Online Appendix

"Migrants, Ancestors, and Foreign Investments"

Konrad B. Burchardi

Thomas Chaney

Tarek A. Hassan

A Data Appendix

Overview

To construct the migration and ancestry data up until the year 2000, we download the 1880, 1900, 1910, 1920, 1930, 1970, 1980, and 2000 waves of the Integrated Public Use Microdata Series (IPUMS) from https://usa.ipums.org/usa-action/samples. For each wave, we select the largest available sample; for example, if a 1% and 10% sample was available for 1880 data, we used the 10% sample. To construct the 2010 data, we used the 2006-2010 American Community Survey (ACS) sample provided on the IPUMS website. For a more detailed overview on the specific waves used, see Appendix Table 1.

For each sample, we obtain the following variables: year, datanum, serial, hhwt, region, statefip, county, cntygp97, cntygp98, puma, gq, pernum, perwt, bpl, mbpl, fbpl, nativity, ancestr1, yrimmig, mtongue, mmtongue, fmtongue, and language.

We construct the number of migrants from origin country o to destination county d in t, $I^t_{o,d}$, as well as the measure of ancestry $A^t_{o,d}$ from 1980 onward. We first aggregate the individual-level census data to counts of respondents at the level of historic US counties (or country groups from 1970 onwards) and foreign countries, and then transform the data into 1990 country-county level using various transition matrices. Details are given in the following sections.

How we create transition matrices

We create a set of transition matrices that transform non-1990 countries to 1990 countries and non-1990 counties/county groups to 1990 counties.

- Birthplace-to-country: The aim is to construct transition matrices that map all the birthplace answers into 1990 countries. In each wave of the US Census, respondents were asked to report their country of birth. All possible answers (across time) are listed

here: https://usa.ipums.org/usa-action/variables/BPL#codes_section. The censuses from 1850-2012 contain roughly 550 possible different answers to the question of birthplace. In each census data set, they are saved in the variable "bpld." What follows is our procedure for building those matrices:

1. We start with a transition matrix of zeros, with all possible answers to the 1990 birthplace question as rows and all 1990 countries as columns. A cell in row r and column c of the transition matrix answers the question, "What is the probability that an individual who claims his/her birthplace as r refers to the area that in 1990 is country c?" So all cells contain values in [0,1], and rows sum up to 1.
2. For each row r in the transition matrix, if r with certainty refers to the area that in 1990 is country c, we simply change the entry in cell (r,c) from 0 to 1; if r does refer to an area that in 1990 is in multiple countries, then we search for the 1990 population of each possible country, and assign probabilities in proportion to the population data. We use the population information from the Worldbank database.[1].

Panel A in Appendix Table 2 lists the distribution of weights that we end up using, and the affected countries and persons.

- Ancestry-to-country: The aim is to construct transition matrices that map all the answers to the ancestry question into 1990 countries. The 1980, 1990, 2000, and 2010 census data provide information on the ancestry (ancestr1, 3-digit version). All possible answers (across time) are listed here: https://usa.ipums.org/usa-action/variables/ANCESTR1/#codes_section. The procedure is the same as in the birthplace-to-country procedure. Panel B in Appendix Table 2 lists the distribution of weights that we end up using, and the affected countries and persons.

- Group-to-county & PUMA-to-county: The aim is to construct transition matrices that map all the county groups/PUMAs into individual counties. For the years 1970 and 1980, the US census data are at the US county group level. A "county group" is an agglomeration of US counties. For the years 2000 and 2010, the census data are at the PUMA level. A "PUMA" is also an agglomeration of US counties.[2] To construct transition matrices from county agglomeration level to county level, we download the corresponding matching files from the IPUMS website. We use data on the population of each county (within each county group/PUMA) to assign a probability that an observation from county group/PUMA g in year t is from county c in year t. This approach gives a transition matrix from year t county groups to year t counties. Appendix Table 3 lists the distribution of weights that we end up using, and the affected counties and persons.

[1] http://data.worldbank.org/indicator/SP.POP.TOTL

[2] Detailed description of "county group" and "PUMA" can be found here: https://usa.ipums.org/usa/volii/tgeotools.shtml.

- County-to-county: The aim is to construct transition matrices that map all the non-1990 counties into 1990 counties. This step is necessary because the list and boundaries of US counties changed over time. Similarly to the birthplace-to-country and ancestry-to-country procedure, we use one transition matrix per census year (1880, 1900, 1910, 1920, 1930, 1970, 1980, 2000, 2010). Such a transition matrix has as rows all US counties, indexed c, in year t, and as columns all 1990 US counties, indexed m. Each cell of the transition matrix takes a value that answers the question, "Which fraction of the area of the county c in year t is in 1990 part of county m?" Appendix Table 3 lists the distribution of weights that we end up using, and the affected counties and persons. More specifically, we build these matrices as follows:

    1. We download the year-specific map files. For 1880 us counties, we obtain the 503MB GIS file from Atlas: http://publications.newberry.org/ahcbp/downloads/united_states.html and extract the 1880 part. For 1900, 1910, 1920, and 1930 counties, we obtain the maps from IPUMS: https://usa.ipums.org/usa/volii/ICPSR.shtml. Finally, for 1970, 1980, and 1990 counties, we obtain the maps from NHGIS: https://data2.nhgis.org/main.
    2. We project non-1990 maps onto 1990 counties. We used the intersect command in ArcGIS to map year-specific counties onto 1990 counties based on area. This approach gives a transition matrix from non-1990 counties to 1990 counties.

Appendix Table 1: Description of each IPUMS wave

Wave	Description
1880	We use the 10% sample with oversamples; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1900	We use the 5% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1910	We use the 1% sample; the sample is unweighted; we use the region identifiers statefip and county.
1920	We use the 1% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1930	We use the 5% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1970	We use the 1% Form 1 Metro sample; the sample is unweighted; we use the region identifiers statefip and cntygp97 (county group 1970); note that only four states can be completely identified because metropolitan areas that straddle state boundaries are not assigned to states; identifies every metropolitan area of 250,000 or more.
1980	We use the 5% State sample; the sample is unweighted; we use the region identifiers statefip and cntygp98 (county group 1980); the sample identifies all states, larger metropolitan areas, and most counties over 100,000 population.
1990	We use the 5% State sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and puma; the sample identifies all states, and within states, most counties or parts of counties with 100,000 or more population.
2000	We use the 5% Census sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use region identifiers statefip and puma; the sample identifies all states, and within states, most counties or parts of counties with 100,000 or more population.
2010	We use the American Community Service (ACS) 5-Year sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use region identifiers statefip and puma, which contain at least 100,000 persons; the 2006-2010 data contains all households and persons from the 1% ACS samples for 2006, 2007, 2008, 2009 and 2010, identifiable by year.

Appendix Table 2: Historical birthplace to current country: transition matrices

Panel A: Birthplace		weights $\in (0,1)$	weight = 1	weights = 0
1880	# of answers	22	258	9
	# of persons	26,301	50,177,184	4,933
	% of persons	0.05%	99.94%	.01%
1900	# of answers	15	131	6
	# of persons	23,345	6,555,140	5,339
	% of persons	0.35%	99.56%	.08%
1910	# of answers	20	99	4
	# of persons	31,072	5,613,136	3,105
	% of persons	0.55%	99.39%	.05%
1920	# of answers	13	174	7
	# of persons	36,070	3,905,455	12,559
	% of persons	0.91%	98.77%	.32%
1930	# of answers	25	194	9
	# of persons	35,930	3,086,341	61,462
	% of persons	1.13%	96.94%	1.93%
1970	# of answers	12	77	3
	# of persons	318,800	6,323,100	230,800
	% of persons	4.64%	92.00%	3.36%
1980	# of answers	32	222	7
	# of persons	491,760	4,774,820	313,300
	% of persons	8.81%	85.57%	5.61%
1990	# of answers	24	209	7
	# of persons	721,595	8,532,585	484,433
	% of persons	7.41%	87.62%	4.97%
2000	# of answers	11	136	0
	# of persons	1,122,532	13,144,632	0
	% of persons	7.87%	92.13%	0%
2010	# of answers	14	137	1
	# of persons	1,302,255	11,131,046	17,148
	% of persons	10.46%	89.40%	.14%
2010*	# of answers	14	188	1
	# of persons	3,512,123	300,415,680	37,469
	% of persons	1.16%	98.83%	.01%
Panel B: Ancestry		weights $\in (0,1)$	weight = 1	weights = 0
1980	# of answers	29	227	143
	# of persons	924,400	198,525,616	27,412,380
	% of persons	0.41%	87.51%	12.08%
1990	# of answers	29	239	9
	# of persons	2,941,941	217,720,512	27,445,182
	% of persons	1.19%	87.75%	11.06%
2000	# of answers	17	137	22
	# of persons	6,000,639	191,300,704	84,120,558
	% of persons	2.13%	67.98%	29.9%
2010	# of answers	19	142	30
	# of persons	8,454,279	229,211,968	66,299,030
	% of persons	2.78%	75.41%	21.81%

The table reports statistics on the transition of data from the 'answer' level to 1990 country level. For each survey wave, and each question – birthplace in Panel A and primary ancestry in Panel B – the table reports the number of answers that can be directly linked to a 1990 country (weight = 1), that are assigned to several 1990 countries using population weights (weights $\in (0,1)$) and that cannot be linked to any modern country with sufficient certainty (weights = 0). The table also reports the number of respondents (scaled from the original data using the person weights provided) in each category. Answers with weights zero essentially consists of "Not Reported" (e.g. 23, 24, 54 and 30 million respondents for the 1980, 1990, 2000 and 2010 ancestry data, respectively) and "African-American" (e.g. 26, 22 and 25 million respondents for the 1990, 2000 and 2010 ancestry data, respectively). The remainders are mostly cases such as "African", "Uncodable", "Bohemian", "Nuevo Mexicano", "Other", etc. In Panel A, all years except 1880 consist of the number of persons that report birthplace since the last Census wave. For the 2010 Census wave the additional entry (denoted by a *) reports the respective numbers for all respondents in that wave.

