SUPPLEMENTARY MATERIAL

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

OA.1. ONLINE APPENDIX: DATA PROCESSING

OA.1.1 Geography Harmonization

In the main analysis, geographic location is defined at the level of **State Economic Areas (SEAs)**. SEAs are groups of contiguous counties within a state delineated by the U.S. Census Bureau prior to the 1950 Census to reflect integrated regional labor and economic markets.

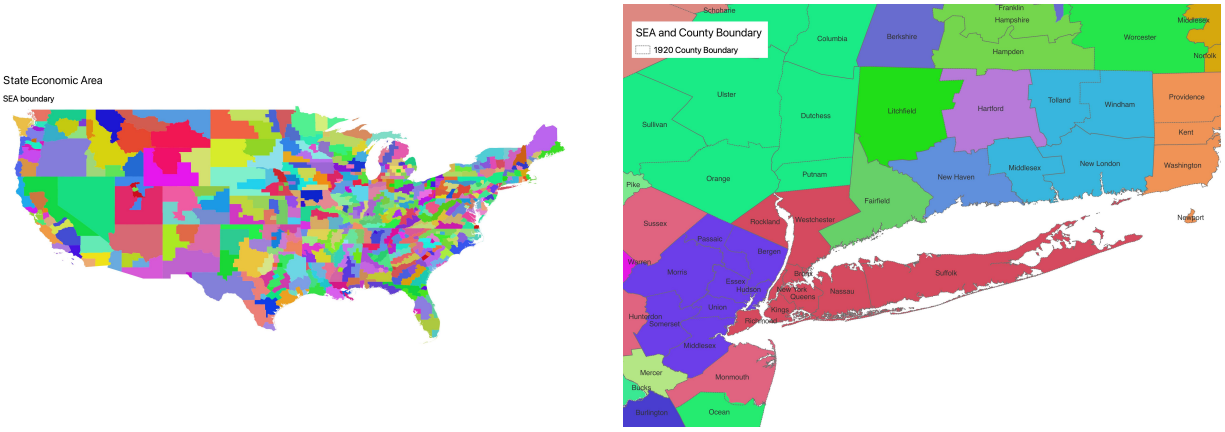
We aggregate data to the SEA level for two reasons. First, SEA boundaries remain consistent over our sample period, which avoids time-varying geographic definitions that would otherwise introduce measurement inconsistencies in panel regressions (county boundaries changed substantially between 1880 and 1920). Second, SEA-level aggregation reduces sparsity in patent counts. At the county level, patenting is frequently zero or missing, which generates instability in log specifications and inflates sampling variance. Aggregation to SEAs ensures strictly positive patent counts for the vast majority of observations.

Figure OA.1 illustrates the geographic construction. The left panel displays the 452 SEA boundaries used in the analysis. We overlay historical county shapefiles with 1940 SEA shapefiles obtained from the National Historical Geographic Information System (Manson et al., 2019) using Geographic Information System (GIS) software. For each county-SEA intersection, we compute area-based weights and use these weights to aggregate county-level variables to the SEA level. All patent and manufacturing data are then expressed at the SEA-year level.[30]

For record linkage between patent and population census data, we implement a hierarchical geographic matching procedure. We first attempt to match individuals within the recorded county of residence. If no match is found, we expand the search to the corresponding SEA. This procedure accommodates short-distance migration across adjacent counties.

[30]The left panel of Figure OA.1 shows the 452 SEA boundaries, with colors indicating distinct SEAs. The right panel of Figure OA.1 shows the New York SEA (in Red), which includes counties covering New York City and its surrounding areas.

FIGURE OA.1: State Economic Areas (SEA) and county boundaries



Notes: In the left panel, we present the State Economic Area (SEA) boundary, adhering to the definition provided by the US Census Bureau. The right panel depicts the New York SEA (in Red), which includes counties covering New York City and its surrounding areas. These overlays are used to construct spatial crosswalks that harmonize geographic identifiers across the Population Census, Census of Manufactures, and U.S. Patent records.

OA.1.2 Data Translation

Immigration Records I: Hamburg Passenger Lists

The Hamburg Passenger Lists (HPL) are recorded in German and stored in a non-structured format. To construct analyzable variables, we implement a structured pipeline consisting of parsing, translation, and harmonization. We focus on three fields central to the analysis: final destination (*Zielort*), nationality (*Nationalität*), and pre-migration occupation (*Beruf*).

Destinations (*"Zielort"*) We translate ship route and destination fields to identify passengers bound for the United States. While most departures from Hamburg were to North America, some passengers traveled to South America (e.g., Buenos Aires, Rio de Janeiro) and other European destinations. We restrict the sample to passengers whose final destination is recorded as the United States.

Nationality (*"Nationalitaet"*) Nationality information is used to facilitate linkage between immigration records and the U.S. Population Census. This field is directly recorded in the HPL beginning in 1898. For earlier records, nationality is imputed using the name-based classification tool *NamePrism* (also used by Diamond et al. (2019)), which predicts nationality and ethnicity based on first and last names. Because NamePrism provides probability-based reliability scores, we restrict the sample to observations with predicted nationality probabilities of at least 50 percent.

Occupation (*"Beruf"*) We translate and harmonize pre-migration occupations into the 1950 Census Bureau occupation classification (OCC1950) to ensure comparability across years. Approximately 53 percent of HPL records report occupation, with near-universal coverage for working-age males. The most commonly reported occupations include farm

laborers, general laborers, managers, and skilled manufacturing workers (e.g., tailors, shoemakers, cabinet makers). The translation from German to English and the subsequent harmonization into OCC1950 codes are described in detail in Section OA.1.3.

OA.1.3 Data Harmonization

Pre-Migration Occupations Both the Castle Garden records and the Hamburg Passenger Lists report immigrants' occupations prior to their arrival in the United States. We refer to these as pre-migration occupations. To ensure comparability across datasets and over time, we harmonize all reported occupations to the U.S. Census occupation classification system (OCC1950). For the Hamburg Passenger Lists, occupational entries are recorded in German and in free-text format. We first translate the original German occupational strings into English. We then standardize spelling variants and resolve ambiguous titles before mapping all occupations to the corresponding OCC1950 codes. This harmonization procedure yields a consistent occupational classification across immigration records and U.S. Census data, enabling direct comparison of pre-migration and post-arrival occupational outcomes.

Classifying pre-migration occupations to *OCC1950* The Castle Garden (CG) data contain 1,290 distinct pre-migration occupation strings, while the Hamburg Passenger Lists (HPL) contain 903 distinct German occupation strings. We construct a crosswalk mapping all pre-migration occupations to the U.S. Census occupation classification system (OCC1950), which consists of 245 occupational categories (we exclude non-occupational responses such as "keeping house" or "in school").

For HPL records, we first translate German occupation strings into English using the Google Translation API. All translated entries are manually reviewed to correct mistranslations and standardize terminology prior to classification.

We then implement a multi-step matching procedure to assign each distinct occupation string (both from CG and HPL) to an OCC1950 category:

1. **String Similarity Matching.** We perform initial matching based on exact and approximate string similarity, including the normalization of spelling variants, the removal of punctuation, and case standardization.

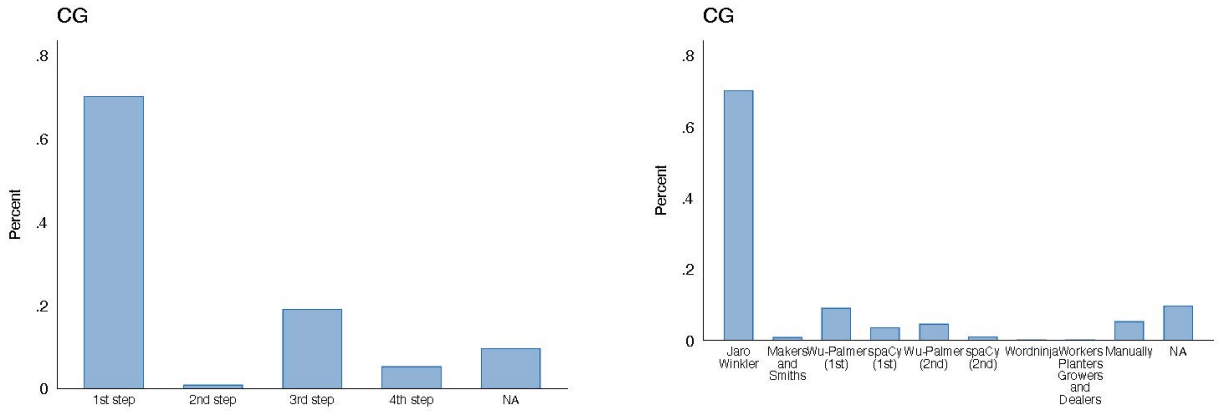
2. **Targeted Group Analysis.** For high-frequency occupational groups containing suffixes such as "maker" (e.g., shoemaker, bookmaker) or "smith," we conduct targeted semantic matching to ensure consistent classification within related occupational families.

3. **Semantic and Lexical Matching.** For unmatched strings, we apply iterative semantic matching using lexical databases (WordNet) and natural language processing tools (spaCy) to identify conceptually related OCC1950 categories.

4. **Manual Verification and Correction.** Remaining unmatched or ambiguous cases are manually reviewed and assigned to the most appropriate OCC1950 category.

