Italian Migration to the United States: The Role of Pioneers' Locations*

Matias Brum†

Job market paper.

The most recent version is available here.

October 17, 2019.

Abstract

This paper investigates the effect of size and location decisions of early migrants' flows on migration and settlement decisions of subsequent migrants from the same communities of origin. Filling a gap in the historical data, I focus on Italian mass migration to the US at the turn of the twentieth century and combine new data sets with a surname matching technique to generate new estimates of the yearly migratory flow from each Italian municipality to each US county. I exploit variation across time, origin municipalities, and destination counties and use an instrumental variables approach. I find that early migrants' location decisions matter: municipalities connected to (on average) more dynamic counties sent more migrants to the US later on. Moreover, municipalities that concentrated more early migrants in a high performing county experienced also higher outmigration to that county later on, and displayed lower concentration of further migrants in that destination.

Keywords: Age of Mass Migration, Italian migration, networks.
JEL Classification Numbers: F22, J15, J61, N31, N33.

*I am highly indebted to my first supervisor, Marco Manacorda, for his guidance and valuable comments, and to my second supervisor, Francesco Fasani, for multiple discussions and comments on earlier versions of the paper. I also thank Gaia Narciso, Javier Ortega, Barbara Petrongolo, Allison Shertzer, Andrea Tesei and Javier Vazquez-Grenno for many insightful comments. My thanks to participants at QMUL Economics Reading Group, RIDGE, 31st EALE Conference, Labour in History and Economics Conference, Fourth Workshop on Migration, Health, and Well-Being, and Migration and Mobility Workshop. Remaining errors are mine.

†Assistant Professor, IECON-FCEA, Universidad de la Republica. Address: Lauro Müller 1921, Office 401, Montevideo 11200, Uruguay. Email: matbrum@gmail.com.

1 Introduction

Do location decisions of pioneering migrants have an effect on size and settlement patterns of subsequent migratory waves from the same communities of origin? This paper focuses on this question for one of the main mass migration episodes history: Italian migration to the US at the turn of the twentieth century.

A first interest in the question stems from a policy perspective, as governments interested in controlling migratory flows may benefit from policies affecting early migrants' economic situation or location within the receiving country. Second, it is also interesting from an academic perspective, given that traditional models in the literature are silent on the topic (Harris & Todaro, 1970; Borjas, 1987). Moreover, although empirical research suggests a variety of potential mechanisms and channels for effects in opposite directions (e.g., McKenzie et al. 2013; Farré & Fasani 2013; Munshi 2003; Edin et al. 2003; Borjas 1994; Beaman 2011; McKenzie & Rapoport 2010), it has not directly answered the question, mostly due to data limitations. A central contribution of this paper lies precisely in adapting a machine-learning surname matching technique to combine unused datasets to fill this data gap, generating a sub-national origin-destination data set for the case under study.

Early migrants' outcomes may affect the size and location of subsequent migratory waves through information and resource provision, with a net effect stemming from type and quality of information, cuts in migration costs, reductions in migrant quality and potential congestion at destination.[1] Moreover, pioneers' sub-national settlement patterns may be an additional factor to consider when studying international migration: randomness in migrants' initial distribution across counties, or location decisions based on short run considerations may have important, long-run consequences.

Previous empirical research has documented a positive effect of the stock of past migrants and of wages in destination countries on further migratory flows from sending countries, at the national level.[2] Nevertheless, research on the effect of first-wave migrants' outcomes and location decisions on further migratory flows, at the two-way sub-national level, is practically non-existent. Lack of empirical studies at this

[1] In terms of information, prospective migrants may over- or under-estimate opportunities abroad (McKenzie et al., 2013; Farré & Fasani, 2013); pioneers' knowledge of local labour markets may attract further migrants (Munshi, 2003; Edin et al., 2003) while networks may be harmful after a size threshold (Borjas, 1994; Beaman, 2011). Pioneers' better locations may affect further migratory waves through economic resources, for instance through remittances, upfront coverage of migratory costs, or food and shelter provision upon arrival. These elements may increase or reduce incentives to migrate, and affect further migrants' characteristics (see McKenzie & Rapoport (2010) and references therein).

[2] See Hatton (2000), Hatton & Williamson (2005), and the review in Hatton (2010); see also Ortega & Peri (2013).

level stems from the difficulty in obtaining an adequate data set with information on migrants' sub-national origins and destinations. Data collected at the destination country usually lacks precise information on migrants' sub-national origin, while data recorded at the origin country usually omits information on migrants' sub-national destination. To overcome this restriction, I use the 1900 US Federal Census Population Schedules to obtain county of residence and surnames of the *full stock* of Italians in the US in 1900, and use the Italians to America Passenger Data File data set (which records ship manifest information for most Italian migrants at the time) to obtain the distribution of Italians by origin municipality and surname. Then, I extend and adapt a machine-learning matching methodology developed by Feigenbaum (2016) to match surnames between the two data sets, using the ship manifest information to impute a municipality of origin to the stock of Italians in the Census.

The combination of these two data sets enables to compute the stock of Italians by county of residence, municipality of origin and year of arrival. This reconstruction of yearly municipality-county flows fills an important gap in the historical data and opens new avenues for empirical research.[3] In this paper, I use this new data to provide causal evidence on the role of (sub-national) location decisions of pioneers on the size and settlement patterns of later migrants from the same municipalities of origin. Although the data is innovative and allows to overcome traditional problems in the literature, it is subject to a variety of measurement error problems. I show that the new data set is consistent with other historical sources, providing partial validation. The empirical estimation is further affected by endogeneity as, for example, unobservable factors could both drive early migrants' location decisions and the size and location decisions of further migratory flows. I address these issues with instrumental variables, using location decisions of non-Italian European migrants to the US to construct a counterfactual distribution of early Italian migrants across the US by municipality.[4]

Results show that an increase in the stock of first-wave migrants from a given municipality in a given county proportionally increases the emigration rate of later-wave migrants from the same origin to the same destination, but has no effect on the degree of concentration of later-wave migrants from that origin in that destination. Moreover, concentration of first-wave migrants from a given municipality in a given county reduces

[3]For example, I use this data to show that Italians from different municipalities located, on average, on counties that differ in income levels and growth rates. Further research may focus on the consequences of these differences on Italian local development.

[4]I depart from the common use in the literature of previous settlement patterns to instrument location decisions of later migrants (e.g. Altonji & Card 1991; Card 2001; Cortes 2008) as, by definition, there are no previous settlers before first-wave migrants.

the overall flow of later-wave migrants from the same origin to that destination, but induces greater concentration of the total flow to the US on that county. Moreover, income shocks at a given destination county barely affects the overall later-wave flow to the affected county and its concentration there.

As the key contribution of the paper, results show that municipalities that concentrated more pioneers in a given county experience higher outmigration to that same destination later on, and display lower concentration of further migrants to the US in that destination. A one SD change in the share of first-wave migrants from a given municipality that settled in a given county amplifies the pull effect of a one standard deviation change in county income, augmenting the later-wave emigration rate from the same origin to the same destination by an additional 0.28 per-thousand. This suggests that whatever changes in resources and information jointly stemming from pioneer concentration and income growth at destination mostly operates in a location-free fashion, as it increases both further migration and cross-county diversification. Results also show that municipalities that (on average) sent more pioneers to counties with higher income growth also sent more followers to the whole US. Results are very similar across a battery of robustness checks, and an extension shows that first-wave emigration rates have positive spill-overs on further migration from neighbouring municipalities, to specific counties and the US as a whole. Taken together, this paper shows that differences in the distribution of pioneers across the destination country matter, and can help explain differences in flows and location decisions of further migrants across sending communities.

This paper is related to four empirical papers. Damm (2009) and Åslund (2005) exploit plausibly exogenous variation in refugee placement in Denmark and Sweden respectively and find that later-wave migrants follow early migrants' settlement patterns, by origin country. Bauer et al. (2005) focus on language proficiency and find that Mexican migrants with better English choose US areas with smaller pre-existing networks. Lafortune & Tessada (2014) investigate the effect of existing networks on location and occupational choice of further migrants for migrants to the US in 1900-1930, and find that later-wave migrants follow the same settlement patterns across US states than previous migrants from the same countries of origin.

I depart from these papers and make three main contributions to the literature. First, I combine new data sets and adapt a matching procedure to construct estimates of yearly Italy-US migratory flows in the second half of the 19th century, at the sub-national level. To the best of my knowledge, there is no precedent for data at this detail level for this historical period, and it can be used for further contributions to many

research strands. Second, armed with this data, I contribute to the economic history and migration literatures as, to the best of my knowledge, there are no studies on the effect of early migrants' outcomes and settlement patterns on the size and location of further flows, at the two-way sub-national level. Focus on the sub-national level is interesting per se and has also the advantage that results can be better ascribed to the role of networks. Third, I focus on the interaction between networks and local economic conditions at destination. I identify the additional contribution to further migration by origin municipality to the US and to specific counties due to income growth, that stems from pioneers' networks. I show that origin communities are connected to different areas in the receiving country (in terms of income), and that this leads to variations in later-wave flows across the sending country.

The rest of the paper is structured as follows. Section 2 reviews the literature, Section 3 briefly presents the historical and institutional context and Section 4 presents the data and descriptive statistics. Section 5 summarizes the surname matching procedure and uses the matched data to characterize migrants' origins and destination patterns. Section 6 presents the estimation strategy and main challenges and proposes an instrumental variable approach to address endogeneity and measurement error concerns. Section 7 presents the main results, Section 8 discusses extensions and robustness checks and Section 9 concludes.

2 Related Literature

The starting point in the literature is to consider wage differentials between origin and destination countries as a driver of migratory flows, borrowing from the seminal paper of Harris & Todaro (1970). This model and subsequent extensions usually do not explicitly distinguish between early and late migrants.[5] Introducing incomplete information allows to interpret these models in a new light: if origin communities have no information at the start of the migration process, pioneers may fill the gap and economic conditions at their destination may be informative of the wage level, affecting the subsequent attractiveness of migration through changes in the expected wage gap.[6] Empirical evidence shows that prospective migrants can over- or under-estimate opportunities abroad: McKenzie et al. (2013) find that prospective Tongan

[5]See Harris & Todaro (1970), Borjas (1987) and the review in Massey et al. (1993).

[6]With a different approach, Burda (1995) and O'Connell (1997) model migration as an investment decision taken under uncertainty. They show that uncertainty on the level and the trajectory of wages at destination can hinder or foster migratory waves. In these models, information provided by first-wave migrants could increase or reduce further migratory flows.

migrants to New Zealand significantly underestimate their income at destination, while Farré & Fasani (2013) suggests that potential Indonesian migrants may overestimate migration gains and react to new information by reducing their propensity to leave.[7]

The literature on network effects argues that migrant settlements in the destination country may reduce costs and increase benefits of migration through, for example, information provision (see (McKenzie & Rapoport, 2010) and references therein).[8] For newly arrived migrants, networks provide information on the local labour market, referring new arrivals, reducing search costs and improving the quality of the matches. Munshi (2003) and Edin et al. (2003) find a positive relationship between network size and labour market outcomes respectively for Mexican migrants to the US and for refugees in Sweden. In contrast, the size of the migrant community may be harmful for new arrivals due to increased competition for jobs. Investigating resettled refugees in the US, Beaman (2011) finds that enlargement of a network harms the currently arriving cohort, but increases income and employment prospects for future ones. Besides, large migrant communities may reduce incentives for human capital investments, as learning the local language (Borjas, 1994). Moreover, if networks provide good quality information to origin communities, then prospective migrants should make better choices (better decide whether to migrate or not and, conditional on migrating, better selecting a sub-national destination).

Networks also reduce migration costs through financial resources, for example sending remittances, covering upfront costs and providing food and shelter upon arrival (McKenzie & Rapoport, 2010; Munshi, 2003; Massey et al., 1993). If networks reduce costs but provide no valuable information, then they would increase further migration with no impact on new arrivals' location decisions: later-wave migrants would follow previous ones into enclaves even if better locations were available. Besides, part of the reduction in costs applies only to early migrants' location at destination (e.g., food and shelter), increasing their attractiveness. Moreover, reductions in migration costs may induce negative selection of further migrants (McKenzie & Rapoport, 2010); lower quality migrants may have more incentives to follow previous waves and less capacity to succeed in different destinations.[9] Finally, remittances may hinder further migration if their effect in origin communities offsets reductions in migration costs.

[7] Relatedly, Borjas & Bratsberg (1996) suggest that erroneous information on opportunities at destination lead to return migration.

[8] Other channels involve reduction of non-monetary (e.g., psychological) costs, provision of a safety net during the initial stages of arrival and integration, etc.

[9] In Brum (2018), I use the Italians to America data set and exploit variation across municipalities to show suggestive evidence of network size negatively affecting migrant quality.

Empirical research has documented a positive effect of the stock of past migrants and of wages in a destination country on further migration from sending countries, at the national level. For historical studies, see Hatton (2000), Hatton & Williamson (2005), and the review in Hatton (2010); see Moretti (1999) for the case of Italy during the Age of Mass Migration. Empirical research has also focused on the role of different characteristics of the destination country, at the sub-national level, as drivers of settlement patterns of different international migrants (see Dunlevy & Gemery (1978) and Dunlevy & Saba (1992) and references therein). However, this strand of research remains silent on whether later arrivals followed location decisions of early migrants. In modern settings, Ortega & Peri (2013) construct bilateral flows for 120 origin countries and 15 OECD destination countries, and document a positive effect of per capita income on further migratory flows.[10]

This paper is related to four previous empirical papers. Damm (2009) and Åslund (2005) exploit plausibly exogenous variation in refugee placement in Denmark and Sweden respectively and find that later-wave migrants follow early migrants' settlement patterns, by country of origin. Bauer et al. (2005) studies the role of networks on location choices of further migrants, focusing on language proficiency of Mexican migrants to the United States, and find that migrants with better English proficiency choose US areas with smaller networks. Closely related to this paper, Lafortune & Tessada (2014) investigate the effect of existing networks on location and occupational choice of further migrants to the US in 1900-1930, and find that later-wave migrants follow the same settlement patterns across US states than previous migrants from the same countries of origin.

I make a contribution to the literature and depart from these related papers in many ways. First, I consider narrow destination locations within the US (counties) and define place of origin at the municipality level instead of country of origin. Focus on a two-way sub-national level has the advantage that results can be better ascribed to the role of networks. For example, municipalities are small enough as to be likely that information or resources sent by first-wave migrants actually reach prospective migrants from the same community, something which is less credible when the unit of analysis is a whole country.[11] Second, I focus on the interaction between networks

[10] See Mayda (2010) and Grogger & Hanson (2011) for other research on the effect of wages at destination on international sorting of migrants. See Andersson et al. (2016) and Karadja & Prawitz (2016) for the effect of outmigration on sending communities, respectively on technological innovation and political change.

[11] Similarly, empirical research at the national level considers wages at the national level; this may not be representative of the conditions taken into account by prospective migrants if the perception of conditions at destination differ across the sending country. For example, Abramitzky et al. (2014)

and local economic conditions at destination, identifying the additional contribution to further migration (to the US and to specific counties) generated by changes in local economic conditions, arising from pioneers' presence.

3 Historical and Institutional Context

In one of the greatest episodes of mass migration in modern history, between 1875 and 1925 16.6 million Italians left their country; by 1924 about 9% of the Italian population resided in the US. Outmigration started with Italian unification (1861), fostered by differences in growth and development trajectories between Italy and the Old and New World, the fall in transport costs, and many local, regional and national negative shocks (Cerase, 1974; Sori, 1979; Cinel, 2002; Del Boca & Venturini, 2005).[12]

During the 19th century, Italy remained a poor and underdeveloped nation, with a majority of the population working in agriculture in predominantly rural areas, especially in the south (Sori, 1979). Woes stemming from low productivity and wages and lack of employment were aggravated by population growth, and the late arrival of the industrial revolution was concentrated in the so-called industrial triangle of the north-west (Milan, Turin, Genoa), with limited national spillovers (Ciccarelli & Fenoaltea, 2013). For those outside of this triangle, real wages, living conditions and future prospects were brighter elsewhere, given industrial growth in Germany, France and the United Kingdom, and abundant natural resources in the United States, Argentina and Brazil. Transatlantic travel costs fell from the 1850s with the transition from sail to steam: vessels gained in capacity and quality and trips became cheaper, safer and shorter (Keeling, 1999b; Sánchez-Alonso, 2007).[13]

Figures C.1 and C.2 in Appendix C show total migration flows over 1870-1945 and the share of migrants coming from the south. Migration grew explosively, reaching 500,000 migrants in 1900 and stabilizing over 1900-1914, when annual outflows averaged 1% of the total population. Migration fell to almost zero during the First World War

find that long-term immigrants from countries with real wages above the European median held higher-paid occupations than US natives upon arrival, while those from countries with wages below the median held occupations equally or lower paid.

[12]Unification -1840s to 1870- was costly, disruptive, and ended in the imposition of Piedmontese taxes, regulations, officials and even constitution across the country. Liberal trade and tariff provisions hurt the south and anti-clerical rules fostered revolts in Sicily and Naples (1860-1870). Other shocks include wars (e.g. Austria, 1866), economic depressions (e.g. Long Depression, 1873-1879) crop failures (late 1840's, 1897), decimation of vineyards (1891-1899), and earthquakes (1857, 1873).

