APPENDIX

"IMMIGRATION, INNOVATION, AND THE GEOGRAPHY OF GROWTH" by C. Arkolakis, S. Lee, and M. Peters

(FOR ONLINE PUBLICATION)

A.1. THEORY

A.1.1 Characterization of Equilibrium

In this section we provide the details of the equilibrium characterization.

Solving the Value Function $V_{rt}(N)$

Consider the value function $V_{rt}(N)$ given in (3.2). Conjecture that $V_{rt}(N)$ is homogeneous in N, i.e. $V_{rt}(N) = v_{rt}N$. Equation (3.2) then implies that

$$v_{rt}N = \max_{I}\left\{\frac{N\pi_{rt}}{P_{rt}} + I\frac{(\pi_{rt}-\chi_{rt})}{P_{rt}} + \beta v_{rt+1}(N+I)\right\} \tag{A.1}$$

$$= N\left(\frac{\pi_{rt}}{P_{rt}} + \beta v_{rt+1}\right) + \max_{I}\left\{I\left(\frac{\pi_{rt}-\chi_{rt}}{P_{rt}} + \beta v_{rt+1}\right)\right\}. \tag{A.2}$$

The optimality condition for I, together with the equilibrium condition that $I < \infty$, implies that $\chi_{rt} = \pi_{rt} + P_{rt}\beta v_{rt+1}$. Hence,

$$v_{rt} = \frac{\pi_{rt}}{P_{rt}} + \beta v_{rt+1}, \tag{A.3}$$

as claimed in (2). This also implies that $\chi_{rt} = P_{rt}v_{rt}$.

Dynamic Equilibrium System

Our theory has an intuitive modularized structure:

1. ***The Growth Block:*** Given the distribution of the population, the stock of knowledge N_{rt-1}, the growth rates of real wages g^{w}_{rt+1}, and the growth rates of innovation cutoffs $g^{\bar{h}}_{rt+1}$, we can solve for the equilibrium innovation cutoff $\bar{h}_{rt}$, and the new stock of knowledge N_{rt}.

2. ***The Trade Block:*** Given the population distribution ℓ^{vs}_{rt}, the stock of local varieties N_{rt} and the cutoff $\bar{h}_{rt}$, we can compute equilibrium wages w_{rt}.

3. ***The Spatial Supply Block:*** Given the initial population distribution ℓ^{vs}_{rt-1}, the distribution of wages w_{rt} and the innovation cutoffs $\overline{h}_{rt}$, we can compute the current population distribution.

The Growth Block The equilibrium cutoff $\overline{h}_{rt}$ is given in (6). Now note that

$$\frac{P_{rt}v_{rt+1}}{w_{rt}} = \frac{P_{rt}v_{rt}}{w_{rt}}\frac{v_{rt+1}}{v_{rt}} = \frac{\chi_{rt}}{w_{rt}}\frac{\chi_{rt+1}/P_{rt+1}}{\chi_{rt}/P_{rt}} = \frac{1}{\overline{h}_{rt}}\frac{w_{rt+1}/P_{rt+1}}{w_{rt}/P_{rt}}\frac{\overline{h}_{rt}}{\overline{h}_{rt+1}} \equiv \frac{1+g^{w}_{rt+1}}{\overline{h}_{rt+1}},$$

where g^{w}_{rt+1} denotes the growth rate of *real* wages. Substituting this expression and the law of motion $N_{rt} = N_{rt-1} + \frac{\theta}{\theta-1}\overline{h}_{rt}^{1-\theta}\mathcal{H}^{\theta}_{rt}\ell_{rt}$ into (6) yields an implicit equation for $\overline{h}_{rt}$:

$$1 = \frac{(\sigma-1)}{\frac{\ell_{rt}\overline{h}_{rt}}{N_{rt-1}+\frac{\theta}{\theta-1}\overline{h}_{rt}^{1-\theta}\mathcal{H}^{\theta}_{rt}\ell_{rt}}\left(1-\left(\frac{\mathcal{H}_{rt}}{\overline{h}_{rt}}\right)^{\theta}\right) + \beta(\sigma-1)\frac{1+g^{w}_{rt+1}}{1+g^{\overline{h}}_{rt+1}}}.$$

Given $\overline{h}_{rt}$, we can compute N_{rt} from the law of motion above.

The Trade Block Given $\overline{h}_{rt}$, $\mathcal{H}_{rt}$ and N_{rt}, we can solve for the distribution of wages $\{w_{rt}\}_r$. Total income in region r is given by

$$Y_{rt} = \frac{\sigma}{\sigma-1}w_{rt}\ell_{rt}\left(1-\left(\mathcal{H}_{rt}/\overline{h}_{rt}\right)^{\theta}\right). \tag{A.4}$$

Labor market clearing requires that

$$Y_{rt} = N_{rt}\left(\frac{w_{rt}}{A_r}\right)^{1-\sigma}\sum_{j=1}^{R}\frac{\tau_{rj}^{1-\sigma}}{\sum_{m=1}^{R}N_{mt}\left(\tau_{mj}\frac{w_{mt}}{A_m}\right)^{1-\sigma}}Y_{jt},$$

where τ_{rj} denote the trade costs between r and j. These equations can be solved for the set of wages w_{rt} together with a choice of numeraire.

The Migration Block Given w_{rt} and $\overline{h}_{rt}$, we can compute the migration probabilities according to (9). Given the pre-determined population distribution ℓ^{vs}_{rt-1}, we can then compute the current population distribution ℓ^{vs}_{rt} from (10) and (11).

A.1.2 Characterization of Spatial BGP

In this section, we characterize the spatial BGP in our economy.

Proposition 1. *Suppose that $\vartheta < 1$, let the aggregate population $\ell_t \equiv \sum_r \ell_{rt}$ grow at rate $\eta > 0$, and labor productivity A_{rt} grow at rate g_A. Let the population distribution be stationary, that is $\varpi^{v}_{r} = \ell^{v}_{rt}/\ell_{rt}$ and λ^{v}_{r} is constant. Then:*

1. *The growth rates for knowledge N_{rt}, human capital stock $\mathcal{H}_{rt}$, and wages w_{rt} are given by*

$$1+g_N=(1+\eta)^{\frac{1}{1-\vartheta}} \quad \textit{and} \quad 1+g_{\mathcal{H}}=(1+\eta)^{\frac{\vartheta}{1-\vartheta}} \quad \textit{and} \quad 1+g_w=(1+g_A)(1+\eta)^{\frac{1}{1-\vartheta}\frac{1}{\sigma-1}},$$

2. *The innovation cutoff $\bar{h}_{rt}$, local wages w_{rt}, and the share of innovators e_{rt} are given by*

$$\bar{h}_{rt} = \alpha\zeta_r^{\frac{1}{1-\vartheta}}\ell_{rt}^{\frac{\vartheta}{1-\vartheta}}\left(\sum_\nu \lambda_r^\nu \varpi_r^\nu\right)^{\frac{1}{\theta}\frac{1}{1-\vartheta}} \tag{A.5}$$

$$w_{rt} = \frac{(N_{rt})^{1/\sigma}A_{rt}^{\frac{\sigma-1}{\sigma}}}{((1-e_r)\ell_{rt})^{1/\sigma}}\left(\sum_{j=1}^R \frac{\tau_{rj}^{1-\sigma}w_{jt}(1-e_j)\ell_{jt}}{\sum_{m=1}^R N_{mt}\left(\tau_{mj}w_{mt}/A_{mt}\right)^{1-\sigma}}\right)^{1/\sigma} \tag{A.6}$$

$$e_{rt}=e^* = \left(\frac{\mathcal{H}_{rt}}{\bar{h}_{rt}}\right)^\theta = \frac{1}{1+\frac{\theta}{\theta-1}\frac{1+g_N}{g_N}(\sigma-1)\left(1-\beta\frac{1+g_w}{1+g_{\mathcal{H}}}\right)}. \tag{A.7}$$

where α is an inconsequential constant, which is common across space.

3. *Let $\ell_t^{s|I}=\sum_r \ell_{rt}^{s|I}$ denote the aggregate mass of immigrants with skill s and ℓ_t^s the aggregate population. Then*

$$\frac{\ell_t^{s|I}}{\ell_t^s}=1-\frac{\eta^B}{\eta+d} \quad \textit{and} \quad \ell_t^\nu=\frac{1}{\eta+d}M_t^\nu. \tag{A.8}$$

Moreover, ℓ_{rt}^ν solves the stationary version of equations (10) and (11) evaluated at the BGP prices.

We now prove these results in turn:

The BGP Growth Rate and the Innovation Cutoff Let $e_{rt}=\left(\mathcal{H}_{rt}/\bar{h}_{rt}\right)^\theta$ denote the share of the population that is engaged in idea creations. The equilibrium cutoff is given by

$$\bar{h}_{rt}=\frac{(\sigma-1)}{\frac{\ell_{rt}}{N_{rt}}(1-e_{rt})+\beta(\sigma-1)\frac{v_{rt+1}}{w_{rt}/P_{rt}}}. \tag{A.9}$$

Along a BGP, the number of ideas N_{rt} grows at a common rate g_N. Hence, $N_{rt}=\frac{1+g_N}{g_N}I_{rt}$. Using (4), we get that

$$\frac{\ell_{rt}}{N_{rt}}=\frac{\ell_{rt}}{I_{rt}}\frac{g_N}{1+g_N}=\frac{\theta-1}{\theta}\frac{1}{\bar{h}_{rt}e_{rt}}\frac{g_N}{1+g_N}$$

Substituting this in (A.9) yields

$$1=\frac{(\sigma-1)}{\frac{g_N}{1+g_N}\frac{\theta-1}{\theta}\left(\frac{1-e_{rt}}{e_{rt}}\right)+\beta(\sigma-1)\frac{v_{rt+1}\bar{h}_{rt}}{w_{rt}/P_{rt}}} \tag{A.10}$$

Now note that $v_{rt+1} = \chi_{rt+1}/P_{rt+1} = w_{rt+1}/(P_{rt+1}\overline{h}_{rt+1})$. Hence,

$$\frac{v_{rt+1}\overline{h}_{rt}}{w_{rt}/P_{rt}} = \frac{\chi_{rt+1}\overline{h}_{rt}/P_{rt+1}}{w_{rt}/P_{rt}} = \frac{w_{rt+1}/P_{rt+1}}{w_{rt}/P_{rt}}\frac{\overline{h}_{rt}}{\overline{h}_{rt+1}} = \frac{1+g}{1+g_{\overline{h}}}, \tag{A.11}$$

where g is the rate of real wage growth and $g_{\overline{h}}$ is the growth rate of the human capital cutoff. Both these growth rates are endogenous and we will solve for them below. However, both of them are also constant along a BGP. The innovator share is therefore also constant across space and given by

$$e^{BGP}_{rt} = e^{BGP} = \frac{1}{1 + \frac{\theta}{\theta-1}\frac{1+g_N}{g_N}(\sigma-1)\left(1-\beta\frac{1+g}{1+g_{\overline{h}}}\right)}. \tag{A.12}$$

The cutoff $\overline{h}_{rt}$ is then given by

$$\overline{h}_{rt} = \left(e^{BGP}\right)^{-1/\theta}\mathcal{H}_{rt} = \left(e^{BGP}\right)^{-1/\theta}\zeta_r N^{\vartheta}_{rt-1}\left(\sum_\nu \lambda^\nu_{rt}\omega^\nu_{rt}\right)^{1/\theta}. \tag{A.13}$$

Rearranging terms yields

$$1 = \left(e^{BGP}\right)^{-1/\theta}\zeta_r\left(\frac{N_{rt-1}}{\ell_{rt}\overline{h}_{rt}}\right)^{\vartheta}\frac{\ell^{\vartheta}_{rt}}{\overline{h}^{1-\vartheta}_{rt}}\left(\sum_\nu \lambda^\nu_{rt}\omega^\nu_{rt}\right)^{1/\theta}. \tag{A.14}$$

Because $\frac{N_{rt-1}}{\overline{h}_{rt}\ell_{rt}} = \frac{\theta}{\theta-1}e^{BGP}\frac{N_{rt}}{I_{rt}}$ is constant along the BGP, this implies that $\overline{h}^{\frac{1-\vartheta}{\vartheta}}_{rt} \propto \ell_{rt}$. Hence, as in the semi-endogenous growth model of Jones (1995), the growth in $\overline{h}_r$ is tied to the rate of population growth.