Appendix Table 3: Historical state-county unit to 1990 state-county unit: transition matrices

Census wave		weights $\in (0,1)$	weight $= 1$	weights $= 0$
1880	# of counties	658	1854	1
	% of persons (birthplace data)	21.54%	78.45%	.01%
1900	# of counties	2211	7	4
	% of persons (birthplace data)	99.09%	0.87%	.05%
1910	# of counties	1517	5	1
	% of persons (birthplace data)	99.00%	0.94%	.05%
1920	# of counties	1355	7	0
	% of persons (birthplace data)	90.80%	9.20%	0%
1930	# of counties	1801	6	0
	% of persons (birthplace data)	90.61%	9.39%	0%
1970	# of countygroups	310	98	0
	% of persons (birthplace data)	34.07%	65.93%	0%
1980	# of countygroups	580	573	0
	% of persons (birthplace data)	17.96%	82.04%	0%
	% of persons (ancestry data)	40.02%	59.98%	0%
1990	# of PUMAs	541	1185	0
	% of persons (birthplace data)	8.97%	91.03%	0%
	% of persons (ancestry data)	32.15%	67.85%	0%
2000	# of PUMAs	620	1451	0
	% of persons (birthplace data)	10.66%	89.34%	0%
	% of persons (ancestry data)	30.36%	69.64%	0%
2010	# of PUMAs	619	1449	1
	% of persons (birthplace data)	12.31%	87.65%	.03%
	% of persons (ancestry data)	30.13%	69.81%	.05%

The table reports statistics on the transition of data from the 'historical spatial area' level to 1990 US county level. For each Census wave the table reports the number of contemporaneous spatial areas that are a subset of a 1990 US county (weight = 1) and the number of contemporaneous spatial areas whose data is transitioned to 1990 US county level using non-degenerate weights (weights $\in (0,1)$). For Census waves 1880 to 1930 the share of their contemporaneous county spatial area in each 1990 US county area is used as weight. For waves 1970 to 2010 there are two steps: In step 1 the share of their contemporaneous countygroup (waves 1970 and 1980) or PUMA (waves 1990 to 2010) population in the contemporaneous county population are used as weights; in step 2 the share of their contemporaneous county spatial area in each 1990 US county area is used as weight. The two-step procedure is necessary because the 1970 to 2010 Census waves do not have a county-level identifier (to protect the privacy of the respondents). The table also reports the share of respondents affected by this transition in the birthplace and ancestry data, respectively.

A.1 Details on the construction of migration and ethnicity data

Details calculation of post-1880 flow of immigrants

For each census wave after 1880, we count the number of individuals in each historic US county d who were born in historic country o (as identified by birthplace variable "bpld" in the raw data) that had immigrated to the United States since the last census wave that contains the immigration variable (not always 10 years earlier). Then we transform these data

- from the non-1990 foreign-country ("bpld") level to the 1990 foreign-country level using bpld-to-country transition matrices.
- from the US-county group/puma level to the US-county level using group/puma-to-county transition matrices.
- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.
- from the post-1990 US-county level to the 1990 US county level. Based on the information from https://www.census.gov/geo/reference/county-changes.html, a new county is either created from part of ONE 1990 county or assigned a new FIPS code after 1990, so we manually change that county's FIPS code to what it was in 1990. A few counties' boundaries have been changed after 1990 but that only involved a tiny change in population, so we ignore these differences.

Details calculation of pre-1880 stock of immigrants

For the year 1880, we calculate for each historic US county d the number of individuals who were born in a historic foreign country o (no matter when they immigrated). We add to those calculations the number of individuals in county d who were born in the United States, but whose parents were born in historic foreign country o. (If the parents were born in different countries, we count the person as half a person from the mother's place of birth, and half a person from the father's place of birth). Then we transform these data

- from the pre-1880 foreign-country ("bpld") level to the 1990 foreign-country level using the pre-1880 country-to-country transition matrix.
- from the pre-1880 US-county level to the 1990 US-county level using the pre-1880 county-to-county transition matrix.

Details calculation of stock of ancestry (1980, 1990, 2000, and 2010)

For the years 1980, 1990, 2000, and 2010, we calculate for each US county group the number of individuals who state as primary ancestry ("ancestr1" variable) some nationality/area. We transform the data

- from the ancestry-answer ("ancestr1") level to the 1990 foreign-country level using ancestry-to-country transition matrices.
- from the US-county group/puma level to the US county-level using group/puma-to-county transition matrices.
- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.
- from the post-1990 US-county to the 1990 US-county level. Based on the information from https://www.census.gov/geo/reference/county-changes.html, a new county is either created from part of ONE 1990 county or assigned a new FIPS code after 1990, so we manually change that county's FIPS code to what it was in 1990. A few counties' boundaries have been changed after 1990 but that only involved a tiny change in population, so we ignore the difference.

A.2 Details on the construction of FDI data

Our FDI data are from the US file of the Bureau van Dijk ORBIS dataset. For each US firm, the raw data set lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. It also provides the number of employees for both US and foreign firms. The steps for building the data follow below.

Clean postcode information

We use firm's postcode as a unique identifier for the county location of the US firm, and then need to ensure that one county uniquely corresponds to one postcode. Vance, NC; Wakulla, FL; Citrus, FL; Rankin, MS; Union, OH; and Du Page, IL share at least one postcode with a neighboring county. In each case we assign that postcode wholly to the county with the larger population (according to Google 2012 population data). In the last step, we hand-coded missing postcodes that we took from main data set. Only one such case existed: 75427 for Dallas.

Build the parent data

We used the following variables from the parent dataset: "Mark" "Company name" "BvD ID number" "Country ISO Code" "City" "Postcode" "NAICS 2007 Core code (4 digits)" "NAICS,

text description" "Number of employees 2013" "Shareholder - Name" "Shareholder - BvD ID number" "Shareholder - City" "Shareholder - Postal code" "Shareholder - NAICS 2007, Core code" "Shareholder - NAICS 2007, text description" "Shareholder - Country ISO code" "Shareholder - Direct %" "Shareholder - Total %" "Shareholder - Number of employees". Here "shareholder" is equivalent to "parent" in our context. The key data-building steps are as follows:

1. Assign numerical values to "Shareholder Direct" and "Shareholder Total":

   - When the stake of a shareholder is described by an acronym rather than a number, we replace it with numerical values as follows: MO, majority owned, is replaced by "75%"; JO, jointly owned, is replaced by "50%"; NG, negligent, is replaced by '0%'; BR, branch and WO, wholly owned are both replaced by "100%".[3]
   - When the stake of a shareholder is described by the following expressions, we replace it with a numerical value as follows: Values with a ">", e.g., " $>$ 25.00" were replaced by the original number plus 10; values with a "<", e.g., " $<$ 34.00", were replaced by the original number minus 10; values with a "±", e.g. "±25.00", were replaced by the original number.

2. Postcode matching: We matched both US firms and US parents (foreign parents were ignored in this step), with our postcode data. Besides the original string variable postcode, we generated new variables postcode5digit and postcodeextension and labeled them "Postal code (5 digit)" and "Postal code (extension)." Similarly, shareholders had shareholderpostcodeUS5digit and shareholderpostcodeUSextension (note the spelling postal code in shareholder variables was unified to postcode).

3. Country-code matching: We matched both companies and their parents. Each firm had four country variables: numerical country code, country name, and 2- and 3- digit ISO country code. Then we adjusted those 2014 country codes to 1990 codes based on the information on post-1990 country changes.

Build the subsidiary data

We used the following variables from the subsidiary dataset: "Mark" "Company name" "BvD ID number" "Country ISO Code" "City" "Postcode" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "Number of employees 2007" "Subsidiary - Name" "Subsidiary - BvD ID number" "Subsidiary - Country ISO code" "Subsidiary - City" "Subsidiary - Postal code" "Subsidiary - NAICS 2007, Core code" "Subsidiary - NAICS 2007, text description" "Subsidiary - Number of employees" "Subsidiary - Direct %" "Subsidiary - Total%" "Branch - Name" "Branch - BvD ID number" "Branch - Country ISO code" "Branch - City" "Branch - Postcode" "Branch - NAICS

[3]See http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2407845 for reference.

2007, Core code" "Branch - NAICS 2007, text description" "Branch - Number employees". The data cleaning process is identical to that of the parent data described above, with the exception that we merged subsidiaries with branches and refer to them collectively as "subsidiaries".

A.3 Details on the construction of other data

International trade.— The data on trade between US states and foreign countries, both at the aggregate level and at the sectoral level, are from the Commodity Flow Survey for the year 2012. The data are collected by the US Census Bureau. A representative sample of establishments are surveyed every five years, and information on their shipments collected. The value of all shipments crossing the US international border are recorded as international trade, along with their foreign origin/destination country. We only used thes readily available data aggregated at the US state and foreign country level. Although they do not cover all of the US foreign trade (the data com from a representative survey, not from the universe of foreign transactions), they are the only publicly available source of international data disaggregated at a geographic level below that of the entire United States. For each origin country and destination state, $Import_{o,d}$ are aggregate imports (in dollars) from country o to US state d in 2012, and $Export_{o,d}$ are aggregate exports (in dollars) from US state d to country o in 2012, where we keep the convention of using o for foreign countries and d for US administrative units, states or counties.

Bilateral distances and latitude differences.— To compute the distance between US counties or states and foreign countries, we used the coordinates for all postal codes within a county or state, and the coordinates of the main city for foreign countries.[4] We define the latitude and longitude of a US county as the unweighted average of the latitudes and longitudes of all postal codes within the county. We define the latitude and longitude of a US state as the unweighted average of the latitude and longitude of all counties within the state. The distance between foreign country o and a US county or state d, $Distance_{o,d}$, is computed as the great circle distance between the two, measured in kms. The latitude difference between a foreign country o and a US county or state d, $Latitude\,Difference_{o,d}$, is the absolute difference between the latitudes of the two, measured in degrees.

Country characteristics.— To shed light on the mechanism through which the presence of foreign ancestry affects the patterns for foreign investment, we constructed several measures of foreign country and US county characteristics. *"Genetic Distance"* is a measure of the genetic distance between a given foreign country and the United States, normalized to take values between 0 and 1. *"Linguistic Distance"* is a measure of the linguistic distance between a given foreign country and the United States; it measures the probability that a randomly selected person in the United States speaks the same language as a randomly selected person from that

[4]The geo-coordinates are downloaded from www.geonames.organd www.cepii.fr, respectively. When a county has multiple postcodes we randomly select one of them and use the geocoordinates for that randomly selected postcode.

country. *"Religious Distance"* measures the religious distance between a given foreign country and the United States, with a similar construction as the linguistic distance.[5] A higher index for *"Genetic Distance"*, *"Linguistic Distance"*, or *"Religious Distance"* corresponds to a greater distance between the United States and that country. *"Judicial Quality"* is a measure of the judicial quality in a given country.[6] A higher index for *"Judicial Quality"* corresponds to a higher-quality judicial system. *"Ethnic Diversity"* is a measure of a country's ethnolinguistic fractionalization.[7]

US county characteristics.— We define three US-county level measures. *"Diversity of Ancestries"* is a measure of the diversity of communities from different ancestries in a given US county.[8] *"Foreign Share"* measures the share of residents in a given county who claim foreign ancestry.