Figures OA.2 and OA.3 report the share of distinct occupation strings matched at each stage for CG and HPL, respectively. Approximately 80 percent of CG occupation strings are matched in the initial string-similarity stage. For HPL, roughly 70 percent are matched through semantic and lexical procedures. Manual assignment accounts for approximately 5 percent of CG occupations and 20 percent of HPL occupations. This procedure yields a complete and consistent crosswalk from pre-migration occupations to OCC1950. Details for each step are discussed below.

FIGURE OA.2: SHARE OF OCCUPATION MATCHES FOR EACH STEP (CG)

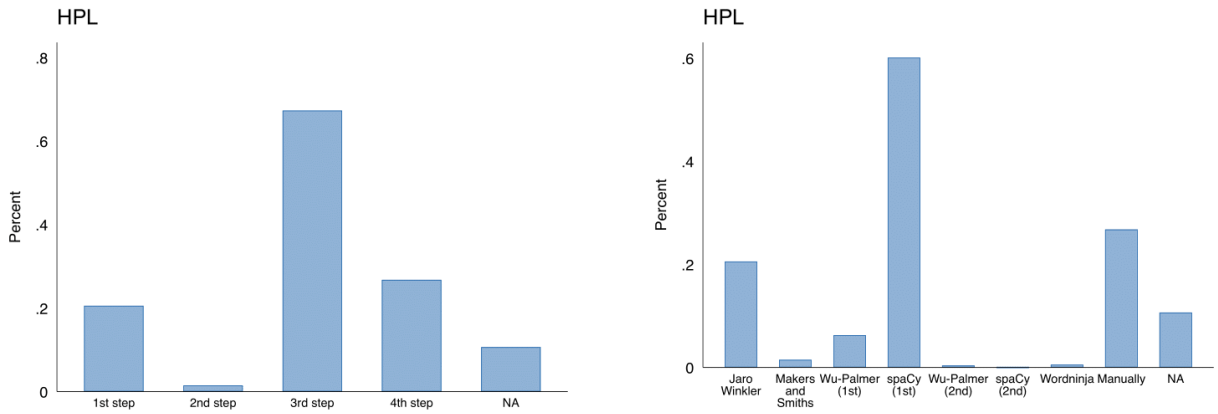


Notes: This graph illustrates the share of CG occupations matched at each step, indicating the contribution of each step to the matching process. The values are derived by summing the frequencies of CG occupation titles that correspond to Census occupation titles at each step. The left panel displays the overall share of occupations matched by four aggregate steps. In the right panel, we further dissect this match rate by different methods within the four steps.

Primer: Data cleaning In the data cleaning process (as the initial step before occupation matching), we employ standard text data pre-processing techniques, including lowercasing, punctuation removal, stop words removal, and lemmatization for occupation strings. Subsequently, to enhance matching rates, we group certain OCC1950 occupations into broader categories. For example, various types of mechanics and repairmen in OCC1950, such as airplane, automobile, and office machine mechanics, are aggregated into the category of "mechanics and repairmen."

To further refine matches, we also break down occupation strings containing multiple occupations. For instance, "mechanics and repairman" is broken down into "mechanics" and "repairmen," both used for linking to improve match accuracy. Additionally, we ensure the order of words within occupations is comparable. For example, OCC1950's "engineers, civil" is reported as "civil engineer" in the Castle Garden data.

FIGURE OA.3: SHARE OF OCCUPATION MATCHES FOR EACH STEP (HPL)



Notes: This graph illustrates the share of HPL occupations matched at each step, indicating the contribution of each step to the matching process. The values are derived by summing the frequencies of HPL occupation titles that correspond to Census occupation titles at each step. The left panel displays the overall share of occupations matched by four aggregate steps. In the right panel, we further dissect this match rate by different methods within the four steps.

Following this cleaning, we are left with 1,290 distinct occupations in the Castle Garden database and 245 in OCC1950.

Step 1: String Similarity Matching In the initial step, we match occupations directly based on the similarity of the occupation strings. This is done by calculating the Jaro–Winkler similarity between the occupational titles in the immigration data and the OCC1950 Census classification (for more details on the Jaro Winkler distance, see **?**, **?**, Winkler (2014)). Matches with a similarity measure below 0.98 are discarded, ensuring a stringent criterion. The OCC1950 classification with the highest Jaro–Winkler similarity measure is then assigned.

This approach is considered accurate due to its strict criteria. For example, the Jaro–Winkler similarity between "book keeper" and "bookkeeper" is 0.982. Out of 1,290 Castle Garden occupations, 154 occupations are matched using Jaro–Winkler similarity. Although this represents a relatively small number, it accounts for almost 50% of all individuals in the data, as such occupations are often larger in scale.

Step 2: Analysis of Specific Groups For occupations not matched in Step 1 using Jaro–Winkler Distance, our focus turns to occupations containing terms like "maker" and "smith." In the Castle Garden data, there are 220 occupations involving "maker" and 23 involving "smith."

To assign these occupations to OCC1950, we compare the occupational descriptions preceding "maker" or "smith" in terms of meaning. We utilize *Wordnet* and *spaCy* for this purpose. *Wordnet*, part of the NLTK library in Python, is a lexical database grouping words into *synsets* (sets of synonyms) based on conceptual-semantic and lexical relationships.

We match occupations by maximizing the Wu-Palmer similarity measure.[31] For instance, "furniture maker" in Castle Garden is matched to "cabinet maker" in OCC1950.

spaCy, an open-source library for Natural Language Processing, vectorizes words and determines similarity using word vectors. As a supplement to WordNet, *spaCy* is employed for occupations not matched using *Wordnet*, benefiting from its extensive word list.

Step 3: Semantic and Lexical Matching For the remaining occupations, we continue with a similar procedure. Initially, we aim to match occupations using *Wordnet* with a Wu-Palmer similarity score of 0.94, resulting in 149 occupation matches. Subsequently, we use *spaCy* for the remaining unmatched occupations, applying a similarity score of 0.6. In this iteration, we achieve 240 occupation matches. This process is repeated with a Wu-Palmer score of 0.9 and a *spaCy* score of 0.55, yielding an additional 120 matches. For the remaining unmatched occupations, we apply the same steps after breaking occupation titles using *Wordninja,* a Python package that separates words without spaces. For example, *Wordninja* separates "laundrymaid" into "laundry maid". This results in 73 additional occupational matches.

Among the unmatched occupations, we encounter multiple instances of "workers", "planters", "growers", and "dealers." We match these occupations to the Census using *spaCy*. For instance, instead of trying to match "cattle dealer", we find the best match for "dealer," which turns out to be "salesman." Similarly, "workers" are matched with "laborers", and "planter" is matched with "farmer."

Step 4: Manual Correction In the last step, we meticulously review the matches generated by the previous procedures, manually correcting any matches that seem incorrect. A total of 532 matches are corrected through this manual review process; this represents approximately 5% of all matches.

OA.2. Features and Measurement of the Patent Data

In this section, we outline the data harmonization and processing procedures for the Patent Database, emphasizing its role in measuring innovation across time and space. While patents may only capture specific aspects of innovative activities, their consistent and extensive coverage makes them a plausible proxy for measuring innovation during the study period.[32]

[31] Wu-Palmer similarity denotes words' pairwise similarity as determined by the depth at which a common ancestor word can be found for the pair's synset trees. In short, this similarity measure calculates the similarity in terms of the meaning of words.

[32] To the best of our knowledge, other innovation-related activities, such as trademark data during our study period, are deemed unsuitable for our study. This is attributed to the trademark database's lack of information on founders' names and its inability to capture the type of knowledge created.

Despite the extensive use of the current US patent database in economics literature, its current format poses challenges for our study due to limited and inconsistent coverage of key information, including patent inventors' and assignees' names, geographic locations, collaboration patterns, and nationality. To overcome this limitation, we collect patent-level documents from Google Patents and employ modern Natural Language Processing (NLP) tools to extract relevant information.[33] While similar efforts have been made by Google Patents (https://patents.google.com/), the PatentCity Database by Bergeaud and Verluise (2022), and HISTPAT (Petralia et al., 2016), we compile our own data tailored to our research questions, augmenting Google's Patent search engine results with data from PatentCity. The resulting unified dataset includes patent inventors' and assignees' names, geographic locations, collaboration patterns, nationality, and a flag for individual versus corporate patents, enabling a systematic study of the geography of innovative activities.

Co-Invention and the Number of Co-Inventors During our study period, the patenting process becomes increasingly organized, leading to a rise in the occurrence of multiple co-inventors for a given patent. To capture this collaborative innovation, we identify all co-inventors' names along with their corresponding residential locations. For a patent with 5 co-inventors, for example, we create a dummy variable indicating the co-invented status of the patent. Additionally, we generate a measure of fractional co-invention; in the mentioned case, the fractional co-invention would be 0.2. This fractional invention information proves valuable in assessing both the collaborative nature of innovative activities and the intensive margin of patenting.

Corporate Patenting During our study period, the organization of patenting became more structured, leading to an increased prevalence of corporations orchestrating innovative activities. In such cases, the patent document reflects a distinction between patent inventors (i.e., individuals who have contributed to the claimed invention) and patent assignees (i.e., organizations and/or individuals that have a legal ownership interest in the rights a patent offers). Often, the assignee is the organization employing the inventor.

By scrutinizing the patterns of patent inventors and assignees, we discern corporate patents throughout our study period. For example, if the assignee information includes terms like "Company", "Corp.," "Ltd", "Mfg", "Property", "System", "Gmbh", "Enterprise", and the patent inventor(s) is not the assignee, we classify such patents as corporate. This analysis helps identify patents associated with organizational entities during the specified period.