Let the aggregate population growth rate be equal to $\dot{\ell}_t/\ell_t = \eta$. Because $N_{rt-1}/\overline{h}_{rt}\ell_{rt}$ is constant, we have that

$$1 + g_N = \frac{N_{rt}}{N_{rt-1}} = \frac{\overline{h}_{rt+1}\ell_{rt+1}}{\overline{h}_{rt}\ell_{rt}} = \left(\frac{\ell_{rt+1}}{\ell_{rt}}\right)^{1+\frac{\vartheta}{1-\vartheta}} = (1+\eta)^{\frac{1}{1-\vartheta}}. \tag{A.15}$$

The human capital cutoff $\overline{h}_{rt}$ and the human capital supply $\mathcal{H}_{rt}$ therefore grow at the rates

$$1 + g_{\overline{h}} = 1 + g_{\mathcal{H}} = (1+\eta)^{\frac{\vartheta}{1-\vartheta}}.$$

Given (A.15) we can solve for the human capital cutoffs $\overline{h}_{rt}$ as

$$g_N = \frac{\theta}{\theta-1}e^{BGP}\frac{\overline{h}_{rt}\ell_{rt}}{N_{rt-1}} \quad \text{and} \quad \overline{h}^{\theta}_{rt} = \frac{1}{e^{BGP}}\mathcal{H}^{\theta}_{rt} = \frac{1}{e^{BGP}}\zeta^{\theta}_r N^{\vartheta\theta}_{rt-1}\sum_\nu \lambda^\nu_{rt}\omega^\nu_{rt}.$$

Combining these equations and noting that ϖ_{rt}^{ν} and λ_{rt}^{ν} are stationary along a BGP yields

$$\bar{h}_{rt} = \left(e^{BGP}\right)^{\frac{\vartheta-\frac{1}{\theta}}{1-\vartheta}} \left(\frac{\theta}{\theta-1}\frac{1}{g_N}\right)^{\frac{\vartheta}{1-\vartheta}} \zeta_r^{\frac{1}{1-\vartheta}} \ell_{rt}^{\frac{\vartheta}{1-\vartheta}} \left(\sum_{\nu} \lambda_r^{\nu} \varpi_r^{\nu}\right)^{\frac{1}{\theta}\frac{1}{1-\vartheta}}. \tag{A.16}$$

As for wages, taking the price of output in a particular region j as the numeraire, the growth rate of wages (and hence income per capita) is given by

$$1 + g_w = (1 + g_A)(1 + \eta)^{\frac{1}{1-\vartheta}\frac{1}{\sigma-1}}. \tag{A.17}$$

Finally, we can solve for the local distribution of spatial knowledge N_{rt} along a BGP. In particular,

$$N_{rt} = \frac{1+g_N}{g_N} I_{rt} = \frac{1+g_N}{g_N}\frac{\theta}{\theta-1}\bar{h}_{rt} e^{BGP} \ell_{rt} \tag{A.18}$$

Labor Supply Along the BGP Given the BGP distribution of wages, we can use the labor supply module to compute the population distribution. Along the BGP, the aggregate population grows at a constant rate and is fully determined from the mass of international inflows. In particular, (10) and (11) imply that

$$\ell_t^{\nu s} = \frac{1}{\eta + d} M_t^{\nu s} \quad \text{and} \quad \ell_t^{USs} = \frac{\eta^B}{\eta - \eta^B + d}\frac{1}{\eta + d}\sum_{\nu \in \mathrm{I}} M_t^{\nu s}. \tag{A.19}$$

Hence, given $M_t^{\nu s}$, which is exogenous in our theory, we can directly compute the mass of migrants and natives by skill group. The distribution across space then follows from (10) and (11) under the restriction that the distribution is stationary, i.e. $\ell_{rt}^{\nu s} = (1+\eta)\ell_{rt-1}^{\nu s}$.

A.2. Additional Data Details

Our analysis combines four primary data sources: individual-level U.S. population census from 1880 to 1920, the complete universe of U.S. patents since 1790, immigration arrival records from Castle Garden, and passenger departure records from the Hamburg Passenger Lists.

Section A.2.1 describes these datasets. Section A.2.2 explains the linkage procedures, and Sections A.2.3 and A.2.4 evaluate linked-data quality. The harmonization procedure is described in the Online Appendix.

A.2.1 Main Datasets for Analysis

U.S. Population Census *(1880—1920)*

The U.S. Population Census covers the full population and reports nativity, immigration status, and employment. We use the restricted-access version from Ruggles et al. (2020), which contains individual-level identifiers such as names and places of birth. These identifiers allow us to link individuals across census waves (intercensal linkage) and to integrate census records with patent and immigration data.[18]

Historical Immigration Records *(1820–1914)*

We construct our immigration database from two sources: the Castle Garden (CG) Immigration Database, which records *inflows* to the United States, and the Hamburg Passenger Lists (HPL), which record *outflows* from Europe to the United States.

The *Castle Garden Immigration Database (1820–1914)* The CG Immigration Database contains records of individuals arriving in New York City between 1820 and 1914. During this period, most immigrants to the United States entered through the port of New York, making the CG database a comprehensive source for immigration in this era. The database includes approximately 11 million individuals and reports name, age, nationality, year of arrival, and, notably, pre-migration occupation. These records were compiled from the Customs Passenger Lists prepared by ship captains and filed upon arrival. To link CG records to the U.S. Population Census, we use name, age, nationality, and arrival year. Additional details are provided in Section A.2.2.

The *Hamburg Passenger Lists (1850–1914)* The HPL provide records of individuals departing from Hamburg, Germany, to destinations worldwide, including the United States, between 1850 and 1939. The database contains approximately six million individuals and reports name, age, nationality, departure date, final destination, and, importantly, occupation. For consistency with the CG database, we use four variables — name, age, nationality, and year of departure — to link HPL records to the U.S. Population Census, following the procedure described in Section A.2.2.

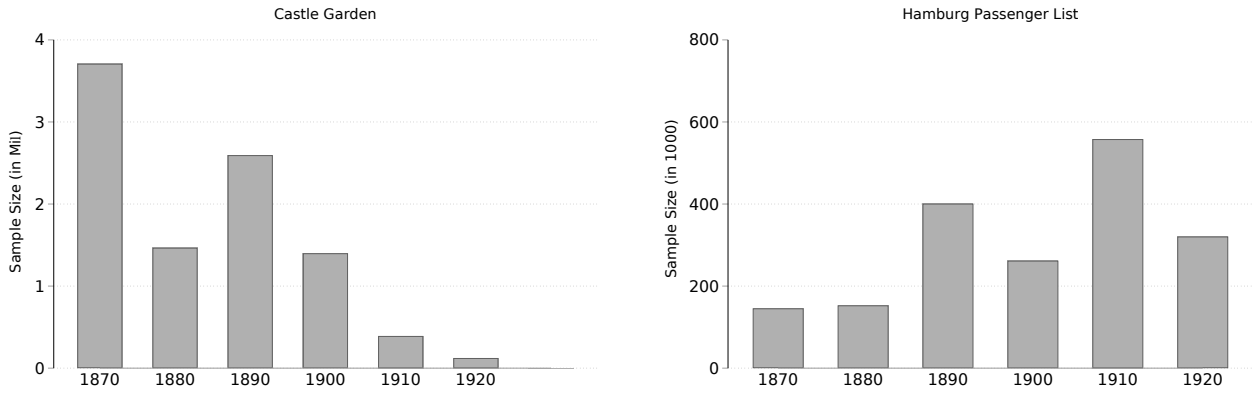
Access to the HPL is granted through a cooperation agreement with the State Archive of Hamburg. Due to privacy restrictions, we use data only through 1914, covering approximately 4.6 million individual records. Of these, approximately 77 percent list the United States as the intended destination. Additional details on data preparation are provided in the Online Appendix.[19]

[18] We additionally link the 1920 and 1930 censuses to study internal migration patterns of natives and immigrants within the United States.

[19] Together, the CG and HPL databases contain approximately 13 million individual records. Our working

In Figure A.1, we present the number of individuals in the CG Database (left panel) and the HPL (right panel) by decade, showing that the two sources provide complementary temporal coverage: the CG database contains more immigrants in earlier decades, whereas the HPL includes more immigrants after 1890.

FIGURE A.1: THE HISTORICAL IMMIGRATION RECORDS



Notes: The figure shows the number of individuals in the Castle Garden (left panel) and Hamburg Passenger Lists (right panel) databases, by decade. Bars report the working sample of male immigrants used for record linkage; for the HPL, this further restricts to records listing the United States as the intended destination. These counts are therefore smaller than the full database totals reported in the text. Records are binned by decade; the final bin (labeled 1920) covers 1910–1914, as both databases end in 1914 with the outbreak of the First World War.

Historical Data on U.S. Patents

To measure innovative activity, we use two primary sources of patent data: Google's Patent Search Engine (https://patents.google.com/) and PatentCity (Bergeaud and Verluise, 2024).

Google Patent Data. Since 1790, the United States Patent and Trademark Office (USPTO) has granted millions of patents, with digitized records available beginning in 1836. In collaboration with Reed Tech, Google applied Optical Character Recognition (OCR) to pre-1975 U.S. patent documents, converting them into machine-readable text. From Google's Patent Search Engine, we extract patent numbers and grant years, inventor and assignee names and locations, invention descriptions, and industry classifications based on the Cooperative Patent Classification (CPC) system.

