τ}). Intuitively: (37) leverages the fact that we expect lots of Italians in region r, if during the time when lots of Italians were coming to the US, lots of immigrants (or all origins) were setting line in region r. It is in this sense that A_{or}^t suffers less from the concern that a particular region might attracts particular immigrants, who might share common aspects of comparative advantage. In Figure 22 we show the distribution of exogenous ancestries and actual ancestries and their correlation.

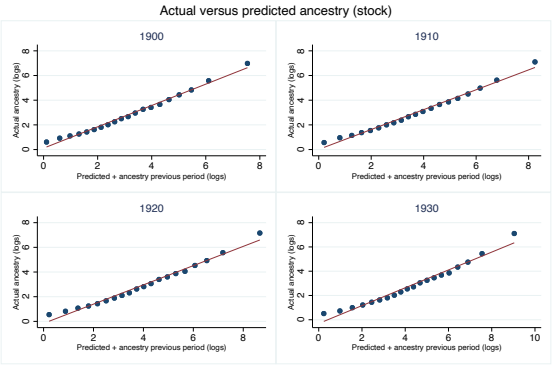


Figure 22: The predictive power of exogenous ancestry