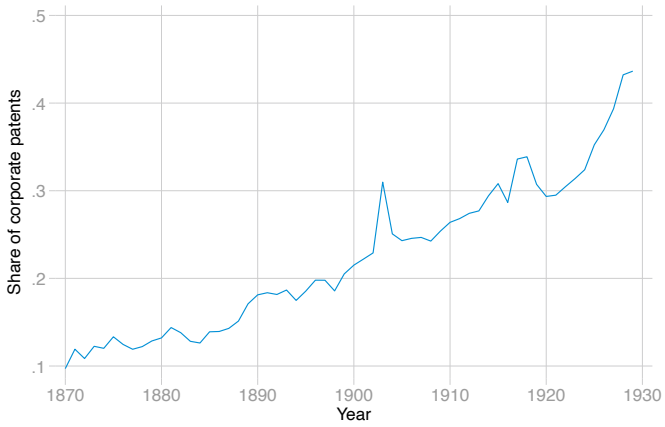
For our patent-census linking process, we include corporate patents by focusing on patent

[33] As discussed in Bergeaud and Verluise (2022), most historical patent documents are available as scanned images rather than a systematically harmonized database. Therefore, converting images into a database necessitates the use of modern techniques such as Natural Language Processing (NLP) and Geographic Information System (GIS).

inventors rather than assignees. In such cases, we utilize the patent inventor's name and residential location to establish links between patent and U.S. Population Census records.

Figure OA.4 plots the share of corporate patents during our study period. Each value is calculated as the number of corporate patents divided by the total number of patents granted in that year. As shown in Figure OA.4, a consistent upward trend in corporate patents is evident throughout the study period. The share of corporate patents starts at 10% in the early 1880s and steadily increases, surpassing 40% by the late 1920s.

FIGURE OA.4: SHARE OF CORPORATE PATENTS



Notes: The figure plots the share of corporate patents during our study period. Each value is calculated as the number of corporate patents divided by the total number of patents granted in that year.

Citizenship/Nationality of Patent Inventors As noted by Bergeaud and Verluise (2022), citizenship information of patentees is not available in standard patent datasets during the study period. To address this gap, we employ both Natural Language Processing (NLP) techniques and record linking methods to extract and enhance inventor citizenship information.

In terms of NLP techniques, we extract inventor citizenship information by analyzing the text of patents obtained from the Google Patents API. However, not all patent inventors declare their citizenship during the study period, with less than 40% of patents (granted between 1880 and 1929) providing such information, consistent with Bergeaud and Verluise (2022). Recognizing the limitations of NLP-based extraction alone, especially in capturing citizenship, which is crucial for our focus on studying immigration and innovation, we augment this information.

We augment patent inventors' citizenship by linking patent inventors (from the patent database) with federal US population censuses from 1880 to 1920. Modern machine-learning-based record-linking techniques are employed for this purpose. This record-linking-based citizenship classification aids in identifying immigrants, even when citizenship was not declared in the original patent documents — an occurrence in over 60% of patents during the study period.

Multiple Patents by the Same Inventor We observe instances where the same inventor holds multiple patents, determined by considering the patent inventor's full name, residential location, patent grant years, and the areas of claimed inventions. To capture this phenomenon, we generate a quasi-inventor ID and quantify the number of patents attributed to each inventor. This information on the "number of patents" proves valuable in measuring the intensive margin of patenting — shedding light on the extent of an inventor's involvement in patent activities.

International Patents Utilizing Natural Language Processing techniques on patent-level documents sourced from Google Patents, we have pinpointed inventors residing abroad who were credited with US patents. We refer to these as "international inventors" and their associated patents as "international patents," a classification that allows us to study their impact on the patent landscape.

International patents constitute between 5 and 10% of the overall patent portfolio, indicating that inventors abroad made a meaningful contribution to US innovation.

OA.3. Coverage, Quality, and Potential Biases of the Linked Records

This section provides additional detail on the linked records used throughout the paper, organized in three parts: (i) match counts and coverage, (ii) the quality of our matches, and (iii) potential biases and selection concerns in the matching procedure. These results are referenced in Appendix Section A.2.

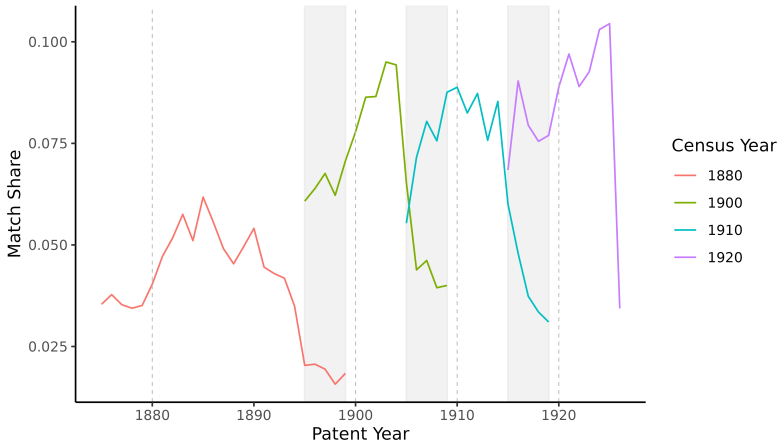
OA.3.1 Match Summary and Coverage

TABLE OA.1: MATCHING SUMMARY BY IMMIGRATION DATASET

		1880	1900	1910	1920
CG: Total	—	127,884	210,076	206,376	147,461
CG: British	No. Used for Matching	470,035	418,675	248,346	115,400
	Match Count	29,643	36,867	28,226	20,307
	Match Rate	0.057	0.055	0.044	0.039
CG: German	No. Used for Matching	1,086,610	1,506,990	1,110,002	675,860
	Match Count	63,169	136,390	114,182	75,719
	Match Rate	0.050	0.062	0.054	0.041
CG: Irish	No. Used for Matching	716,614	415,929	200,713	83,477
	Match Count	25,181	1,727	23,818	14,731
	Match Rate	0.031	0.002	0.029	0.023
CG: Italian	No. Used for Matching	8,901	500,057	608,741	492,375
	Match Count	260	16,676	23,708	22,780
	Match Rate	0.028	0.032	0.034	0.032
HPL: Total	—	30,879	84,072	112,593	94,404
HPL: British	No. Used for Matching	30,281	113,293	99,626	68,058
	Match Count	928	1,441	1,759	1,540
	Match Rate	0.029	0.011	0.013	0.012
HPL: German	No. Used for Matching	207,901	525,007	553,280	467,915
	Match Count	27,094	73,970	84,545	57,869
	Match Rate	0.121	0.116	0.111	0.072
HPL: Italian	No. Used for Matching	652	3,451	5,020	4,372
	Match Count	15	68	285	336
	Match Rate	0.022	0.018	0.048	0.051

Notes: This table reports the number of uniquely matched observations between the immigration datasets (CG and HPL) and the Census, along with the corresponding match rate by nationality over the study period.

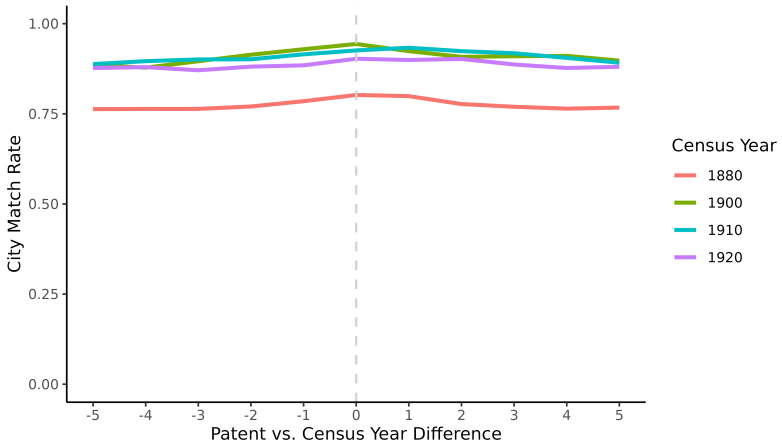
FIGURE OA.5: PATENT YEAR DISTRIBUTIONS



Note: The above figure illustrates the distribution of patent grant years of patents uniquely matched to each census. For example, of the 120,298 patents matched to the 1910 Census, approximately 8% were granted in 1907. Vertical dotted lines indicate census years. Shaded regions denote patent years eligible for matching to both adjacent censuses (1895–1899, 1905–1909, and 1915–1919). For patents granted in 1907, 4,379 records are matched to the 1900 Census and 9,673 records to the 1910 Census.

OA.3.2 Match Quality

FIGURE OA.6: CITY MATCH RATES BY CENSUS YEAR



Notes: The above figure illustrates the match rates between city information for census individuals and inventor city information. The x-axis represents the difference between patent year and the matched census year.

Potential Biases and Selection Concerns

We next examine whether systematic selection arises along geographic, demographic, or name-based dimensions that could affect match probabilities.

Geographic Bias in Patent-Census Matches Table OA.2 reports the distribution of patents across State Economic Areas (SEAs) in both the full patent dataset and the patent-census matched dataset. Match rates are lower in more populous SEAs, consistent with

increased name collisions in densely populated regions. All human-capital regressions therefore include SEA fixed effects to account for geographic variation in match probability (e.g., see Table OA.4).