PatentCity Data. Bergeaud and Verluise (2024) extend this digitization effort by collecting original patent document images, applying OCR, and using Named Entity Recognition (NER) to identify inventors, assignees, and geographic information. PatentCity systematically records patent numbers, grant years, inventor and assignee names, and detailed location information (city, county, and state).

sample is restricted to male immigrants: we exclude women because last-name changes at marriage make record linkage unreliable.

Together, these sources provide complementary strengths: Google offers detailed classification and descriptive textual information, while PatentCity provides enhanced disambiguation of inventors, assignees, and locations.

We construct a unified patent-level dataset by combining patent descriptions and industry classifications from Google with inventor and location information from PatentCity. We harmonize CPC-based industry classifications to the 1950 Census industry classification (IND1950). We then link patent records to the population census using inventor names, locations, and patent grant years. Additional details on the linkage procedure are provided in Section A.2.2.

A.2.2 Record Linking

This section describes (i) linking immigrants from the CG and HPL databases to the U.S. Population Census, (ii) linking patent inventors from U.S. patent records to the census, and (iii) linking census records across years.

Linking with Random Forest Classifiers: Supervised Machine-Learning

We implement a supervised machine-learning approach—random forest classification (henceforth, RFC)—to link records. RFC is a bagging method that trains multiple decision trees on subsamples of the training data. This approach reduces overfitting and improves predictive stability (Lee, 2019).

Before proceeding with RFC, we compared its performance to two commonly used machine-learning record-linking methods: support vector machines (SVM) and XGBoost (XGB). Relative to SVM, RFC produced fewer false positives and a larger number of unique matches. RFC and XGBoost performed similarly, but we select RFC due to its interpretability, scalability, and ease of tuning.

The linking procedure proceeds in two stages: identifying potential matches using RFC and selecting unique matches from this candidate set.

Training the Classifier We train the classifier using existing linked data. Because RFC is supervised, training requires labeled training data. We use hand-labeled and SVM-labeled matches provided by IPUMS, which contain census individuals linked across waves. The training data include a binary match indicator equal to 1 for matches and -1 for non-matches.

From the labeled data, we construct features that allow RFC to distinguish matches from non-matches. To improve model tractability, we apply recursive feature elimination (RFE) to select the final feature set.[20]

[20] RFE iteratively removes the least important variable until a user-specified number of features remains. We measure feature importance using logistic regression coefficients and select the final feature set via grid search to maximize classifier accuracy.

The selected features are: **name similarity**, the Jaro–Winkler similarity for given name and surname;[21] **Double Metaphone match**, an indicator equal to one if records share at least one phonetic code under the Double Metaphone algorithm, evaluated for given name and surname;[22] **NYSIIS surname match**, an indicator equal to one if surnames share the same NYSIIS phonetic code (Rajkovic and Jankovic, 2007; Koneru et al., 2016); **initial match**, an indicator equal to one if the first letters of given and surnames match; **initial only**, an indicator set to true if either record has a single-letter given name; **age difference**, the absolute age difference in years (CG–Census and HPL–Census only); and **percentage age difference**, the age difference divided by mean age (CG–Census and HPL–Census only).

We then tune RFC hyperparameters using k-fold cross-validation (Hastie et al., 2009). The dataset is partitioned into folds, with each fold serving as a validation set while the remaining folds are used for training. The average validation accuracy estimates the out-of-sample predictive performance (Fushiki, 2011). We conduct a grid search over RFC hyperparameter values and select the specification with the highest cross-validated accuracy (Pedregosa et al., 2011).

Identifying Potential Record Matches After training, the classifier generates predicted match probabilities for record pairs across datasets.

The predicted probability equals the mean predicted match probability for all decision trees in the random forest. We retain high-probability pairs as potential matches (See Section A.2.2).

When linking immigration records to the Census, we link all immigration records observed up to the census year. For example, when linking the 1900 Census to CG data, we consider all CG records through 1900. To reduce false positives and computational burden, we restrict comparisons to census records with matching country of birth. When the immigration year is observed, we require that the difference between the immigration year in CG or HPL and the census-reported immigration year is at most two years.[23]

When linking patents to the Census, we match patents within a 15-year window spanning the five years before to the nine years after the census year.[24] We restrict comparisons to

[21] Jaro–Winkler similarity enters the procedure in two ways: as a blocking criterion, candidate pairs with Jaro–Winkler similarity below 0.8 in either given name or surname are excluded prior to classification; and as a continuous feature supplied to the random forest. The 0.8 blocking floor follows prior literature (Jaro, 1989, 1995; Winkler, 2014); Feigenbaum (2016) reports that 95% of IPUMS-linked records exceed 0.8 similarity, and Abramitzky et al. (2021) describe a 0.9 threshold as conservative.

[22] The Double Metaphone is an algorithm that codes English words (and foreign words often heard in the United States) phonetically by reducing them to a combination of 12 consonant sounds, reducing matching problems from incorrect spelling.

[23] The 1900, 1910, and 1920 censuses report immigration year for foreign-born individuals. The 1880 Census does not; for 1880, matches are generated using the remaining identifiers—birthplace, birth year, name agreement, and middle initial—without the immigration-year filter. Because 1880 relies on a weaker constraint set, we place greater weight on the remaining identifiers and verify 1880 match quality using the departure-port validation described below.

[24] For example, we match the 1900 Census to patents from 1895 to 1909. Because the 1890 Census was

literate, English-speaking census individuals residing in the same county as the patent inventor.[25] The classifier then evaluates candidate pairs using the features described above.

Table A.1 summarizes the search constraints for each dataset pair.

Identifying Unique Record Matches After generating potential matches, we apply a pruning procedure to obtain uniquely matched pairs. A candidate pair qualifies as a potential match only if its predicted match probability is at least 0.85, as described in Section 2; the pruning procedure below therefore operates exclusively on pairs that already satisfy this threshold. Let r_A denote a record in dataset A and r_B a record in dataset B. Among potential matches, we retain the match (r_A, r_B) if either: (i) r_A matches only r_B and r_B matches only r_A; or (ii) the predicted match probability for (r_A, r_B) exceeds all alternative matches involving r_A or r_B by at least 15%.

In Section A.2.3, we evaluate the performance of our RFC-based methodology relative to alternative linking approaches.

Linking Immigrants to the U.S. Census

This subsection reports match counts and match rates for the linked immigration datasets.

Linking the CG Immigration Records to the U.S. Census The upper panel of Table OA.1 (reported in the Supplementary Material) reports match counts and match rates by nationality for the CG data linked to the Census. German immigrants account for approximately 50% of CG-Census matches and are slightly overrepresented in the final linked sample. In contrast, match rates for Irish and British immigrants are systematically lower. The table also shows a sharp increase in Italian immigration around the turn of the twentieth century.

Linking the HPL Immigration Records to the U.S. Census The lower panel of Table OA.1 reports match counts and match rates for HPL-Census links by nationality and period. Because Hamburg is a German port city, German immigrants constitute the majority of recorded passengers. Italian and Irish passengers are underrepresented, as they typically departed for the United States from other ports.

Section A.2.3 discusses potential sources of variations in match rates across nationalities and over time and evaluates implications for sample representativeness.

destroyed, we match the 1880 Census to patents from 1875-1899.

[25] Following Abramitzky et al. (2020), restricting to county may exclude individuals who moved between patenting and census enumeration. We therefore also test broader geographic constraints using State Economic Areas (SEAs).

TABLE A.1: SEARCH CONSTRAINTS FOR RECORD LINKING

Datasets Linked	Search Constraints
CG–Census	– Matching birthplace country in CG and census – $\lvert$CG arrival year $-$ census immigration year$\rvert \leq 2$ (applied when census immigration year is observed; see note) – $\lvert$CG birthyear $-$ census birthyear$\rvert \leq 2$ – Matching middle initial (if exists in both records)
HPL–Census	– Matching birthplace country in HPL and census – $\lvert$HPL departure year $-$ census immigration year$\rvert \leq 2$ (applied when census immigration year is observed; see note) – $\lvert$HPL birthyear $-$ census birthyear$\rvert \leq 2$ – Matching middle initial (if exists in both records)
Patent–Census	– Literate and English-speaking US residents only – Matching US residence county in patent and census records – Matching middle initial (if exists in both records)
Census–Census[†]	– Matching birthplace (birth country if foreign-born; birth state if US-born) – $\lvert$census birthyear difference$\rvert \leq 2$ – Close agreement in given name, surname, and parental birthplaces

Notes: This table enumerates the search constraints used to generate the RFC matches. We only perform random forest classification on record pairs that meet the search constraints. When pruning, we apply the rules in the order specified. Before applying each rule, we save the uniquely matched record pairs and ignore them when applying subsequent pruning rules. The immigration-year constraint for CG– and HPL–Census links applies only to censuses that report immigration year (1900, 1910, and 1920); the 1880 Census does not record immigration year, so for 1880 candidate pairs are screened on the remaining constraints—birthplace country, birth year (± 2), name agreement, and middle initial—without the immigration-year filter. [†] For census-to-census linkage, geographic location and occupation are deliberately excluded from the search constraints to avoid mechanically reducing match rates for individuals who migrate or change occupation or industry.

Linking Patents to the U.S. Census

This subsection reports match counts and rates between patent records and Census individuals and describes the temporal matching strategy.

Table A.2 reports the number of matches between patents and census individuals by year. The first row reports matched patent–inventor records, while subsequent rows report the number of distinct matched inventors. The mapping between patents and inventors is many-to-many: a single patent may list multiple inventors, each matched separately to the Census, and a single inventor may hold many patents. For this reason, the patent–inventor record count differs from the count of unique matched patents reported in Table 1, and the two need not coincide in any given year.

Figure OA.5, reported in the Supplementary Material, presents the distribution of patent grant years among matched inventor records. Patents are matched within a 15-year

TABLE A.2: MATCH COUNTS, PATENTS TO CENSUS

	1880	1900	1910	1920
Patent Match Count	115,100	94,867	120,298	120,652
Inventor Count: Total	57,027	53,349	68,561	67,852
Inventor Count: Native	44,145	42,296	55,093	54,167
Inventor Count: Foreign	12,882	11,053	13,468	13,685

Notes: The first row reports the number of patent–inventor match records linking patents to Census individuals; because a single patent may list multiple inventors, each matched separately to the Census, this count exceeds the number of unique matched patents reported in Table 1. Subsequent rows report the number of distinct Census individuals (inventors) matched to at least one patent; a single individual may hold multiple patents.

window spanning the five years before to the nine years after each census year; the 1880 window is extended through 1899 because the 1890 Census was destroyed. Shaded regions indicate patent years that fall within the windows of two adjacent censuses and are therefore eligible for matching to either (1895–1899, 1905–1909, and 1915–1919). When an inventor is matched in both censuses, we retain the match associated with the census year closest to the patent grant year. This rule minimizes temporal misalignment between inventive activity and observed residence and improves internal consistency, as shown in Section A.2.3.