[13]Declines in costs were irregular and took longer to reach the longer routes: fares fell from the 1850s for the UK-US routes and from the 1870s for the Spain-Southern Cone ones (Sánchez-Alonso, 2007). Passage prices varied sharply due to shipping cartels (Keeling, 1999a).

and never recovered, dropping again due to restrictions imposed in the late 1920s and the Great Depression. In terms of regional composition, in the 1870s mostly northern Italians left for neighbouring countries; in the 1880s northern flows diversified and went also to South America, while southern migration started to grow. In the 1890s the share of southern migrants doubled (20% to 40%), and reached parity with the north in the 1900s; about two-fifths (pre-1900) to two-thirds or over (1900 onwards) of southern migrants went to the US. Italian migrants to the US were mostly working-age men, illiterate and unskilled or low-skilled agricultural labourers or peasants from the south and center-south. They settled predominantly in large urban areas in the north-east, and worked in non-agricultural activities, as construction, railways, and mining (Del Boca & Venturini, 2005).

Through most of the Age of Mass Migration Italy and the US maintained open borders. In Italy, migration started to be recorded only in the 1870s and a *Commissariat of Emigration* was created in 1901, mostly to protect, train and support migrants.[14] Outmigration was partly seen as a solution to southern woes; this liberal approach was overturned by Mussolini only in 1926.[15] In the US, positive economic conditions together with favourable naturalization and land availability policies attracted massive inflows of north and west European migrants, during most of the 19th century.[16] Though anti-immigration sentiment grew in the second part of the century, the first bill to successfully restrict European migration was passed as late as 1921.[17] US restrictions on entrance, exit restrictions of the Italian Fascist government and the Great Depression meant a sudden stop to Italian flows to the US in the late 1920s and 1930s.

[14] This body enforced fixed ticket costs, granted licenses to carriers, regulated remittance transfers, and dealt with foreign governments protecting Italian migrants abroad, among others.

[15] In 1926 a law invalidated all passports held by Italians in Italy; strict eligibility criteria for new passports were passed next year. In 1930, penalties and fines were introduced for clandestine migrants or those assisting them.

[16] Naturalization was simple and straightforward and managed at the state level until 1906, requiring only five years of residence. The Homestead Act (1862) granted about 65 hectares of public land to any US citizen (or migrant that had filed to become one) willing to settle for at least five years and improve the land, at no cost.

[17] The Literacy Act (1917) barred entry to illiterate migrants but was unsuccessful due to its laxity, exemptions, and literacy improvements in the sending countries (Daniels, 2005). The Emergency Immigration Act (1921) restricted annual migration from any country to 3% of the stock of residents from that country in the 1910 Census; the Immigration Act (1924) lowered the quota to 2% of the 1920 stock and imposed a cap on total immigration at 150,000 individuals from 1927 onwards.

4 Data

4.1 Data sources

1900 United States Federal Census Population Schedules

The key data source is the 1900 US Federal Census Population Schedules (hereafter Full US Schedules). Digitalized by the Church of the Latter-day Saints, it includes information for *all* US residents in June 1st, 1900, for a small set of demographic variables: age, gender, year of arrival, marital status, position in the household, name, surname, place of birth and county and state of residence. The data allows the identification of the full stock of Italian-born individuals residing in the US in 1900, their surnames, and their county of residence. Of the about 76 million individuals in the data set, 489,041 declare being Italian born. This figure matches a different historical source, suggesting the data set effectively captures the full stock of Italians in the US at the time.[18]

Table 1 presents descriptive statistics for Italians in the data set. It shows that men are almost two-thirds of all Italian residents and that the bulk of migrants arrived in the last years of the 20th century (median arrival at 1893). Italians are 31.2 years old on average, mostly married (57.8%), and after Heads (29.9%) and Wives (22.2%), the most common position in the household is Boarder (19.4%).[19]

Table 1 here

United States Census IPUMS samples

Further information for the US comes from the 100% IPUMS samples of the US Census for 1880 and 1900 (Ruggles et al., 2015). These include several socio-demographic characteristics and a variable called Occupational Score Index, which assigns to each individual the annual median income of his occupational category in 1950. Though this variable takes the same value for all individuals with the same occupation, Abramitzky et al. (2012) argue that it is a good proxy for lifetime earnings, and it is among the best sources of information on income for the period. Averaged at the county level, this variable proxies economic conditions by county of destination.[20]

[18] 484,027 Italians according to Carter et al. (2006).

[19] Boarding was a cheap mean of housing that also provided access to job networks; boarders clustered by ethnicity or origin and provided additional income to household heads. In my data, 95.3% of boarders are men and one-third of them are married (their wives most likely in Italy).

[20] The US Census does not report income until 1940. In this year, average county income and average Occupational Score Index are highly correlated. As Average Occupational Score Index reflects

Table 2 shows descriptive statistics for the US population in general and living in counties with at least one Italian. The table shows a predominantly young and rural population (average age of 24.1; 73.6% residing in rural areas). With almost half of workers employed in agriculture and more than half of children attending school, the literacy rate was high (83%) and migrants were a sizeable proportion of the population (13.3%). The average US resident in 1880 had an annual income of 18,353 (1950) US dollars, although the standard deviation is large (10,158). Note that the average income in counties with at least one Italian is slightly higher (19,675 USD). Moreover, Italians settle in counties with lower proportion of rural and black population, higher proportion of immigrants, and lower proportion of the work-force employed in agriculture, compared to the rest of the population.

Table 2 here

Italians to America Passenger Data File

The Italians to America Passenger Data File data set (ITA from here onwards) records information on Italian migrants to the US from ship manifests over 1855-1900; the almost 850,000 individuals represent about 82% of total Italian arrivals in the period.[21] Crucially, this data set reports place of last residence, name and surname.[22] Table C.1 in Appendix C reports descriptive statistics. Consistent with historical studies, three-quarters of migrants are men, a majority of them employed in unskilled or low-skilled agricultural activities before departure. Migrants are young (average age of 26.7) and come from the port of Naples (41.4%) and Genoa (40.3%).[23]

Unfortunately, place of last residence is missing or invalid for about 60% of the sample.[24] To address this issue, I compute the distribution of migrants by surname

counties' occupational structure, cross-county variation responds to differences in economic structures.

[21] According to United States official arrival statistics (Carter et al., 2006).

[22] Legislative requirements lead to surnames being missing only for 0.6% of the sample, which I drop. I also drop migrants with US citizenship or last residence as they should have been registered in a prior trip, losing an additional 3.8% of the sample.

[23] 3.4% of the sample reports more than one port of departure, probably reflecting the ship's route (e.g., 'Marseilles & Naples'). I took the first mention as the port of origin. Some individuals may report the *last* port of departure instead of the actual port of embarkation, which could explain the figure for Havre (France), which at the time was the last continental port for ships coming from Italy. More than 99% of individuals arrive at the port of New York.

[24] Immigration was regulated at the state level up to the Immigration Act of 1891, which unified regulations; in 1893 the federal government set a new standard on mandatory information in ship manifests. Prior to 1898, missing data mostly corresponds to lack of information for all passengers in a trip: 87.3% of individuals with missing last residence come in trips with no residence information for *any* passenger. Table A.1 in Appendix A presents the incidence of missing residence year by year.

and last residence for individuals with non-missing information, and use it to impute a municipality to individuals with missing or invalid origin. This procedure is detailed in Appendix A. This exercise assigns a set of potential municipalities of origin and corresponding weights reflecting the probability that an individual comes from a given municipality, to 83.7% of the problematic individuals in the ITA data set.

1881 Italian Census Summary

The Italian Census Summary of 1881 reports information on several variables at the national, regional and provincial level.[25] Although population data is available for all municipalities, information on the demographic structure of the population or the occupational distribution of working-age individuals is available for only a very small sub-set of municipalities. In Appendix C, Table C.2 summarizes the main variables of interest taken from the 1881 Census and the number of municipalities with data per variable; Table C.3 presents basic descriptive statistics for selected variables.

This data shows that in 1881 Italy had 29.3 million inhabitants unevenly distributed across the country: median and mean population per municipality were below 4,000, while 5% of the municipalities held 35.8% of the population. Provincial capitals were larger than other municipalities and hosted 17.1% of the population. Agriculture was the main sector of occupation, employing 45.5% and 27.0% of working men and women respectively.[26] Furthermore, Italy was lagging behind in terms of educational attainment: the national literacy rate stood at 38.1%, about half of the value for Germany, France, England and other north and west European countries at the time (Johansson, 1977).[27] Already in 1881, economic development was unevenly distributed across the country, which is apparent from literacy rates. Literacy surpassed 70% in the north (close to north and west European countries) and fell to two thirds to one third of that level in the south.[28]

[25] *Censimento Della Popolazione del Regno d'Italia al 31 Dicembre 1881*, (Ministero di Agricoltura, Industria e Commercio, 1884). In 1881 Italy was divided in 16 Regions, 69 Provinces, numerous Districts (*Circondario*) and Subdistricts (*Mandamento*), which comprised around 8,000 Municipalities (*Comune*). In the ITA data set, flows over 1855-1880 represent 1.7% of flows over 1855-1900, hence the 1881 Census provides information at the start of the migratory wave.

[26] Occupational data in the census covers individuals aged 9 to 65.

[27] Literacy is recorded from ages 6 and older.

[28] Table C.4 in Appendix C reports literacy rates by province, for men, women and in total. Note also how rates for men are two to three times higher than for women in the south, while the gap is much lower in the north.

Other historical data

Two other sources provide information on Italian migration to the US. First, historical records compiled by the Italian government (referred as *Annuario* data onwards) provide data on regional migration to the US over 1876-1900.[29] The *Annuario* is the best source of information on Italian migration for the period, and is still used by scholars today. The main shortcoming is that migration data is based on *requests* of the first document needed for passport *applications*, and may misstate actual regional flows (albeit not by much).[30] Second, historical records compiled by the United States government report annual arrivals of Italian migrants from 1820 onwards.[31]

I use these sources to contrast migration flows observed in the ITA data set and ease concerns over the imputation process. A detailed explanation of this exercise is in Appendix A. Results in Table A.4 show that: i) at the national level the ITA data set accurately tracks US and Italian historical records; ii) both historical series behave similarly; iii) Italians' outflows *by region* observed in the ITA data set are in line with Italian historical records. These results reduce concerns over measurement error, and suggest that the lack of coverage of the ITA data is not concentrated in given years, and that the imputation exercise does not lead to severe mis-estimation of the regional composition of migrants.[32] Section 6.2 discusses measurement error; next, I use all the data to present origin and destination patterns of Italian migrants to the US.

4.2 Italians in the US: Origin and destination patterns

First, note that the Full US Schedules provides information on the *stock* of Italians observed in 1900, while the ITA data set captures the *inflow* of Italians over 1855-1900. The cumulative flow from the ITA data set is higher than the stock in the Full US Schedules due to mortality and return migration. US historical statistics show 1,040,446 Italian arrivals over 1820-1900 (Carter et al., 2006), suggesting a raw rate of return

[29] The *Annuario Statistico della Emigrazione Italiana dal 1876 al 1925* (Commissariato Generale Dell'Emigrazione, 1927) was commissioned by Mussolini to systematize and harmonize the information registered by many institutions at different administrative levels.

[30] The *nulla obsta* was issued at the municipal or provincial level and stated government's non-opposition to the intention of obtaining a passport. Opposition could stem from arbitrary to sensible reasons (e.g. past criminal convictions). *Nulla obsta* issuance inaccurately captures true migration: passport applications could be rejected at any point of the process, individuals could desist from migrating, leave in other years or to destinations other than declared, or migrate without a passport.

[31] In *Historical Statistics of the United States: Millennial Edition* (Carter et al., 2006).

[32] Imputation errors concerning municipalities from the same region do not affect the observed regional distribution, only mis-imputations that cross over regional borders do. Results then cannot fully ease concerns on the imputation exercise, but suggest it does not distort the regional distributions too much. Also, different errors could cancel each other out.

(not accounting for mortality) of 53%, in line with rates for Italians and other southern European migrants in the literature (Bandiera et al., 2013; Hatton & Williamson, 2005). Note also that Italians in the Full US Schedules in 1900 may not reside in their original settlement location due to internal migration.[33] In strict terms, the Full US Schedules captures behaviour and characteristics of *stayers alive in 1900*. Potential problems arising from mortality, internal and return migration are discussed in Section 6.2.

Settlement patterns in the US

Using the Full US Schedules, Figure 1 displays the distribution of Italians by US state, showing strong concentration in the northeast: New York (37.7%), Pennsylvania (13.6%), New Jersey (8.6%) and Massachusetts (6.0%) hold 66% of Italians. California (4.7%) and Louisiana (3.6%) were the main receivers beyond the northeast. Italians cluster in counties with large urban centers, as New York, Brooklyn, Philadelphia, Chicago, Boston, Essex, Pittsburgh, Jersey City, Providence, San Francisco and New Orleans. Heavy concentration of Italians by county reflects in a Gini Index of 0.7615 and a Dissimilarity Index of 0.6167; this last figure implies that almost 62% of Italians would need to relocate to obtain the same county distribution of non-Italians.

Figure 1 here

Figure C.3 in Appendix C shows the distribution of arrivals by state and year of arrival for the top eight receiving states (holding 83.1% of all Italians).[34] California and Louisiana were the main recipients of very early settlers although their importance declines from the 1860s onwards, in favour of New York and other northeastern states: in the late 1890s New York received about 40% of all Italian migrants.[35] To better visualize the changing pattern of Italian settlements, I compute median latitude and longitude of Italians' destination by arrival year; Figure 2 maps the trajectory followed by the median migrant.[36] The median Italian always located in the northeast; close to Chicago early on, close to Philadelphia in the early 1870s and in New York City and surrounding areas from 1875 onwards.

[33] Nevertheless, Boustan et al. (2010) suggest low internal mobility of international migrants, during the Great Depression.

[34] A regression-based exercise compares the distribution of total flows by year observed in the Full US Schedules with other historical sources for the US and Italy; Table B.2 in Appendix B shows that census data accurately captures the structure of Italian migration in the period.

[35] See Table C.5 in Appendix C for the distribution across all states by five-year sub-periods.

[36] I take the coordinates of the centroid of each county as the latitude and longitude of each migrant's destination. The mean migrant follows a similar trajectory.

Figure 2 here

Outmigration patterns from Italy

Using the ITA data set, Figure 3 presents the distribution of migrants by province of origin and shows that Italians come mostly from the south (the so-called *mezzogiorno*). The regions of Campania (29%), Sicily (16.7%), Calabria (10.9%) Basilicata (7.0%), Molise (6.5%) and Abruzzo (5.6%) represent 75.8% of total outmigration, consistent with Italian historical sources. Within the south, migrants come mostly from the west coast of continental Italy (Cosenza, Potenza, Naples, Salerno, Avellino) and northern Sicily (Palermo, Messina). Figure C.4 in Appendix C shows the distribution of migrants by region of origin and year of departure for the top ten sending regions, responsible for 88% of all Italian migrants.[37] It shows no clear pattern up to 1880 and stable trends onwards: Campania was always the main sender and Sicily's importance increased in relative terms especially in the last five years.[38]

Figure 3 here

This brief characterization suggests differences in destination choices of early and late migrants, and similarities in origin regions and provinces. To better study the way in which location decisions of early migrants affected those of later ones, a key piece of information is missing: the place of origin of Italians in the US. The next section applies a surname matching technique to bridge this gap.

5 Surname Matching and Municipality Imputation

This section explains the machine-learning approach followed to match Italian surnames in the ITA data set and the Full US Schedules, describes the subsequent municipality imputation process and presents a comparison of the resulting data with other sources as a validation exercise. A detailed explanation is in Appendix B. Note that Abramitzky et al. (2019) evaluate different methods for record linkage and conclude that coefficient estimates and parameters of interest are similar when using linked samples based on each method.

[37] The data records year of arrival to the US; I refer to year of departure for simplicity.
[38] See Table C.6 in Appendix C for the distribution for all regions across the period.

5.1 Surname matching

In a nutshell, I use the distribution of municipalities of origin across surnames observed in the ITA data set to assign a municipality of origin to Italians observed in the Full US Schedules. This exploits the geographical concentration of Italian surnames, which allows a relatively precise identification of municipality of origin by surname.[39] Table C.7 in Appendix C presents basic descriptive statistics on the surname distribution observed in the ITA data set and the Full US Schedules. The table shows that Italian surnames are highly concentrated by origin municipality and destination county.

The surname matching process faces two challenges. First, 86.4% of married Italian women carry their husbands' surname, while at least 72.5% got married before migrating and should have a different surname (in Italy women keep their surname after marriage).[40] To address this issue I drop women (35.2%) of the sample.[41] Second, surnames are registered with error: only 37.7% of all Italian men in the Full US Schedules can be matched to the ITA data set when using original surnames. To address this issue, I adapt a machine-learning matching methodology developed by Feigenbaum (2016), originally used to link individual records in two samples. The process is summarized below, details are in Appendix B.