TABLE OA.2: MATCH FREQUENCY BY SEA REGION

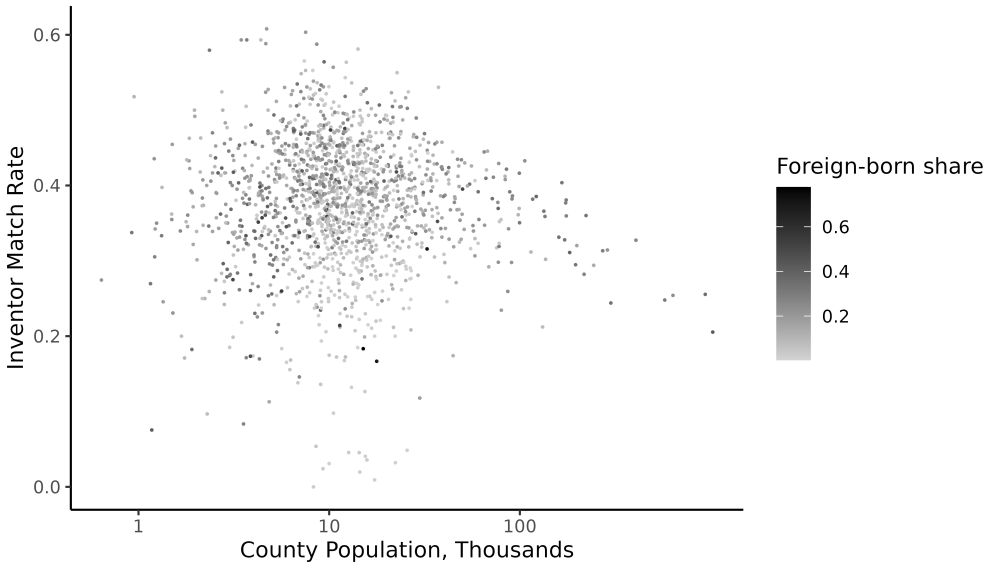
SEA	Record Count	Match Count	% All Patents	% Matched Patents
NYC Metro	128,075	30,558	8.9	6.8
Chicago Metro	92,512	24,802	6.5	5.5
Boston Metro	59,783	20,328	4.2	4.5
Jersey City/Newark	52,300	19,361	3.6	4.3
Philadelphia Metro	48,035	13,909	3.4	3.1
Pittsburgh Metro	27,668	8,129	1.9	1.8
San Francisco Metro	22,606	7,459	1.6	1.7
St. Louis Metro	21,483	7,084	1.5	1.6
Cleveland Metro	21,415	7,583	1.5	1.7
Los Angeles Metro	16,409	5,521	1.1	1.2

Notes: The table reports patent shares by SEA in the full and matched datasets. SEAs are ordered by total patent filings.

County Population, Immigrant Share, and Match Rate Figure OA.7 plots inventor match rates by county against county population. Match rates decline with population size, again consistent with increased name collisions in large counties. Substantial dispersion remains conditional on population. Larger counties also tend to have higher immigrant shares and lower match rates, suggesting that differences in name distributions across nativity groups may affect match probabilities.

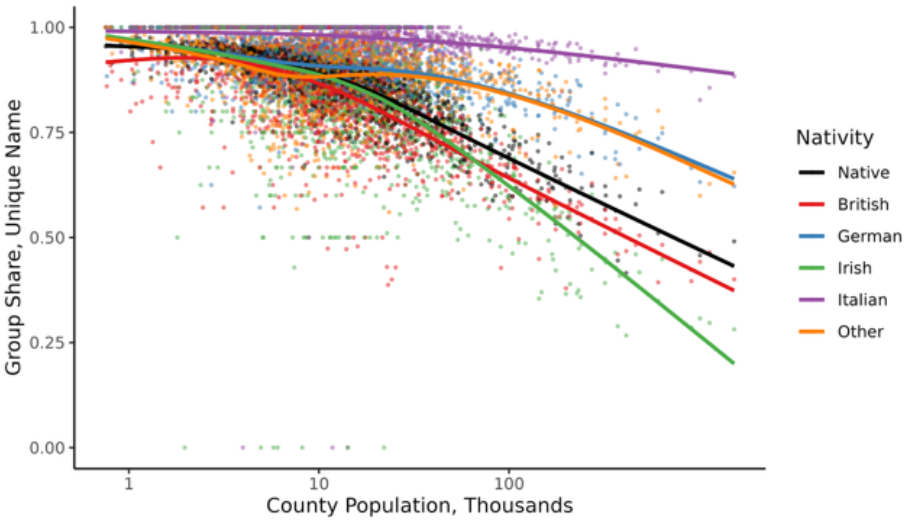
Name Uniqueness by Nativity A potential concern is that immigrants may have more distinctive names, mechanically increasing their match probability relative to natives. To evaluate this channel, Figure OA.8 plots name uniqueness against county population for patent-matching-eligible individuals in the 1920 Census. Within each county, a name is defined as unique if no other individual shares the same standardized given name and surname combination. For counties with fewer than approximately 22,000 residents (about 90 percent of all counties), name uniqueness is broadly similar across nativity groups, with differences emerging primarily in the largest counties. These patterns indicate that differential name uniqueness across nativity groups is unlikely to explain match-rate differences in the majority of counties.

FIGURE OA.7: MATCH RATES FOR INVENTOR RECORDS BY COUNTY POPULATION



Notes: This figure plots inventor match rate by county against county population. Each point represents a county. Darker shading indicates a higher share of foreign-born residents.

FIGURE OA.8: NAME UNIQUENESS BY COUNTY POPULATION



Notes: The figure reports the share of unique standardized given name–surname combinations by county and nativity group for patent-matching-eligible individuals in the 1920 Census. The sample is restricted to counties with at least 100 eligible individuals.

We next examine whether inventor name characteristics predict the likelihood of a unique match. Table OA.3 reports summary statistics for four name-based measures constructed from patent records: surname frequency, given name frequency (per 10,000 individuals), and surname and given name length. Name frequencies are computed at the county level using patent-matching eligible individuals from the 1900 Population Census. Given names are standardized prior to computing frequency and length to align with the matching procedure. These name characteristics are then used to assess their explanatory power in

matching probability regressions.

TABLE OA.3: SUMMARY OF NAME CHARACTERISTICS

	Mean	SD	Min.	Max.
Last Name Frequency (per 10k)	0.070	0.530	0.017	74.050
Given Name Frequency (per 10k, stnd.)	0.450	11.471	0.017	986.904
Surname Length	6.613	1.779	1	16
Given Name Length (stnd.)	5.954	1.402	1	15

Notes: Summary statistics for name frequency and length. Given names are standardized prior to computation.

To quantify these patterns formally, we estimate regressions of match probability on name characteristics and nativity composition.

In Table OA.4, we regress a unique match indicator variable for each inventor record against the characteristics of the inventor's given name and surname. The shares reported for natives and immigrants in the table are calculated from the share of potential matches with the associated birthplace. The first column reports the result when regressing only on nativity shares from potential matches. We find that records with a relatively large share of native potential matches are more likely to be matched. The second column reports results when regressing on name frequency and county population. The difference in explanatory power between the models in columns 1 and 2 suggests that match likelihoods are better explained by name characteristics rather than nativity. In addition, we note that the relative importance of county population is very high compared to name frequency, given the low magnitude of the name frequency variables listed in Table OA.3. Finally, in column 3, we observe that highly native names are still more likely to be matched than immigrant names, even when controlling for further name characteristics such as name length.

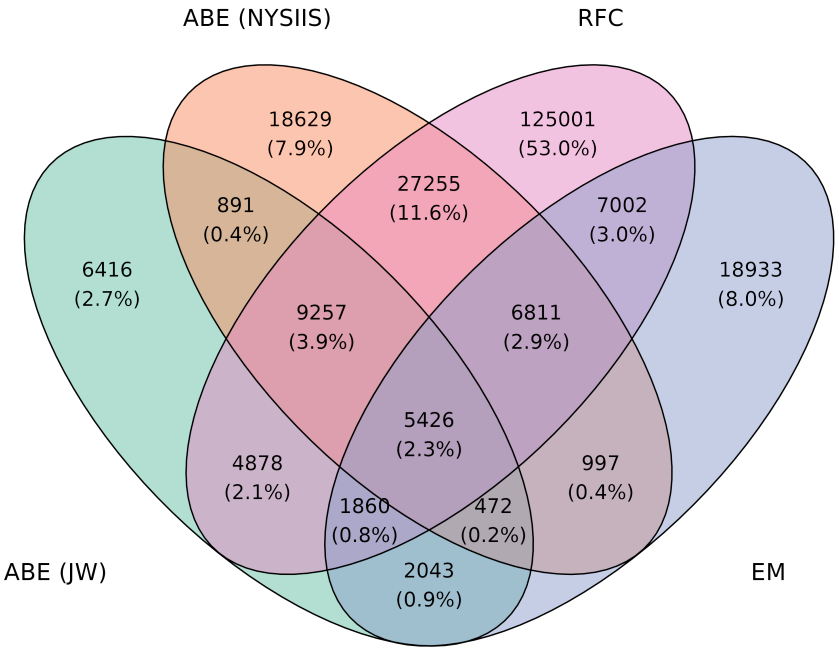
TABLE OA.4: MATCH LIKELIHOODS BY INVENTOR CHARACTERISTICS

	Dependent variable:		
	Inventor Match		
	(1)	(2)	(3)
Share British	−0.232***		−0.086***
	(0.002)		(0.002)
Share German	−0.179***		−0.064***
	(0.001)		(0.001)
Share Irish	−0.392***		−0.126***
	(0.002)		(0.002)
Share Italian	−0.219***		−0.245***
	(0.009)		(0.009)
Share Other Non-Native	−0.165***		−0.044***
	(0.001)		(0.001)
Given Name Frequency (per 10k, stnd.)		−0.0002***	−0.0002***
		(0.00000)	(0.00000)
Last Name Frequency (per 10k)		−0.003***	−0.003***
		(0.00001)	(0.00001)
Given Name Length (stnd.)			0.006***
			(0.0001)
Surname Length			0.015***
			(0.0001)
Log County Population		−0.094***	−0.087***
		(0.0001)	(0.0002)
Observations	2,465,446	2,465,446	2,465,446
Adjusted R^2	0.031	0.174	0.186

*p<0.05; **p<0.01; ***p<0.001

Notes: This table reports the results of regressing a unique match indicator variable on characteristics of the inventor record. Immigrant shares represent the share of potential census matches with the listed birthplace from 0 to 1. For example, a British share of 1 indicates that all potential matches are British. If all shares are 0, then all potential matches are native. Because given names are standardized before matching, we use the standardized given names to compute name frequency and length. We note that records with low shares of potential immigrant matches would be most likely to match.