Sectoral characteristics.— We separated sectors into final consumption goods and intermediate inputs. To do so, we use the measure of upstreamness from Antràs et al. (2012). We classified 4-digit NAICS sectors as "final goods" if their upstreamness index is below 2, and as "intermediates" if their upstreamness index is above 2.

A.4 Details on the construction of information demand indices

The Information Demand Index is based on data gathered from Google and created in three steps. In the first step we identify five prominent individuals from country o in category p, where $p \in \{\text{actors, athletes, musicians, politicians}\}$. In the second step we utilise Google Trends to obtain data on the spatial variation in the relative frequency of search queries related to these individuals. In the last step we construct indices of the search intensity related to country o in destination d.

Step 1: To identify the top five prominent individuals from o in each category p, we utilise a tool called Google's featured snippet box. Google's featured snippet box is a response to a search query that is generated by Google and pushed to the top of the result list. Google generates these answers by scraping its top results and using an algorithm to provide what it determines to be the most relevant answer.[9] For our purposes we record the top five names in Google's featured snippet box in response to the query "notable [country] [p]", where [country] is one of the 100 largest countries by 2015 population. For example, searching for "notable Belgium actors" yields

[5] Both genetic and religious distance measures come from Spolaore and Wacziarg (2015).

[6] The measure of judicial quality comes from Kaufmann et al. (2003) and is used in Nunn (2007). It is based on a weighted average of variables measuring perceptions of the effectiveness of the judiciary and the enforcement of contracts.

[7] The measure of fractionalization comes from Alesina et al. (2003). It is equal to 1 minus the Herfindahl index of ethnolinguistic group shares.

[8] It is equal to 1 minus the Herfindahl index of ancestry, measured as the sum of squared fractions of all possible ancestry among people who report foreign ancestry within that US county

[9] See https://support.google.com/webmasters/answer/6229325?hl=en

Google's featured snippet box with an ordered list of Belgian actors. We save the top five names from left to right as the set of search queries $q(o,p)$. If Google's featured snippet box does not give a response for a country, we record a missing entry.[10]

Step 2: Google Trends provides historical and cross-sectional information about the relative importance of a search query. For the United States, the cross-sectional information with the highest granularity is at the level of a Designated Market Area (DMA).[11] Google Trends expresses the relative importance of a search query in a given DMA as an integer value from 0 to 100. This integer value is calculated as follows. First, find the number of searches for the query at hand relative to the total number of searches, and define the maximum search market share of any DMA to 100. Second, divide each search market share by the maximum, and express it as a rounded percentage. If the result does not exceed an unreported threshold, set it to zero (Liang, 2017; Stephens-Davidowitz and Varian, 2015). Formally,

$$G(i,d) = \left\lfloor 100 \frac{share_{i,d}}{\max_{\delta}\{share_{i,\delta}\}} \mathbf{1}[\#(i,d) \geq T] \right\rceil$$

where $\lfloor x \rceil$ is the integer round function, $share_{i,d}$ is the search market share of search query i in DMA d, and T is the unreported search volume threshold. Note that T is defined on the absolute number of searches, rather than the search market share. This implies that DMAs with a larger population will tend to report more data than those with smaller populations. Note also that in addition to $G(i,d)$ being reported as zero for some i and d, we set its value equal to zero if there is no search result from Google's featured snippet box, or if there is no result from Google Trend.

Step 3: We define the p-specific Index for each DMA-country pair as

$$I(p,o,d) = \frac{1}{5} \sum_{i \in q(o,p)} G(i,d)$$

We define the Information Demand Index as the average over the p-specific indices:

$$\text{IDI}_{o,d} = \frac{1}{5} \sum_{p} I(p,o,d).$$

to the query "notable [country] [p]", where [country] is one of the 100 largest countries in 2015.

[10] This is the case for about 4-10% of our sample, depending on the category.

[11] Google Trends also breaks the information down by major city; however, we would lose non-city data.

A.5 Details on the construction of crop suitability measures

The crop suitability index for each origin country o and destination county d is taken from the Food and Agriculture Organization of the United Nations Global Agro-Ecological Zones (FAO-GAEZ) data. We estimate potential agricultural similarity between each origin country o and county d by constructing a distance measure based on the difference in crop suitability of the country and county for a select group of crops. The following outlines the steps taken in order to create a crop suitability distance measure for each country-county pair.

Step 1: To identify the crops to be used in constructing the crop suitability distance measure, we compare data from FAOSTAT on the top crops produced by the U.S. in 2014 with data available from FAO-GAEZ on crop suitability. We then select the top 10 crops, based on value of agricultural production, for which there is data in FAO-GAEZ were used: rice, maize, wheat, soybeans, tomatoes, white potatoes, sugar cane, cotton, yams, and cassava. For each crop, we extract the crop suitability data by selecting the total production capacity data (located within the attainable yield/agro-ecological suitability data) and setting the water-supply to rainfed, input level to high, and time period to baseline (1961-1990).

Step 2: To take the global crop suitability data for each crop and define crop suitability for each county and country, we utilise ArcMap software. For counties, the U.S. 1990 counties border map is used. For countries, we extract data on 1990 country borders from a dataset including all country borders for the period post-WWII to 2015.[12] Then, for each crop, we utilise the ArcMap software to calculate the average of crop suitability for each county and country.[13]

Step 3: We rescale the crop suitability data to a 0 to 1 scale and then calculate the crop suitability distance measure for each pair, country o and county d, as

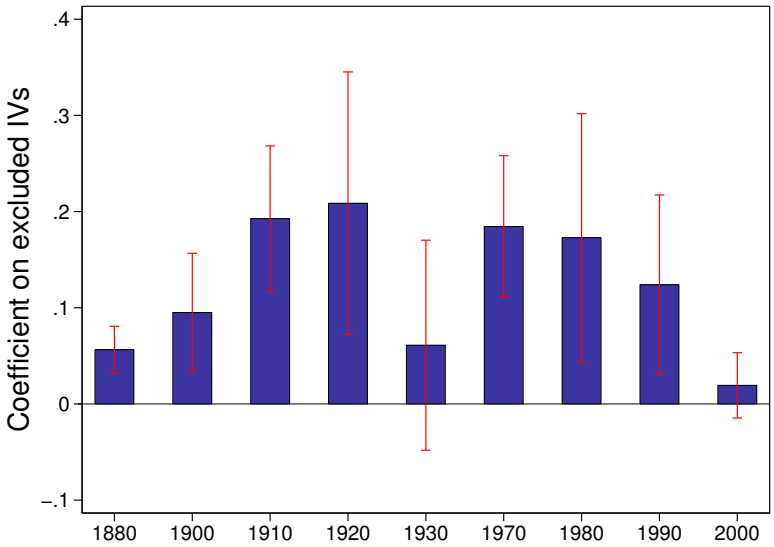
$$DistanceCS_{o,d} = \frac{\sum_{n=1}^{10} s_o^n s_d^n}{\sqrt{\sum_{n=1}^{10}(s_o^n)^2}\sqrt{\sum_{n=1}^{10}(s_d^n)^2}}$$

where s^n is the measure of crop suitability for crop n.

[12] The original shapefile is version 0.6 (updated November 30, 2016) from Weidmann, Nils B., Doreen Kuse, and Kristian Skrede Gleditsch. "The Geography of the International System: The CShapes Dataset." International Interactions 36, no. 1 (2010).

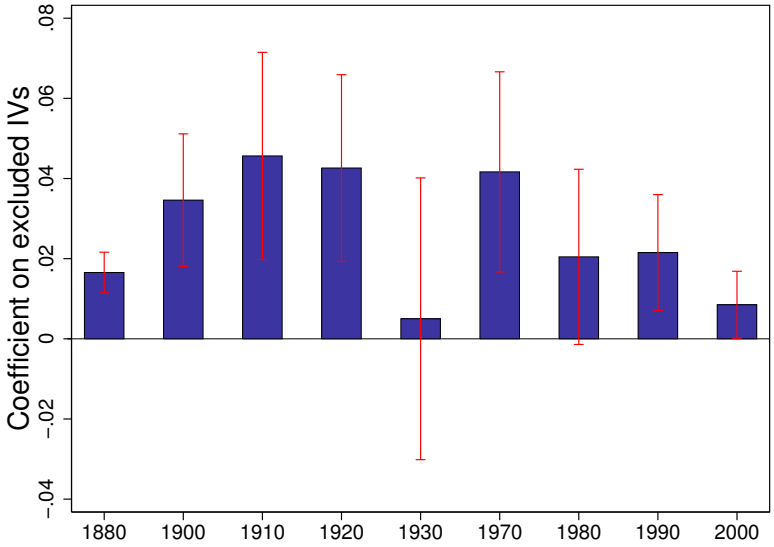
[13] Data is missing for 23 counties and 35 countries due to issues with overlapping polygons as well as missing 1990 boundaries data for certain countries.