In Step 1 I consider all possible combinations of surnames in both samples, use the Bi-gram and Jaro-winkler similarity algorithms to construct measures of match quality, and use conservative thresholds to reduce the space of potential matches. The resulting data set (circa 4.8 million surname pairs) is the basis of the machine-learning process. In Step 2 I generate *training data*: a set of surname pairs manually coded as correct matches. Identical surnames in both data sets are classified as a correct match. I code further pairs as correct based on rules regarding differences in surnames' last letters, correction of duplicate/triplicate letters, and treatment of blanks and apostrophes in compound surnames. Surname pairs with high values of the similarities indexes are also coded as correct. These criteria yields a *training data set* of about 350,000 surname pairs, manually coded into correct (30.4%) or incorrect (69.4%) matches.

In Step 3 I run a logit regression of the manual coding against a characteristics of

[39] This links to further research that relies on surname matching, in the migration literature (Spitzer & Zimran, 2014; Abramitzky et al., 2012), and beyond (Güell et al., 2017; Feigenbaum, 2014). In particular, see Gagliarducci & Manacorda (2016) and Mastrobuoni (2015) for research exploiting geographical concentration of Italian surnames.

[40] These represent 20.4% and 14.8% respectively of all Italians in the Full US Schedules.

[41] This may be an advantage for the analysis, if migratory flows of women respond to family reunification decisions (taken by the male household head) (see for example Hatton & Williamson (1998)). If women purely follow a household head and all women come from the same municipality as the household head (an extreme case), their exclusion should not have substantial effects.

the surnames and surname pairs, using the *training data set*. Step 4 uses regression results to predict the probability of a surname pair being manually coded as correct, in the same data set. In Step 5 I adapt a method proposed by Feigenbaum (2016) to transform these predictions into matches and evaluate their quality. This involves a grid search for parameter values that optimize two measures of match quality, built based on a comparison between the predicted match and the actual, manual coding.

In Step 6 I use regression results to predict match probabilities for uncoded surname pairs, and use the parameters from the grid search to transform predicted probabilities into matches. Then, I manually audit a sub-sample of these predicted matches, incorporate the audited pairs to the training set, and repeat the process. This way, the subsequent logit coefficients and grid parameters incorporate any new information from the new observations. I repeat the process eight times in total and take the last set of predicted matches as definitive. To remain cautious, I keep surname pairs for which the average between the Bi-gram and Jaro-winkler indices is above 0.75 and, for each surname in the Full US Schedules, keep only the surname in the ITA data with the highest average. At the end of the process, 68% of the Italian men observed in the Full US Schedules can be matched to the ITA data set.

5.2 Municipality imputation

Next, I use the matched surname pairs to assign origin municipalities to Italians in the Full US Schedules. Appendix B describes this imputation procedure. The sample comprises about ten million observations, corresponding to 260,119 individuals. Each observation is an individual-municipality pair with a weight given by the proportion of Italians in the ITA data set with that surname that come from that municipality.

As an example of the uses of the data, I aggregate it into a Provinces×Counties matrix capturing the share of the stock of Italians by origin province in each county in 1900.[42] I use these proportions as weights, and compute the average income at destination of the mean migrant by province of origin. Figure 4 plots this on a map of Italy, which shows differences of up to a thousand dollars across provinces. Campania and surrounding provinces have higher levels than the rest, which means that migrants from different provinces located, on average, on sufficiently different counties in the United States.

[42]The data allows the construction of matrixes capturing the importance of each origin geographical unit as a source of the stock of migrants by destination geographical unit, and viceversa. For example, Table B.1 in Appendix B shows municipality-county connections; a link is established if a matrix cell takes a positive value (positive probability that an Italian in that county has a surname from that municipality). The table shows that most municipalities are linked to few counties, and viceversa.

Figure 4 here

To further motivate this paper, I focus on migrants arriving before 1885 and compute the average occupational score index and its average growth over 1880-1900 by origin municipality. The maps below show wide disparities across municipalities both in terms of starting conditions and performance across time.[43] If migrants from different origin communities experienced different levels of economic wellbeing and dynamism, this may have lead to differences in remittances and information flows, and thus on size and distribution of further migratory flows. This is precisely what is investigated in the rest of this paper.

Figure 5 here

5.3 Validation exercise

Here I contrast this new data set with other sources; Appendix B details the procedure. Results in Table B.2 show that the yearly migratory flows observed in this matched data set fit very well with other historical sources. Moreover, the yearly distribution of Italian migrants *by region* in the matched data are in line with the distribution observed in the unmatched ITA data set (shown in Appendix A to track the regional distribution seen in historical records). These results suggest that the sub-sample of matched men in the Full US Schedules is representative of the yearly Italy-US flows during the period, and that the surname matching and municipality imputation procedures do not lead to severe mis-estimation of the regional composition of yearly migratory flows, reducing concern over measurement error.[44] Finally, note that the detailed data set on two-way sub-national yearly migration flows is entirely new: it is a contribution to economic history in general and an input for further studies on mass migration.

[43] The municipality corresponding to the first percentile of the distribution of average occupational score index (*Boretto*, 18,326 USD) is about 5,000 dollars below the one in the ninety nine percentile (*Gottolengo*, 23,879 USD). In dynamic terms, the municipality in the first percentile of the growth distribution (*Follina*) experienced a reduction in mean occupational score index of 1.68% while that on the ninety nine percentile (*Vestone*) witnessed growth by 8.5%.

[44] As caveats, mis-imputation of municipalities from the same region does not affect the observed regional distribution (only those crossing regional borders do). Results then cannot fully ease concerns on the surname matching and municipality imputation, but suggest that errors are not severe enough as to affect the yearly regional distribution. Also, different errors could cancel each other out.

6 Empirical Strategy

6.1 Estimating equations

I start with a simplified view of migratory decisions at the municipality-county level, onwards referred to as the Municipality-County Model. In period $t = 0$, a first wave of migrants -the pioneers- leave from municipality m to county c. Later, prospective migrants in m in period $t > 0$, decide whether to migrate to c, or not. This decision depends on municipality and county characteristics, on the (expected) earnings at destination and on the stock of pioneers from m already settled in c, among others. The baseline estimating equation is the following:

$$Y_{mct} = \alpha + \beta_3 I_{ct} + \beta_4 \frac{N_{mc0}}{Pop_{m0}} + \beta_5 I_{ct} \times \frac{N_{mc0}}{N_{m0}} + \beta_6 \frac{N_{mc0}}{N_{m0}} + d_t + \varepsilon_{mct} \tag{1}$$

Y_{mct} is the outcome variable, defined below. I_{ct} represents average county income. $\frac{N_{mc0}}{Pop_{m0}}$ represents the initial ($t = 0$) stock of migrants from m residing in c, as a proportion of m's baseline population, capturing municipality m's first-wave outmigration rate, a proxy for network effects. $\frac{N_{mc0}}{N_{m0}}$ reflects pioneers' location decisions: it captures the proportion of first-wave migrants from m to the whole US that settled in county c.[45] The interaction term ($I_{ct} \times \frac{N_{mc0}}{N_{m0}}$) measures the additional effect of a change in average county income on the outcome, arising from a change in the share of pioneers from m in c. d_t is a time fixed effect that controls for time-specific shocks common to all counties and municipalities.

The main outcome of interest is the ratio of migrants from m to c in t as a share of m's baseline population (i.e., the outmigration rate from m to c in t, $\frac{N_{mct}}{Pop_{m0}}$). In this case, β_3 captures the effect of county income on the flow of migrants from m to c in t, reflecting the 'pull effect' of contemporaneous economic conditions at destination. β_4 captures the effect of changes in first-wave network size on the probability of outmigration from the same municipality to the same county (in $t > 0$). β_6 captures the effect of changes in the degree of pioneer concentration on county c on further flows from the same municipality to the same county, and approximates the relative importance of county c among all other sources (counties) of information and resources for individuals in m. The coefficient on the interaction term, β_5, is the main parameter of interest. It captures the additional effect of county income on further migration to a given county that stems from greater concentration of pioneers from m in that county.

[45] This term can be thought to capture the 'intensity' of an information signal: it approximates the relative importance of county c among all other sources (counties) of information for individuals in m. A similar interpretation holds for the case of remittances and resource provision in general.

A positive sign indicates that the same increment in county income leads to greater increases in the probability of migrating from m to c in t, the larger the proportion of pioneers from m that initially settled in c.

Note that any effect on the outcome, the later-wave outmigration rate from m to c, may stem from a combination of changes in the total outmigration rate from m to the whole US and changes in the distribution of later-wave migrants across US counties.[46] To disentangle these effects, I use as another outcome variable, the concentration of later-wave migrants from m to the US in county c, that is, the proportion of all later-wave migrants from m to the whole US that settles in county c ($\frac{N_{mct}}{N_{mt}}$. Now the coefficients reflect the effect of the variables of interest on the distribution of followers across the US. In particular, the coefficient on the interaction term, β_5, captures the additional effect of county income on further concentration of migrants in county c that stems from greater concentration of pioneers from m in the same county. A positive sign indicates that the same increment in county income leads to greater increases in the probability of locating in c in t when migrating to the US, the larger the proportion of pioneers from m that initially settled in c.

I estimate four different specifications of the Municipality-County Model. Specification 1 corresponds to Equation 1 above. Specification 2 adds county characteristics (X_{2ct}): distance to New York, proportion of men, average age, proportion of blacks, school attendance rate (ages 6-15), literacy rate (ages 10 and older), and proportion of rural population.[47] It also adds interactions between time fixed effects and baseline municipality characteristics ($X_{1m0} \times d_t$): distance to main port, shares of employed men and women working in agriculture, men and women employment and literacy rates, and share of population in five age groups by gender. This specification filters out the effect of other (contemporaneous) county characteristics that may affect migration rates, and approximates the same for municipality characteristics with the interactions; all additions control for other drivers of later-wave migration rates.

Specification 3 adds municipality and county fixed effects (d_m, d_c), which control for municipality- and county-specific (observable and unobservable) characteristics that may structurally affect both the size and distribution of both first- and later-wave migrants to each county. Specification 4, my preferred one, adds two-way interactions between municipality, county and time fixed effects (d_{mt}, d_{mc}, d_{ct}). These fixed effects control for municipality-specific time shocks, county-specific time shocks, and for com-

[46] For example, an increase in later-wave outmigration from m to c could reflect a reduction in total outmigration from m to the US that is more than compensated by increased concentration of (the reduced flow of) followers in c.

[47] All county characteristics are time-varying except for distance to New York.

mon (time-invariant) municipality-county idiosyncratic effects. In this specification, corresponding to equation 2 below, only the main parameter of interest (β_5) can be identified. Identification comes from municipality-county-time variation.

$$Y_{mct} = \alpha + \beta_5 I_{ct} \times \frac{N_{mc0}}{N_{m0}} + d_{mt} + d_{mc} + d_{ct} + \varepsilon_{mct} \tag{2}$$

Next, I aggregate the Municipality-County Model by time and municipality; summing over c I obtain the following baseline estimating equation:

$$Y_{mt} = \alpha + \beta_4 \frac{N_{m0}}{Pop_{m0}} + \beta_5 \sum_c I_{ct} \times \frac{N_{mc0}}{N_{m0}} + \beta_6 \sum_c \frac{N_{mc0}}{N_{m0}} + d_t + \varepsilon_m \tag{3}$$

Y_{mt} is the main outcome variable: the migration rate of later-wave migrants from m to the US in $t > 0$ ($\frac{N_{mt}}{Pop_{m0}}$). $\frac{N_{m0}}{Pop_{m0}}$ measures m's first-wave outmigration rate to the US, and $\sum_c \frac{N_{mc0}}{N_{m0}}$ is the sum across counties of the proportion of pioneers' from m across counties.[48] d_t is a time fixed effect that controls for time-specific shocks common to all municipalities.[49] β_4 measures the effect of initial network size on further migration from m to the US: a positive sign indicates that municipalities with larger initial stocks of pioneers in the US send more later-wave migrants to that country. $\sum_c I_{ct} \times \frac{N_{mc0}}{N_{m0}}$ is a weighted average of average county income in t across the US, with weights given by the proportion of first-wave migrants from m that located in each county at baseline. β_5 is the main coefficient of interest and captures the effect of the mean income that the average first-wave migrant from m to the US would obtain in t, on further migratory flows from m to the US. A positive sign indicates that pioneers' better (average) initial location decisions increase the flow of migrants from m to the US.

I estimate six specifications of the model, following the same control addition pattern used before. Specification 1 corresponds to Equation 3 above. Specification 2 adds interactions between time fixed effects and municipality characteristics ($X_{1mt} \times d_t$), which control for other observable factors that may affect later-wave migration rates to the US.[50] Specification 3, incorporates municipality fixed effects (d_m), which account for time-invariant (observable and unobservable) municipality specificities that may affect both first- and later-wave migration rates to the US. Specifications 4, 5 and 6 are the same as specifications 1, 2 and 3 but with province-time fixed effects (d_{pt}). Specification 6 is the most saturated one and my preferred one. The main parameter

[48] It is equal to 1 for municipalities with positive first-wave migration to the US and zero otherwise.

[49] The specification should include $\overline{I}_t$ (US average income) but the coefficient cannot be identified as it is absorbed by d_t.

[50] It should add county characteristics averaged at the US level ($\overline{X}_2 t$) but the coefficients cannot be identified as they are absorbed by d_t.

of interest is β_5, identified in this specification from municipality-time variation, after controlling for province-specific time-varying shocks.

I estimate all specifications through OLS and IV with standard errors clustered at the municipality level and weighted by baseline municipality population. Based on the descriptive evidence presented before, I set 1885 as the threshold to separate the first wave of Italian migrants from later migrants.[51] I aggregate data for later migrants into three five-year periods. I take baseline municipality population and other characteristics from the 1881 Italian Census summary. I use the 1880 and 1900 IPUMS 100% US Census samples for county characteristics, interpolating to obtain values for intermediate years. In particular, I compute average county income considering all non-Italian county residents.[52]

The estimation sample for the Municipality-County Model comprises all municipality-county pairs and the estimation sample for the Municipality-US Model comprises all municipalities of origin; in both cases the estimation is based on the extensive and intensive margins. The following subsection discusses how causal interpretation of the coefficients of interest is threatened by measurement error and endogeneity, while subsection 6.3 introduces instrumental variables to address them.

6.2 Empirical challenges

Measurement error

A first source of concern is the ITA data set. First, it underestimates true network size since it covers about 82% of Italian arrivals over 1855-1900. If it is a representative sample of the full flow of Italians in the period, then the composition by place of origin is unaffected and measurement error leads to attenuation bias; non-random biases in geographical coverage lead to biases in the parameters of interest. Second, migratory flows may be overestimated, as second or higher order trips to the US can only be partially filtered. If the incidence of second-timers is uncorrelated with the outcome of interest the problem boils down to attenuation bias.[53] Third, place of origin is missing or invalid for 60.9% of the sample, and cannot be re-imputed in all cases. The imputation exercise implicitly assumes that the distribution of origin places by surname

[51] Arrivals up to 1885 recorded in the Full US Schedules represent 16.4% of the full stock Italians in that data set. The corresponding figure for the sample of Italian men with surnames that can be matched to the ITA data set is 14.9%.

[52] This measure is highly correlated with others, computed with all residents, only natives, only men, only non-Italian men. In robustness checks available upon request, I show that results do not depend on how average county income is calculated.

[53] For example, if pioneers from municipalities with higher later-wave outmigration rates are more (less) prone to circular migration, β_5 will be biased upwards (downwards).

observed for unproblematic individuals is the same for problematic ones. Differences in these distributions that are uncorrelated with the error term lead to attenuation bias, while more complex patterns of correlation between missing or invalid data and the error term will bias coefficients in unknown directions.

Fourth, the imputation exercise cannot attribute place of origin to 9.9% of the ITA data set. 73.5% of these cases correspond to migrants with unique surnames (i.e. that appear in the data only once); the rest are migrants with surnames shared by others but with missing information for all of them. If the inability to impute a place of origin is randomly distributed across municipalities then measurement error leads to attenuation bias, while if it is correlated with the error term, then coefficients will be biased. In summary, all these problems lead to measurement error (in N_{mc0} and N_{m0}); attenuation bias arises if these errors are randomly distributed while non-random errors lead to biases in unknown directions.

A second source of concern are errors in surnames. The matching procedure aims to pair truly identical surnames affected by spelling or transcription mistakes, but some matches will inevitably be incorrect. Random errors in surnames lead to attenuation bias; non-random ones will further bias coefficients in any direction.[54] Note that the imputation of municipality of origin to Italians observed in the Full US Schedules introduces noise in N_{mc0}, even if surnames are error-free and if the surname distribution in the ITA data set is correct.[55] Nevertheless, it should be noted that its not clear, a priori, why and what type of relationship may exist between surname errors and migration behavior. Moreover, the recent literature using record linkage techniques tends to find little correlation between matching errors and characteristics and outcomes (see for instance Abramitzky et al. (2019)).

A third source of concern is the Full US Schedules, as county of residence in 1900 may differ from pioneers' initial location. Yearly flows by county seen in the data may be inaccurate due to mortality, internal and return migration.[56] If these

[54]For example, suppose that surnames from municipality m_l are more likely to be recorded with error. If pioneers from m_l locate in more dynamic counties, and m_l sends larger migratory flows later on, then β_5 in the Municipality-US Model will be biased upwards. Note that correlations between surnames' propensity to be recorded with error and migrants' settlement patterns are possible but unlikely. Besides, similar surnames often have the same root and come from neighbouring municipalities, which reduces the effect of errors.