Linking Censuses Across Years

We match male individuals across the decennial U.S. Population Censuses between 1880 and 1920. The linkage criteria require (i) identical birthplaces (i.e., birth country for foreign-born individuals; birth state for U.S.-born individuals), (ii) age differences of at most two years, and (iii) close agreement in names and parental birthplaces. We deliberately exclude geographic location (e.g., county or state of residence) and occupation from the linking criteria to avoid mechanically reducing match rates for individuals who migrate or change occupation/industry.

Other Approaches to Linking Historical Records

We compare our supervised machine-learning approach (RFC-based) to commonly used linkage procedures in the literature. Section A.2.4 compares match datasets generated by alternative approaches. Replication is facilitated by publicly available documentation and code.[26] We apply each method to the same immigration datasets and the U.S. population

[26]See https://ranabr.people.stanford.edu/historical-record-linking for code and descriptions of commonly used historical linkage procedures.

census.[27]

Abramitzky, Boustan, and Eriksson (ABE): Iterative Linking

We implement the iterative procedure of Abramitzky et al. (2012, 2014). For datasets A and B, we (i) restrict to records in A unique on the matching key (given name, surname, age, birthplace) within a five-year band (± 2 years); (ii) retain only one-to-one candidates in B, discarding multiple-candidate cases; (iii) repeat allowing age differences of one then two years for unmatched records; and (iv) run the reverse pass ($B \to A$) and keep the intersection. We use both the phonetic (NYSIIS) and Jaro–Winkler variants, denoted *ABE-NYSIIS* and *ABE-JW*; the JW variant matches on first and last initials and birth years within a five-year band and retains pairs with JW distance < 0.1.

Expectation Maximization (EM)

We also implement the unsupervised EM procedure of Abramitzky, Mill, and Pérez (2020), which needs no hand-linked training data. For datasets A and B, we (i) block candidates in B on birthplace, first and last initials, and age differences ≤ 2 years; (ii) compute Jaro–Winkler distances for given names and surnames; (iii) estimate the posterior probability of a true match via EM; and (iv) retain pairs with match probability $\geq p_m$ whose next-best candidate falls below L, setting $p_m = 0.5$ and $L = 0.45$.

A.2.3 Quality of Matches

While the previous section compares algorithm outputs, this section evaluates the accuracy of our RFC-based linking procedure using independent validation information, assessing whether it introduces systematic bias or disproportionately represents particular groups relative to alternative methods.

Match Accuracy

We evaluate RFC linking accuracy along two dimensions. First, we use departure-port information in the immigration datasets to assess immigration-census linkage quality. Second, we use city-of-residence information in the patent data to evaluate patent-census linkage accuracy.

Immigration–Census Accuracy: Departure-Port Validation The Castle Garden (CG) dataset records the departure port of all arriving passengers. By construction, all Hamburg

[27] Inventors frequently file multiple patents, which implies a many-to-one structure for patent-census linkage. Many existing linkage procedures (including iterative linking and EM) are designed for one-to-one matches. We therefore restrict method comparisons to immigration-census linkage to ensure comparability across approaches.

Passenger Lists (HPL) records originate from Hamburg. Therefore, for census individuals matched to both CG and HPL, the corresponding CG record should list Hamburg as the departure port. While the ground truth of match status of individuals is unobserved, this implication provides a natural validation criterion.

We construct three validation sets of HPL-census matches: (i) matches generated solely by RFC; (ii) unique matches identified by at least two independent linking methods; and (iii) unique matches identified by at least three independent linking methods.

Treating individuals in each validation set as true Hamburg-origin individuals, we define precision as the share of CG matches whose recorded departure port includes Hamburg (either directly or via transit). A classifier with perfect precision would assign Hamburg as the departure port for all CG-HPL matched individuals.

Table A.3 reports the five most common departure ports for CG-HPL-matched individuals across the three validation sets. Across specifications, between 77 and 85 percent of matched individuals list a Hamburg-related departure port — Hamburg directly (roughly 40–43%), Hamburg–Havre (31–35%), or Hamburg–Southampton (7–8%) — consistent with our precision definition. The remaining matches list non-Hamburg ports such as Bremen (7%) and Liverpool–Queenstown (2%), which likely reflect a small share of transit passengers or false matches. Taken together, these results indicate high within-class accuracy across all three validation sets.

TABLE A.3: DEPARTURE PORT VALIDATION

	% of HPL-Matched Individuals		
CG Departure Port	RFC only	≥ 2 methods	≥ 3 methods
Hamburg-related ports			
Hamburg	40.0	39.9	43.3
Hamburg–Havre	30.7	30.9	34.5
Hamburg–Southampton	6.9	6.9	7.5
Hamburg-related total	**77.6**	**77.7**	**85.3**
Other ports			
Bremen	7.1	7.3	3.1
Liverpool–Queenstown	2.7	2.3	2.1

Notes: The above table reports the frequency of the top five departure ports for HPL-matched individuals and matched with RFC to CG records.

In Table A.4, we also compare the precision of each linking method, measured by the consistency of departure-port information. We report two RFC specifications: the baseline, and a more conservative variant that imposes the ABE uniqueness restriction, limiting matches to CG and Census records with unique names within a five-year (± 2) band. On

the consensus validation sets—matches agreed upon by at least two, or at least three, independent methods—RFC attains precision comparable to the alternative procedures (roughly 79–80% for the two-method overlap and 87% for the three-method overlap) while generating substantially more matches overall. Importantly, RFC's non-consensus matches achieve comparable precision: as shown in the "RFC only" column of Table A.3, matches identified by RFC alone exhibit a Hamburg-related departure-port share of 77.6%, essentially identical to the 77.7% share for matches agreed upon by at least two methods. This indicates that RFC's larger match volume does not come at the cost of lower precision on the additional matches it recovers.

TABLE A.4: COMPARISON OF LINKING METHOD PRECISION

	Match Volume		Precision (% Hamburg port)	
Linking Method	Total	Rel. to RFC	≥ 2 methods	≥ 3 methods
RFC	149,590	1.00	79.2	86.8
RFC (unique, ± 2 yrs)	120,726	0.81	80.3	87.7
ABE (NYSIIS)	70,108	0.47	79.9	87.3
EM	53,160	0.36	80.4	86.1
ABE (JW)	28,806	0.19	80.9	86.6

Notes: This table compares Hamburg departure frequency for each linking method. In this table, we construct the RFC dataset with and without the name uniqueness constraints imposed by ABE, where the procedure only matches CG and Census records with unique names in a five year band. We apply immigration year filters for all methods.

Patent–Census Accuracy: City-of-Residence Validation To assess the quality of patent-census matches, we compare inventor residence information reported in the census to residence information extracted from patent texts. As described in Section A.2.2, the linkage procedure restricts candidate matches to individuals residing in the same U.S. county in both patent and census records. Sub-county geographic information (e.g., city or town) is not used in the matching algorithm. For example, if a patent lists the inventor as residing in ‘‘`Ann Arbor, in the county of Washtenaw, State of Michigan`’’, the search space is restricted to census individuals in Washtenaw County, Michigan. We define a *city match* as a case in which the sub-county location (e.g., "Ann Arbor") coincides in both the patent text and the census record, even though city information was not used in the matching procedure. Figure OA.6, in the Supplementary Material, reports city match rates as a function of the difference between patent publication year and the census year. City match rates are highest when the patent publication year is closest to the census year, consistent with within-county mobility over time. Nevertheless, match rates remain high across specifications. For the 1880 Census, city match rates exceed 75 percent across year gaps, and for later census waves, they exceed 90 percent when patents are published near the

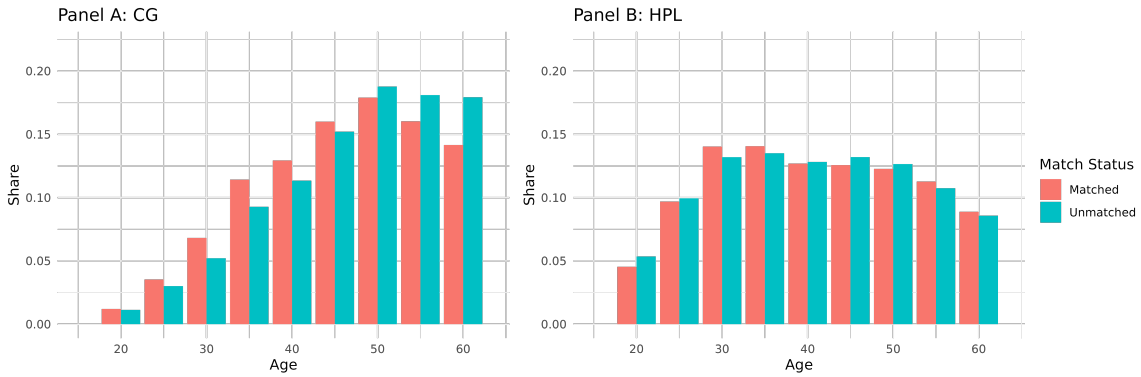
census year.

Representativeness

Beyond match accuracy, we assess representativeness by comparing observable characteristics of matched and unmatched records along two dimensions: the age distribution of immigration-census matches and the patent classification distribution of patent-census matches. Two patterns emerge: for Castle Garden, matched immigrants are somewhat younger on average than unmatched immigrants, while for the Hamburg Passenger Lists the matched and unmatched age distributions are very similar; matched patent records preserve the distribution of classifications observed in the full dataset. Taken together, these patterns indicate that the linked samples retain the central demographic and technological structure of the original populations.

Immigration–Census Representativeness: Age Distributions We begin by comparing matched and unmatched individuals in the immigration datasets. Figure A.2 reports age distributions for matched and unmatched male immigrants aged 16-60 linked to the 1910 Census. For Castle Garden (CG), matched individuals are younger on average than unmatched individuals. For the Hamburg Passenger Lists (HPL), the matched and unmatched age distributions are very similar, with no appreciable difference in mean age. The CG pattern is consistent with life-cycle bias and age misreporting among older individuals, which can reduce match probabilities; for HPL, matched and unmatched individuals are comparable in age.[28]

FIGURE A.2: AGE DISTRIBUTIONS FOR IMMIGRATION RECORD MATCHES



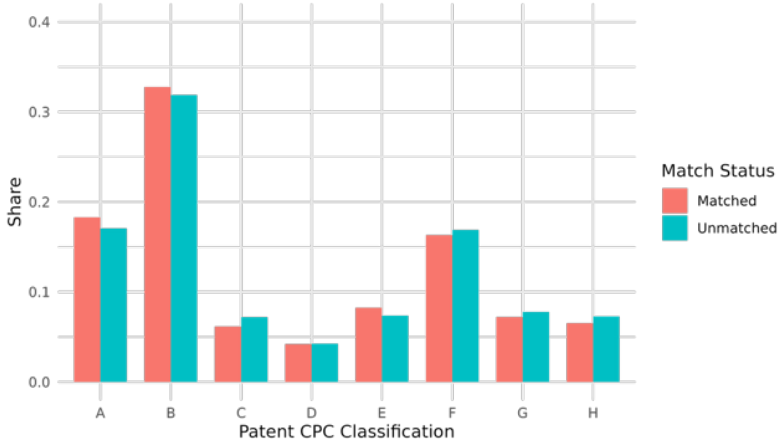
Notes: The above figures show the age distribution across matched and unmatched observations from the CG Data (left panel) and HPL (right panel) against 1910 Census. The matched group includes males aged between 16 and 60 years who are uniquely matched with 1910 Census. The unmatched group includes males aged between 16 and 60 years who are unmatched with 1910 Census.