B Additional figures and tables



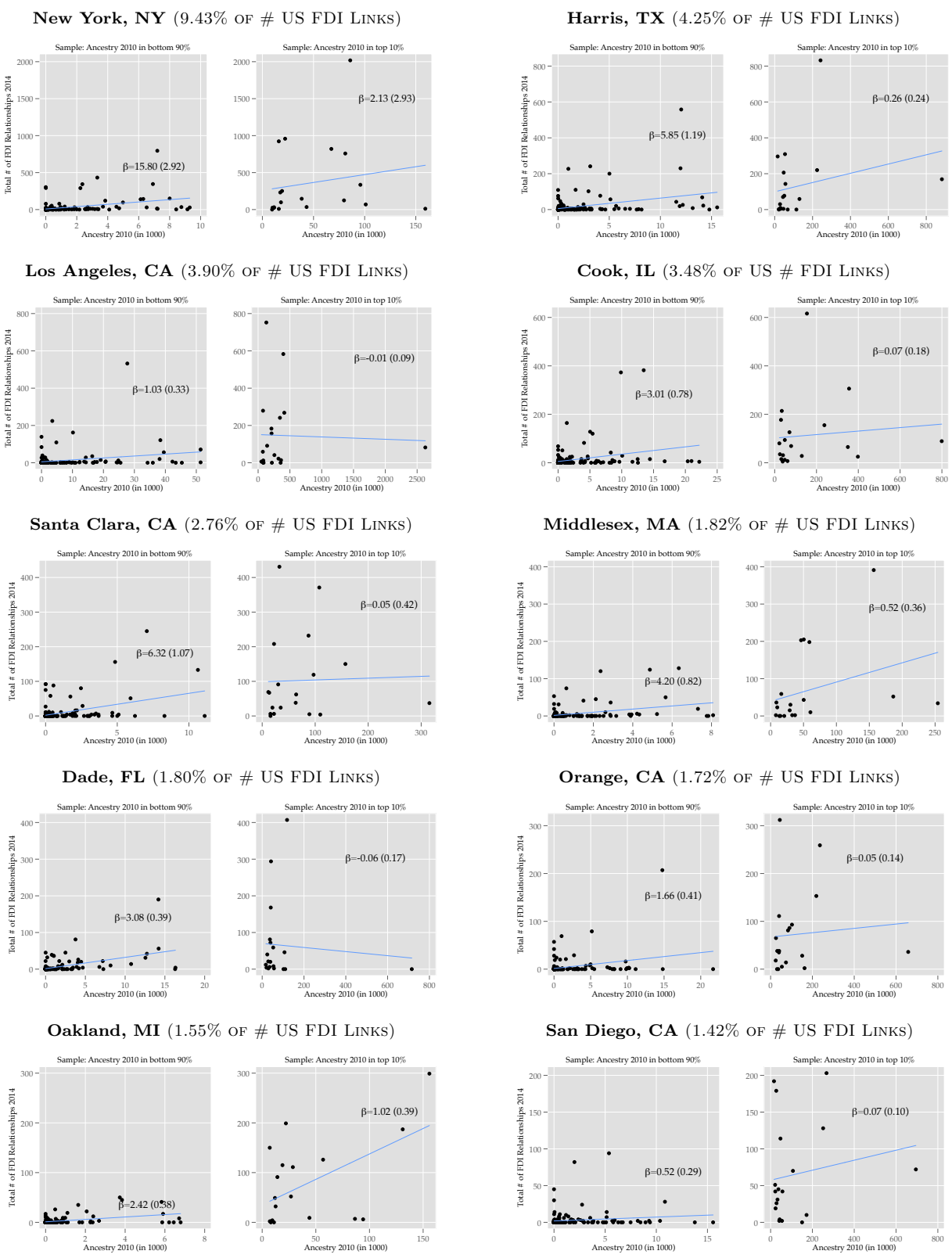
Appendix Figure 1: First-Stage Coefficients

Notes: Coefficient estimates (bars) and 95% confidence intervals (lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ from Table 2, column 2. The dependent variable is Log Ancestry 2010. Robust standard errors are clustered at the origin country level.



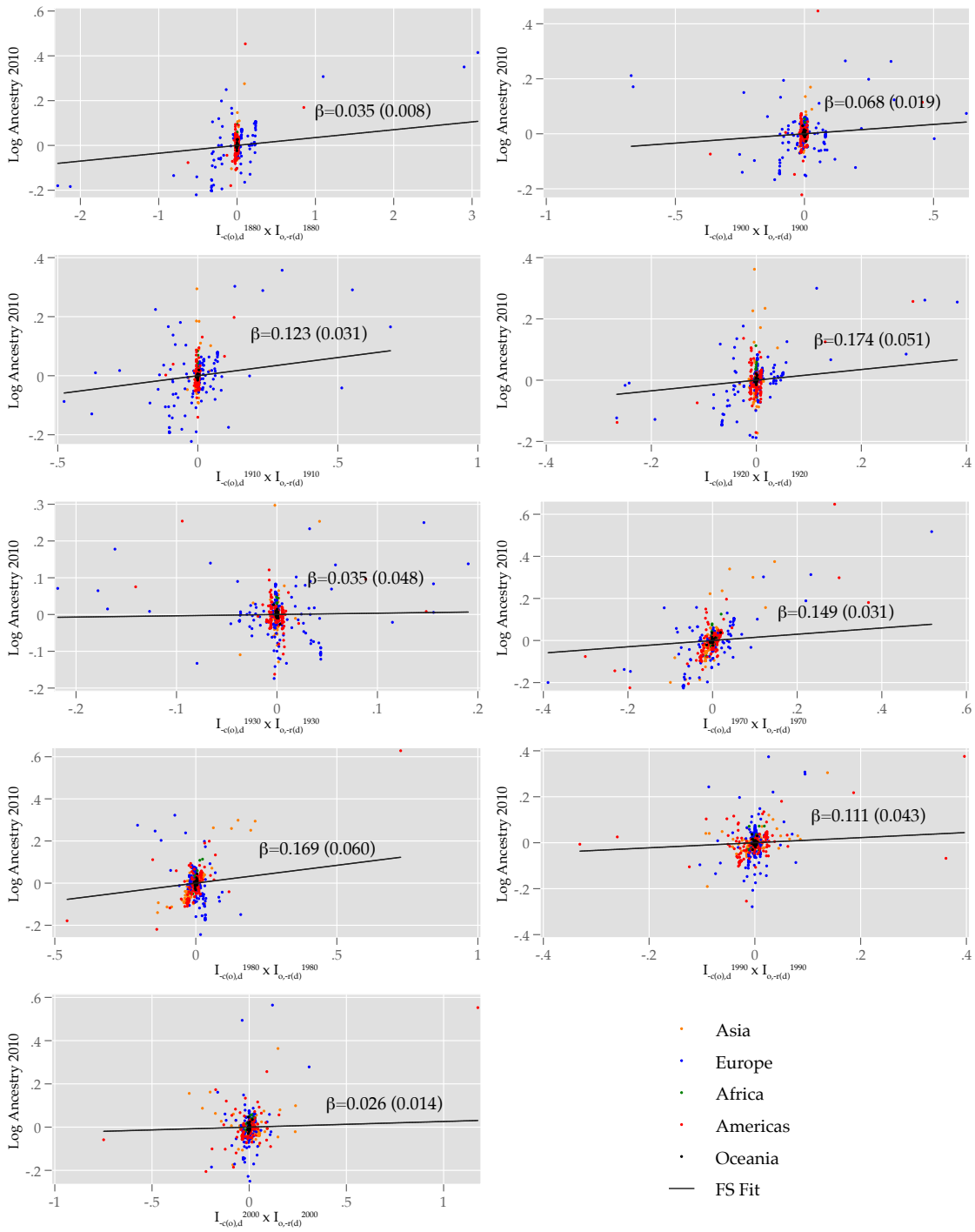
Appendix Figure 2: Reduced-Form Coefficients

Notes: Coefficient estimates (bars) and 95% confidence intervals (lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ from a reduced form regression corresponding to the specification in column 2 of Table 2, using the 2014 FDI dummy as dependent variable. Robust standard errors are clustered at the origin country level. The R^2 of this regression is 0.218.



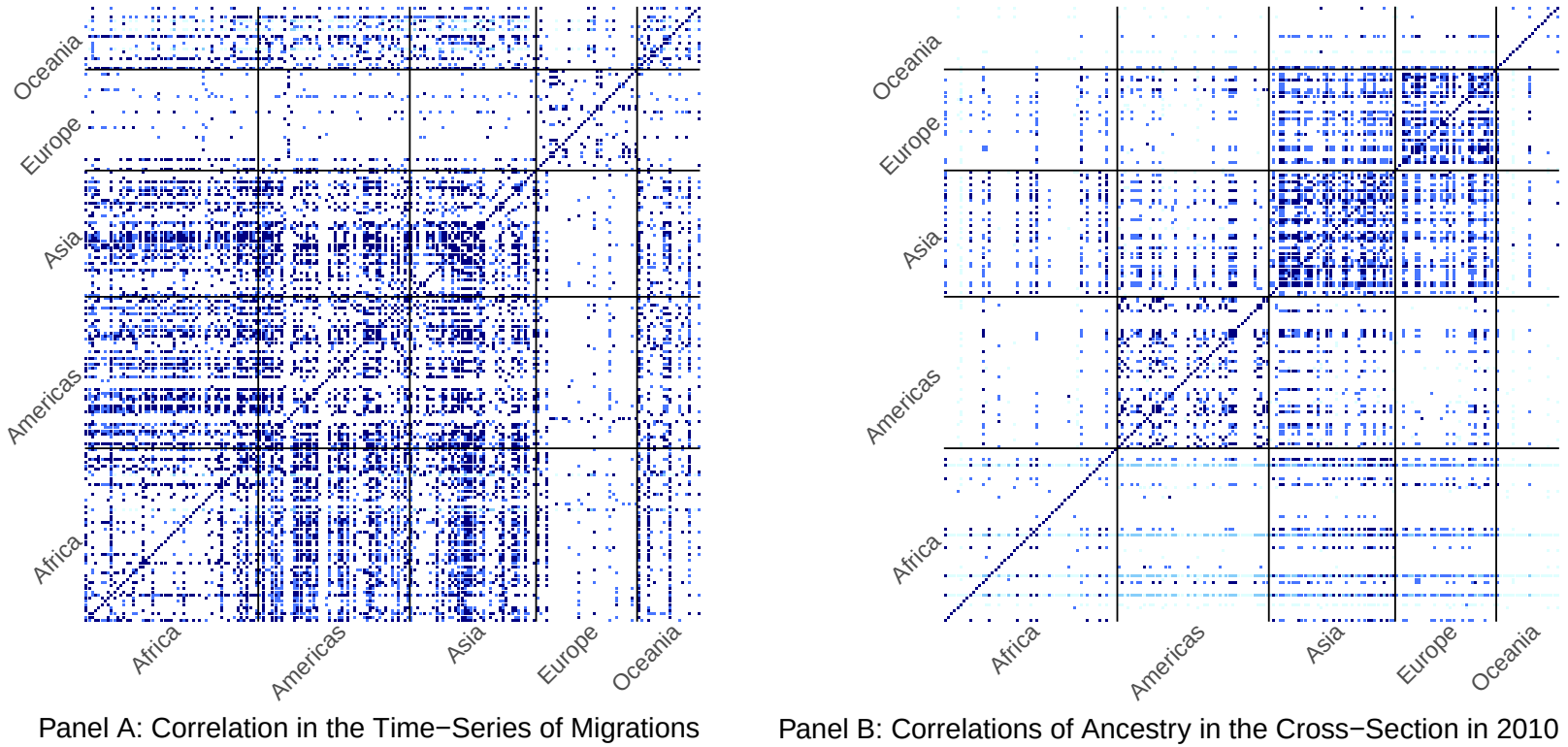
Appendix Figure 3: Ancestry and Total # of FDI relationships (Raw Data)

Notes: The figure presents scatter plots of the raw data for *Ancestry 2010* and *Total # of FDI relationships 2014* for the 10 largest US counties in terms of # of FDI relationships (counties' share of total US FDI relationships indicated in title). For each county, data is shown separately by origins with ancestry share in the bottom 90% of ancestries in d, and in the top 10% of ancestries in d. Linear regressions are fitted separately for each subfigure; coefficient estimates and standard errors (in parentheses) are provided.