[55]Suppose a Chicago resident with the surname *Benedetti* that comes from Florence. If at least one *Benedetti* in the ITA data set comes from a different municipality (e.g. Naples), then N_{mc0} will be greater than zero for the Chicago-Naples pair (even if no Neapolitan ever settled there), and lower than the true number for the Chicago-Florence pair. Moreover, if errors in the ITA data set are not randomly distributed, and the surname distribution used for the imputation process is biased, the resulting N_{mc0} will also be affected, biasing coefficients in both models.

[56]The literature suggests return migration rates around 50% for Italians and southern Europeans

elements are unrelated to municipalities and counties, then coefficients will be affected by attenuation bias. If they are related (e.g. to specific municipality-county pairs), then the spatial distribution of Italians by origin municipality across counties observed in 1900 will differ from the original one, biasing coefficients in unknown directions.[57]

Recall that validation exercises reduce concerns over some sources of measurement error. For example, the yearly distribution of Italian migrants by region observed in the matched data are in line with Italian historical records, and the ITA data set adequately covers the yearly flows of Italian migrants to the US. Moreover, Boustan et al. (2010) finds evidence of very low levels of internal migration for international migrants in the US during the Great Depression, while Cinel (2002) suggests this is also the case for Italians during the Age of Mass Migration.[58] Note also that the battery of controls and fixed effects included in my preferred specifications (especially the two-way fixed effects in the Municipality-County Model) also control for some potential sources of measurement error, in particular those with differential incidence precisely by municipality and county.

To sum up, under classical measurement error previously discussed issues lead to attenuation bias. It is likely that some of these errors are not random (and interact with each other), affecting the coefficients in unknown ways. Section 6.3 discusses an instrumental variable that also addresses measurement error concerns.

Endogeneity

Any (omitted) factor that simultaneously affects initial and further migratory flows and location decisions from a given municipality introduces bias. As a simple example, if municipality income leads to pioneers settling in dynamic counties and causes low further migratory flows, the coefficients of interest will pick up this relationship even if pioneers had no true effect on flows. A spurious relationship also arises if pioneers themselves affect county growth: if migrants from a municipality locate randomly

(Bandiera et al., 2013; Hatton & Williamson, 2005).

[57] Let us discuss three examples. Suppose that pioneers from some municipalities locate in stagnant counties, and that this reduces the size of later flows. First, if mortality rates are higher for these pioneers (e.g. due to bad economic conditions at destination), then poor (rich) counties will be under (over) represented in the data for these municipalities. Second, if return migration is higher for these pioneers, then poor (rich) counties will be under (over) represented in the data for these municipalities. In these two cases the coefficients of interest will be biased. Third, suppose that pioneers from municipalities that send greater migratory flows later on moved from stagnant to dynamic counties. Then, β_5 will be biased upwards when estimating both models.

[58] In Boustan et al. (2010), foreigners represent only two percent of cross-county moves at the national level and in a sample of large metropolitan areas; Cinel (2002) suggests that bad economic outcomes at destination triggered return rather than internal migration.

across the US but cause high growth at destination, and the municipality (unrelatedly) sends large later-wave migratory flows to the same destinations, then regressions will find a spuriously positive β_5. Note that the battery of controls and fixed effects included in the estimations ease some endogeneity concerns. In the Municipality-County Model, two-way fixed effects in my preferred specification account for many types of unobservables that may affect both first- and second-wave migration rates and location decisions. Municipality and province-time fixed effects similarly ease concerns in the Municipality-US Model.[59] Nevertheless, in the next sub-section I propose an instrument to fully address endogeneity concerns.

6.3 Instrumental variables

A successful instrument has to be (sufficiently) correlated with the endogenous variables and must affect the outcome only through them: it must be relevant and fulfill the exclusion restriction. Endogeneity affects $\frac{N_{mc0}}{Pop_m}$, $\frac{N_{mc0}}{N_{m0}}$, $I_{ct} \times \frac{N_{mc0}}{N_{m0}}$ and $\sum_c I_{ct} \times \frac{N_{mc0}}{N_{m0}}$. I address concerns over the first two terms by using the distribution of non-Italian migrants across counties by year of arrival to construct counterfactual versions of them. The underlying hypothesis is that location decisions made upon arrival are (partly) driven by limited information and short-term opportunities. For example, immigrants arriving in a specific year may be more prone to locating in places with substantially higher labour demand at that time, that are known or thought to be momentarily of easier (or cheaper) settlement for migrants, that are recruiting immigrants for a particular activity (e.g., construction), etc. In this case, Italian migrants' location decisions within each year of arrival will be correlated with those of other migrants.[60] Then, using only arrivals up to 1885 (pioneers), I construct the following counterfactual:

$$Z_{mc0} = \sum_{y=0}^{y=1885} N_{my} \frac{F_{cy}}{F_y} \tag{4}$$

$\frac{F_{cy}}{F_y}$ is the proportion of all non-Italian first-wave foreigners that arrived in year y that settled in county c (from the Full US Schedules). N_{my} represents Italian pioneers from m that arrived in y. Equation 4 assigns all arrivals from m in y to different counties according to location decisions made by other, non-Italian migrants that arrived at the same time, and adds them up to obtain a counterfactual number of pioneers from m that settled in c (Z_{mc0}). Note that this term should also be less affected by measurement

[59] Note that endogeneity problems are more severe in the Municipality-US Model, as municipality-specific time shocks cannot be accounted for.

[60] This pattern holds in the data, as shown by first-stage regression results later on.

error if it approximates the original distribution of Italian pioneers, absent differential mortality rates and return or internal migration.

To construct Z_{mc0} I use non-Italian migrants from continental Europe excluding Germany. I exclude English-speaking migrants since they may be in better condition to integrate and choose where to settle, and may suffer less penalties in the labour market and lower costs in general (e.g., no language barrier). I exclude migrants from Asia and Africa as they may be too distant culturally and face additional restrictions and barriers (e.g. skin colour).[61] I exclude Germans given that their migration to the US started thirty to fourty years early than Italians, and hence German migrants arriving at the same time than Italians may be more likely to follow their own previous networks and less likely to be driven by short term or random elements. Then I use Z_{mc0} to construct counterfactual versions of remaining endogenous variables. First-stage regression results indicating instruments' relevance are reported in tables C.9 and C.10 in Appendix C.

The exclusion restriction holds, first, if non-Italians' yearly location decisions are uncorrelated with the error term, which boils down to two requirements or assumptions. First and as it can be seen in the data, there must be variation in outmigration flows by municipality across (pre-1885) years.[62] Second, the variation in the yearly distribution of migrants by municipality of origin must be uncorrelated with non-Italian migrants' yearly location patterns. If non-Italians and pioneers arriving in the same year chose (exactly) the same destination counties, then $Z_{mc0} = N_{mc0}$.[63] Note however that history shows a myriad of shocks that affected the outmigration timing by geographical area, providing anecdotal evidence supporting the random timing assumption.[64]

[61] Continental Europe includes France, Spain, Portugal, Austria, Belgium, Switzerland, Hungary, Bulgaria, Romania, Denmark, Czech Republic, Netherlands, Norway, Sweden, Finland, Russia, Greece, Croatia, Poland, Ukraine, among others. Main English-speaking origin countries are England, Scotland, Ireland, and Canada. Section 8 tests robustness of results to instruments constructed considering all non-Italian migrants, and non-Italian non-English speaking migrants.

[62] I show variation across municipalities of origin in the composition of yearly Italy-US migratory flows in three ways. First, Table C.8 in Appendix C presents mean, median and standard deviation of year of arrival by province of origin, which shows variation in arrival years by province of origin. Second, Figure C.5 in Appendix C plots the mean and standard deviation of arrival year, by municipality of origin, pointing towards differences in migration timing. Third, I regress arrival year on municipality of origin dummies; statistical tests reject that the coefficients are jointly equal to zero ($F = 10.4$, $p = 0$) and to each other ($F = 10.6$, $p = 0$).

[63] For example, suppose a poor municipality m_p, with high incentives to migrate. Suppose that information on dynamic US counties arrives in t_0 and triggers the first migratory wave towards them, and that poverty in m_p leads to large later-wave migratory flows to the US, independently of pioneers' location in the US. If non-Italians arriving in t_0 chose the same dynamic counties, then $Z_{mc0} = N_{mc0}$.

[64] Italian unification (1840's to 1870) was costly and disruptive, with uneven effects across the country. Liberal trade and tariff provisions hurt the South and anti-clerical provisions fostered revolts in Sicily and Naples (1860-1870). Other shocks include wars (e.g. Austria, 1866), economic depressions

The exclusion restriction holds, second, if the size and distribution of non-Italian pioneers has no effect on average county income (I_{ct}). In this direction, Sequeira et al. (2017) find positive effects of large immigrant inflows at the county level (during 1850-1920) in subsequent growth and development in the short and long run, which generate endogeneity concerns. Let us present two counterpoints. First, the present paper considers only Italian migrants up to 1900 and, particularly, defines pioneers as those arriving prior to 1885. The total inflow of Italians up to 1885 represents a mere 0.4% of the US population in 1880, and the flow up to 1900 represents 2.3% of the 1880 population as well. On average, immigrants from all countries represented an annual average inflow of 0.3% of the 1880 population. Then, the inflow of Italian pioneers seems small enough and the period under consideration short enough as to ease concerns over (long-term) effects on county growth. Second, what is strictly required for coefficients to be biased is for pioneers to have differential effect on county income by municipality of origin, and these differentials to be correlated with later-wave migration by municipality. That is, if Italian migrants do have an effect on county income but it is the same across origin municipalities, then the estimation is not compromised.

Under the above assumptions, instrumental variable estimates of the parameters of interest are unbiased. Recall also that controls and fixed effects included in several specifications partially address endogeneity concerns.[65].

7 Main results

This section presents and discusses empirical results. Variables are re-scaled for ease of interpretation: migration rates are expressed in thousands (i.e., a 0.001 migration rate takes the value of 1), average county income and the proportion of first-wave migrants are re-scaled by their respective standard deviations (SD). These transformations are carried onwards to all interactions.

Table 3 presents results for the Municipality-County Model, for the first outcome (municipality-county later-wave outmigration rate, N_{mct}/Pop_{m0}). Column 1 includes the outmigration rate of first-wave migrants from m to c, average county income,

(e.g. Long Depression, 1873-1879) crop failures (late 1840's, 1897), decimation of vineyards (1891-1899), and earthquakes (1857, 1873). Still, to fully overcome the endogeneity problem, an instrument would need to exploit both yearly settlement patterns of non-Italian migrants and *fully exogenous* changes in yearly flows of first-wave migrants by municipality of origin. For example, severe weather shocks are an ideal candidate, and this is currently work in progress.

[65]For example, municipality-county fixed effects control for structural (i.e. time invariant) effects of migrants from specific municipalities over specific counties, county-time fixed effects controls for differential county trajectories (common across sending municipalities), and municipality-time fixed effects capture county-independent trends in first- and later-wave outmigration rates.

the share of pioneers from m settling in c, the interaction of both, and time fixed effects. Column 2 adds time-varying county characteristics and interactions between municipality baseline characteristics and time fixed effects. Column 3 adds municipality and county fixed effects, and Column 4 adds two-way interactions between time, county and municipality fixed effects. Note that only the interaction term can be identified in all specifications, as the other terms are absorbed by the two-way fixed effects, and can be separately identified only in Columns 1-3.

Results are significant at 1% levels in almost all cases and first-stage F-statistics are above conventional levels and point to strong instruments. Results confirm a positive and strong pure network effect: a one per-thousand increase in the first-wave migration rate from m to c leads to an about 0.9 (OLS) to 1.1 (IV) per-thousand increase in m's post-1885 emigration rate to c.

Turning to the interaction term, in my preferred specification (Column 4) OLS estimates show that a one SD change in the share of first-wave migrants from m that settled in c *amplifies* the pull effect of a one SD change in average county income, augmenting the later-wave emigration rate from m to c by an additional 0.009 per-thousand. IV estimates show a much larger effect, of about 0.28 per-thousand. To interpret this effect, note that for municipalities with positive first- or later-wave migration to the US, the mean and standard deviation of the dependent variable are respectively 0.012 and 0.101. Then, given a one SD increase in c's income, a one SD increase in the share of pioneers from m in c increases the later-wave emigration rate from m to c by 2.7 SD, a strong effect. Moreover, the difference in the interaction term between columns 3 and 4 suggests that the controls included in the first case (and in particular, county and municipality characteristics) do not fully account for all elements that affect the outcome variable (better captured by the two-way fixed effects in Column 4).

Turning to the main effects, note that the effect of average county income is negligible, and the share of pioneers from m that settled in c has a negative effect. Evaluated at the mean value of county income, IV results in Column 3 indicate that a one SD change in the share of pioneers from m that settled in c *reduce* later-wave emigration from m to c by about 0.16 per-thousand, about 1.6 SD of the later-wave outmigration rates from m to c. Finally, recall that the substantial difference between IV and OLS coefficients could be due to the unbiased estimation of a larger true effect (i.e., addressing endogeneity) or to the estimation of the LATE (or a combination of both). Nevertheless, a ten-fold jump of IV estimates over OLS ones suggests that endogeneity may be at least part of the problem. In other words, the original distribution of pioneers across the US may indeed be endogenous to a certain extent

(a result with important implications for the literature using initial settlement as an instrument for later migration).

Table 3 here

Recall that the effects on the outcome ($N_{mct}/Popm0$) discussed above can be due to a combination of changes in the overall later-wave migration rate from m to the US as a whole and of changes in the proportion of followers choosing c over c'. To further shed light on this, Table 4 presents results of the Municipality-County Model, for the second outcome (proportion of later-wave migrants from m to the US settling in c, N_{mct}/N_{m0}). In other terms, previous results estimate the effect of the right hand side variables on the (joint) probability of emigrating from m *and* locating in c after 1885, while the following results estimate the effect of the same variables on the probability of locating in c (vis-a-vis c'), conditional on migrating from m to the US.

Turning to results, note that IV estimates in Column 3 show no effect of first-wave migration rates from m to c on later-wave location decisions conditional on migrating to the US (coefficients are close to zero and not significant). All other coefficients are significant at 1% levels while first-stage F-statistics again point to strong instruments. The main effect of income is negligible, suggesting little pull of local economic conditions on the distribution of later-wave migrants. More interestingly, the original distribution of pioneers across the US has a positive effect on later location decisions: IV results in Column 3 show that, evaluated at the mean value of county income, a one SD change in the share of first-wave migrants from m that settled in c *increases* the share of later-wave migrants from m locating in c, by about 6.3 percentage points. Then, results suggest that it is not the size of the stock of pioneers in a county what matters for followers' location decisions, but the relative importance (for pioneers) of that destination vis-a-vis others.

More importantly, results also show that the positive feedback between past and further location decisions is ameliorated by local economic conditions: the IV coefficient in Column 4 shows that a one SD change in average county income *mitigates* the pull effect of a one SD change in the share of first-wave migrants from m that settled in c, *reducing* the share of later-wave migrants from m settling in c by 7.8 percentage points. Note that for municipalities with positive first- or later-wave migration, the standard deviation of the dependent variable is about 0.028, hence the effect amounts to a reduction of 2.8 SD.

In summary, results suggest that, given municipality m and county c: 1) an increase

in the stock of first-wave migrants from a m in a c proportionally increases the emigration rate of later-wave migrants from m to c, but has no effect on the distribution across counties of later-wave migrants from m to the US; 2) c's income has a negligible effect on later-wave migration from a m to that c and on the distribution of further migrants from m to the US across counties; 3) concentration of first-wave migrants from m in c reduces the overall flow of later-wave migrants from m to c but induces further concentration of this flow in c; 4) the positive effect of changes in c's income on later-wave migration from m to c is amplified by the share of pioneers from m in c (alternatively, the negative effect of pioneer concentration on further migration is ameliorated by income growth); 5) but the positive effect of greater income in c on followers' concentration in c is ameliorated by the proportion of pioneers from m that settled in c (alternatively, the negative effect of pioneers' concentration in c on followers' concentration in c is countered by c's income growth). To better understand these dynamics I turn now to the Municipality-US model.

Table 4 here

Table 5 presents results for the Municipality-US Model, with the later-wave outmigration rate from m to c as outcome (N_{mt}/Pop_{m0}). Column 1 includes the emigration rate of first-wave migrants from m to the US and time fixed effects. Column 2 adds interactions between time fixed effects and municipality baseline characteristics, and Column 3 adds municipality fixed effects. Columns 4-6 repeat the estimations in columns 1-3 incorporating province-time fixed effects. Note that the coefficient for the first-wave emigration rate can only be identified in columns 1, 2, 4 and 5.

Both OLS and IV estimates are positive, significant at 1% levels, very similar across specifications, and confirm a positive and strong network effect: a one per-thousand increase in the first-wave migration rate from m to the US leads to an about 1.2 per-thousand increase in the later-wave migration rate from m to the US. Note however that first-stage F-statistics in these cases are close to but below the conventional level of 10, suggesting weak instruments, hence these results should be taken with caution. F-statistics are well above conventional levels in columns 3 and 6. IV results for my preferred specification (Column 6) show that a one SD change in mean income of the average pioneer leads to an increase in later-wave migration rates from m to the US of about 2.3 per-thousand. Note that for municipalities with positive first- or later-wave migration, the standard deviation of the dependent variable is about 4.89; hence the effect amounts to an about 0.46 SD change in the outcome.