FIGURE OA.9: MATCH OVERLAP BY LINKING METHOD



Notes: The Venn diagram shows the overlap of record pairs across linking methods for 1910 CG-Census linking. We color RFC-linked matches in pink.

OA.4. ROBUSTNESS

In this section we provide additional details for the robustness exercises contained in Section 6.

Robustness with respect to structural parameters

In the first six rows of Table 8 we test the robustness of our results with respect to different structural parameters. In rows 2 and 3 we allow for aggregate knowledge spillovers, that is we assume that $\psi_{rt} = \zeta_r N_{rt-1}^{\vartheta} (\sum_j N_{jt-1})^{\gamma}$. The higher γ, the stronger the sensitivity of local human capital ϕ_{rt} to the aggregate amount of knowledge, $\sum_j N_{jt-1}$. As such, γ operates as an additional scale elasticity and long-run growth is given by $g_y = g_A + \frac{1}{\sigma-1}\frac{1}{1-\vartheta-\gamma}\eta$. The baseline model is nested for the case of $\gamma = 0$. Because $\sum_j N_{jt-1}$ is a pure aggregate spillover that is common across locations, γ is not identified from cross-sectional data. In fact, our estimate of ϑ is independent of γ, because $\sum_j N_{jt-1}$ would be subsumed in the time fixed effect. To test to the importance of this channel, we consider two alternatives: $\gamma = 0.15$ and $\gamma = 0.3$.

In rows 4 and 5 we consider the elasticity of substitution across varieties σ. σ is an important parameter, because it determines the strength of variety gains. And because long-run growth is given by $g_y = g_A + \frac{1}{\sigma-1}\frac{1}{1-\vartheta}\eta$, a lower value of σ amplifies the effects

of population growth on growth. We consider two alterative values: $\sigma = 4$ and $\sigma = 7$.

Finally, in row 6 we turn to the capitalists' discount rate β. Rather then assuming that capitalists are long-lived, we assume that they are myopic and set $\beta = 0$. By putting more value on current profits, innovation incentives are more sensitive to current changes in policy.

For all cases, we fully recalibrate our model, including all location fundamentals, by matching the exact same moments as in our baseline model.[34] We then perform the counterfactuals of Section 5 in each model. In Table 8 we report a summary of these results.

Robustness: Dynamic Migration

In the last row of Table 8 we report the results of an estimation where we allow workers to be forward-looking. Consider an agent of type ν who currently resides in area o at the beginning of period t. The agent chooses destination $j' \in J$ to solve

$$v^{\nu}_{o,t} = \max_{j \in J} \ln\left(u_{j,t} B^{\nu}_{j,t}\right) - \ln \mu_{jo} + \beta V^{\nu}_{j,t+1} + \tfrac{1}{\varepsilon^D} \varepsilon_{j,t+1}, \tag{OA.1}$$

where $u_{j,t}$ is the indirect utility flow from real income in j, $B^{\nu}_{j,t}$ is the local amenity, μ_{jo} is the moving cost from o to j, and β is the discount factor. The preference shock $\varepsilon_{j,t+1}$ follows a Type-I Extreme Value distribution, and ε^D denotes the dynamic migration elasticity. The model delivers the closed-form choice probability

$$m^{\nu}_{jrt} = \frac{\exp\left(\ln u_{j,t} B^{\nu}_{j,t} - \ln \mu_{jr} + \beta V^{\nu}_{j,t+1}\right)^{\varepsilon^D}}{\sum_{j' \in J} \exp\left(\ln u_{j',t} B^{\nu}_{j',t} - \ln \mu_{j'r} + \beta V^{\nu}_{j',t+1}\right)^{\varepsilon^D}}. \tag{OA.2}$$

To bring Equation (OA.2) to the data, we follow Traiberman (2019) and combine the contemporaneous log-flow $\ln m^{\nu}_{jrt}$ with the one-period-ahead log-flow $\ln m^{\nu}_{oj,t+1}$ across pairs of destinations (j, j'). Using the closed-form expression of $V^{\nu}_{j,t+1}$, the value terms cancel and we obtain

$$\ln \frac{m^{\nu}_{jrt}}{m^{\nu}_{j'rt}} + \beta \ln \frac{m^{\nu}_{oj,t+1}}{m^{\nu}_{oj',t+1}} = \varepsilon^D \left[\ln u_{j,t} B^{\nu}_{j,t} - \ln u_{j',t} B^{\nu}_{j',t}\right] - \varepsilon^D \left(\ln \mu_{jr} - \ln \mu_{j'r}\right) - \beta \varepsilon^D \left(\ln \mu_{oj} - \ln \mu_{oj'}\right). \tag{OA.3}$$

Replacing $B^{\nu}_{j,t}$ with $\mathcal{B}_j \ell^{-\varrho}_{j,t} \left(\varpi^{\nu}_{j,t}\right)^{\iota}$, and using the equilibrium expression for the indirect-

[34] The calibration results for these different specifications are available upon request.

utility flow as in Equation (23), the final form of the equation can be written as

$$\begin{aligned}(\text{OA.4}) \quad \ln\frac{m^{\nu}_{jrt}}{m^{\nu}_{j'rt}} + \beta\ln\frac{m^{\nu}_{oj,t+1}}{m^{\nu}_{oj',t+1}} &= (C_j - C_{j'}) + \iota\varepsilon^D\left(\ln\varpi^{\nu}_{j,t} - \ln\varpi^{\nu}_{j',t}\right) \\ &+ \varepsilon^D\left[\ln\left(1+\tfrac{1}{\theta-1}e^{\nu}_{j,t}\right) - \ln\left(1+\tfrac{1}{\theta-1}e^{\nu}_{j',t}\right)\right] - \varepsilon^D\left(\ln\mu_{jrt} - \ln\mu_{j'rt}\right) \\ &- \beta\varepsilon^D\left(\ln\mu_{oj,t+1} - \ln\mu_{oj',t+1}\right),\end{aligned}$$

where the destination-specific component are defined as

$$C_j \equiv \varepsilon^D\ln(w_j/P_j) - \varrho\varepsilon^D\ln\ell_j + \varepsilon^D\ln\mathcal{B}_j.$$

Estimation Results The only structurally new parameter introduced in the dynamic model is the labor discount factor, β, which we set to 0.35, the same value as the capitalist discount factor in the main text.[35] We hold the migration elasticity ε^D and the labor-congestion elasticity ϱ fixed at their values in the static model reported in Table 6. The parameter of interest is the own-nationality homophily ι, which enters through the coefficient $\iota\varepsilon^D$ on the local nationality share difference $\ln\varpi^{\nu}_{j,t} - \ln\varpi^{\nu}_{j',t}$. To address the endogeneity of local nationality shares, we instrument ϖ^{ν}_{rt} using (i) the shift-share instrument of Card (2001), and (ii) the predicted ancestry distribution instrument following Burchardi et al. (2026).

Table OA.5 reports the estimates of Equation (OA.4). The second column using Card instrument is our preferred specification, since the pre-1920 sample is too short for a reliable first stage of the instrument proposed in Burchardi et al. (2026). We get an estimate of 0.115 for ι, which is not significantly different from the static estimation 0.13. Therefore, we keep ι the same value, 0.13, as in the static migration model for the further analysis.

Calibration of $\mathcal{B}$ Because the dynamic specification introduces a future expectation value term into the original Equation (23), the cross-sectional calibration of $\mathcal{B}_j$ used in the static model can no longer be applied; we therefore exploit the panel structure of the data together with the estimated fixed effects in Equation (OA.4) to recover $\mathcal{B}_j$ in an internally consistent way. From the definition of C_j, the local amenity is recovered as [36]

$$\mathcal{B}^{\text{DYN}}_j = \exp\left(\tfrac{1}{\varepsilon^D}C_j - \ln(w_j/P_j) + \varrho\ln\ell_j\right),$$

where superscript DYN indicates that $\mathcal{B}_j$ is consistent with agents' dynamic migration problem. In Figure OA.10 we compare $\mathcal{B}^{\text{DYN}}_j$ with its counterpart from the static model

[35] We also consider an alternative calibration with $\beta = 0.6$. After recalibrating all other parameters, the results are quantitatively similar. Interested readers may request these results from the author or reproduce them by modifying the relevant parameter values in the replication package.

[36] This recovers $\mathcal{B}$ as a residual that is, by construction, common across types ν. In the counterfactual exercises we allow the differential urban pull of immigrants and skilled workers as in Equation (16).