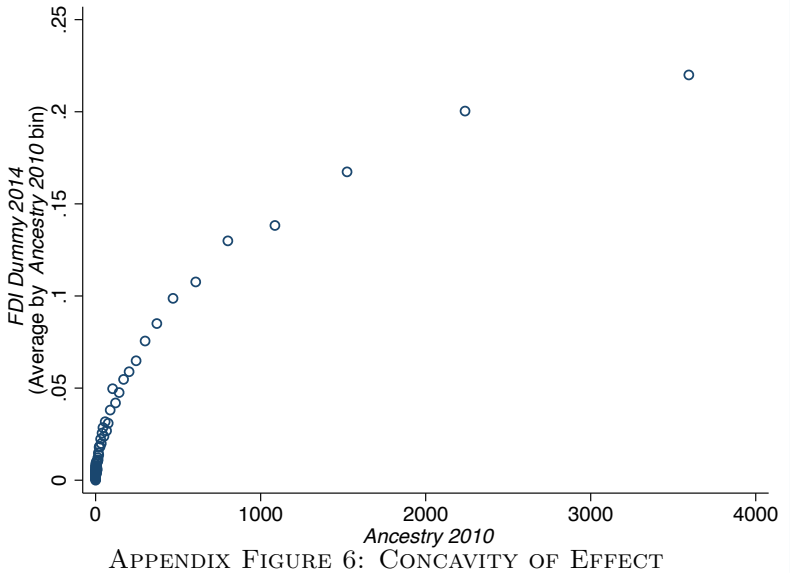
Appendix Figure 4: First Stage Fit

Notes: The figure shows conditional scatter plots of *Log Ancestry 2010* versus each of the interacted instruments $I^t_{-c(o),d} \times I^t_{o,-r(d)}$. Each subfigure is constructed as follows: both *Log Ancestry 2010* and $I^t_{-c(o),d} \times I^t_{o,-r(d)}$ are regressed on destination × continent-of-origin fixed effects, origin × destination-census-region fixed effects, distance, and latitude difference, as well as the interacted instruments for all time periods except t; for visual clarity, residuals of both regressions are binned, separately for each origin o, by quintiles of the residuals of the former regression; the binned data is scattered, colour-coded by the continent of o. Note that each graph shows the partial correlation that identifies the nine coefficients of interest in our standard specification of the first stage in column 4 of Table 2. The first stage – corresponding to a linear least squares fit of the data before binning – is shown as black line, and the respective first stage coefficient estimates (and standard errors in brackets) are shown.



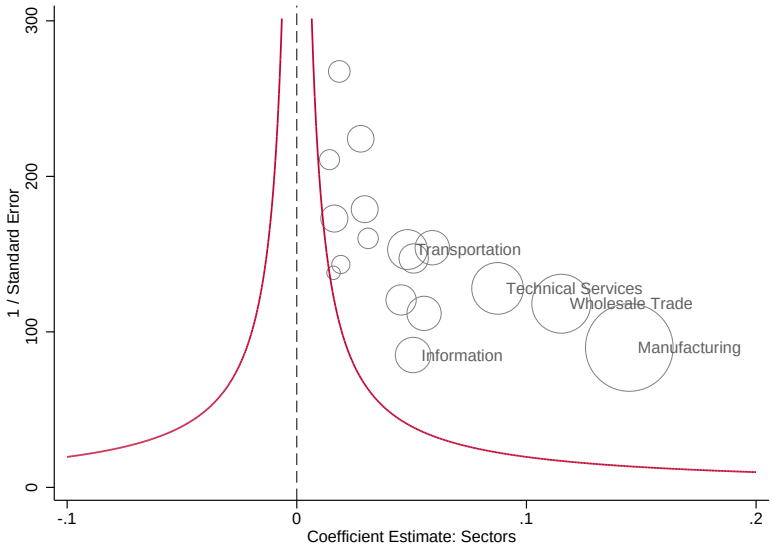
Appendix Figure 5: Correlation Between Countries of Immigration Waves to the US and 2010 Ancestry Across the US

Notes: These correlation plots display the magnitude of the time-series correlation of total migration to the US between each pair of countries (Panel A) and the magnitude of the correlation of ancestry in 2010 across the US counties between each pair of origin countries (Panel B). A pair of origin countries for which the correlation is negative or not significant at the 5% level is displayed with a white dot. For country pairs with positively correlated total migration significant at the 5% level, a darker shaded dot indicates a higher correlation value.



Appendix Figure 6: Concavity of Effect

Notes: This figure plots of the mean of *FDI Dummy 2014* within bins of *Ancestry 2010*. The *Ancestry 2010* bins are constructed as centiles of the conditional distribution of *Ancestry 2010*|*Ancestry 2010*>0. The lowest bin corresponds to *Ancestry 2010*$=0$. We do not plot the mean of *FDI Dummy 2014* in the 99th and 100th centile *Ancestry 2010* bin for visual clarity; the overall concave pattern extends to these observations.



Appendix Figure 7: Heterogeneous Effects across Sectors

Notes: This figure shows funnel plots of the estimated coefficients and standard errors from separate IV regressions of the FDI dummy on Log 2010 Ancestry for each sector. In all regressions, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as excluded instruments, and control for log distance as well as latitude difference. We plot the estimated coefficients (x axis) against the reciprocal of estimated standard errors on ancestry. The size of the circle is proportional to the size of the sector. The imposed curve is $y = 1.96/x$ for positive x region and $y = -1.96/x$ for negative x region. Circles above the curve indicate statistically significant coefficients. See section 3.5 for details.

Appendix Table 4: Summary Statistics on the Intensive Margin of FDI

Origin-destination pairs	(1)	(2)	(3)
Ancestry 2010 (in thousands)	10.038	16.502	10.861
	(40.989)	(62.950)	(43.593)
# of FDI Relationships	11.043		
	(39.738)		
# of Parents in Destination		2.282	
		(4.336)	
# of Parents in Origin		8.063	
		(26.132)	
# of Workers Employed at Subsidiary in Destination (in thousands)		6.328	
		(32.695)	
# of Subsidiaries in Origin			1.797
			(10.671)
# of Parents in Destination			0.761
			(3.105)
# of Workers Employed at Subsidiary in Origin (in thousands)			2.435
			(40.360)
N	10851	4065	9082

Notes: The table presents means (and standard deviations). Variables refer to our sample of country-county pairs used in Appendix Table 19. Column 1 shows data for observations that have at least one FDI link. Column 2 shows data for observations that have at least one subsidiary in the origin. Column 3 shows data for observations pairs that have at least one subsidiary in the destination.

Appendix Table 5: Assignment of States to Census Regions

Census Region	State Names
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
Middle Atlantic	New Jersey, New York, Pennsylvania
East North Central	Illinois, Indiana, Michigan, Ohio, Wisconsin
West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
South Atlantic	Delaware, District Of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia
East South Central	Alabama, Kentucky, Mississippi, Tennessee
West South Central	Arkansas, Louisiana, Oklahoma, Texas
Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming
Pacific	Alaska, California, Hawaii, Oregon, Washington

Appendix Table 6: Alternative Instruments based on Immigration and Ancestry Correlation

	FDI 2014 (Dummy)			
	(1)	(2)	(3)	(4)
Panel A: Time-Series Correlation of Total Migration to the US				
Log Ancestry 2010	0.176***	0.177***	0.177***	0.181***
	(0.027)	(0.028)	(0.027)	(0.027)
Log Distance	0.021	0.021	0.021	0.022
	(0.029)	(0.028)	(0.028)	(0.028)
N	612495	612495	612495	612495
Panel B: Cross-Section Correlation of 2010 Ancestry Across the US				
Log Ancestry 2010	0.217***	0.186***	0.217***	0.217***
	(0.031)	(0.028)	(0.031)	(0.031)
Log Distance	0.031	0.023	0.031	0.031
	(0.031)	(0.029)	(0.031)	(0.031)
N	612495	612495	612495	612495
Correlation Cutoff	.5	.75	.5	.5
Significance Cutoff	NA	NA	.1	.05

Notes: The table displays estimates of the specification from column 3 of Panel A in Table 3 removing or not alternative sets of migrations from the construction of pull factors. In Panel A, we exclude from the pull factor all countries for which the time-series correlation of total migration to the US with o's migration to the US is greater than .5, greater than .75, greater than .5 and significant at the 10% level, and greater than .5 and significant at the 5% level, respectively. In Panel B, we exclude from the pull factor all countries for which the correlation of 2010 ancestry across destination counties with o's ancestry across destination counties is greater than .5, greater than .75, greater than .5 and significant at the 10% level, and greater than .5 and significant at the 5% level, respectively.