In summary, results suggest that, given municipality m: 1) an increase in the stock of first-wave migrants from m anywhere in the US proportionally increases the share of later-wave migrants from m to the US; 2) changes in the (average) performance of first-wave migrants from m increase later-wave migration rates from m to the US (though a one SD change in income leads to less than a half SD change in further migration rates).

Table 5 here

Discussion

In this subsection I seek to better interpret previous results, in terms of potential mechanisms at play. I propose to think of pioneers as providers of information and resources, which may be partially tied to a specific location. With respect to resources, remittances can be freely used by the origin community, for instance to finance further migration to anywhere in the US (location-free). In contrast, room and board, pre-paid tickets, or other forms of credit and support upon arrival, are (much more) tied up to the actual location of pioneers (location-specific). Likewise, though clearly part of the information sent by pioneers is location-free and may affect outmigration in general (e.g. advice on cost and procedure of the migration process, costumes and institutions of the destination country, etc.), another part of the information is location-specific (e.g. on wages, unemployment, available industries and activities for employment, etc.). Also, pioneers' knowledge of the local labour market and capacity to operate as a reference for newcomers most likely is also location-specific. In this setting, changes in local economic conditions at destination, in the number of pioneers and their distribution across counties, all translate into changes in the provision of (location-free and location-specific) information and resources. Though this paper does not investigate the role of each mechanism or channel (mostly due to data shortcomings), the results obtained so far do shed light on how they may have operated.

First, recall that results for the pure network effect suggest low levels of cross-county spillovers. This means that, keeping constant local economic conditions at destination and the distribution of pioneers across counties, the change in resources and information arising from a greater stock of pioneers from m in c has little to no effect on the concentration patterns of further migrants. For example, any additional flow of remittances stemming from the greater stock of pioneers is not used by further migrants to locate in different destination counties; this suggests that the pure network effect operates mainly through location-specific channels.

Second, results show that increased concentration of pioneers from m in c: i) increases followers' agglomeration in the same destination; ii) reduces total further flows to that destination; iii) reduces overall migration from m to the US as a whole.[66] This suggests that increased pioneer concentration in c has a cost for the sending municipality, in terms of total further outflows both to the US and to c. Pioneer diversification may help to support a greater total flow of followers through resources provided across more destinations, while pioneer concentration limits the total number of followers that can be supported in a given place and harms overall further migration. For example, greater concentration of pioneers in c improves the safety net available for the existing community and for further migrants, but sustaining the safety net may reduce the resources (remittances) sent home. Alternatively, pioneer diversification may increase further migration to the US if it allows the origin community to hedge risks. Taken together, results suggest that the pure concentration effect may also operate mainly through location-specific channels.

Third, results show that changes in the mean income of the average pioneer lead to a large increase in the emigration rate from m to the US, while a larger (in dollars) increase in income in only one county has no effect on the later-wave migration rate from m to that county. This suggests that although pioneers may inform their origin communities of income changesin c, this information by itself may not induce a (large) response from the origin community. In an unconstrained setting, a sharp increase in c's income could increase further migration to c due to higher expected earnings there; results show that this is not the case and suggest financial constraints are binding. In fact, a back-of-the-envelope calculation shows that a one SD deviation change in c's average income most likely translates into a small increase in remittances, which may help explain the result.[67] Moreover, even if the flow of followers does not respond to a positive income shock in c due to financial constraints, it could locate in greater proportion in the positively affected county; results show that this is not the case and suggest that there is a limit to pioneers' ability to help and support further migrants at their destination. Taken together, results suggest that positive income shocks in c imply small increases both in location-free and location-specific resources. Still, results

[66] Recall that changes in followers' flows from m to c are a combination of changes in flows from m to the US and changes in followers' distribution across counties. Then, an increase in followers' concentration in a given county coupled with a reduction in the total flow to that destination results from a reduction in the flow to the destination county as a whole.

[67] Recall that the effect of changes in county income is estimated holding income of remaining counties constant. Municipalities with positive migration to the US at any point in time display an average concentration of pioneers in c of 1.14%; then, a 2830 dollar increase (one SD) in county income increases the average income of the mean pioneer by only 32.3 dollars (less than 0.04 SD change).

show that differences in the distribution of pioneers across the destination country matter, and can help explain differences in migratory flows and location decisions of further migrants across sending communities. In particular, municipalities that sent more (less) pioneers to dynamic (stagnant) counties will send more (less) followers to the whole US, which will locate in a greater (smaller) set of destinations.

Fourth, results show that the interplay between pioneer concentration and local income shocks at destination plays an important role. Though positive income shocks in c have little effect on migration and location decisions of later-wave migrants, the interaction terms in the Municipality-County models show that municipalities that concentrated more pioneers in c experience higher outmigration to that same destination later on, and display lower concentration of all further migrants to the US in that destination. This is observed if the increase in total migration from m to the US more than compensates a reduction in the concentration of further migrants from m to the US in c, which is confirmed by a back-of-the-envelope calculation.[68] This suggests that the joint effect of pioneer concentration and income growth operates in greater proportion through location-free channels, as it tends to increase both further migration and follower diversification across the US.

Moreover, the interaction term in the Municipality-County models sheds further light on how the mechanisms may be operating. Recall that pioneer concentration increases follower concentration, but that this is ameliorated by positive income shocks in c. Conditional on a fixed number of pioneers in c, increased pioneer concentration implies mainly changes through the information channel, which seem to be location-specific given the negative effect on further migration to the US and the positive effect on further concentration of followers in c. However, these effects are countered when increased pioneer concentration is coupled with greater county income. This involves both increases in resources (both location-free, as remittances, and location-specific, as room and board) and information, and results show that the main element at play is location-free. In other terms, the additional resources flowing to the origin community stemming both from greater county income and greater pioneer concentration are used to increase and diversify the later-wave of migrants, even when there should be additional resources available specifically in c.[69]

[68] A one SD increase in county income (2830 dollars) together with a one SD increase in pioneer concentration (about 4.9 percentage points) leads to a 0.28 per-thousand increase in further migration from m to c and amounts to an increase in the mean income of the average pioneer of about 170.5 dollars. Results for the Municipality-US model show that this leads to an about 0.4 per-thousand increase in further migration from m to the US. Moreover, the fraction of this new wave of followers that concentrates in c is lower than for the pioneers.

[69] Of course, it could be the case that the response to increased county income by an increased

Finally, note that different models and specifications exploit different sources of variation, and some may still be vulnerable to endogeneity (e.g., in the Municipality US-Model, municipality-specific time trends cannot be controlled for). Specification 4 of the Municipality-County model, the most demanding one, allows only the identification of the interaction between county income and pioneer concentration. Then, the most conservative interpretation of the results, and this paper main contribution, posits that pioneer concentration: i) amplifies the effect of county income shocks on further flows from the same origin communities to the same destination counties; ii) ameliorates the effect of county income shocks on concentration of further flows from the same origin communities to the US in the same destination counties. Alternatively, income growth i) ameliorates the negative effect of pioneer concentration on further flows from the same origin communities in the same destination counties; ii) counteracts the concentrating effect on further flows from the same origin communities to the US in the same destination counties.

These results are consistent with a setting in which additional resources stemming from greater county income and greater pioneer concentration are used by the origin community to increase and diversify the subsequent wave of migrants to the US, that is, a setting in which these resources operate mostly as location-free. Still, as this paper does not investigate channels and mechanisms at play, the previous discussion can be consistent with alternative explanations. To further probe into the results, the following section presents extensions and a variety of robustness checks.

8 Extensions and Robustness checks

This section first investigates the existence of spill-overs to neighbouring municipalities, and then tests robustness of results to using a Bartik instrument for county income, considering a different threshold for first- and later-wave migration, using four alternative definitions of the instrument for pioneers' locations, disregarding weights, trimming the sample, and focusing on Southern Italy only.

Extensions

Size and location of first-wave migratory flows may also have an effect on neighbouring communities, for instance due to information diffusion or remittance spill-overs. Given an origin municipality (m_o), I re-compute the outcome variables considering a subset

proportion of pioneers at a given destination is to severely reduce resource provision for followers at that destination, to further increase remittances.

of neighbouring municipalities (m_d) and repeat the empirical exercise.[70] I use distances of 5, 10, 15 and 20 kilometres. Table 6 presents results for the Municipality-County Model, for the first outcome (later-wave migration rate from neighbours of m to c). Columns 1, 3, 5 and 7 report regression results for specification 3, columns 2, 4, 6 and 8 use specification 4. OLS and IV estimates are positive and significant at 1% levels, showing positive spill-overs. Estimates are between 50% to 70% of the main results: pioneers' concentration also magnify the pull effect of changes in a given county's income on emigration from neighbouring municipalities, albeit to a lesser extent than on the origin municipality.

Table 6 here

Table 7 presents results for the Municipality-County Model, for the second outcome (share of later-wave migrants from neighbours of m settling in c). Columns 1, 3, 5 and 7 report regression results for specification 3, columns 2, 4, 6 and 8 use specification 4. Note first that IV estimates for the interaction term in my preferred specification (columns 2, 4, 6 and 8) are close to zero and not significant. This suggests no additional effect of county income on the pull effect exerted by pioneers' location decisions, on neighbouring municipalities later-wave migrants' settling patterns. IV results for specifications 1, 3, 5 and 7 suggest that the agglomeration effects of pioneer concentration also extends to neighbouring municipalities, as the coefficients are positive, significant at 1% levels and at about 80% levels of the main ones.

Table 8 presents results for the Municipality-US Model. Columns 1, 3, 5 and 7 report regression results for specification 2, columns 2, 4, 6 and 8 use specification 3. Coefficients for my preferred specification for the interaction term show a positive and significant effect at 1% levels, about 20% to 40% in size of the main effect. This implies positive spill-overs: a one SD deviation in mean income of the average pioneer of municipality m_o increases later-wave outmigration rates from neighbouring municipalities to the US. Coefficients are positive and significant for the main network effect, also showing positive spill-overs: a one per-thousand increase in first-wave emigration rates from m_o to the US increases later-wave outmigration rates from neighbouring municipalities, by 20% to 30% of the main effect. Both spill-overs fall with distance to m_o.

[70] For municipality m_o, neighbouring municipalities (m_n) are those for which m_o's centroid is up to d kilometres from m_o's centroid. Coordinates come from ISTAT; I use geodetic distances to account for earths' curvature.

Table 8 here

Robustness checks

First of all, in order to further address potential endogeneity concerns over the relationship between Italian migrants and county income, I construct two counterfactual versions of income based on the procedure developed in Bartik (1991). Recall that county income is the average of individuals' Occupational Score Index, which itself is the median income associated with each individual's occupation. For the first counterfactual, I interact the national growth rate of the number of individuals in each occupation (266 categories) with counties' 1880 occupation composition. For the second counterfactual I aggregate by industry (147 categories). I repeat the IV estimations using also the counterfactual measure of income as instrument; Tables C.11, C.12 and C.13 in Appendix C present regression results. Results are very similar to the main ones and first-stage F-statistics show strong instruments in all cases, easing concerns over identification issues.

Second, I consider four alternative definitions of the instrument for pioneers' location. The main estimations used the distribution of non-Italian continental Europeans excluding Germans to assign a counterfactual destination for Italian migrants, by year of arrival. Here I repeat the exercise, using as reference group: i) all non-Italian Europeans (including Great Britain, that is, English speakers), ii) all non-Italians (the less restrictive reference group, including Africans, Asians, Latin Americans, etc.), iii) non-Italian Continental Europeans (including Germans), iv) non-Italian Continental Europeans excluding Germans, Swedes and Norwegians (the most restrictive reference group, leaving aside other migrants that arrived mostly earlier than Italians). Results in Tables C.14, C.15 and C.16 of Appendix C are very similar to the main estimations, and first-stage F-statistics again confirm instrument strength.

Third, I use 1890 as an alternative threshold to divide first- and later-wave migrants.[71] For the sample of Italians with surnames that can be matched to the ITA data set, those arriving up to 1890 represent 31.6% of arrivals up to 1900, which means that Italians in the US approximately doubled between 1885 and 1900. Results are also very similar to the main estimates and first-stage F-statistics show strong instruments.

Finally, I conduct three additional robustness checks, results are not reported here for brevity and are available upon request. First, I repeat the estimation with no weights and find very similar results. Second, I drop the top and bottom 1% of

[71] See Appendix C, Tables C.17, C.18 and C.19 for regression results.

municipalities in the population distribution.[72] This implies dropping the three main ports of departure (Naples, Genoa and Palermo), the national capital (Rome), and most regional (14 out of 16) and provincial (50 out of 69) capitals. Coefficients are very similar to the main estimations, albeit about 15% larger in absolute value in the case of Municipality-County models. Third, I consider only municipalities in Southern Italy.[73] Results are again very similar to the main estimations.

In summary, the main findings of the paper are robust to a battery of robustness checks, while an extension shows that first-wave emigration rates to specific counties and to the US have positive spillovers on further migration from neighbouring municipalities to those counties and to the whole US, respectively. Moreover, pioneer concentration also magnifies the pull effect of changes in county income on emigration from neighbouring municipalities, though this spill-over falls with distance.

9 Final Remarks

This paper investigates the relationship between the size and location patterns of early waves of Italian migrants to the US, on migration and settlement decisions of Italian later-wave migrants from the same communities of origin. The reviewed literature suggests two main transmission channels, information and economic resources, with potentially ambiguous effects. Thus, early migrants' better location decisions in terms of income growth at destination counties could lead to greater or lower emigration rates in the future, to the same counties or to the destination country as a whole. The sign and existence of an effect remains an empirical problem, that has received little attention from the empirical literature mainly due to data restrictions. Likewise, evidence on the relationship between early and later waves of migrans' distribution decisions across the destination, and the interplay between settlement patterns and local economic growth is lacking.

In this paper, I combine new data sets through a surname matching technique to generate new estimates of the yearly flow of migrants from each Italian municipality to each US county in the second half of the 19^{th} century, filling a gap in the historical data. The data set is innovative and allows to overcome traditional problems in the migration literature. Though it is subject to a variety of measurement error problems, validation exercises ease some concerns over them. Moreover, causal interpretation of simple OLS estimates may be affected by endogeneity, for example as pioneers'

[72] 158 municipalities with population in 1881 below or equal to 319, or above or equal to 27562.

[73] 3240 municipalities in the regions of Abruzzi, Basilicata, Calabria, Campania, Lazio, Marche, Molise, Puglia, Sardegna, Sicilia and Umbria.

location decisions may be correlated with unobservable municipality traits linked to later-wave emigration patterns. These issues are addressed with two strategies. First, I take advantage of the sources of variation allowed by the data set and incorporate a battery of controls and fixed effects that mitigate many potential issues. Second and more importantly, I resort to instrumental variables and use location decisions of other European migrants to the US to construct a counterfactual distribution of early Italian migrants across the US, by municipality of origin.

First, results show that increased first-wave municipality-county migration rates increase later-wave migration rates from the same origin municipalities to the same destination counties approximately in the same proportion. A similar result holds for first- and later-wave municipality-US migration rates, and suggests little cross-county migratory spill-overs from the same sending municipality: an additional first-wave migrant from a given municipality to a given county increases further migration mainly to the same destination and not to the whole US. Moreover, concentration of first-wave migrants from a given municipality to the US in a specific county reduces further migration from the same origin to the whole US, increases concentration of later-wave migrants from the same municipality to the US in that destination, with a net negative effect on the overall further flow from the municipality to the county.

Second, results show that positive income shocks at a given county have little effect on further migration from municipality linked to that county. Still, the distribution of first-wave migrants across the receiving country matters: municipalities connected (on average) to counties with higher income send more later-wave migrants to the whole US (who locate in a larger set of destination counties), and viceversa.

Third and as the main contribution of the paper, results show that the interplay between first-wave migrants' location decisions and local income shocks at destination plays an important role. In particular, a one SD change in average county income mitigates the pull effect of a one SD change in the share of first-wave migrants from a given municipality that settled in that county, reducing the share of later-wave migrants from the same origin settling in that destination by 7.8 percentage points. This suggests that the joint effect of pioneer concentration and income growth operates increasing further migration from the same origin to the whole US and reducing concentration of followers in the same destination county. This is consistent with a setting in which additional resources stemming from greater county income and greater pioneer concentration operate as if they were not tied up or specific to pioneers' specific location. Finally, an extension shows positive spill-overs to neighbouring municipalities, and results are robust to disregarding weights, using alternative definitions of the

instrument, considering a Bartik instrument for county income, trimming the sample, and focusing on Southern Italy only.

To sum up, this paper surpasses traditional data limitations in the international migration literature and contributes with a first exploration of the role of pioneers' location decisions on size and location patterns of later wave migrants from the same communities of origin. Taken together, results are consistent with income at destination attracting additional later-wave migrants from the same municipalities of origin, the larger the proportion of pioneers in the destination county. Nevertheless, note that substantial measurement error problems and endogeneity issues mean that results can also be consistent with complex selection patterns.

Tables and Figures in the text

Tables

Table 1. Descriptive statistics. Italians, 1900 US Census Population Schedules.