TABLE OA.5: Estimation of Own-nationality Homophily ι in dynamic model

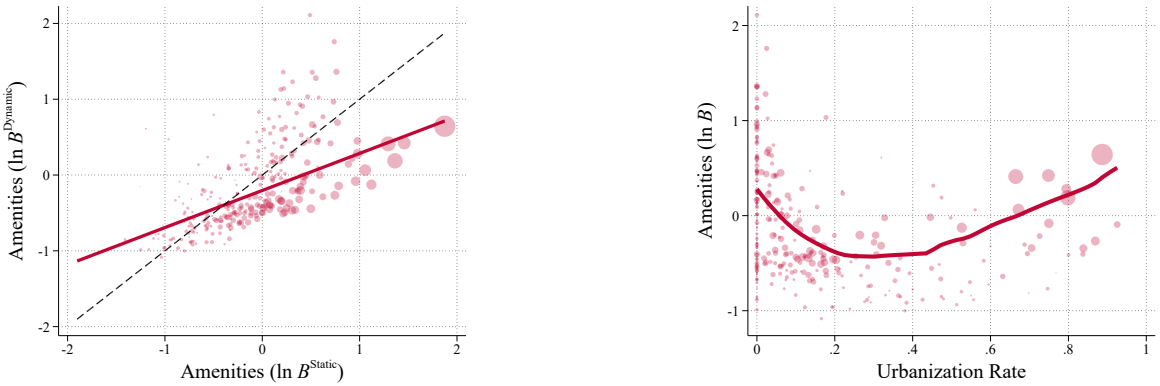
	(1)	(2)	(3)
Log Nationality Share Diff	0.167***	0.220***	0.388***
	(0.000)	(0.000)	(0.000)
Innovator Share Diff	52.038***	60.780***	90.370***
	(0.094)	(0.105)	(0.119)
Log Trade Cost Diff	-0.766***	-0.778***	-0.763***
	(0.000)	(0.000)	(0.000)
Forward Log Trade Cost Diff	-0.559***	-0.575***	-0.557***
	(0.000)	(0.000)	(0.000)
J-K FE	Yes	Yes	Yes
IV	No	Card	Burchardi et al.
R^2		.307	.296
N	104,004,134	99,838,374	102,468,612

Notes: This table reports estimates of equation (OA.4). All specifications include origin-destination (J-K) fixed effects. Column (1) presents OLS estimates. Column (2) presents estimates using the instrument from Card (2001). Column (3) presents estimates using the Burchardi et al. (2026) instrument, extended to include the contemporaneous year (no leave-one-out). Column (4) uses the Burchardi instrument in leave-one-out form, which excludes the own-country share when constructing predicted immigrant stocks. Column (5) uses the Burchardi leave-one-out instrument but omits the recursive term in its construction. Standard errors are reported in parentheses.

and plot it as a function of the urbanization rate. Panel A shows that the dynamic estimates are positively correlated with the static ones. Panel B shows that they remain positively related to urbanization, even though amenities are a bit larger at the very low levels of urbanization.

FIGURE OA.10: LOCAL AMENITIES: COMPARISON TO THE STATIC MODEL

(A) Dynamic vs. Static $\mathcal{B}$ (B) $\mathcal{B}$ vs. Urbanization Rate

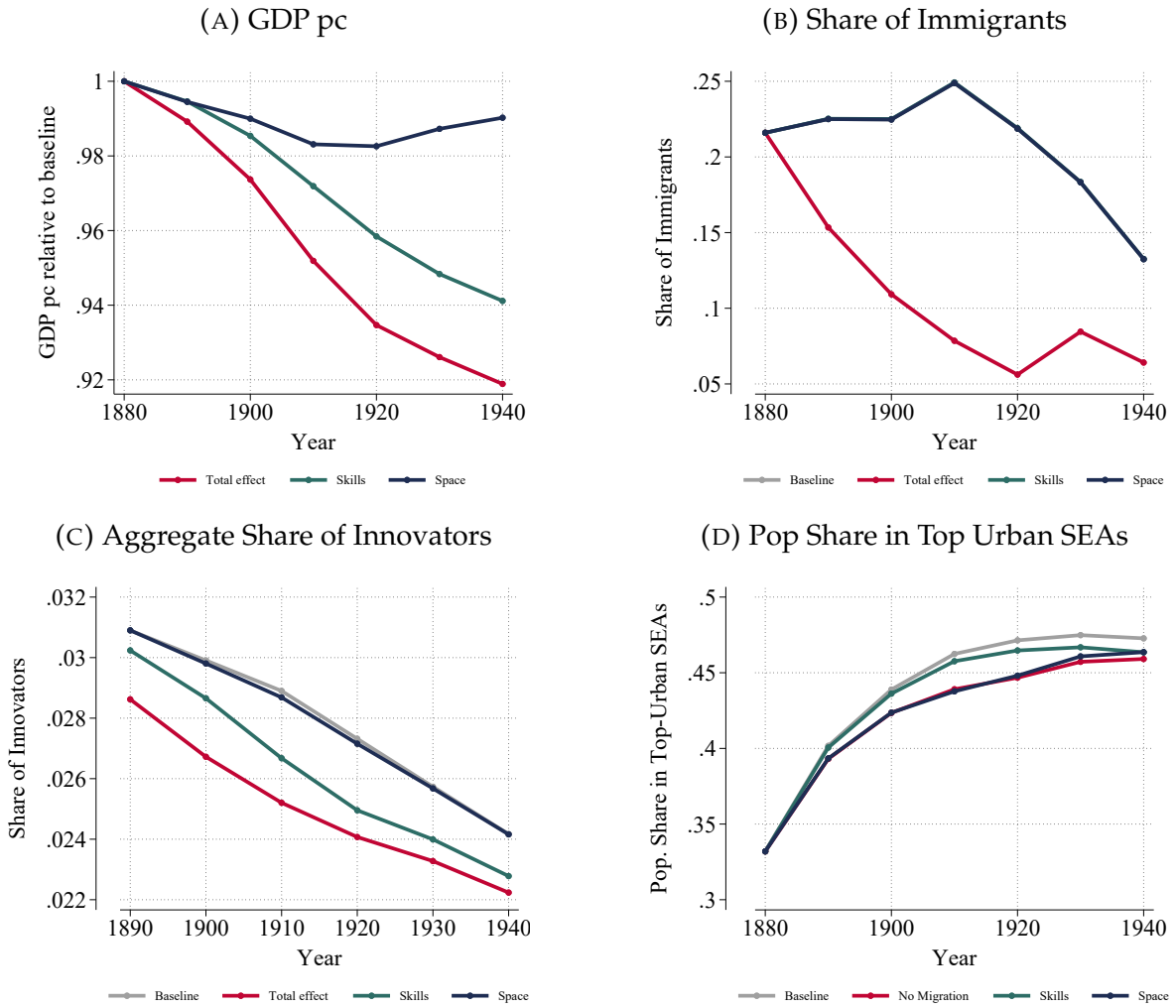


Notes: Each marker is an SEA. We normalize the average value of $\mathcal{B}$ to be 1. In panel (A) the horizontal axis reports $\ln \mathcal{B}$ from the static model and the vertical axis reports $\ln \mathcal{B}$ recovered from (OA.4); the dash line is the 45-degree line. In panel (B) the horizontal axis is the urbanization rate and the vertical axis reports the $\ln \mathcal{B}$ from (OA.4).

Counterfactual Analysis We then replicate the counterfactual exercises of the main text using the parameters calibrated in the dynamic model. As Figure OA.11 shows, the dynamic results closely mirror their static counterparts across all four aggregate margins: GDP per capita, the immigrant share, the share of innovators, and the population share in

top-urbanized SEAs. This is consistent with the last row of Table 8, that also showed very similar results between our baseline model and the model with forward-looking workers.

FIGURE OA.11: THE AGGREGATE IMPACT OF INTERNATIONAL MIGRATION (DYNAMIC MIGRATION)



Notes: The figure shows the change in aggregate income per capita, relative to the baseline economy (Panel A), the aggregate share of immigrants (Panel B), the overall share of people engaged in innovation (Panel C), and the share of the population living in the most urbanized locations. 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 yellow and the "no urban bias" counterfactual, where immigrants initially settle in proportion to the spatial allocation of natives in blue.

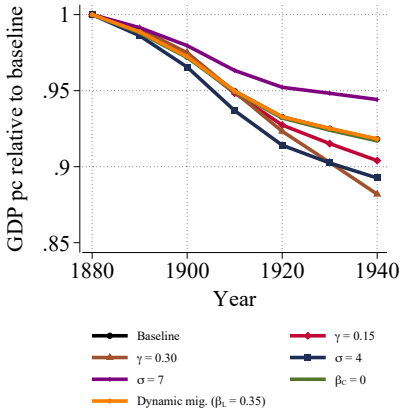
OA.4.1 Robustness Summary

Figure OA.12 examines the robustness of the counterfactual GDP per capita effects across alternative parameters. The qualitative patterns are stable across specifications: shutting down migration or skilled migration lowers GDP per capita in the migration-restriction counterfactuals, while removing the 1921/24 quotas raises GDP per capita over the longer horizon. The magnitudes are more sensitive to the parameters governing global knowledge spillovers, γ, and the elasticity of substitution, σ, than to the two discount-

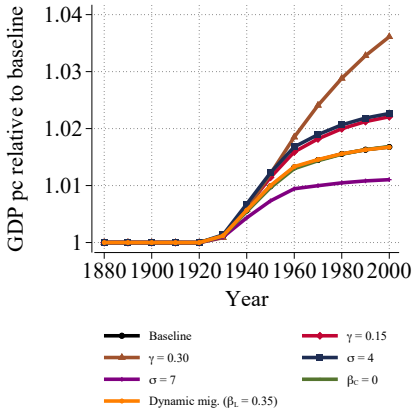
factor assumptions. In particular, changing γ or σ generates visibly larger variation in the counterfactual paths, whereas the no-capitalist-discounting case and the forward-looking migration specification remain close to the baseline in most panels.