[28] Age misreporting among older individuals is well documented in the demographic literature and can reduce linkage accuracy.

Patent–Census Representativeness: Patent-Classification Distributions We next examine whether the RFC linkage alters the distribution of patent classifications. Figure A.3 compares the distribution of patent classes in the full patent dataset to the distribution among matched patents.

The distributions are highly similar, indicating that RFC-based linkage does not disproportionately select particular technological categories. This suggests that the matching procedure preserves the composition of patent classifications.

FIGURE A.3: PATENT CLASSIFICATION DISTRIBUTIONS FOR PATENT MATCHES



Notes: This figure shows the distribution of patent classifications across matched and unmatched patent observations.

A.2.4 Comparing Match Datasets

Having described the linking procedures and alternative algorithms, we next compare the datasets generated by each method to assess robustness across specifications.

We compare the outputs of all linkage methods described in Section A.2.2 using CG-1910 Census matches.[29] For all methods, we impose identical immigration-year filters (an absolute difference of at most two years between immigration records and census reports). Table A.5 reports match counts by method and birthplace (four major immigrant-sending countries). RFC generates substantially more matches than alternative procedures.

Taken together, these comparisons indicate that RFC increases match volume without materially altering observable age or birthplace distributions: as the right panel of Figure A.4 shows, the share of matched CG records by place of birth under RFC closely tracks both the alternative algorithms and the full CG cross-section.

[29] ABE and EM produce one-to-one matches; therefore, comparisons are restricted to immigration-census linkage rather than patent-census linkage.

TABLE A.5: CG-CENSUS MATCH COUNTS BY LINKING METHOD

Birthplace	RFC	ABE (NYSIIS)	EM	ABE (JW)	RFC vs. best alt.
All matches	187,490	69,738	43,544	31,243	2.7×
British	26,682	7,676	3,168	7,249	3.5×
German	102,595	39,356	22,153	15,235	2.6×
Irish	21,377	9,869	6,750	4,134	2.2×
Italian	22,083	4,188	1,578	1,354	5.3×

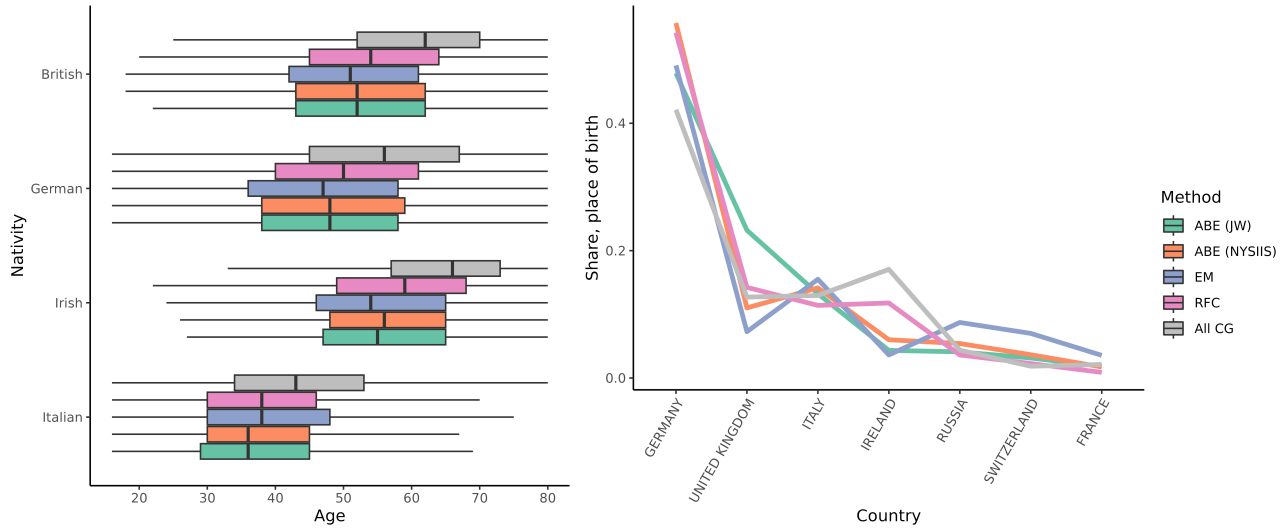
Notes: The above table shows match counts for the 1910 CG and Census, categorized by linking method and by birthplace (four major immigrant-sending countries during our study period). RFC-based linking produces a significantly larger number of unique matches than other methods, and the distribution of birthplaces falls within those of ABE and EM.

Distribution Comparisons

We next compare age and birthplace distributions across methods. The left panel of Figure A.4 plots age distributions for matches generated by ABE (JW and NYSIIS variants), EM, and RFC, along with the full 1910 CG cross-section. The RFC age distribution closely mirrors those produced by alternative methods. The original cross-sectional CG age distribution skews older, reflecting the mechanical absence of deceased immigrants in census matches.

The right panel reports birthplace distributions for each linking procedure. RFC preserves nationality shares that are similar to those generated by ABE and EM, and closely aligned with the original CG distribution.

FIGURE A.4: Age and Birthplace Distributions by Linking Method



Notes: The left panel (boxplots) shows age distributions for CG records linked by ABE-JW, ABE-NYSIIS, EM, and RFC, as well as the age distribution for all CG records. The right panel reports the share of matched records by place of birth for each linking method, alongside the distribution for all CG cross-sectional records.

Match Overlap

Figure OA.9 in the Supplementary Material illustrates the overlap of matched record pairs across methods. Each procedure identifies a substantial number of unique matches. Approximately 33 percent of RFC matches overlap with at least one alternative method. After imposing immigration-year constraints, RFC produces a larger number of unique matches without reducing accuracy, as measured by departure-port validation.

A.3. Quantitative Analysis

In this section we provide more details for our quantitative analysis. Section A.3.1 contains details for our estimation strategy. Section A.3.2 contains additional results for our counterfactual analysis of immigration restrictions. Finally, in Section OA.4 we discuss the robustness analysis reported in Section 6 in more detail.

A.3.1 Estimation Details

Matched Immigration Records As described in Section 4.1, we match immigration records from Castle Garden (CG) and the Hamburg Passenger Lists (HPL) to the US Population Census. In Table A.6 we report the number of matches and the size of the respective immigration record data. We are able to match between 85,000 and 150,000 CG records to the Census and between 25,000 and 50,000 records from the HPL.

Table A.6: Immigration Records: 1880–1920

	1880	1900	1910	1920
Number of records in CG	1105819	996906	226073	18578
Number of CG records matched to Census	136581	152276	86939	84973
Number of records in HPL	112146	205996	490550	286296
Number of HPL records matched to Census	25095	50434	35076	49034
Number of matches in both CG and HPL	7868	14922	7752	6984

Notes: The table reports the number of immigration records in the Castle Garden (CG) and Hamburg Passenger Lists (HPL) databases for each census year, along with the corresponding number of records successfully linked to the US Population Census. The final row reports the number of immigrants who are matched to the Census in both CG and HPL.

Aggregate Growth We measure the growth rate of real aggregate GDP per capita in our model using the Fisher chained index. The Fisher Index is defined by

$$g_t^F = \sqrt{\frac{P_{t-1}Y_t/L_t}{P_{t-1}Y_{t-1}/L_{t-1}} \times \frac{P_tY_t/L_t}{P_tY_{t-1}/L_{t-1}}},$$

where P_t is the price of the consumption good at time t, Y_t is the total quantity of production, and L_t is the total population.

In the context of our model with R regions $r = 1, ..., R$, we construct the following auxiliary indices of output per capita, evaluated at the respective prices:

$$Y_{t-1}(P_{t-1}) = \frac{\sum_{r=1}^{R} P_{rt-1} Y_{rt-1}}{\sum_{r=1}^{R} L_{rt-1}} \quad \text{and} \quad Y_{t-1}(P_t) = \frac{\sum_{r=1}^{R} P_{rt} Y_{rt-1}}{\sum_{r=1}^{R} L_{rt-1}}.$$

and

$$Y_t(P_{t-1}) = \frac{\sum_{r=1}^{R} P_{rt-1} Y_{rt}}{\sum_{r=1}^{R} L_{rt}} \quad \text{and} \quad Y_t(P_t) = \frac{\sum_{r=1}^{R} P_{rt} Y_{rt}}{\sum_{r=1}^{R} L_{rt}}.$$

where P_{rt} denotes the price of total output in region r at time t, and Y_{rt} denotes total physical output in region r at time t.