Appendix Table 7: The Effect of Ancestry on FDI: Variations of Leave-Out Instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI Dummy (2014)						
Panel A: baseline specification	$\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.231***	0.190***	0.187***	0.187***	0.189***	0.198***	0.191***
	(0.023)	(0.024)	(0.024)	(0.024)	(0.030)	(0.023)	(0.024)
N	612495	612495	612495	612495	459150	612495	612300
Panel B: no leave-out	$\{I^t_o(I^t_d/I^t)\}$ excluded						
Log Ancestry 2010	0.204***	0.202***	0.174***	0.174***	0.173***	0.183***	0.215***
	(0.020)	(0.019)	(0.022)	(0.022)	(0.027)	(0.022)	(0.017)
N	612495	612495	612495	612495	459150	612495	612300
Panel C: single country/county leave-out	$\{I^t_{o,-d}(I^t_{-o,d}/I^t_{-o})\}$ excluded						
Log Ancestry 2010	0.212***	0.204***	0.172***	0.171***	0.173***	0.185***	0.216***
	(0.020)	(0.019)	(0.024)	(0.024)	(0.030)	(0.024)	(0.017)
N	612495	612495	612495	612495	459150	612495	612300
Panel D: county/continent leave-out	$\{I^t_{o,-d}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.223***	0.217***	0.183***	0.183***	0.186***	0.200***	0.227***
	(0.022)	(0.021)	(0.024)	(0.024)	(0.030)	(0.024)	(0.018)
N	612495	612495	612495	612495	459150	612495	612300
Panel E: adjacent state leave-out	$\{I^t_{o,-adj(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.232***	0.204***	0.192***	0.192***	0.181***	0.206***	0.237***
	(0.024)	(0.022)	(0.022)	(0.022)	(0.027)	(0.021)	(0.019)
N	640764	640764	640764	640764	459150	640764	640560
Panel F: correlated migrations leave-out	$\{I^t_{o,-r(d)}(I^t_{-s^1(o),d})(I^t_{-s^1(o)})\}$ excluded						
Log Ancestry 2010	0.229***	0.188***	0.181***	0.181***	0.173***	0.182***	0.182***
	(0.023)	(0.024)	(0.027)	(0.027)	(0.031)	(0.026)	(0.027)
N	612495	612495	612495	612495	459150	612495	612300
Panel G: correlated ancestry leave-out	$\{I^t_{o,-r(d)}(I^t_{-s^2(o),d})(I^t_{-s^2(o)})\}$ excluded						
Log Ancestry 2010	0.268***	0.200***	0.217***	0.217***	0.222***	0.221***	0.220***
	(0.031)	(0.030)	(0.030)	(0.030)	(0.036)	(0.024)	(0.029)
N	612495	612495	612495	612495	459150	612495	612300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	Yes	No	No
Agricultural Similarity (Cosine)	No	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	No	Yes

Notes: The table repeats the estimates from Panel A in Table 3 in Panel A and then shows variations in the following panels, removing or not different sets of migrants from the interaction of pull and push factors. The construction of the interaction is indicated above each panel. In Panel D, $adj\,(d)$ refers to the adjacent states for the state of county d; thus we exclude from the push factor of o migrations to any state adjacent to the state of d, including the state itself. In Panel E and Panel F, "s" refers to similar countries; that is, in Panel E we exclude from a given pull factor of o to d all countries for which the time correlation of total migration to the US with o's migration to the US is greater than .5 and significant at the 5% level while for Panel F we exclude from a given pull factor of o to d all countries for which the correlation of 2010 ancestry across the US with o's ancestry across the US is greater than .5 and significant at the 5% level.

Appendix Table 8: The effect of Ancestry in 2000 on FDI in 2007

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: IV	*FDI 2007 (Dummy)*					
Log Ancestry 2000	0.250***	0.184***	0.182***	0.182***	0.188***	0.184***
	(0.018)	(0.020)	(0.020)	(0.020)	(0.019)	(0.021)
KP F-stat on excluded IV's	12.06	10.32	167.32	165.46	156.29	189.21
Stock-Yogo 5% critical values	20.25	20.25	21.10	21.10	21.18	21.10
Stock-Yogo 10% critical values	11.39	11.39	11.52	11.52	11.52	11.52
N	612,495	612,495	612,495	612,495	612,495	612,300
Panel B: OLS	*FDI 2007 (Dummy)*					
Log Ancestry 2000	0.216***	0.184***	0.184***	0.184***	0.184***	0.200***
	(0.015)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
N	612,495	612,495	612,495	612,495	612,495	612,300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	Yes	Yes	Yes	Yes	Yes
Principal Components	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2000}_{o,-r(d)}(I^{2000}_{-c(o),d}/I^{2000}_{-c(o)})$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from IV (Panel A) and OLS (Panel B) regressions of equation (1) at the country-county level. The dependent variable in all panels is a dummy indicating an FDI relationship between origin o and destination d in 2007. The main variable of interest is *Log Ancestry 2000*, instrumented using various specifications of equation (4). In all columns in Panel A, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,1990}$ as excluded instruments. Columns 3-6 also include the first five principal components of the higher-order interactions of push and pull factors as instruments. Column 5 also includes the interaction of the push and pull factor constructed using data from the 1990-2000 wave. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 9: Nonlinear Least Squares Estimation

β	π
0.1683***	0.0010***
(0.0012)	(0.0000)

Notes: The table presents coefficient estimates from a nonlinear least squares regression at the country-county level. The dependent variable is the dummy for FDI in 2014. It shows (un-adjusted) NLS standard errors. We obtain the optimal β and π by solving the nonlinear least squares problem in equation (6), excluding the fixed effects. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 10: Alternative Functional Forms

	FDI 2014 (Dummy)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestry 2010	0.002***					
	(0.001)					
Log Ancestry 2010 (-1 for $-\infty$)		0.186**				
		(0.080)				
(Ancestry 2010)$^{1/3}$			0.191***			
			(0.022)			
Log Ancestry 1980				0.218***		
				(0.034)		
Log Ancestry 1990					0.203***	
					(0.028)	
Log Ancestry 2000						0.193***
						(0.024)
N	612495	612495	612495	612495	612495	612495

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable is the dummy for FDI in 2014. The main variable of interest in each column is the measure of ancestry indicated by the first column of the table. In the second row, we use Log(Ancestry/1000) instead of Log(1+Ancestry/1000), and replace Log(0) with -1. All specifications are the same as that in column 3 of Table 3, except that principal components are excluded. Standard errors are given in parentheses and are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 11: Varying ownership cutoffs

Panel A: FDI dummy on ancestry (IV)	*FDI 2014 (Dummy)*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.189***	0.190***	0.190***	0.157***
	(0.024)	(0.024)	(0.024)	(0.029)
R^2	0.352	0.352	0.352	0.318
N	612495	612495	612495	612495
Panel B: # of FDI relationships on ancestry (IV)	*Log Total # of FDI relationships*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.408***	0.394***	0.402***	0.075
	(0.042)	(0.045)	(0.046)	(0.062)
R^2	0.750	0.749	0.749	0.770
N	10445	10393	10365	6981
Ownership cutoff	keep $\geq 5\%$	keep $\geq 25\%$	keep $\geq 50\%$	keep $< 50\%$
Destination × Continent FE	Yes	Yes	Yes	Yes
Origin × Census Region FE	Yes	Yes	Yes	Yes
Principal Components	Yes	Yes	Yes	Yes

Notes: This table presents coefficient estimates from variations of the IV regression in column 3 of Table 3 (Panel A) and in column 2 of Appendix Table 19 (Panel A). We vary the ownership cutoff across columns: In columns 1, 2, and 3 we keep all shareholder-subsidiary pairs with ownership $\geq 5\%$, $\geq 25\%$, $\geq 50\%$, respectively. The number of origin-destination pairs with any FDI under these cutoffs are 10445, 10393, and 10365. In column 4 we keep all shareholder-subsidiary pairs with ownership $< 50\%$, which results in 6981 origin-destination pairs with any FDI. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 12: Alternative Standard Errors: Main Specification

Panel A: Analytical	
Robust	0.0092
Cluster by county	0.0171
Cluster by country†	0.0243
Cluster by county and country	0.0280
Cluster by state and country	0.0285
Cluster by state	0.0189
Cluster by continent	0.0070
Cluster by state*country	0.0114
Panel B: Bootstrap	
Robust	0.0090
Cluster by county	0.0152
Cluster by country	0.0284

Notes: This table shows various standard errors on Log Ancestry 2010 based on our standard specification (column 3 of Table 3). The bootstrapped standard errors in Panel B are obtained using 1,000 draws with replacement. † denotes our standard specification.

Appendix Table 13: Alternative Standard Errors: Other Specifications

	Standard specification	Communist natural experiment	Intensive margin	Immigration 1990-2000
	(1)	(2)	(3)	(4)
Outcome variable	FDI Dummy (2014)		Log total # of FDI Relationships	Immigration 1990-2000
Panel A: Clustered by Country (Standard)				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.024)	(0.032)	(0.056)	
Log Ancestry 1990				9.662**
				(4.455)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.358)
Panel B: S.E. Clustered by County				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.017)	(0.032)	(0.077)	
Log Ancestry 1990				9.662**
				(4.327)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.230)
Panel C: S.E. Clustered by State				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.019)	(0.029)	(0.071)	
Log Ancestry 1990				9.662**
				(3.942)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.267)
Panel D: Clustered by County and Country				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.028)	(0.051)	(0.120)	
Log Ancestry 1990				9.662**
				(4.800)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.105)
Panel E: Clustered by State and Country				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.028)	(0.048)	(0.116)	
Log Ancestry 1990				9.662**
				(4.308)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.044)

Notes: This table shows variations based on four main regressions: standard specification (column 3 in Table 3), communist natural experiment (column 5 in Table 4), intensive margin (based on column 2 in Panel A of Appendix Table 19), and immigration 1990-2000 (column 1 in Table 6). In Panel A, we reproduce the standard error clustering in our main tables; in Panel B, we cluster by county; in Panel C, we cluster by country, in Panel D we double cluster by county and country, and in Panel E we double cluster by state and country.

Appendix Table 14: Placebo Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI 2014 (Dummy)						
Panel A	*Assign to alphabet neighbor*						
Log Ancestry 2010	-0.012	-0.007	0.009	0.009	0.006	0.010	0.012
	(0.019)	(0.014)	(0.027)	(0.027)	(0.033)	(0.027)	(0.030)
N	612495	612495	612495	612495	459150	612495	612300
Panel B	*Assign to alphabet neighbor on a different continent*						
Log Ancestry 2010	-0.025	-0.020	0.010	0.010	0.004	0.014	0.013
	(0.021)	(0.014)	(0.033)	(0.033)	(0.038)	(0.037)	(0.037)
N	612495	612495	612495	612495	459150	612495	612300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	Yes	No	No
Agricultural Similarity (Cosine)	No	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from placebo regressions corresponding the the specifications in Table 3. In Panel A, we assign the outcomes (FDI 2014 Dummy) for each origin country to the next country in the alphabet. In Panel B, we assign the outcomes (FDI 2014 Dummy) for each origin country to the next country in the alphabet that is from another continent.