FIGURE OA.12: ROBUSTNESS OF THE COUNTERFACTUAL GDP PER CAPITA EFFECTS

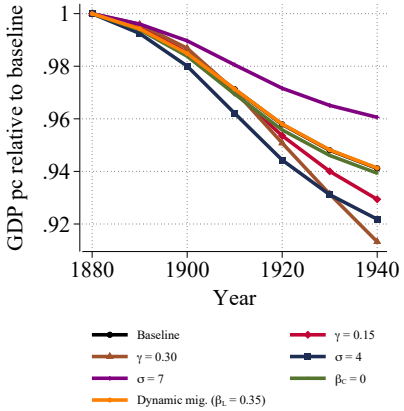
(A) No Migration (Counterfactual I)



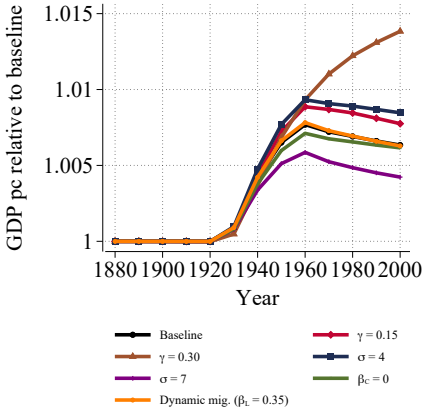
(B) No Migration (Counterfactual II)



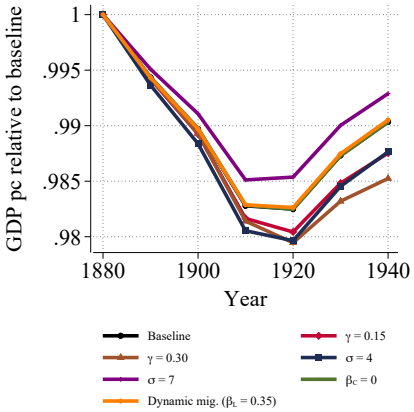
(C) No Skilled Migration (Counterfactual I)



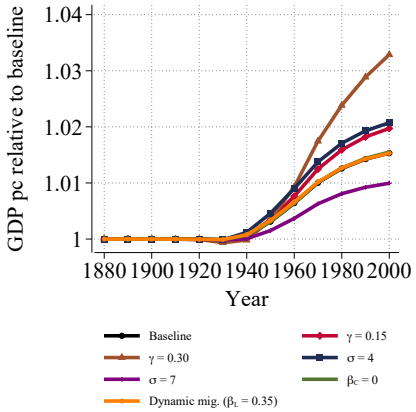
(D) No Skilled Migration (Counterfactual II)



(E) No Urban Bias (Counterfactual I)



(F) No Urban Bias (Counterfactual II)



Notes: Each panel plots model GDP per capita under a counterfactual relative to the baseline. Rows correspond to the three counterfactual families: shutting down all immigration (Panels A–B), shutting down skilled immigration (Panels C–D), and removing the urban settlement bias of immigrants (Panels E–F). The left column (Panels A, C, E) covers the 1880–1940 migration-restriction counterfactuals; the right column (Panels B, D, F) covers the 1880–2000 horizon that removes the 1921/24 quotas. Within each panel we overlay the baseline specification together with all robustness specifications: global knowledge spillovers $\gamma \in \{0.15, 0.30\}$, elasticity of substitution $\sigma \in \{4, 7\}$, no capitalist discounting $\beta_C = 0$, and forward-looking labor migration with $\beta_L = 0.35$.

OA.5. Construction of Instruments for Nationality Shares

This appendix describes the construction of the instruments for $\ln \varpi_{rt}^n$ in equation (17). The parameter of interest is $\iota\varepsilon$, which is the coefficient on $\ln \varpi_{rt}^n$, the share of nationality n in the immigrant stock of region r at time t. Since this variable may be endogenous to local conditions, we use two alternative shift-share instruments: a Card-style instrument and a Burchardi-style predicted-ancestry instrument. The Card instrument is our preferred specification, while the Burchardi-style instrument is used as a robustness check.

Card instrument. Let A_{rt}^n denote the immigrant stock of nationality n in region r at time t. Since in the region-level data we observe immigrant stocks rather than flows, we construct immigrant flows as

$$M_{rt}^n = A_{rt}^n - A_{r,t-1}^n. \tag{OA.5}$$

For 1880, we assume that the stock before 1880 is zero, so that

$$M_{r1880}^n = A_{r1880}^n. \tag{OA.6}$$

The total inflow of nationality n to the United States at time t is

$$M_t^n = \sum_r M_{rt}^n. \tag{OA.7}$$

The Card shift-share term is constructed as

$$Z_{rt}^{n,C} = M_t^n \frac{A_{r,t-1}^n}{A_{t-1}^n}, \qquad A_{t-1}^n = \sum_r A_{r,t-1}^n. \tag{OA.8}$$

The first-stage equation is

$$A_{rt}^n = \delta_t + \delta_{rt} + \delta_{nt} + \beta_1 A_{r,t-1}^n + \beta_2 M_t^n \frac{A_{r,t-1}^n}{A_{t-1}^n} + \epsilon_{rt}^n. \tag{OA.9}$$

Using the fitted value $\widehat{A}_{rt}^n$ from (OA.9), we construct the predicted nationality share as

$$\widehat{\varpi}_{rt}^n = \frac{\widehat{A}_{rt}^n}{\sum_{n'} \widehat{A}_{rt}^{n'}}. \tag{OA.10}$$

The instrument for $\ln \varpi_{rt}^n$ is then $\ln \widehat{\varpi}_{rt}^n$.

Burchardi-style instrument. The Burchardi-style instrument, proposed by Burchardi et al. (2026), uses the interaction between an origin-specific push factor and a destination-

specific pull factor to generate predicted immigrant stocks. In the original construction, the push factor from o to d is the total number of migrants arriving in the United States from o at time t, excluding migrants from o who settled in the same region as d. The pull factor from o to d is the fraction of migrants from other origins who settle in d at time t, excluding migrants from the same continent as o.

The starting point is the following linear formulation of immigrant stock:

$$A_{od}^{t} = a_t + a_{ot} + a_{dt} + b_t A_{od}^{t-1} + I_o^t \left(c_t \frac{I_d^t}{I^t} + d_t \frac{A_{od}^{t-1}}{A_o^{t-1}} \right) + \nu_{od}^t. \tag{OA.11}$$

Solving this equation recursively gives

$$A_{od}^{t'} = \sum_{t=initial}^{t'} \left(a_t + a_{ot} + a_{dt} + c_t I_o^t \frac{I_d^t}{I^t} + \nu_{od}^t \right) \prod_{\tau=t+1}^{t'} (b_\tau + d_{o\tau} I_o^\tau), \tag{OA.12}$$

with zero initial stock for all origin-destination pairs.

Following the original leave-out logic, the aggregate origin push and destination pull terms are replaced by leave-out versions:

$$I_o^t \quad \text{is replaced by} \quad I_{o,-r(d)}^t, \tag{OA.13}$$

and

$$\frac{I_d^t}{I^t} \quad \text{is replaced by} \quad \frac{I_{-c(o),d}^t}{I_{-c(o)}^t}. \tag{OA.14}$$

This adjustment avoids constructing the push or pull factor from flows that may be directly affected by the same origin-destination shocks.

In our implementation, to produce the shift-share, we use a slightly different leave-out structure:

$$Z_{rt}^{n,B} = M_{(-r)t}^{n} * \frac{M_{rt}^{-n}}{M_t^{-n}}. \tag{OA.15}$$

That is, for the push factor we exclude the same destination region r rather than the broader region containing r, and for the pull factor we exclude the same nationality n rather than the same continent as n. This is different from the original construction and may be problematic if there are common shocks for a broad origin group or for a particular U.S. region.

As in the original paper, we assume that the stocks of immigrants before 1880 were zero, and we treat the immigrant stock in 1880 as the immigrant flow for 1880.

Let $\mathcal{R}^n_{rt}$ denote the vector of recursive interaction terms generated by the recursive solution of the immigrant-stock equation. In our specification, the first-stage equation for the decade 1920 is given by

$$A^n_{r1920} = \beta_0 + \delta_{rt} + \delta_n + \beta_1 Z^{n,B}_{r1880} + \beta_2 Z^{n,B}_{r1900} + \beta_3 Z^{n,B}_{r1910} + \Gamma'_{1920}\mathcal{R}^n_{r1920} + \epsilon^n_{rt}, \tag{OA.16}$$

where $\mathcal{R}^n_{r1920}$ contains the 11 recursive interaction terms.

For 1910, the first-stage equation is

$$A^n_{r1910} = \beta_0 + \delta_{rt} + \delta_n + \beta_1 Z^{n,B}_{r1880} + \beta_2 Z^{n,B}_{r1900} + \Gamma'_{1910}\mathcal{R}^n_{r1910} + \epsilon^n_{rt}, \tag{OA.17}$$

where $\mathcal{R}^n_{r1910}$ contains the 4 recursive interaction terms.