We then combine these expressions into the corresponding Fisher Index as follows:

$$g_t^F = \sqrt{\frac{Y_t(P_{t-1})}{Y_{t-1}(P_{t-1})} \times \frac{Y_t(P_t)}{Y_{t-1}(P_t)}}.$$

Total real GDP per capita at time t relative to 1880 is then simply given by

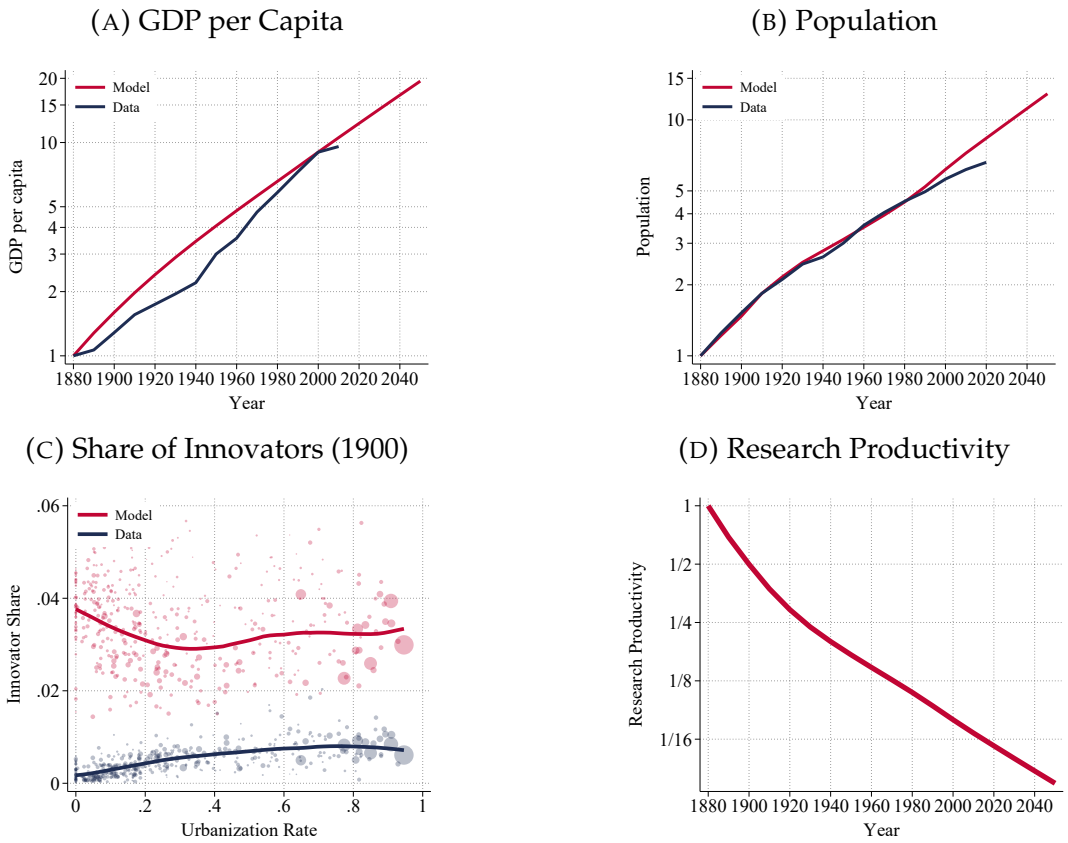
$$y_t / y_{1880} = \prod_j g_j^F \tag{A.20}$$

In Panel A of Figure A.5, we plot GDP per capita as observed in the data (blue line) and in our model as computed according to (A.20).

Estimation of θ In Figure 2 we displayed the regression in (14) graphically. In Table A.7 we report the actual regression results. In the first 5 columns, we (as suggested by our theory) only control for region-year fixed effects, where a region is defined as an "SEA-urban cell". In the different specifications, we impose different thresholds for the maximum number of patents we consider. In all specifications, we estimate θ to be around 1.3. In the last two columns, we explicitly control for a full set of region-type-year interactions, where a type is a nationality-skill combination. Hence, this specification includes fixed effects for say skilled British immigrants living in an urban location in a particular SEA in a given decade. Doing so does not affect our estimate of the slope coefficient.

Estimation of Demographics: η_b and d We discipline the birth rate η_b and the death rate d from the rate of population growth and the share of immigrants between 1880 and 2000 given the observed inflow of immigrants (M_t). Our primary source for immigrant inflows is census microdata, where we identify decadal arrivals using the information on year of immigration. Since this information is not contained in the 1940, 1950, and 1960 censuses, we fill these gaps using administrative records on legal immigrant arrivals from the Department of Homeland Security (DHS). The DHS records differ conceptually from

FIGURE A.5: MODEL FIT: ADDITIONAL IMPLICATIONS



Notes: In this figure we display GDP per capita (Panel A), the overall population (Panel B), the correlation between the urbanization rate (as measured in the data) and the share of innovators in 1900 (Panel C), and aggregate research productivity (Panel D).

TABLE A.7: ESTIMATION OF PARETO TAIL: θ

	(1)	(2)	(3)	(4)	(5)
log Number of patents	-1.305***	-1.306***	-1.309***	-1.317***	-1.356***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Region-Year FE	yes	yes	yes	yes	yes
Max num of patents	-	200	100	50	20
R^2	.785	.785	.785	.784	.778
N	40,422	40,418	40,407	40,320	39,593

Notes: The table reports estimates of equation (14). In columns 1-5 we control for region-year FE and consider different cutoffs for the number of patents in each cell. In column 6 and 7 we control for a full set of region-year-nationality interactions.

the census data because they focus on legal immigration. We thus assume that the total number of immigrants reported in the DHS data (M_t^{DHS}) is related to the immigrant inflow in the Census and our model (M_t) according to $M_t^{DHS} = \kappa^{DHS} M_t$. We then calibrate the three numbers $(\eta_b, d, \kappa^{DHS})$ to jointly minimize a loss function with two components: the squared deviation of model-implied total population growth from the data (1880–2000), and the mean squared error between model-implied and observed immigrant shares across all available years. As seen in Figure 6, we fit the aggregate share of immigrants well. As seen in Panel B of Figure A.5, our model also replicates the overall rate of population growth well.

Spatial Sorting of Recent Arrivals We estimate equation (19) via pooled regression and report the results in Table A.8. Across all specifications, population and nationality share enter with positive and significant coefficients, with population elasticities ranging from 0.519 to 0.734 and nationality share coefficients ranging from 0.346 to 0.530, indicating that immigrants systematically sort toward larger destinations and destinations with a bigger population share of their own nationality. The interaction between urban status and immigrants' skill is also positive and statistically significant throughout, suggesting that skilled immigrants disproportionately select into urban areas upon arrival. We take the estimates from Column (5), which absorbs the most comprehensive set of fixed effects via nationality × decade × skill interactions, as our preferred specification and use these coefficients to compute immigrants' arrival according to (18).

Estimation of Relative Skill Supply In Figure 3 we have shown the relative skill supplies of immigrants of different nationalities and of the native population. As explained in the main text, these results are computed from $\lambda_{rt}^{\nu} = \kappa^{\nu} \lambda_{rt}^{US}$, where κ^{ν} is estimated from (21), which we for convenience replicate here:

$$\ln e_{rt}^{\nu} = \ln\left(\lambda_{rt}^{\nu}\left(\frac{\psi_{rt}}{\bar{h}_{rt}}\right)^{\theta}\right) = \ln\kappa^{\nu} + \ln\left(\lambda_{rt}^{US}\left(\frac{\psi_{rt}}{\bar{h}_{rt}}\right)^{\theta}\right) = \ln\kappa^{\nu} + \delta_{rt}. \tag{A.21}$$

In Table A.9 we report the results of estimating this regression. Column 1 contains our baseline estimation, where we control for (as implied by our theory) a year-region fixed effect δ_{rt}. In columns 2 and 3 we only consider region-year-nationality cells with at least 200 or 500 observations. Finally, in column 4 we use the entire sample by adding a small positive number to e_{rt}^{ν} in case $e_{rt}^{\nu} = 0$. Table A.9 shows that the estimates of κ^{ν} are insensitive to these choices.

Time-invariant fundamentals As explained in the main text, our calibration strategy delivers estimates of local fundamentals $\{\zeta_{rt}, A_{rt}, \mathcal{B}_{rt}\}$ for the decades 1880, 1900, 1910, and 1920. Because such fundamentals are time-invariant in our theory, we estimate a region fixed effect and then use this estimate in our calibrated model. For example, we estimate the regression $\ln\zeta_{rt} = \delta_r + u_{rt}$ and then use $\zeta_r = \exp\left(\widehat{\delta_r}\right)$. We proceed in the

TABLE A.8: POOLED REGRESSION FOR ARRIVALS

	(1)	(2)	(3)	(4)	(5)
Urban	-0.177	0.187	0.217*	0.160	0.228*
	(0.111)	(0.114)	(0.121)	(0.117)	(0.119)
Urban × High Skill	0.355***	0.303***	0.347***	0.362***	0.219*
	(0.111)	(0.110)	(0.110)	(0.114)	(0.122)
Population	0.519***	0.636***	0.689***	0.733***	0.734***
	(0.0547)	(0.0412)	(0.0450)	(0.0436)	(0.0439)
Nationality Share	0.346***	0.410***	0.456***	0.527***	0.530***
	(0.0186)	(0.0183)	(0.0172)	(0.0191)	(0.0192)
Skill FE	Yes	Yes	Yes	No	No
Decade FE	No	Yes	Yes	No	No
Nationality FE	No	No	Yes	No	No
Nat× Decade FE	No	No	No	Yes	No
Nat× Skill FE	No	No	No	Yes	No
Nat× Decade× Skill FE	No	No	No	No	Yes
N	3,782	3,782	3,782	3,782	3,781
R^2	.287	.454	.543	.647	.654

Notes: This table reports the estimates of equation (19) via pooled regression. Column (1) includes Skill fixed effects. Columns (2) and (3) additionally control for Decade fixed effects. Column (3) further includes Nationality fixed effects. Column (4) replaces these with Nationality × Decade and Nationality × Skill fixed effects. Column (5) absorbs the most comprehensive set of fixed effects via Nationality × Decade × Skill interactions. The preferred specification is Column (5), whose coefficients serve as baseline parameters for model prediction. Standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE A.9: ESTIMATION OF RELATIVE SKILL SUPPLY: κ^{ν}

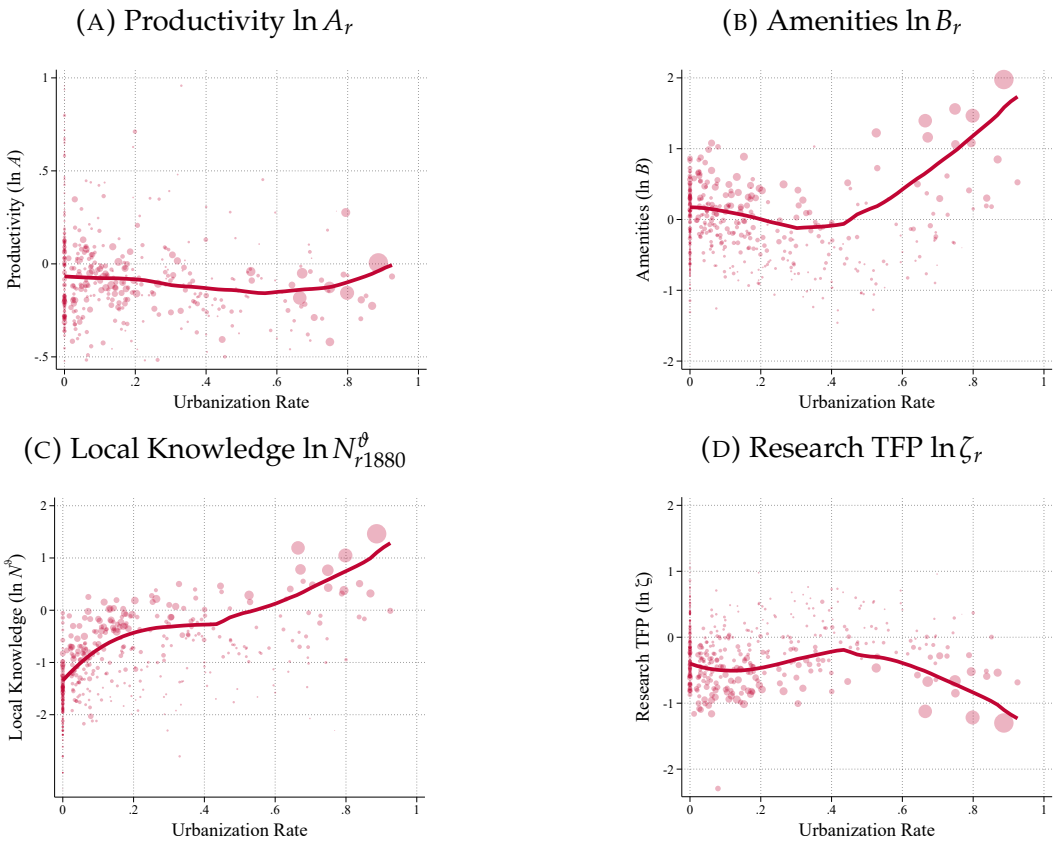
	(1)	(2)	(3)	(4)
British	0.279***	0.259***	0.222***	0.228***
	(0.025)	(0.025)	(0.026)	(0.024)
German	-0.269***	-0.277***	-0.294***	-0.271***
	(0.057)	(0.058)	(0.059)	(0.056)
Irish	-1.239***	-1.260***	-1.309***	-1.220***
	(0.093)	(0.091)	(0.087)	(0.088)
Italian	-1.472***	-1.482***	-1.506***	-1.409***
	(0.055)	(0.056)	(0.059)	(0.048)
Others	-0.524***	-0.530***	-0.541***	-0.521***
	(0.035)	(0.035)	(0.036)	(0.034)
Year-SEA FE	yes	yes	yes	yes
Cell size cutoff	-	200	500	-
Impute if zero?	no	no	no	yes
R^2	.887	.866	.854	.955
N	6,074	5,363	4,538	11,008