Variable	Mean
Age	31.2
Year of arrival	1892.2
	% in sample
Gender	
Men	64.8%
Marital status	
Married	57.8%
Single	38.8%
Widowed	3.3%
Position in the household	
Head	29.9%
Wife	22.2%
Boarder	19.4%
Son/daughter	18.9%
Employee	2.1%
Other	7.5%
Observations	489,041

Notes: The table reports descriptive statistics for all Italians found in the 1900 US Full Count Census Index. Household position consists of 32 categories; the table reports those with a frequency above than 2 percentage points.

Table 2. Descriptive statistics. US population, 1880 IPUMS 100% Census sample.

Variable	All counties % in sample	Counties with Italians % in sample
% Men	50.9%	50.8%
% Foreign born	13.3%	16.8%
% Black	13.1%	9.6%
% Attends school (age 6 to 14)	59.1%	63.0%
% Lives in rural areas	73.6%	63.0%
% In the labour force (age 16+)	54.7%	54.9%
Proportion of the labour force employed in:		
Agriculture	43.8%	33.3%
Mining	0.7%	0.9%
Construction	4.7%	5.7%
Manufacturing (durables)	5.3%	6.7%
Manufacturing (non-durables)	6.7%	8.6%
Transportation	3.6%	4.4%
Telecommunications	0.1%	0.2%
Utilities and sanitary services	0.1%	0.1%
Wholesale trade	0.6%	0.7%
Retail trade	7.2%	9.0%
Finance, insurance and real estate	0.4%	0.5%
Business and repair services	1.4%	1.6%
Personal services	8.6%	10.1%
Entertainment and recreation services	0.1%	0.2%
Professional and related services	3.0%	3.2%
Public administration	0.7%	0.8%
Other/unknown	13.0%	14.0%
Observations	50,140,482	34,158,584

	All counties Mean (Std. dev.)	Counties with Italians Mean (Std. dev.)
Age	24.1	25.0
	(18.6)	(18.7)
Occupational Score Index	18,353	19,675
	(10,158)	(10,435)
Literacy (age 10+)	0.83	0.87
	(0.37)	(0.34)

Notes: The table reports the proportion of individuals in the sample with a set of characteristics, and the mean and standard deviation for further selected variables. Figures are reported for the full US population, and for the sub-sample of individuals living in counties with at least one Italian resident, and are computed from the IPUMS Census 100% sample. The literacy rate is defined for individuals aged 10 or older. The Occupational Score Index is an IPUMS constructed variable that imputes to each individual the median income within an individuals' occupation in 1950 and is measured in constant 1950 US dollars. Industries are defined according to the 1950 classification used by IPUMS. Participation in the labour force is defined for individuals aged 16 or older. School attendance is computed for individuals aged 6 to 14.

Table 3. Main results, Municipality - County model. Later-wave municipality-county outmigration rates.

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	OLS			
First-wave municipality-county outmigration rate	0.9360***	0.9348***	0.9040***	
	(0.0311)	(0.0311)	(0.0306)	
% of first-wave municipality-US migrants in county	-0.1138***	-0.1117***	-0.0651***	
	(0.0087)	(0.0086)	(0.0068)	
Average county income	0.0006***	0.0006***	0.0002***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0135***	0.0132***	0.0064***	0.0086***
	(0.0011)	(0.0011)	(0.0008)	(0.0014)
Observations	46727679	46727679	46727679	46727679
R-squared	0.483	0.484	0.504	0.832
	IV			
First-wave municipality-county outmigration rate	1.1590***	1.1583***	1.1600***	
	(0.0434)	(0.0434)	(0.0435)	
% of first-wave municipality-US migrants in county	-0.1359***	-0.0382	-0.5981***	
	(0.0297)	(0.0294)	(0.0529)	
Average county income	0.0001***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0175***	0.0060*	0.0716***	0.2773***
	(0.0034)	(0.0034)	(0.0062)	(0.0167)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	738.1	710.2	170.5	875.4
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the municipality-county per-thousand migration rate in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. The table also reports mean and standard deviation of the outcome variables for municipalities with positive outmigration to the US in any period. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 4. Main results, Municipality - County model. County concentration of later-wave municipality-US migrants.

Variables	(1) % of later-wave municipality-US migrants in county	(2) % of later-wave municipality-US migrants in county	(3) % of later-wave municipality-US migrants in county	(4) % of later-wave municipality-US migrants in county
	OLS			
First-wave municipality-county outmigration rate	0.0438***	0.0431***	0.0224***	
	(0.0029)	(0.0029)	(0.0016)	
% of first-wave municipality-US migrants in county	-0.0556***	-0.0544***	-0.0219***	
	(0.0031)	(0.0031)	(0.0019)	
Average county income	0.0004***	0.0003***	0.0000***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0085***	0.0083***	0.0035***	-0.0076***
	(0.0004)	(0.0004)	(0.0003)	(0.0014)
Observations	46727679	46727679	46727679	46727679
R-squared	0.138	0.141	0.234	0.582
	IV			
First-wave municipality-county outmigration rate	0.0007	0.0006	0.0014	
	(0.0021)	(0.0022)	(0.0020)	
% of first-wave municipality-US migrants in county	0.0787***	0.1110***	0.1165***	
	(0.0161)	(0.0168)	(0.0242)	
Average county income	0.0000***	0.0002***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0040**	-0.0078***	-0.0089***	-0.0785***
	(0.0019)	(0.0019)	(0.0028)	(0.0096)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	738.1	710.2	170.5	875.4
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the proportion of later-wave migrants from a municipality that settled in a county, in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 5. Main results, Municipality - US model. Later-wave municipality-US outmigration rates.

Variables	(1) Later-wave municipality-US outmigration rate	(2) Later-wave municipality-US outmigration rate	(3) Later-wave municipality-US outmigration rate	(4) Later-wave municipality-US outmigration rate	(5) Later-wave municipality-US outmigration rate	(6) Later-wave municipality-US outmigration rate
			OLS			
First-wave municipality-US outmigration rate	1.2455***	1.2360***		1.1885***	1.1878***	
	(0.0478)	(0.0491)		(0.0529)	(0.0524)	
Mean income of average first-wave migrant	0.0723***	0.0530***	2.0939***	-0.0068	-0.0017	1.1301***
	(0.0161)	(0.0154)	(0.2053)	(0.0153)	(0.0150)	(0.1124)
Observations	23556	23556	23556	23556	23556	23556
R-squared	0.796	0.801	0.935	0.839	0.840	0.950
			IV			
First-wave municipality-US outmigration rate	1.2672***	1.2646***		1.2236***	1.2234***	
	(0.0488)	(0.0495)		(0.0572)	(0.0569)	
Mean income of average first-wave migrant	0.9683***	0.6560*	4.3050***	-0.0040	0.1191	2.2557***
	(0.3381)	(0.3432)	(0.2068)	(0.2309)	(0.2681)	(0.1387)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1^{st} stage)	9.7	7.5	3498.3	10.6	8.2	3323.5
Time FE	X	X	X			
Time FE × Munic. baseline controls		X	X		X	X
Municipality FE			X			X
Province × Time FE				X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1890. The dependent variable is the municipality-US per-thousand later-wave migration rate, in three five-year intervals after 1885. The main independent variables are: the first-wave migration rate from each municipality to the United States, and a weighted average of county income, with weights equal to the proportion of first-wave migrants in each county. Column 1 includes time fixed effects, Column 2 adds municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds municipality fixed effects. Columns 4-6 repeat the estimations in 1-3 incluiding province-year fixed effects. In the bottom panel, the key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 6. Municipality spillover results, Municipality - County model. Later-wave neighbouring municipalities-county outmigration rates.

Variables	(1) Later-wave municipality-county outmigration rate 5 km radius	(2) Later-wave municipality-county outmigration rate 5 km radius	(3) Later-wave municipality-county outmigration rate 10 km radius	(4) Later-wave municipality-county outmigration rate 10 km radius	(5) Later-wave municipality-county outmigration rate 15 km radius	(6) Later-wave municipality-county outmigration rate 15 km radius	(7) Later-wave municipality-county outmigration rate 20 km radius	(8) Later-wave municipality-county outmigration rate 20 km radius
	OLS							
First-wave municipality-county outmigration rate	0.3338***		0.4098***		0.3696***		0.3470***	
	(0.0332)		(0.0346)		(0.0234)		(0.0211)	
% of first-wave municipality-US migrants in county	-0.0367***		-0.0523***		-0.0504***		-0.0513***	
	(0.0046)		(0.0058)		(0.0046)		(0.0046)	
Average county income	0.0002***		0.0002***		0.0002***		0.0002***	
	(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Interaction: average county income × % of first-wave municipality-US migrants in county	0.0041***	0.0058***	0.0061***	0.0073***	0.0060***	0.0078***	0.0062***	0.0077***
	(0.0006)	(0.0010)	(0.0008)	(0.0011)	(0.0006)	(0.0010)	(0.0006)	(0.0010)
Observations	46727679	46727679	46727679	46727679	46727679	46727679	46727679	46727679
R-squared	0.168	0.837	0.333	0.911	0.403	0.937	0.429	0.9488
	IV							
First-wave municipality-county outmigration rate	0.4466***		0.5261***		0.4754***		0.4454***	
	(0.0537)		(0.0470)		(0.0335)		(0.0306)	
% of first-wave migrants in county	-0.3437***		-0.4402***		-0.4659***		-0.4537***	
	(0.0339)		(0.0383)		(0.0352)		(0.0373)	
Average county income	0.0001***		0.0001***		0.0001***		0.0001***	
	(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0422***	0.1597***	0.0552***	0.2038***	0.0583***	0.1965***	0.0572***	0.1937***
	(0.0040)	(0.0203)	(0.0046)	(0.0177)	(0.0041)	(0.0131)	(0.0044)	(0.0139)
Observations	46727679	46727679	46727679	46727679	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	170.5	875.4	170.5	875.4	170.5	875.4	170.5	875.4
Time FE	X		X		X		X	
County controls	X		X		X		X	
Time FE × Munic. baseline controls	X		X		X		X	
County and Municipality FE	X		X		X		X	
Municipality × County FE		X		X		X		X
Municipality × Time FE		X		X		X		X
Time × County FE		X		X		X		X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the neighbouring municipalities-county per-thousand migration rate in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Columns 1, 3, 5 and 7 include time, county and municipality fixed effects, time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population), municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Columns 2, 4, 6 and 8 add two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. Neighbours are defined as municipalities with centroids up to 5km, 10km, 15km and 20km from the centroid of the municipality of reference, for which the right-hand side variables are defined. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 7. Municipality spillover results, Municipality - County model. County concentration of later-wave neighbouring municipality-US migrants.

Variables	(1) Later-wave outmigration rate per county 5 km radius	(2) Later-wave outmigration rate per county 5 km radius	(3) Later-wave outmigration rate per county 10 km radius	(4) Later-wave outmigration rate per county 10 km radius	(5) Later-wave outmigration rate per county 15 km radius	(6) Later-wave outmigration rate per county 15 km radius	(7) Later-wave outmigration rate per county 20 km radius	(8) Later-wave outmigration rate per county 20 km radius
	OLS							
Pioneers/Pop	0.0170***		0.0115***		0.0079***		0.0067***	
	(0.0023)		(0.0015)		(0.0010)		(0.0007)	
% of pioneers in county	-0.0000		-0.0013		-0.0027***		-0.0025***	
	(0.0010)		(0.0009)		(0.0007)		(0.0006)	
County income	0.0000***		0.0001***		0.0001***		0.0001***	
	(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Income × % of pioneers	-0.0001	-0.0004	0.0002	-0.0006*	0.0004***	-0.0003	0.0004***	0.0001
	(0.0001)	(0.0003)	(0.0001)	(0.0003)	(0.0001)	(0.0002)	(0.0001)	(0.0002)
Observations	46727679	46727679	46727679	46727679	46727679	46727679	46727679	46727679
R-squared	0.161	0.592	0.417	0.690	0.579	0.771	0.723	0.8614
	IV							
Pioneers/Pop	0.0227***		0.0103***		0.0047***		0.0034***	
	(0.0032)		(0.0022)		(0.0013)		(0.0008)	
% of pioneers in county	0.0621***		0.0926***		0.0802***		0.0957***	
	(0.0099)		(0.0152)		(0.0142)		(0.0100)	
County income	0.0001***		0.0001***		0.0001***		0.0001***	
	(0.0000)		(0.0000)		(0.0000)		(0.0000)	
Income × % of pioneers	-0.0076***	-0.0054	-0.0107***	-0.0061	-0.0090***	-0.0026	-0.0109***	-0.0028
	(0.0011)	(0.0069)	(0.0018)	(0.0084)	(0.0017)	(0.0066)	(0.0012)	(0.0054)
Observations	46727679	46727679	46727679	46727679	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	170.5	875.4	170.5	875.4	170.5	875.4	170.5	
Time FE	X		X		X		X	
County controls	X		X		X		X	
Time FE × Munic. baseline controls	X		X		X		X	
County and Municipality FE	X		X		X		X	
Municipality × County FE		X		X		X		X
Municipality × Time FE		X		X		X		X
Time × County FE		X		X		X		X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the proportion of later-wave migrants from neighbouring municipalities that settled in a county, in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Columns 1, 3, 5 and 7 include time, county and municipality fixed effects, time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population), municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Columns 2, 4, 6 and 8 add two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. Neighbours are defined as municipalities with centroids up to 5km, 10km, 15km and 20km from the centroid of the municipality of reference, for which the right-hand side variables are defined. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table 8. Municipality spillover results, Municipality - US model. Later-wave neighbouring municipalities-US outmigration rates.

Variables	(1) Later-wave outmigration rate to the US 5km	(2) Later-wave outmigration rate to the US 5km	(3) Later-wave outmigration rate to the US 10km	(4) Later-wave outmigration rate to the US 10km	(5) Later-wave outmigration rate to the US 15km	(6) Later-wave outmigration rate to the US 15km	(7) Later-wave outmigration rate to the US 20km	(8) Later-wave outmigration rate to the US 20km
				OLS				
Pioneers/Pop	0.2952***		0.3011***		0.2311***		0.1836***	
	(0.0555)		(0.0404)		(0.0265)		(0.0210)	
Income × % of pioneers	0.0149	0.3105***	0.0085	0.2809***	-0.0087	0.2841***	0.0068	0.2412***
	(0.0528)	(0.1030)	(0.0417)	(0.0851)	(0.0328)	(0.0664)	(0.0267)	(0.0538)
Observations	23556	23556	23556	23556	23556	23556		
R-squared	0.306	0.940	0.544	0.967	0.659	0.974	0.729	0.980
				IV				
Pioneers/Pop	0.3063***		0.3108***		0.2349***		0.1858***	
	(0.0664)		(0.0462)		(0.0289)		(0.0230)	
Income × % of pioneers	-0.1617	0.8446***	-0.9125**	0.8022***	-1.1846***	0.6787***	-1.0487***	0.5615***
	(0.5305)	(0.1372)	(0.4283)	(0.1329)	(0.3900)	(0.0928)	(0.3562)	(0.0741)
Observations	23556	23556	23556	23556	23556	23556		
F-stat. (1^{st} stage)	8.2	3323.5	8.2	3323.5	8.2	3323.5	8.2	3323.5
Time FE × Munic. baseline controls			X	X	X	X		
Municipality FE							X	X
Province × Time FE	X	X	X	X	X	X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the neighbouring municipalities-US per-thousand later-wave migration rate in three five-year intervals after 1885. The main independent variables are: the first-wave migration rate from each municipality to the United States, and a weighted average of county income, with weights equal to the proportion of first-wave migrants in each county. Columns 1, 3, 5 and 7 include time fixed effects, municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects and province-time fixed effects. Columns 2, 4, 6 and 8 add municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. Neighbours are defined as municipalities with centroids up to 5km, 10km, 15km and 20km from the centroid of the municipality of reference, for which the right-hand side variables are defined. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Figures

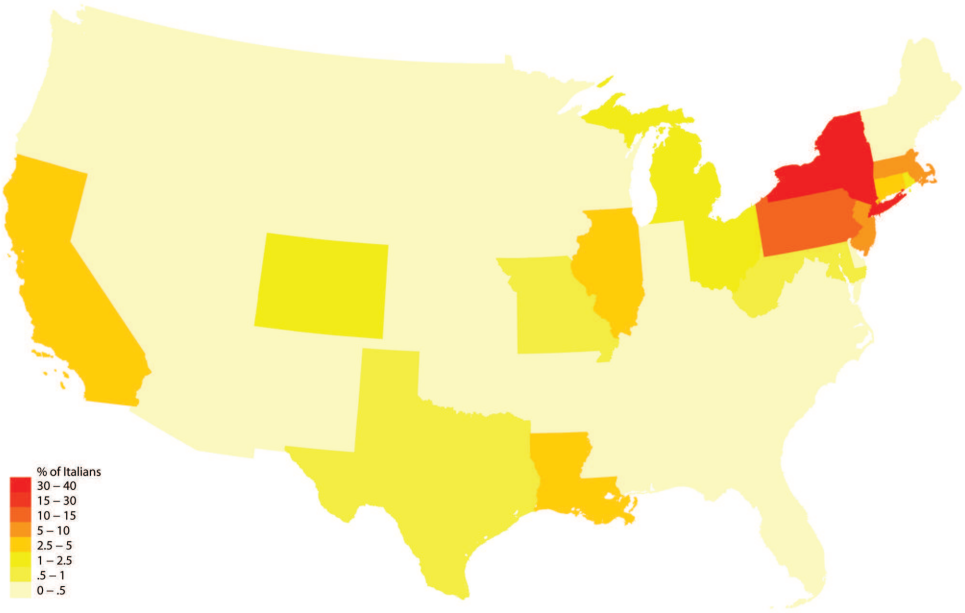


Figure 1. Distribution of Italian migrants by state of residence in 1900.