For 1900, the first-stage equation is

$$A^n_{r1900} = \beta_0 + \delta_{rt} + \delta_n + \beta_1 Z^{n,B}_{r1880} + \Gamma'_{1900}\mathcal{R}^n_{r1900} + \epsilon^n_{rt}, \tag{OA.18}$$

where $\mathcal{R}^n_{r1900}$ contains the recursive interaction term

$$M^n_{(-r)1900} M^n_{(-r)1880} * \frac{M^{-n}_{r1880}}{M^{-n}_{1880}}. \tag{OA.19}$$

Similar to the Card IV, we use the fitted value $\widehat{A}^n_{rt}$ from the first stage to construct the predicted nationality share:

$$\widehat{\varpi}^n_{rt} = \frac{\widehat{A}^n_{rt}}{\sum_{n'} \widehat{A}^{n'}_{rt}}. \tag{OA.20}$$

The instrument for $\ln \varpi^n_{rt}$ is then $\ln \widehat{\varpi}^n_{rt}$.

REFERENCES

ABRAMITZKY, R. AND L. BOUSTAN (2022): *Streets of gold: America's untold story of immigrant success*, Hachette UK.

ABRAMITZKY, R., L. BOUSTAN, K. ERIKSSON, J. FEIGENBAUM, AND S. PÉREZ (2021): "Automated Linking of Historical Data," *Journal of Economic Literature*, 59, 865–918.

ABRAMITZKY, R., L. BOUSTAN, K. ERIKSSON, S. PÉREZ, AND M. RASHID (2025): "Census Linking Project: Version 3.0 [dataset]," https://censuslinkingproject.org.

ABRAMITZKY, R., L. P. BOUSTAN, AND K. ERIKSSON (2012): "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102, 1832–56.

——— (2014): "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 122, 467–506.

ABRAMITZKY, R., R. MILL, AND S. PÉREZ (2020): "Linking individuals across historical sources: A fully automated approach," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53, 92–111.

AKCIGIT, U. (2017): "Economic Growth: The Past, the Present, and the Future," *Journal of Political Economy*, 125, 1736–1747.

AKCIGIT, U., J. GRIGSBY, AND T. NICHOLAS (2017): "Immigration and the Rise of American Ingenuity," *American Economic Review*, 107, 327–31.

ALLEN, T. AND C. ARKOLAKIS (2014): "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 129, 1085–1140.

ALTONJI, J. G. AND D. CARD (1991): "The effects of immigration on the labor market outcomes of less-skilled natives," in *Immigration, trade, and the labor market*, University of Chicago Press, 201–234.

BERGEAUD, A. AND C. VERLUISE (2022): "PatentCity: a dataset to study the location of patents since the 19th century," .

——— (2024): "A New Dataset to Study a Century of Innovation in Europe and in the US," *Research Policy*, 53, 104903.

BERNSTEIN, S., R. DIAMOND, A. JIRANAPHAWIBOON, T. MCQUADE, AND B. POUSADA (2022): "The contribution of high-skilled immigrants to innovation in the United States," Tech. rep., National Bureau of Economic Research.

BILAL, A. (2023): "The geography of unemployment," *The Quarterly Journal of Economics*, 138, 1507–1576.

BILAL, A. AND E. ROSSI-HANSBERG (2021): "Location as an Asset," *Econometrica*, 89, 2459–2495.

BLOOM, N., C. I. JONES, J. VAN REENEN, AND M. WEBB (2020): "Are Ideas Getting Harder to Find?" *American Economic Review*, 110, 1104–1144.

BRODA, C. AND D. WEINSTEIN (2006): "Globalization and the Gains from Variety," *Quarterly Journal of Economics*, 121, 541–585.

BURCHARDI, K. B., T. CHANEY, T. A. HASSAN, L. TARQUINIO, AND S. J. TERRY (2026): "Immigration, Innovation, and Growth," *American Economic Review*, 116, 828–861.

BURSTEIN, A., G. HANSON, L. TIAN, AND J. VOGEL (2020): "Tradability and the Labor-Market Impact of Immigration: Theory and Evidence From the United States," *Econometrica*, 88, 1071–1112.

CARD, D. (1990): "The Impact of the Marial Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

——— (2001): "Immigrant inflows, native outflows, and the local labor market impacts of higher immigration," *Journal of Labor Economics*, 19, 22–64.

CREWS, L. G. (2023): "A Dynamic Spatial Knowledge Economy," Working Paper.

DESMET, K., D. K. NAGY, AND E. ROSSI-HANSBERG (2018): "The Geography of Development," *Journal of Political Economy*, 126, 903–983.

DIAMOND, R., T. MCQUADE, AND F. QIAN (2019): "The Effects of Rent Control Expansion on Tenants, Landlords, and Inequality: Evidence from San Francisco," *American Economic Review*, 109, 3365–94.

DUSTMANN, C., U. SCHÖNBERG, AND J. STUHLER (2017): "Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment," *The Quarterly Journal of Economics*, 132, 435–483.

ECKERT, F. AND M. PETERS (2022): "Spatial Structural Change," Working Paper.

FAJGELBAUM, P. D., E. MORALES, J. C. SUÁREZ SERRATO, AND O. M. ZIDAR (2019): "State Taxes and Spatial Misallocation," *The Review of Economic Studies*, 86, 333–376.

FEIGENBAUM, J. J. (2016): "Machine Learning Approach to Census Record Linking," Tech. rep.

FUSHIKI, T. (2011): "Estimation of prediction error by using K-fold cross-validation," *Statistics and Computing*, 21, 137–146.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): in *The elements of statistical learning*, Berlin, Germany: Springer.

HORNBECK, R. AND M. ROTEMBERG (2024): "Growth Off the Rails: Aggregate Productivity Growth in Distorted Economies," *Journal of Political Economy*, 132, 3547–3602.

JARO, M. A. (1989): "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

——— (1995): "Probabilistic linkage of large public health data files," *Statistics in medicine*, 14, 491–498.

JONES, C. (2005): "Growth and Ideas," in *Handbook of Economic Growth*, ed. by P. Aghion and S. Durlauf, Elsevier, vol. 1, Part B, chap. 16, 1063–1111, 1 ed.

JONES, C. I. (1995): "R&D-based Models of Economic Growth," *Journal of Political Economy*, 103, 759–784.

——— (2022): "The past and future of economic growth: A semi-endogenous perspective," *Annual Review of Economics*, 14, 125–152.

KERR, S. P., W. KERR, ÇAĞLAR ÖZDEN, AND C. PARSONS (2016): "Global Talent Flows," *Journal of Economic Perspectives*, 30, 83–106.

KERR, W. (2018): *The Gift of Global Talent: How Migration Shapes Business, Economy & Society*, Stanford University Press.

KLEINMAN, B., E. LIU, AND S. J. REDDING (2023): "Dynamic spatial general equilibrium," *Econometrica*, 91, 385–424.

KONERU, K., V. S. V. PULLA, AND C. VAROL (2016): "Performance Evaluation of Phonetic Matching Algorithms on English Words and Street Names - Comparison and Correlation," in *Proceedings of the 5th International Conference on Data Management Technologies and Applications - Volume 1: DATA*, Insticc, SciTePress, 57–64.

KORTUM, S. (1997): "Research, Patenting, and Technological Change," *Econometrica*, 65, 1389–1419.

KRUGMAN, P. (1980): "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, 70, 950–959.

LEE, S. K. (2019): "Essays in History and Spatial Economics with Big Data," Working Paper.

MANSON, S., J. SCHROEDER, D. VAN RIPER, AND S. RUGGLES (2019): "IPUMS National Historical Geographic Information System: Version 14.0 [Database]," .

NAGY, D. K. (2023): "Hinterlands, City Formation and Growth: Evidence from the U.S. Westward Expansion," *The Review of Economic Studies*, 90, 3238–3281.

OBERFIELD, E. AND D. RAVAL (2021): "Micro Data and Macro Technology," *Econometrica*, 89, 703–732.

PEDREGOSA, F., G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, AND E. DUCHESNAY (2011): "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

PERI, G. (2016): "Immigrants, Productivity, and Labor Markets," *The Journal of Economic Perspectives*, 30, 3–30.

PETERS, M. (2022): "Market size and spatial growth—evidence from Germany's post-war population expulsions," *Econometrica*, 90, 2357–2396.

PETRALIA, S., P.-A. BALLAND, AND D. RIGBY (2016): "HistPat Dataset," .

RAJKOVIC, P. AND D. JANKOVIC (2007): "Adaptation and Application of Daitch-Mokotoff Soundex Algorithm on Serbian Names," Tech. rep.

REDDING, S. J. AND E. ROSSI-HANSBERG (2017): "Quantitative Spatial Economics," *Annual Review of Economics*, 9, 21–58.

ROMER, P. M. (1990): "Endogenous Technological Change," *The Journal of Political Economy*, 98, S71–s102.

RUGGLES, S., S. FLOOD, R. GOEKEN, J. GROVER, E. MEYER, J. PACAS, AND M. SOBEK (2020): "Integrated Public Use Microdata Series: Version 10.0 [Machine-readable database]," .

SEQUEIRA, S., N. NUNN, AND N. QIAN (2020): "Immigrants and the Making of America," *The Review of Economic Studies*, 87, 382–419.

TRAIBERMAN, S. (2019): "Occupations and import competition: Evidence from Denmark," *American Economic Review*, 109, 4260–4301.

WALSH, C. (2025): "The Entry Multiplier," Working paper, available at SSRN: https://ssrn.com/abstract=5468409.

WINKLER, W. E. (2014): "Advanced Methods for Record Linkage," Tech. rep.