Notes: The table reports the results of estimating (21). Columns 1 uses the entire sample where $e^{\nu}_{rt} > 0$. In columns 2 and 3 we require all cells to have at least 200 or 500 observations respectively. In column 4, we add a small positive number to e^{ν}_{rt} in case $e^{\nu}_{rt} = 0$. All regressions are weighted by the number of observations in each cell.

same way for A_{rt} and $\mathcal{B}_{rt}$.

In Figure 5 we already plotted the resulting fundamentals for different deciles of urbanization. In Figure A.6 we plot the raw data. As also discussed around Figure 5, local productivity does not vary much with a region's urbanization rate, amenities are increasing at the very top, there is a steep urban gradient of local knowledge, and research TFP is relatively flat, except at the very top where the very large knowledge stock dictates a relatively low level of research TFP to rationalize the estimated innovation human capital ψ_r.

FIGURE A.6: LOCAL FUNDAMENTALS AND THE URBANIZATION RATE



Notes: Each panel plots a local fundamental against the urbanization rate in 1880: location productivity ($\ln A_r$), the exogenous component of amenities ($\ln B_r$), the knowledge component of local human capital ($\ln N^{\theta}_{r1880}$), and research TFP ($\ln \zeta_r$). Each dot is a State Economic Area (SEA) and the size of the dots reflects the SEA's population, overlaid with a local-linear fit.

Additional Results Figure A.5 also contains two additional implications of our estimated model. In Panel C we report the implied share of innovators as a function of the urbanization rate. As implied by Figure 6, our model predicts a higher share of innovators, reflecting the fact that not all innovation is patented. In terms of the cross-sectional predictions, our model predicts a non-monotone relationship. This is in contrast to the empirically observed pattern, which is increasing. Intuitively, our model implies that the return to innovation in very rural areas is quite high, given how little local knowledge (N_{rt}) these locations have accumulated. Between an urbanization rate of around 20%

and 100%, our model replicates the empirical relationship between the urbanization rate and the innovator share well. In Panel D we report our model's implication for research productivity, which (following Bloom et al. 2020) we define as idea growth per researcher, $g_{Nt}/\ell_{rt}e_{rt}$. Our model implies that

$$\frac{g_{Nt}}{\ell_{rt}e_{rt}} = \frac{\theta}{\theta-1}\frac{\zeta_r\left(\sum_\nu \lambda_{rt}^\nu \varpi_{rt}^\nu\right)^{1/\theta}}{N_{rt-1}^{1-\vartheta}e_{rt}^{\frac{1}{\theta}}}.$$

Research productivity is declining in the stock of knowledge N_{rt} (given that $\vartheta < 1$), reflecting the traditional logic of ideas getting harder to find. It is also decreasing in the share of innovators, because (given the aggregate skill supply and research TFP ζ_r) an increase in the share of innovators cuts deeper in the talent pool and reduces the average talent of researchers.

A.3.2 Quantification: Immigration and US Growth

A.3.2.1 Counterfactual I: The Era of Mass Migration: 1880-1920

This section contains additional results for our first quantitative exercise discussed in Section 5.1. Figure A.7 complements Figure 7 in the main text in that it shows the counterfactual path of the aggregate population (Panel A) and the aggregate flow of patents (Panel B), relative to the baseline for the three counterfactual migration paths considered in the main text. In the absence of international migration between 1880 and 1920, the overall population in the US would have declined by almost 30%. The decline in the flow of patents would have had almost the same magnitude. In Panels C and D, we depict the aggregate share of innovators and the population share in the urban SEAs (i.e., the SEAs in the top quintile of the urbanization distribution in 1880).

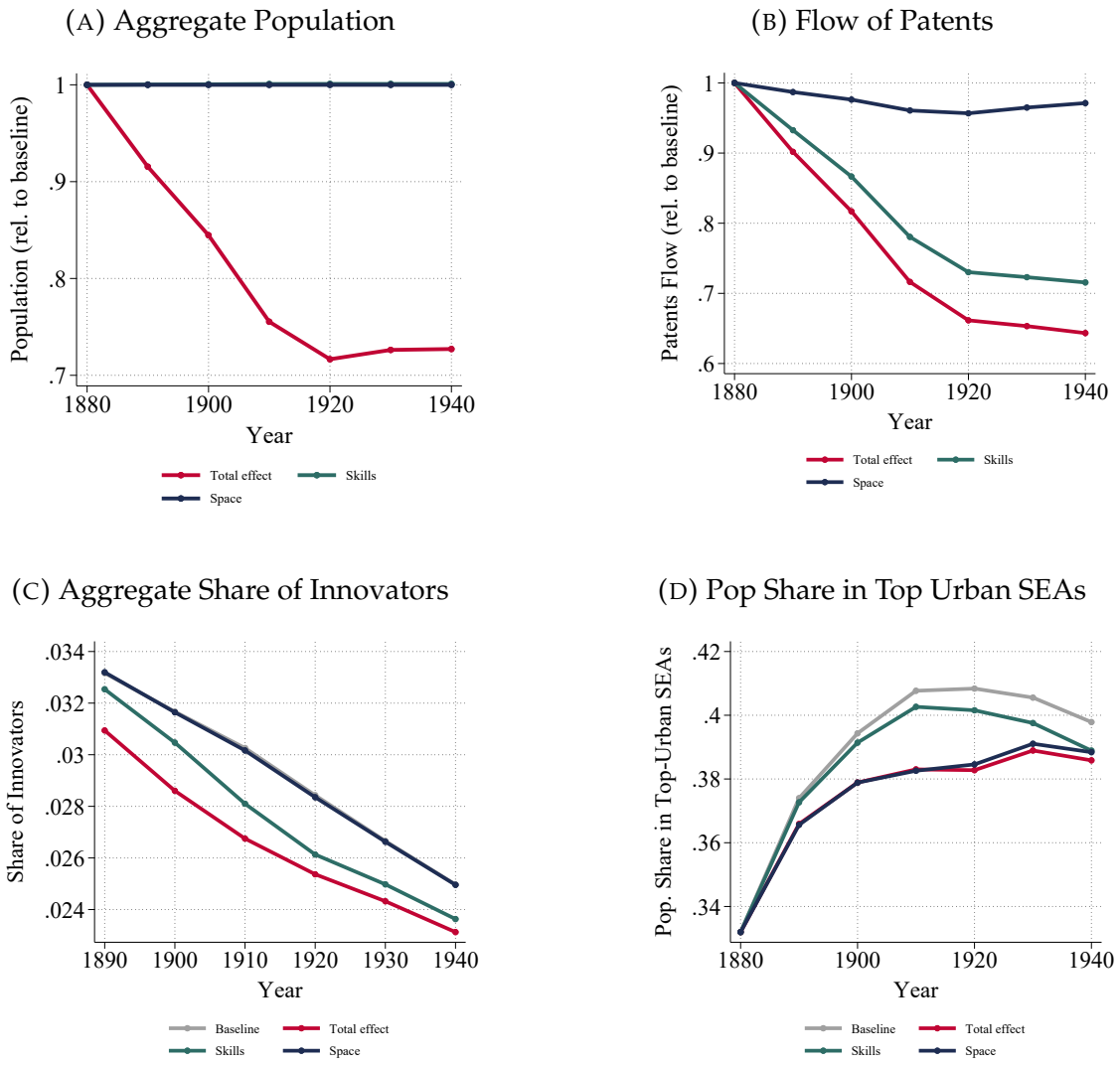
Figure A.8 complements Figure 8 and reports local population growth, relative to the baseline economy, in the absence of international arrivals. Consistent with migrants' urban bias, the population decline is particularly large in cities.

A.3.2.2 Counterfactual II: Removing the Immigration Restrictions in 1920s

In this section, we explain in more detail how we implement our second counterfactual of removing the immigration quotas in the early 20th century (Section 5.2).

Background The 1921 Emergency Quota Act and the subsequent 1924 Johnson–Reed Act represented a watershed in American immigration history. Prior to these statutes, the United States operated under a largely open-door regime: between 1880 and 1914, annual arrivals routinely exceeded one million, with Southern and Eastern Europeans accounting for the majority of the inflows. The 1921 Act imposed the first official ceiling on immigration, capping each nationality at three percent of its foreign-born population

FIGURE A.7: THE AGGREGATE IMPACT OF INTERNATIONAL MIGRATION

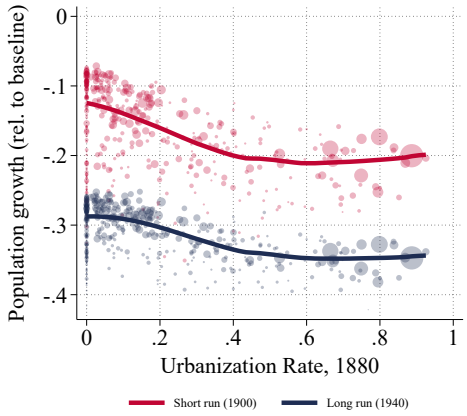


Notes: The figure shows the change in the aggregate population (panel A), the change in the flow of patents (panel B), the aggregate share of innovators (panel C), and the population share in the urban SEAs (panel D) relative to the baseline economy. We depict the baseline calibration in grey, the counterfactual with no immigrant inflows between 1880 and 1920 in red, the "no skilled migrants" counterfactual in teal and the "no urban bias" counterfactual in blue.

recorded in the 1910 Census. The 1924 Act further tightened these quotas, reducing them to two percent of the 1890 Census count, a benchmark chosen deliberately to favor Western European source countries. As a consequence, gross inflows from Italy, Poland, Russia, and much of Southern and Eastern Europe fell by more than 85 percent between the early 1920s and the 1930s, remaining suppressed until the Hart–Celler Act of 1965 dismantled the national-origins framework.