Appendix Table 15: The Effect of Ancestry on FDI: Five Largest Countries and Counties

	FDI 2014 (Dummy)
Panel A: Top 5 Ancestries	*Log Ancestry 2010*
Germany	0.216***
	(0.009)
Britain	0.271***
	(0.009)
Mexico	0.171***
	(0.011)
Ireland	0.202***
	(0.010)
Italy	0.219***
	(0.007)
Panel B: Largest 5 Counties	*Log Ancestry 2010*
Los Angeles, California	0.137***
	(0.019)
Cook, Illinois	0.146***
	(0.020)
Harris, Texas	0.169***
	(0.023)
San Diego, California	0.164***
	(0.024)
Orange, California	0.160***
	(0.020)

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable in all panels is the dummy for FDI in 2014. Panel A presents the coefficient on *Log Ancestry 2010* when we run our estimation separately for each of the largest five origin countries. Panel B presents the coefficient on *Log Ancestry 2010* when we run our estimation separately for each of the five US counties with the largest population in 2010. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as IVs. All specifications control for log distance and latitude difference. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 16: The Effect of Ancestry on FDI: Country Specific Effects

	Point Estimate	Standard Error	*FDI 2014 (Dummy)* > 0
United Arab Emirates	11.875***	(2.712)	60
Kuwait	6.098***	(2.120)	22
Finland	4.113***	(0.513)	180
New Zealand	2.980***	(0.511)	107
Oman	2.481	(1.597)	6
British Virgin Islands	2.467***	(0.604)	100
Australia	2.201***	(0.384)	369
Malaysia	2.005***	(0.406)	90
South Africa	1.832***	(0.247)	80
Tunisia	1.438***	(0.345)	9
Iceland	1.359***	(0.276)	25
Saudi Arabia	1.144***	(0.158)	29
Belgium and Luxembourg	1.086***	(0.087)	354
Puerto Rico	1.034***	(0.240)	26
Israel	0.944***	(0.156)	137
Bahamas	0.943***	(0.308)	44
Switzerland	0.814***	(0.048)	371
Denmark	0.684***	(0.043)	278
Thailand	0.583***	(0.070)	68
Japan	0.566***	(0.051)	575
Uruguay	0.541***	(0.115)	21
Austria	0.531***	(0.042)	148
Chile	0.502***	(0.078)	73
Brazil	0.496***	(0.047)	140
Barbados	0.462**	(0.234)	38
Canada	0.461***	(0.024)	809
Norway	0.459***	(0.028)	239
Malta	0.451	(0.281)	11
Costa Rica	0.447***	(0.140)	30
Turkey	0.444***	(0.067)	48
Netherlands	0.442***	(0.019)	398
Panama	0.439***	(0.115)	44
Indonesia	0.413***	(0.076)	29
Argentina	0.412***	(0.056)	64
Sweden	0.405***	(0.018)	323
Senegal	0.383	(0.314)	2
France	0.346***	(0.013)	528
South Korea	0.346***	(0.023)	155
Liberia	0.341*	(0.190)	6

Spain	0.335***	(0.014)	300
India	0.320***	(0.018)	233
China	0.299***	(0.015)	248
Kenya	0.292*	(0.175)	5
Venezuela	0.275***	(0.046)	32
Britain	0.271***	(0.009)	664
Egypt	0.259***	(0.051)	23
Belize	0.255***	(0.086)	14
Hungary	0.240***	(0.033)	52
Colombia	0.237***	(0.028)	45
Italy	0.219***	(0.007)	489
Peru	0.218***	(0.033)	30
Germany	0.216***	(0.009)	608
Portugal	0.206***	(0.028)	85
Samoa	0.204**	(0.086)	5
Ireland	0.202***	(0.010)	247
Morocco	0.197**	(0.078)	11
Nigeria	0.190***	(0.055)	18
Sri Lanka	0.180	(0.120)	6
Czechoslovakia	0.177***	(0.029)	54
Romania	0.173***	(0.041)	23
Mexico	0.171***	(0.011)	259
Pakistan	0.168***	(0.039)	23
USSR	0.165***	(0.015)	97
Ghana	0.156	(0.095)	6
Bulgaria	0.156**	(0.064)	11
Philippines	0.154***	(0.019)	50
Lebanon	0.150***	(0.047)	20
Bolivia	0.142**	(0.066)	8
Greece	0.131***	(0.028)	42
Trinidad and Tobago	0.130*	(0.067)	15
Socialist Yugoslav	0.121***	(0.028)	29
Jamaica	0.114***	(0.032)	15
Honduras	0.103***	(0.032)	14
Algeria	0.099	(0.076)	3
Guatemala	0.097***	(0.033)	14
Poland	0.092***	(0.015)	63
Viet Nam	0.091***	(0.025)	18
Jordan	0.090	(0.063)	7
Cameroon	0.085	(0.065)	2
Dominican Republic	0.082***	(0.025)	16

Ecuador	0.081**	(0.032)	15
Paraguay	0.079	(0.056)	4
Nicaragua	0.069*	(0.036)	7
Albania	0.069	(0.046)	3
North Korea	0.068	(0.072)	1
El Salvador	0.066**	(0.026)	13
Sudan	0.065	(0.065)	1
Fiji	0.065	(0.046)	5
Bangladesh	0.040	(0.032)	2
Cambodia	0.039	(0.028)	3
Haiti	0.026	(0.019)	2
Ethiopia	0.026	(0.025)	1
Syria	0.016	(0.016)	1
Myanmar	0.007	(0.007)	1
Afghanistan	0.003	(0.003)	1
Guyana	0.002	(0.002)	1
Iraq	0.002	(0.002)	1
Cuba	-0.000***	(0.000)	1
Libya	-0.022	(0.024)	1
Grenada	n/a	n/a	0
Sierra Leone	n/a	n/a	0
Somalia	n/a	n/a	0
Iran	n/a	n/a	0
Tonga	n/a	n/a	0
Cape Verde	n/a	n/a	0
Mauritania	n/a	n/a	0
Nepal	n/a	n/a	0
Greenland	n/a	n/a	0
Yemen	n/a	n/a	0
Equatorial Guinea	n/a	n/a	0
Mongolia	n/a	n/a	0
State of Palestine	n/a	n/a	0
Lao	n/a	n/a	0

Notes: The table is an extension of Appendix Table 15 Panel A, where we only show the results for top five ancestries. Results are sorted on the point estimate. The last column shows the number of US counties that have an FDI link with the corresponding country. All countries with ancestry < 1 are discarded.

Appendix Table 17: The Effect of Ancestry on FDI: Sector-Specific Effects

20 Sectors Based on 2007 NAICS code	Point Estimate	Standard Error	*FDI 2014 (Dummy)* > 0
Manufacturing	0.165***	(0.024)	5,549
Wholesale Trade	0.141***	(0.026)	2,513
Professional, Scientific, and Technical Services	0.122***	(0.024)	1,925
Retail Trade	0.085***	(0.020)	846
Information	0.084***	(0.018)	906
Transportation and Warehousing	0.084***	(0.016)	620
Administrative and Support and Waste Management and Remediation Services	0.083***	(0.018)	855
Real Estate and Rental and Leasing	0.077***	(0.020)	662
Finance and Insurance	0.071***	(0.019)	1,143
Other Services (except Public Administration)	0.053***	(0.014)	301
Management of Companies and Enterprises	0.049***	(0.014)	524
Construction	0.040**	(0.016)	510
Accommodation and Food Services	0.035***	(0.010)	239
Arts, Entertainment, and Recreation	0.030***	(0.006)	131
Mining, Quarrying, and Oil and Gas Extraction	0.028***	(0.009)	528
Health Care and Social Assistance	0.024**	(0.012)	291
Utilities	0.022*	(0.012)	338
Educational Services	0.009	(0.006)	111
Agriculture, Forestry, Fishing and Hunting	0.007**	(0.003)	149
Public Administration	0.001	(0.001)	10

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions for each of the 20 2-digit NAICS sectors at the country-county level. Each row of the table corresponds to one regression. The dependent variable in each row is a dummy variable for FDI in 2014 in the sector indicated. The last column shows the number of country-county pairs that have an FDI link with the corresponding country. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 18: Heterogeneous Effects across Sectors and Firms

FDI 2014 (Dummy)	*Log Ancestry 2010* (1)	*FDI 2014 (Dummy)* > 0 (2)
Panel A: Individual Sectors		
Manufacturing	0.165***	
	(0.024)	
Trade	0.152***	
	(0.026)	
Information, Finance, Management, and other Services	0.143***	
	(0.024)	
Construction, Real Estate, Accomodation, Recreation	0.125***	
	(0.021)	
Health, Education, Utilities, and other Public Services	0.042**	
	(0.019)	
Natural Resources	0.035***	
	(0.009)	
Panel B: Small vs. Large Firm Size		
Above Median	0.112***	1,840
	(0.018)	
Below Median	0.051**	723
	(0.024)	
p-value of χ^2 test, H_0: equality of coefficients	0.000	

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions at the country-county level. Each row of the table corresponds to a separate regression. The dependent variables in all rows are dummy variables that are one if any firm within the indicated subset of firms in destination county d has a parent or subsidiary in origin country o. These subsets of firms are five sector groups (panel A) and for small versus large firms (panel B). The cutoff value between small and large firms is the median employee number, which is 1380 for US firms that are subsidiaries and 1057 for US firms that are parents. Throughout, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as intrumental variables. "*FDI 2014 (Dummy)* > 0" refers to the number of country-county pairs that have an (non-zero) FDI link in the corresponding sector. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 19: The Effect of Ancestry on the Intensive Margin of FDI

	OLS	IV/GMM	IV/GMM	IV/GMM
	(1)	(2)	(3)	(4)
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.245***	0.356***	0.292***	0.147***
	(0.048)	(0.056)	(0.021)	(0.031)
N	10,851	10,851	10,851	10,851
Panel B	*Log # of subsidiaries in destination with shareholders in origin*			
Log Ancestry 2010	0.275***	0.339***	0.288***	0.242***
	(0.050)	(0.059)	(0.016)	(0.045)
N	9,082	9,082	9,082	9,082
Panel C	*Log # of workers employed at subsidiaries in destination*			
Log Ancestry 2010	0.304*	0.077	0.326***	0.192
	(0.175)	(0.236)	(0.051)	(0.139)
N	9,082	9,082	9,082	9,082
Destination FE	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes
Destination × Continent FE	Yes	Yes	No	No
Origin × Census Region FE	Yes	Yes	No	No
Principal Components	No	Yes	Yes	Yes
Heckman Correction	No	No	No	Yes

Notes: The table presents OLS (column 1) and IV/GMM (columns 2-4) estimates of equation (7). The dependent variables are specified for each panel in the table. The main variable of interest is *Log Ancestry 2010*. All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin, and destination fixed effects. The coefficient estimates on these controls are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 20: The Effect of Ancestry on the Intensive Margin of Trade (State Level)