Notes: The figure plots the proportion of Italian residents in the US in 1900 by state. Alaska and Hawaii are omitted for ease of presentation.

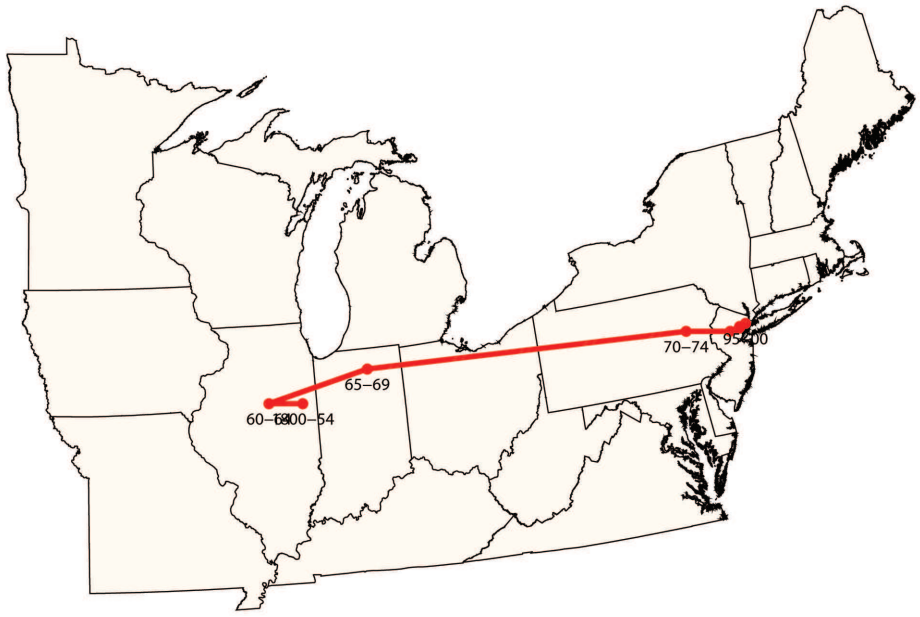


Figure 2. Location of median migrant by year of arrival.

Notes: The figure plots the median latitude and longitude for Italian residents in the US by year of arrival intervals. Individuals' are assigned the latitude and longitude corresponding to the centroid of the county in which they reside. The map includes only the northeast of the US for ease of viewing. The red line joins the median locations chronologically.

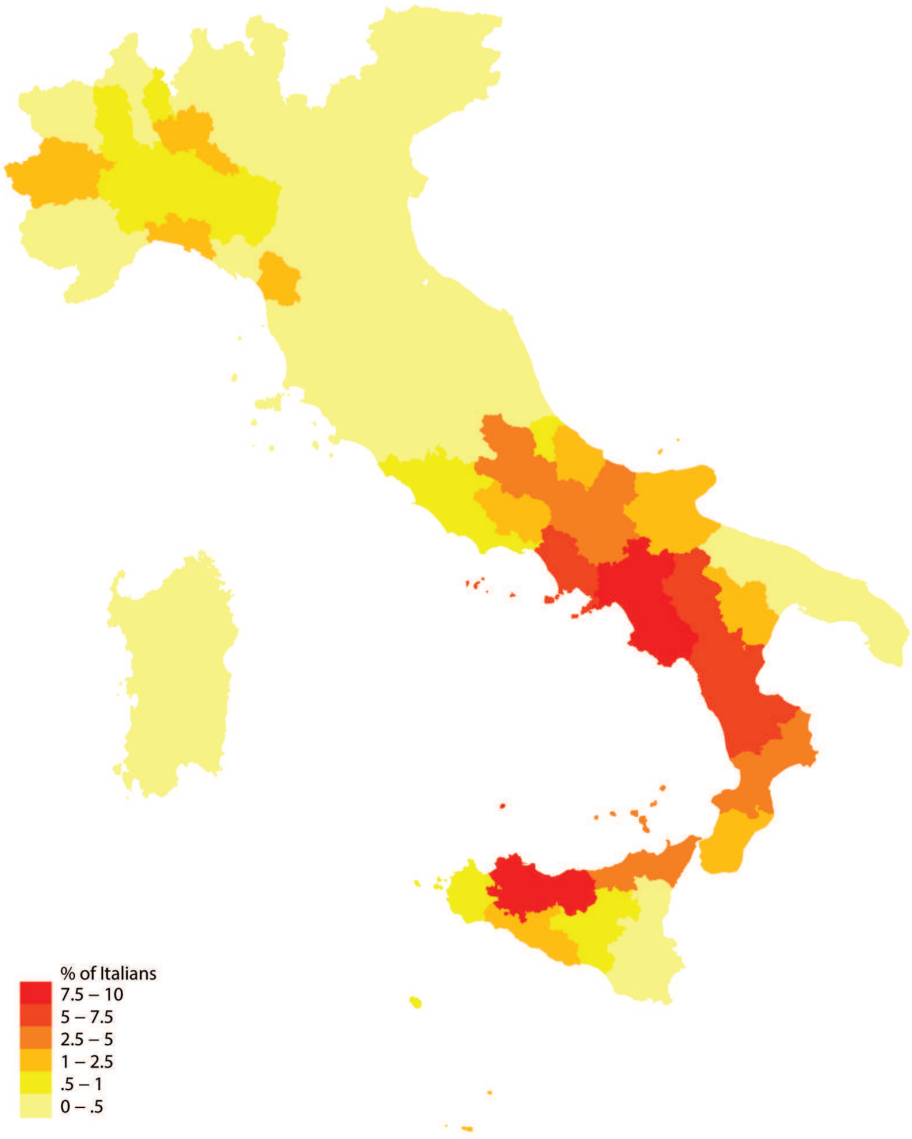


Figure 3. Distribution of Italian migrants by province of origin.

Notes: The figure plots the shares of Italian arrivals to the United States (1855-1900) computed from ITA data by province of origin.

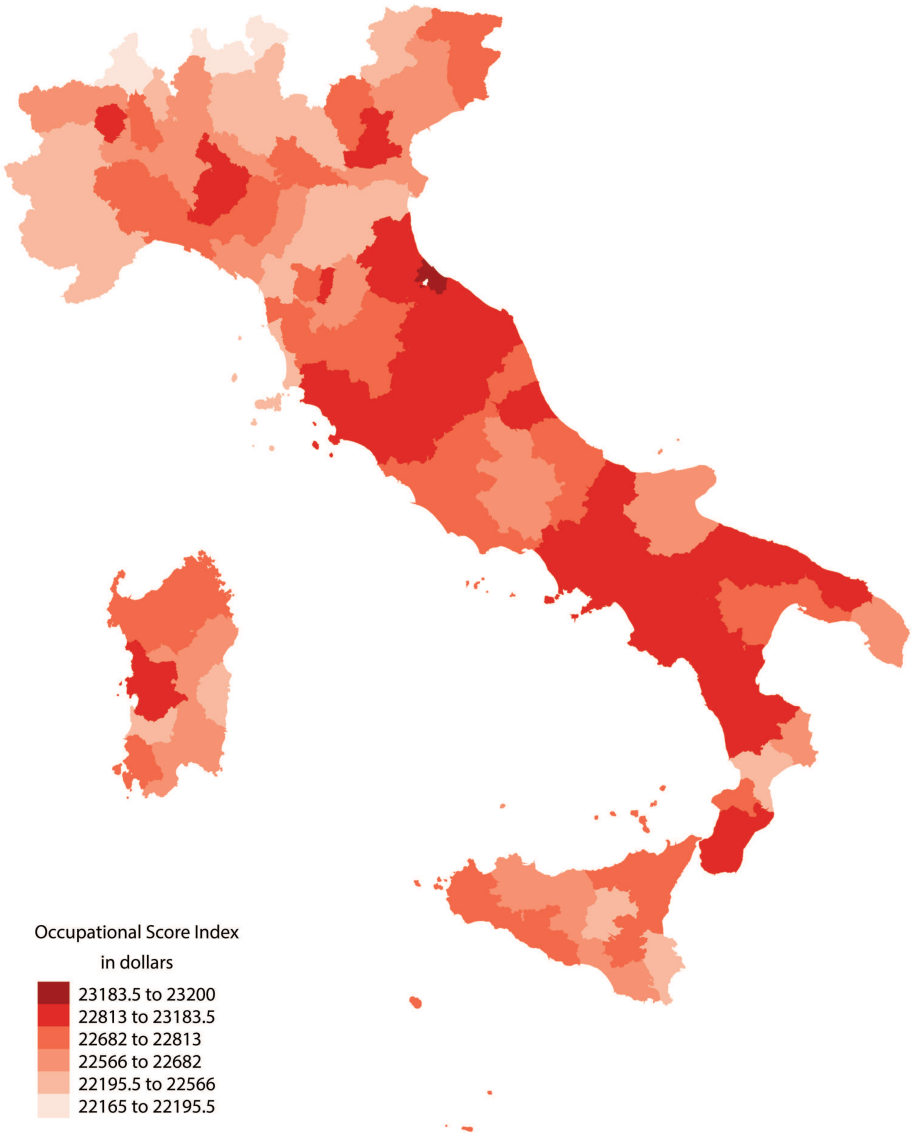
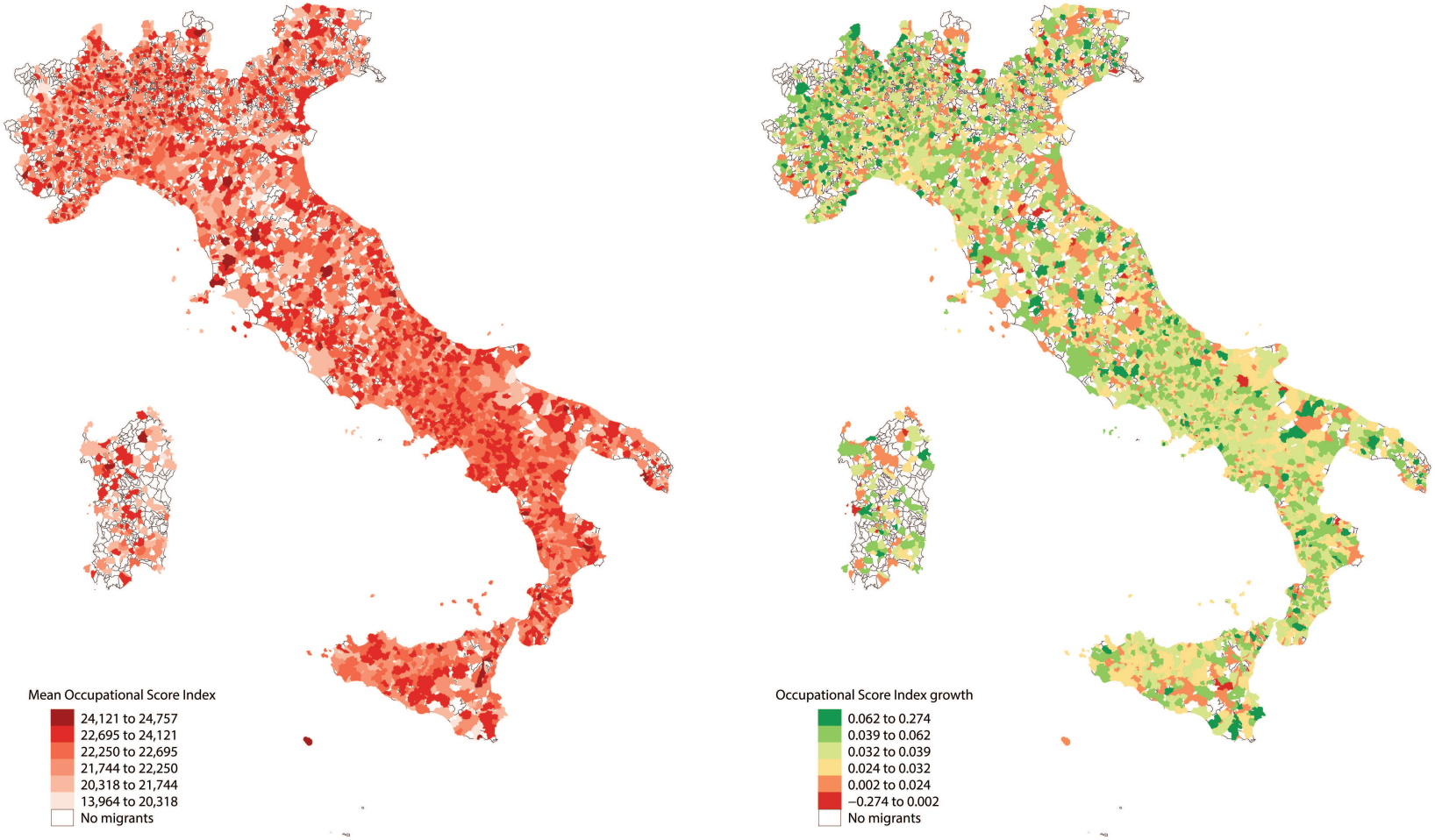


Figure 4. Average Occupational Score Index by province of origin.

Notes: The figure plots the average Occupational Score Index obtained from the imputation exercise for each province of origin.



(a) Mean of 1880's values

(b) Growth rate 1880-1900

Figure 5. Occupational Score Index by sending municipality.

Notes: The map on the left shows the average occupational score index at destination for each sending municipality; the proportion of migrants to the US up to 1885 settling in each county are used as weights. The map on the right shows the growth between 1880 and 1900, by sending municipality.

References

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the Age of Mass Migration. *The American Economic Review*, *102*(5), 1832–1856.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the Age of Mass Migration. *Journal of Political Economy*, *122*(3), 467–506.

Abramitzky, R., Boustan, L. P., Eriksson, K., Feigenbaum, J. J., & Pérez, S. (2019). Automated linking of historical data. Technical report, National Bureau of Economic Research.

Altonji, J. G. & Card, D. (1991). The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, trade, and the labor market* (pp. 201–234). University of Chicago Press.

Andersson, D., Karadja, M., & Prawitz, E. (2016). Mass migration, cheap labor, and innovation. *Unpublished paper, Uppsala University*.

Bandiera, O., Rasul, I., & Viarengo, M. (2013). The making of modern America: Migratory flows in the Age of Mass Migration. *Journal of Development Economics*, *102*, 23–47.

Bartik, T. J. (1991). Boon or boondoggle? the debate over state and local economic development policies. *In Who Benefits from State and Local Economic Development Policies? Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, pp. 1-16*, 1–16.

Bauer, T., Epstein, G. S., & Gang, I. N. (2005). Enclaves, language, and the location choice of migrants. *Journal of Population Economics*, *18*(4), 649–662.

Beaman, L. A. (2011). Social networks and the dynamics of labour market outcomes: Evidence from refugees resettled in the US. *The Review of Economic Studies*, *79*(1), 128–161.

Borjas, G. (1987). Self-selection and the earnings of immigrants. *The American Economic Review*, *77*(4), 531–53.

Borjas, G. (1994). The economics of immigration. *Journal of Economic Literature*, *32*(4), 1667–1717.

Borjas, G. J. & Bratsberg, B. (1996). Who leaves? the outmigration of the foreign-born. *The Review of Economics and Statistics*, 165–176.

Boustan, L. P., Fishback, P. V., & Kantor, S. (2010). The effect of internal migration on local labor markets: American cities during the great depression. *Journal of Labor Economics*, *28*(4), 719–746.

Brum, M. (2018). Italian migration to the United States: The role of migrant networks. Unpublished manuscript.

Burda, M. C. (1995). Migration and the option value of waiting. *Economic and Social Review*, *27*(1), 1.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414–427.

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

Carter, S. B., Gartner, S. S., Haines, M. R., Olmstead, A. L., Sutch, R., & Wright, G. (2006). *Historical Statistics of the United States: Millennial Edition*, volume 1. Cambridge: Cambridge University Press.

Cerase, F. P. (1974). Expectations and reality: a case study of return migration from the United States to southern Italy. *The International migration review*, *8*(2), 245–262.

Ciccarelli, C. & Fenoaltea, S. (2013). Through the magnifying glass: provincial aspects of industrial growth in post-unification Italy. *The Economic History Review*, *66*(1), 57–85.

Cinel, D. (2002). *The national integration of Italian return migration, 1870-1929*. Cambridge University Press.

Commissariato Generale Dell'Emigrazione (1927). *Annuario Statistico Della Emigrazione Italiana dal 1876 al 1925*. Edizione del Commissariato Dell'Emigrazione.

Cortes, P. (2008). The effect of low-skilled immigration on us prices: evidence from cpi data. *Journal of political Economy*, *116*(3), 381–422.

Damm, A. P. (2009). Determinants of recent immigrants' location choices: quasi-experimental evidence. *Journal of Population Economics*, *22*(1), 145–174.

Daniels, R. (2005). *Guarding the golden door: American immigration policy and immigrants since 1882*. Macmillan.

Del Boca, D. & Venturini, A. (2005). Italian migration. In K. Zimmermann (Ed.), *European migration: What do we know?* chapter 9, (pp. 303–352). Oxford: Oxford University Press.