Constructing Counterfactual Inflows To quantify the growth consequences of these immigration quotas, we need to construct the counterfactual inflows in the absence of the policy. To do so, we exploit that the immigration restrictions fell unevenly across nationality groups. Drawing on the pre-restriction census, we decompose the group of "other immigrants" into three sub-populations: approximately 70% were other Europeans

FIGURE A.8: INTERNATIONAL MIGRATION AND LOCAL POPULATION GROWTH



Notes: The figure shows the change in the local population ℓ_{rt} of the economy without the arrival of international immigrants between 1880 and 1920 relative to the baseline economy. We show both the short-run effect in 1890 and the long-run effects in 1940.

(including Southern and Eastern Europeans), who faced the same severe quota restrictions as Italians; roughly 5% were Asians; and the remaining 25% were Western Hemisphere migrants (e.g., Canadians, Mexicans, and Latin Americans), who were explicitly exempt from numerical quotas and therefore continued to arrive largely unimpeded. Hence, approximately 75% of the group of "other immigrants" was subject to restrictions.

We anchor the total immigrant inflow for each decade in $t \in \{1930, \ldots, 1960\}$ to the 1920 aggregate level of $\bar{M}_{1920}$, because this year represents the last pre-restriction period unaffected by the quota legislation.

Western European groups. The British, German, and Irish groups received quotas that were never binding under either Act, so their observed inflows were not directly distorted by the restrictions. We therefore retain their actual historical values:

$$\widetilde{M}_{n,t} \;=\; M_{n,t}, \qquad n \in \{\text{Bri, Ger, Iri}\}, \quad t \in \{1930, \ldots, 1960\}. \tag{A.22}$$

Italians and the Other group. For the two restricted groups—Italians and Others—we allocate the residual between the counterfactual aggregate and the observed Western European total, preserving the relative shares of the two groups at their 1920 levels. Formally, define the residual available to restricted groups in decade t as

$$R_t \;=\; \bar{M}_{1920} \;-\; \sum_{n \in \{\text{Bri, Ger, Iri}\}} M_{n,t}. \tag{A.23}$$

The 1920 share of Italians among all restricted-group arrivals is $s_{\text{Ita}} = M_{\text{Ita},1920} \,/\, (M_{\text{Ita},1920} + M_{\text{Oth},1920})$, with the complementary share $s_{\text{Oth}} = 1 - s_{\text{Ita}}$ going to the Other group. Counterfactual inflows are then $\widetilde{M}_{\text{Ita},t} = s_{\text{Ita}} \times R_t$ and $\widetilde{M}_{\text{Oth},t} = s_{\text{Oth}} \times R_t$ for $t \in \{1930, \ldots, 1960\}$.

Post-1960 inflows revert to their actual values for all nationality groups, reflecting the

structural break introduced by the Hart–Celler Act.

Skill Composition of Counterfactual Immigrants Knowing the volume of counterfactual immigrants is insufficient; we also need to know their skill content, that is, κ^{ν}. We therefore extend the regression analysis of Section 4.4 to a finer nationality decomposition, splitting the Other group into four constituent sub-groups: *Other European*, *American*, *Asian*, and *African and Other*. Estimating the model on this expanded classification yields sub-group-specific κ^{ν} coefficients — see Table A.10. We then assign the counterfactual immigrant inflows the average of *Other Europeans* and *Asians*, because both groups were affected. In terms of their spatial allocation, we hold fixed the initial arrival probabilities at their 1920 levels.

TABLE A.10: ESTIMATION OF RELATIVE SKILL SUPPLY WITH NON-EUROPEAN: κ^{ν}

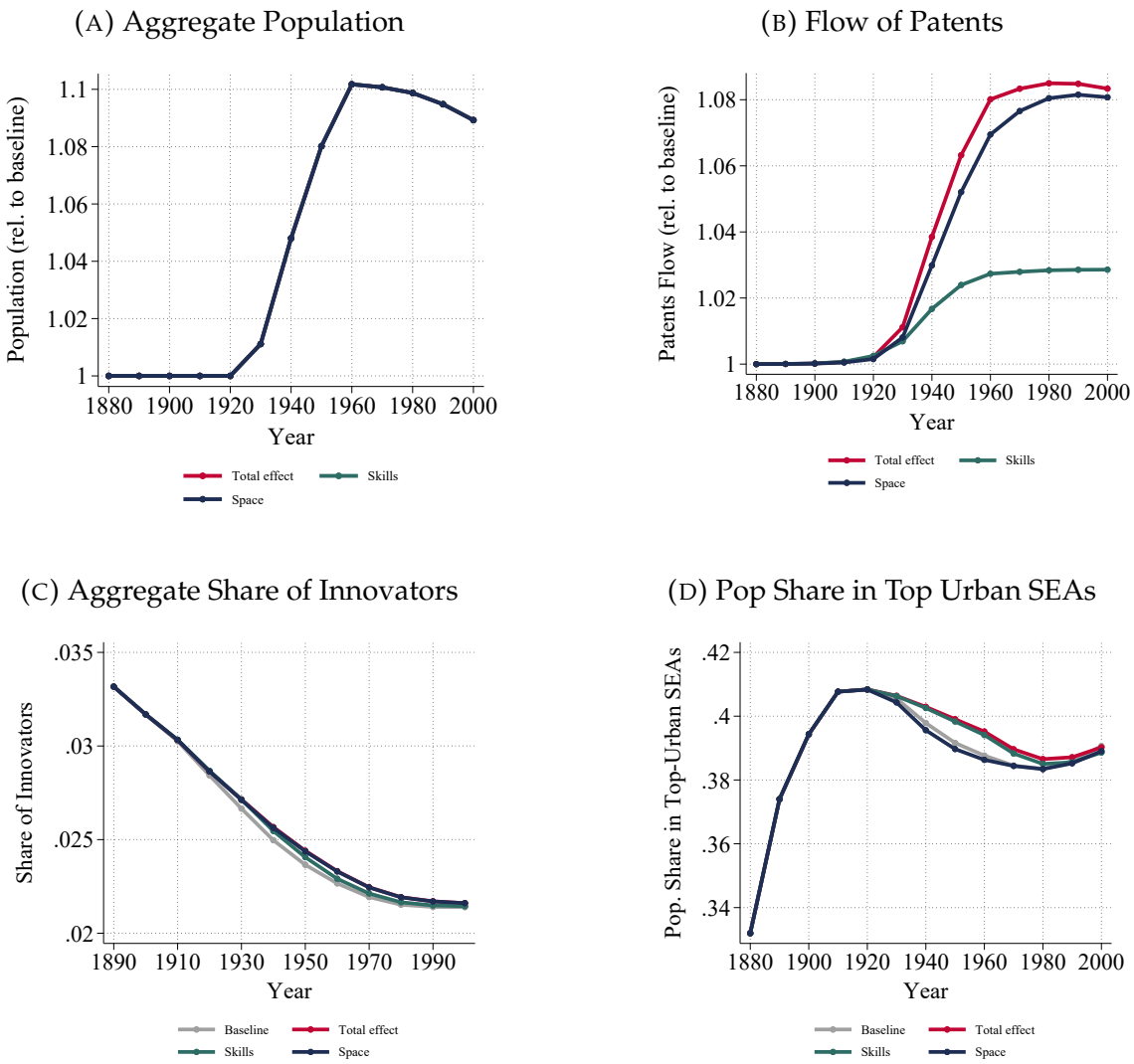
	(1)	(2)	(3)	(4)
British	0.281***	0.260***	0.223***	0.234***
	(0.025)	(0.025)	(0.026)	(0.024)
German	-0.268***	-0.277***	-0.294***	-0.268***
	(0.057)	(0.057)	(0.059)	(0.056)
Irish	-1.235***	-1.258***	-1.308***	-1.207***
	(0.092)	(0.090)	(0.086)	(0.087)
Italian	-1.479***	-1.491***	-1.516***	-1.399***
	(0.054)	(0.055)	(0.058)	(0.048)
Other European	-0.623***	-0.633***	-0.649***	-0.623***
	(0.041)	(0.042)	(0.043)	(0.041)
American	-0.215***	-0.235***	-0.272***	-0.255***
	(0.073)	(0.073)	(0.076)	(0.067)
Asian	-1.784***	-1.834***	-1.900***	-1.517***
	(0.417)	(0.415)	(0.417)	(0.262)
African and Other	0.539**	0.125	0.099	-0.607***
	(0.248)	(0.170)	(0.223)	(0.094)
Year-SEA FE	yes	yes	yes	yes
Cell size cutoff	-	200	500	-
Impute if zero?	no	no	no	yes
R^2	.871	.837	.828	.956
N	7,024	5,967	4,863	16,249

Notes: This table reports estimates of κ^{ν} from an extended nationality decomposition, splitting the Other group into four sub-groups: *Other European*, *American*, *Asian*, and *African and Other*. All specifications include Year-SEA fixed effects. In columns 2 and 3 we require all cells to have at least 200 or 500 observations respectively. In column 4, we add a small positive number to Θ_{rt}^{ns} in case $\Theta_{rt}^{ns} = 0$. All regressions are weighted by the number of observations in each cell.

Supplementary Results Figure A.9 reports additional aggregate outcomes for the counterfactual exercise that removes immigration restrictions. Panels A and B show that removing the restrictions raises aggregate population and patenting after 1920, with effects increasing through mid-century and remaining positive thereafter. The patenting effect is much smaller when the skill-composition channel is shut down, but remains close to the benchmark removing-ban counterfactual when the urban-bias channel is eliminated.

Panels C and D show that the aggregate share of innovators and the population share in the most urban SEAs change little relative to the baseline. Overall, the effects operate mainly through the scale of population and innovation rather than through large changes in innovator shares or urban concentration.

FIGURE A.9: THE AGGREGATE IMPACT OF REMOVING RESTRICTIONS



Notes: The figure shows the change in the aggregate population (panel A), the change in the flow of patents (panel B), the aggregate share of innovators (panel C), and the population share in the urban SEAs (panel D) relative to the baseline economy. We depict the baseline calibration in grey, the counterfactual removing the ban in red, the counterfactual where all previously banned migrants are unskilled in teal ("Skills") and the counterfactual where immigrants do not have an urban bias in blue ("Space").