	OLS	IV	IV
	(1)	(2)	(3)
Panel A	*Log Total # of FDI relationships*		
Log Ancestry 2010	1.001***	1.374***	0.079***
	(0.077)	(0.183)	(0.025)
R^2	0.659	0.626	0.847
N	2,208	2,202	2,191
Panel B	*Log Aggregate Exports*		
Log Ancestry 2010	1.519***	2.993***	-0.149
	(0.173)	(0.357)	(0.138)
R^2	0.416	0.374	0.665
N	4,799	4,783	4,739
Panel C	*Log Aggregate Imports*		
Log Ancestry 2010	1.927***	3.447***	0.003
	(0.148)	(0.497)	(0.150)
R^2	0.419	0.360	0.576
N	3,823	3,764	3,815
Origin FE	Yes	Yes	Yes
Destination FE	No	No	Yes
Heckman Correction	Yes	Yes	Yes
Panel D	*Log Exports to Vietnam*		
Log Ancestry 2010	1.169***	1.230***	
	(0.124)	(0.124)	
R^2	0.680	0.678	
N	51	51	
Panel E	*Log Exports to Japan*		
Log Ancestry 2010	0.898***	1.107***	
	(0.197)	(0.128)	
R^2	0.442	0.419	
N	51	51	
Origin FE	Yes	Yes	
Destination FE	No	No	

Notes: The table presents OLS and IV estimates of equation (7) at the state level for FDI and trade. The dependent variables are the log number of total FDI links in 2014 (Panel A), the log of aggregate exports (from the US state) (Panel B), aggregate imports (Panel C), exports to Vietnam (Panel D), and exports to Japan (Panel E). Exports and imports are measured in US dollars in 2011. In all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,...,2000}$ and principal components as excluded instruments. All specifications control for log distance, latitude difference, and origin fixed effects. Standard errors are given in parentheses and are double clustered at the destination state and origin country. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 21: Thought Experiment: A Gold Rush in Los Angeles in 1880

			Predicted Counterfactual Change	
	Ancestry 2010	FDI #	*Ancestry 2010*	*FDI # (in %, IV)*
	(1)	(2)	(3)	(4)
Germany	343,276	241	+65,344	+62.55
Ireland	256,621	40	+61,701	+58.21
UK	396,439	582	+26,645	+21.91
Norway	39,515	55	+4,657	+3.52
Sweden	51,395	71	+4,010	+3.03
France	77,372	278	+3,293	+2.48
Canada	27,722	531	+3,132	+2.36
Switzerland	10,156	162	+2,456	+1.84
Czechoslovakia	17,905	4	+2,140	+1.60
Netherlands	38,392	121	+1,638	+1.23

Notes: The table presents the number of individuals of selected ancestries living in Los Angeles County (column 1), the number of FDI links between Los Angeles County and the countries of origin (column 2), and the predicted changes in these variables under a counterfactual scenario where the pre-1880 pull factor of Los Angeles is 5 times as large as in reality (columns 3 and 4). Column 3 shows the predicted absolute change in ancestry based on a regression analogous to column 9 of Table 2 with *Ancestry 2010* (in levels) as dependent variable, again excluding the principal components. Column 4 shows the predicted change of *Total # of FDI relationships* (in percent) based on the IV regression of *Log Total # of FDI relationships* on *Log Ancestry 2010*, instrumented for by $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$, similar to column 2 of Appendix Table 19 without the principal components as instruments. All regressions control for log distance and latitude difference and include a origin × destination-census-region, and destination × continent-of-origin fixed effects. Only the 10 countries with the highest absolute change in ancestry are shown in the interest of space. The details for the construction of this thought experiment are presented in section 3.6.

Appendix Table 22: Search terms for Germany and Italy

Germany	Italy
Politicians	
Angela Merkel	Aldo Moro
Helmut Kohl	Benito Mussolini
Willy Brandt	Alessandra Mussolini
Joseph Goebbels	Amintore Fanfani
Karl Marx	Angelino Alfano
Actors	
Jürgen Prochnow	Isabella Rossellini
Til Schweiger	Robert De Niro
Franka Potente	John Turturro
Udo Kier	Roberto Rossellini
Daniel Brühl	Roberto Benigni
Athletes	
Katarina Witt	Mario Andretti
Dirk Nowitzki	Armin Zoggeler
Boris Becker	Roberto Baggio
Steffi Graf	Andrea Barzagli
Franz Beckenbauer	Gerhard Plankensteiner
Musicians	
Ludwig van Beethoven	Antonio Vivaldi
Nena	Gioachino Rossini
Johann Sebastian Bach	Giacomo Puccini
Nina Hagen	Ennio Morricone
Felix Mendelssohn	Luciano Pavarotti

This table shows the top five results from Google's Answer Box for each category for the countries Germany and Italy when typing "notable [country] [category]" into Google.

Appendix Table 23: The Effect of Ancestry on Language: Language Specific Effects

	Point Estimate	Standard Error	N	# of US-born in d that speak o at home in 2010
Aleut	1.608***	(0.028)	3,137	116
Malay	1.376	(0.897)	9,411	176
Arabic	1.222***	(0.171)	78,376	45,953
Spanish	1.172***	(0.380)	65,877	65,877
French	0.212***	(0.018)	87,836	87,416
Haitian Creole	0.198***	(0.015)	3,137	595
Greek	0.196***	(0.033)	6,273	1,849
Vietnamese	0.188***	(0.006)	3,137	1,739
Portuguese	0.174***	(0.030)	28,233	13,095
Korean	0.170***	(0.041)	6,274	3,040
Mon-Khmer	0.167***	(0.005)	3,136	316
Urdu	0.159***	(0.020)	3,137	499
Bengali	0.153***	(0.015)	3,137	191
Japanese	0.142***	(0.007)	3,137	1,972
Persian	0.102***	(0.005)	3,137	538
Chinese	0.085***	(0.005)	3,137	1,745
Thai	0.077***	(0.010)	3,137	619
Polish	0.061***	(0.015)	3,137	1,670
Filipino	0.059***	(0.002)	3,137	1,229
Laotian	0.050***	(0.009)	3,137	592
Albanian	0.049***	(0.014)	3,137	181
Italian	0.041***	(0.005)	6,274	5,068
Samoan	0.036	(0.031)	3,137	261
Amharic	0.035**	(0.015)	3,137	123
Tongan	0.034	(0.027)	3,137	115
Russian	0.027***	(0.008)	3,137	53
German	0.022***	(0.002)	15,685	15,513
Hindi	0.015***	(0.003)	3,137	558
Rumanian	0.014**	(0.007)	3,137	417
Turkish	0.012**	(0.005)	3,137	263
Croatian	0.012**	(0.006)	3,137	65
Swahili	0.006	(0.005)	6,274	606
Finnish	0.005	(0.012)	3,137	373
Magyar	0.004**	(0.002)	3,137	578
Indonesian	0.002	(0.003)	3,137	130
Swedish	0.002**	(0.001)	3,137	888
Dutch	0.002**	(0.001)	9,411	4,746
Norwegian	0.002**	(0.001)	3,137	965

Pashto	0.002	(0.002)	3,137	26
Czech	0.001***	(0.000)	3,137	367
Burmese	0.001	(0.001)	3,137	19
Sinhalese	0.000	(0.001)	3,137	9
Danish	0.000***	(0.000)	6,274	1,200
Irish	0.000***	(0.000)	3,137	459
Afrikaans	-0.000***	(0.000)	3,137	6
Nepali	-0.000***	(0.000)	3,137	52
Bulgarian	-0.000***	(0.000)	3,137	85
Creole	n/a	n/a	3,136	2,252
Bantu	n/a	n/a	9,411	432

Notes: The table is an extension of Table 8, where we only show the results for a set of selected languages. The table is sorted on the size of the point estimate. The last column shows the # of US-born residents in d that speak the language of o at home.

Appendix Table 24: Accounting For the Effect of Ancestry

	FDI Dummy (2014)				
	(1)	(2)	(3)	(4)	(5)
Log Ancestry 2010	0.222***	0.213***	0.212***	0.213***	-0.025
	(0.021)	(0.068)	(0.068)	(0.025)	(0.028)
Sector Similarity (Rank Correlation)		0.012			
		(0.020)			
Sector Similarity (Cosine Correlation)			0.022		
			(0.023)		
Log # of residents in d that speak language of o at home				0.005	
				(0.006)	
Information Demand Index (standardized)					0.078***
					(0.013)
N	612,495	23,708	23,708	454,812	19,110
Destination FE	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes
Principal Components	Yes	Yes	Yes	Yes	Yes

Notes: This table shows IV regressions at the county-country (columns 1-4) and DMA-country (column 5) level. Each column is a variation of the simple specification (column 2 in Table 3) that has origin and destination fixed effects. All variables are defined as in the previous tables. The relatively low number of observations in columns 2 and 3 is due to data availability in the industry share of employment: When calculating the correlation between industries' share of employment in county d and country o, the correlation coefficient is missing for those country-county pairs that have at least one missing share of employment. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 25: Generational Effects

	FDI 2014 (Dummy)					
	IV	IV	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Log Ancestry 2010	0.187***		0.155***	0.242***		0.163***
	(0.024)		(0.022)	(0.043)		(0.014)
Log Foreign-born 2010		0.207***	-0.012	-0.082*		
		(0.014)	(0.031)	(0.049)		
Log Foreign-born 1970					0.286***	0.046
					(0.025)	(0.034)
N	612,495	612,495	612,495	612,495	612,495	612,495

Notes: The table presents the OLS (column 3) and IV (all other columns) estimates of equation (1), contrasting the effect of ancestry and first-generation immigrants (foreign-born) on FDI. The dependent variable is the dummy for FDI in 2014. All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. The coefficient estimates on these control variables are not reported in the interest of space. Standard errors are given in parentheses and clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. For column 4, the Kleinbergen-Paap rk LM statistic on the excluded instruments is 18.211 with a p-value of 0.150. We are thus unable to reject the null that our instruments do not induce differential variation in the two endogenous variables, and interpret any difference in the coefficient estimates with caution. The Kleinbergen-Paap rk LM statistic on the excluded instruments is 29.04 with a p-value of 0.007. We thus have sufficient power to detect differences between the coefficient estimates on the two endogenous variables.

References

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Antràs, P., D. Chor, T. Fally, and R. Hillberry (2012): "Measuring the Upstreamness of Production and Trade Flows," *American Economic Review Papers and Proceedings*, 102, 412–416.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2003): "Governance Matters III: Governance Indicators for 1996–2002," Working Paper No. 3106, World Bank.

Liang, J. (2017): "Cultural Similarity – Measurement using Google Trends," *mimeo University of Chicago*.

Nunn, N. (2007): "Relationship-Specificity, Incomplete Contracts, and the Pattern of Trade," *Quarterly Journal of Economics*, 122, 569–600.

Spolaore, E. and R. Wacziarg (2015): "War and Relatedness," *The Review of Economics and Statistics*, forthcoming.

Stephens-Davidowitz, S. and Varian (2015): "A Hands-on Guide to Google Data," *Working Paper*.