Dunlevy, J. A. & Gemery, H. A. (1978). Economic opportunity and the responses of 'old' and 'new' migrants to the United States. *The Journal of Economic History*, *38*(4), 901–917.

Dunlevy, J. A. & Saba, R. P. (1992). The role of nationality-specific characteristics on the settlement patterns of late nineteenth century immigrants. *Explorations in Economic History*, *29*(2), 228–249.

Edin, P.-A., Fredriksson, P., & Åslund, O. (2003). Ethnic enclaves and the economic success of immigrants—evidence from a natural experiment. *The Quarterly Journal of Economics*, *118*(1), 329–357.

Farré, L. & Fasani, F. (2013). Media exposure and internal migration—evidence from Indonesia. *Journal of Development Economics*, *102*, 48–61.

Feigenbaum, J. J. (2014). A new old measure of intergenerational mobility: Iowa 1915 to 1940. Unpublished manuscript.

Feigenbaum, J. J. (2016). A machine learning approach to census record linking. Unpublished manuscript.

Gagliarducci, S. & Manacorda, M. (2016). Politics in the family: Nepotism and the hiring decisions of Italian firms.

Grogger, J. & Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics*, *95*(1), 42–57.

Güell, M., Pellizzari, M., Pica, G., & Rodríguez Mora, J. V. (2017). Correlating social mobility and economic outcomes. *Forthcoming at the Economic Journal*.

Harris, J. R. & Todaro, M. P. (1970). Migration, unemployment and development: a two-sector analysis. *The American economic review*, *60*(1), 126–142.

Hatton, T. J. (2000). How much did immigrant "quality" decline in late nineteenth century America? *Journal of Population Economics*, *13*(3), 509–525.

Hatton, T. J. (2010). The cliometrics of international migration: a survey. *Journal of Economic Surveys*, *24*(5), 941–969.

Hatton, T. J. & Williamson, J. G. (1998). *The Age of Mass Migration: Causes and economic impact*. Oxford University Press on Demand.

Hatton, T. J. & Williamson, J. G. (2005). *Global migration and the world economy: Two Centuries of Policy and Performance*. MIT Press, Cambridge, Massachusetts.

Istituto Nazionale di Statistica (2011). *L'Italia in 150 anni: sommario di statistiche storiche, 1861-2010*. ISTAT.

Johansson, E. (1977). *The history of literacy in Sweden: in comparison with some other countries*. Umeå universitet.

Karadja, M. & Prawitz, E. (2016). Exit, voice, and political change: Evidence from swedish mass migration to the united states.

Keeling, D. (1999a). Transatlantic shipping cartels and migration between Europe and America, 1880-1914. *Business and Economic History*, *17*, 195–213.

Keeling, D. (1999b). The transport revolution and transatlantic migration. *Economic History*, *19*, 40.

Lafortune, J. & Tessada, J. (2014). Smooth (er) landing? the dynamic role of networks in the location and occupational choice of immigrants. Unpublished manuscript.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1993). Theories of international migration: A review and appraisal. *Population and development review*, 431–466.

Mastrobuoni, G. (2015). The value of connections: Evidence from the Italian-American mafia. *The Economic Journal*, *125*(586), F256–F288.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, *23*(4), 1249–1274.

McKenzie, D., Gibson, J., & Stillman, S. (2013). A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad? *Journal of Development Economics*, *102*, 116–127.

McKenzie, D. & Rapoport, H. (2010). Self-selection patterns in Mexico-US migration: the role of migration networks. *The Review of Economics and Statistics*, *92*(4), 811–821.

Ministero di Agricoltura, Industria e Commercio (1884). *Censimento Della Popolazione del Regno d'Italia al 31 Dicembre 1881*. Ministero di Agricoltura, Industria e Commercio.

Moretti, E. (1999). Social networks and migrations: Italy 1876-1913. *International Migration Review*, 640–657.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *The Quarterly Journal of Economics*, *118*(2), 549–599.

O'Connell, P. G. (1997). Migration under uncertainty: "try your luck" or "wait and see". *Journal of Regional Science*, *37*(2), 331–347.

Ortega, F. & Peri, G. (2013). The effect of income and immigration policies on international migration. *Migration Studies*, *1*(1), 47–74.

Ruggles, S., Genadek, K., Goeken, R., Grover, J., & Sobek, M. (2015). Integrated public use microdata series: Version 6.0:[machine-readable database]. Minneapolis: University of Minnesota, 2015.

Sánchez-Alonso, B. (2007). The other Europeans: immigration into Latin America and the international labour market (1870–1930). *Revista de Historia Economica-Journal of Iberian and Latin American Economic History*, *25*(3), 395–426.

Sequeira, S., Nunn, N., & Qian, N. (2017). Migrants and the making of America: The short-and long-run effects of immigration during the age of mass migration. *NBER Working Paper No. 23289*.

Åslund, O. (2005). Now and forever? initial and subsequent location choices of immigrants. *Regional Science and Urban Economics*, *35*(2), 141–165.

Sori, E. (1979). *L'emigrazione Italiana dall'Unità alla seconda guerra mondiale*, volume 188. Il mulino.

Spitzer, Y. & Zimran, A. (2014). Migrant self-selection: Anthropometric evidence from the mass migration of Italians to the United States, 1907–1925. *mimeo*.

Appendix A. ITA data set: Municipality of origin imputation and validation exercise

The first sub-section details the procedure to impute municipality of origin to individuals with missing or invalid information observed in the Italians to America Passenger Data File data set. The second sub-section shows auxiliary regressions that contrast the resulting data set (after the imputation) with other historical sources, validating the imputation method.

A.1 Imputation procedure

Recall that the raw data set includes surnames and place of last residence. I drop migrants without a surname and those that report US citizenship or US-based last residence, as they should have been registered in a prior trip (4.9% of the sample). Table C.1 in Appendix C reports descriptive statistics for this (cleaned) full data set; Table A.1 below presents the incidence of missing place of last residence by year.

To address missing last residence I use information of individuals with no missing data to impute a place of origin to those with missing information. The imputation procedure takes advantage of the geographical concentration of Italian surnames, allowing a relatively precise identification of municipality of origin by surname. This links to further research that relies on surname matching, in the migration literature (Spitzer & Zimran, 2014; Abramitzky et al., 2012), and beyond (Güell et al., 2017; Feigenbaum, 2014). See Gagliarducci & Manacorda (2016) and Mastrobuoni (2015) for research exploiting geographical concentration of Italian surnames.

The key behind the imputation exercise is that Italian surnames are rare and concentrated, as on average a surname comes only from a very small sub-set of Italian municipalities. To check if this holds in this paper, I construct descriptive statistics reporting the distribution of individuals, surnames and municipalities for the clean ITA data set and two sub-samples. Table A.2 reports the number of individuals, surnames and municipalities for each of these three samples. The top panel in Table A.3 reports information on the distribution of individuals across surnames, for the clean ITA data set. This sample comprises 808,066 individuals with 159,753 surnames, with an average of 5.1 persons per surname (standard deviation of 27.9). Note that 16 surnames (0.01% of all the surnames in this sample) identify 25,180 individuals (3.1% of this sample), while 1% of the surnames cover 36.7% of the sample. These figures and the percentiles reported in the table show a skewed distribution of individuals across surnames.

Next, I restrict the sample to individuals with valid surname and municipality of

origin, which represents 42.9% of the clean ITA data set or 346,322 individuals, with 78,784 distinct surnames, coming from 5,285 municipalities. The middle panel in Table A.3 describes the distribution of individuals across surnames and municipalities, and the concentration of surnames by municipality, for this restricted sample. It shows that on average a surname is shared by 4.4 individuals across 4.8 municipalities, and that 90% of the individuals in this restricted sample share their surname with at most 7 individuals; 90% of the surnames come from at most 8 municipalities.

Note that in the clean ITA data set, municipality of origin is entirely missing for 80,969 surnames (126,760 individuals or 15.7% of the sample).[74] Moreover, municipality of origin is entirely available for 47,163 surnames (68,469 individuals or 8.1% of the sample), which makes imputation unnecessary.[75] Then, the imputation applies over cases in which a surname is shared both by individuals with and without information on municipality of origin (common support). Concretely, the imputation exercise uses information for 277,854 individuals sharing 31,621 surnames coming from 5,057 municipalities to impute an origin place to 334,983 individuals with those same surnames (but no origin information). The bottom panel of Table A.3 describes the distribution of individuals across surnames and the distribution of surnames and municipalities, for this particular sub-sample. It shows that on average a surname is shared by 8.8 individuals across 8.8 municipalities, and that 90% of the individuals in this restricted sample share their surname with at most 17 individuals; 90% of the surnames come from at most 22 municipalities. Though a random individual could come from any municipality out of 5,057, this number drops dramatically once we know her surname.

Using these 277,854 individuals, I compute the proportion of individuals with surname s that come from municipality m, for all municipalities in this sample ($Share_{s,m} = \frac{M_{s,m}}{\sum_m M_{s,m}}$). I use these proportions as weights, to assign potential municipalities of origin to the 334,983 individuals that have a valid surname but no municipality of origin. Thus, each observation in the post-imputation sample represents an individual, a potential municipality of origin and a weight capturing the probability that the individuals' surname comes from that municipality (equal to 1 for individuals with no missing information).[76]

[74] For example, this is the case of 42 individuals that share the surname *Bottinelli*; all of them lack information on municipality of origin, rendering the imputation impossible.

[75] For example, this is the case of 26 individuals with surname *Sorci*; all of them have valid municipalities of origin, making the imputation unnecessary.

[76] Consider the following case, seen in the data. The clean ITA data set features 21 individuals with the surname *Puzo*: 10 have missing or invalid municipality of origin, 2 come from *Ferrazzano*, 2 from *Gildone*, 5 from *Montemiletto*, 1 from *Pietradefusi*, and 1 from *Venticano*. I use the 11 individuals

Note that this imputation exercise: i) imposes the distribution of municipalities by surname of individuals with non-missing information to the rest of the sample; ii) drops a sizeable portion of observations. Both potentially generate measurement error, discussed in detail in Section 6.2. The severity of measurement error depends, among others, on the number of observations available used for the imputation, by surname: intuitively, the imputation should be more precise the larger the number of observations available per surname. A complementary measure is the proportion of individuals with a surname for which municipality is missing: intuitively, a situation in which 20 observations are used to impute an origin to 5 individuals is less prone to measurement error than if they are used to impute origin to 100 individuals. Figure A.1 plots the cumulative proportion of individuals that receive an imputed municipality of origin against the number of observations used (left) and the proportion of missing information among each surname (right).

The figure shows that 25% of the imputed individuals have surnames for which the proportion of missing information is lower or equal to 34.7%. Another quarter of the imputed have surnames for which the share of missing ranges from 34.7% to 46.5%, and another quarter from that figure up to 62.9%. The figure also shows that for 25% of the imputed sample, the imputation is based on information from at least 178 individuals (up to 2782). Another quarter of the imputed have surnames for which the number of observations available with non-missing data ranges from 41 to 178, and another quarter from 8 to 41. These graphs ease concerns over how much the imputation exercise can be trusted.

Applying this methodology allows to impute a municipality of origin to 83.7% of the individuals with valid surname but missing or problematic municipality of origin; I drop the remaining observations. In the next sub-section I present a validation exercise that further reduces concerns over the methodology.

A.2 Validation exercise

First, I compare the migration trends at the national level observed in the ITA data set with those recorded in historical sources of the United States and Italy. I first compute the proportion of Italian migrants recorded in the ITA data set that arrived in each

with non problematic municipality of origin to compute the proportion of migrants with surname *Puzo* coming from each municipality of origin. Then, I assign to the 10 individuals with missing place of origin these 5 potential municipalities of origin, with corresponding weights: *Ferrazano*, 2/11; *Gildone*, 2/11; *Montemiletto*, 5/11; *Pietradefusi*, 1/11; *Venticano*, 1/11. The 10 observations originally corresponding to single individuals with surname *Puzo* and missing place of origin now turn to 50 observations (10 individuals $\times$ 5 potential municipalities of origin each) representing 'fractions' of individuals.

different year: $Share_{ITA,y} = \frac{M_{ITA,y}}{\sum_y M_{ITA,y}}$; M is the stock of migrants, ITA denotes the source of the data, y indicates year, and $\sum_y M_{ITA,y}$ is the stock of migrants in a given period (e.g. 1855 to 1900). That is, all the Italians arriving in year y divided by the stock of Italians arriving in the period covered by the data set. Then I compute the same measure ($Share_{hUS,y} = \frac{M_{hUS,y}}{\sum_y M_{hUS,y}}$) using US historical records (Carter et al., 2006), and Italian ($Share_{hIT,y} = \frac{M_{hIT,y}}{\sum_y M_{hIT,y}}$) ones (Commissariato Generale Dell'Emigrazione, 1927). In all cases $1855 \leq y \leq 1900$, to allow comparability. I then estimate the following equations, through OLS and using robust standard errors:

$$Share_{ITA,y} = \alpha + \beta Share_{hUS,y} + \varepsilon_y \tag{A.1}$$

$$Share_{ITA,y} = \alpha + \beta Share_{hIT,y} + \varepsilon_y \tag{A.2}$$

Equation A.1 comprises ITA data and US historical records, using data for the period 1855 to 1900 (full coverage of the ITA data set). Unfortunately, Italian historical records cover only the period 1876-1900, then the estimation of Equation A.2 uses only data for that sub-period.

Second, I compare the proportion of yearly migrants by Italian region to the US recorded in the ITA data set with Italian historical records. Information on Italian migration to the US by region of origin from the ITA data set results from aggregating flows by municipality (both truly reported and imputed through surname) at the regional level. First, I compute $Share_{ITA,r,y} = \frac{M_{ITA,r,y}}{M_{ITA,y}}$; r indicates each Italian region of origin. Then $Share_{ITA,r,y}$ measures the proportion of all Italians arriving in a given year, that came from each region. Though limited to 1876-1900, Italian historical records do have information on regional flows, hence I construct the same measure with historical data: $Share_{hIT,r,y} = \frac{M_{hIT,r,y}}{M_{hIT,y}}$. Armed with this variables, I estimate the following equation, through OLS and clustering standard errors at the regional level:[77]

$$Share_{hIT,r,y} = \alpha + \beta Share_{ITA,r,y} + d_r + d_y + \varepsilon_{r,y} \tag{A.3}$$

Here d_r and d_y are region and year fixed effects, respectively. Estimation of Equation A.3 uses only data for 1876-1900, as available in Italian historical records.

Finally, as an additional check, I compare the distribution of yearly arrivals reported in US historical records against the one captured by Italian historical records. I estimate

[77] Because of the small number of clusters, I correct the standard errors using 'wild bootstrap' as developed by Cameron et al. (2008).

the following equation, through OLS and using robust standard errors:

$$Share_{hIT,y} = \alpha + \beta Share_{hUS,y} + \varepsilon_y \tag{A.4}$$

Equation A.4 can only be estimated using data for 1876-1900 given availability of Italian historical records. Regression results are presented in Table A.4 at the end. Columns 1-4 correspond to the estimation of equations 1-4 above. A good fit between the variables under study should yield a high R-squared and a coefficient close to one. Results in Column 4 show that the distribution of arrivals by year reported in US and Italian historical statistics are very similar. Results in columns 1 and 2 show that at the national level the ITA data set accurately tracks both the US and Italian historical records. This suggests that the ITA data set adequately covers the true pattern of yearly arrivals, and it is not missing migrants from particular years.

Results in Column 3 show that the flows of Italian migrants by region observed in the ITA data set are also in line with Italian historical records. This is important given that the regional flows observed in the ITA data set partly stem from the municipality of origin imputation. Though measurement error at the municipality level surely exists, this result suggests that errors are not big enough as to signifficantly differ from the true regional pattern of arrivals (or from the one captured in Italian historical records).

Tables for Appendix A

Table A.1. Descriptive statistics. Missing place of last residence by year, 1855-1900 ITA data set.

Year of arrival	Observations	% of total arrivals	% Missing or invalid place of last residence	Year of arrival	Observations	% of total arrivals	% Missing or invalid place of last residence
1855	319	0.04%	100.0%	1879	714	0.08%	94.8%
1856	249	0.03%	96.4%	1880	9,409	1.11%	99.2%
1857	330	0.04%	97.6%	1881	8,186	0.97%	88.7%
1858	165	0.02%	100.0%	1882	13,350	1.58%	79.9%
1859	119	0.01%	100.0%	1883	23,305	2.76%	82.2%
1860	256	0.03%	100.0%	1884	11,690	1.38%	90.4%
1861	230	0.03%	100.0%	1885	13,453	1.59%	87.2%
1862	221	0.03%	96.4%	1886	25,900	3.06%	82.8%
1863	90	0.01%	96.7%	1887	27,855	3.30%	81.6%
1864	53	0.01%	98.1%	1888	36,852	4.36%	75.6%
1865	144	0.02%	92.4%	1889	26,762	3.17%	87.2%
1866	336	0.04%	97.0%	1890	57,318	6.78%	71.3%
1867	2	0.00%	100.0%	1891	58,117	6.87%	80.0%
1868	3	0.00%	100.0%	1892	49,455	5.85%	76.6%
1869	2	0.00%	100.0%	1893	57,483	6.80%	48.2%
1870	9	0.00%	100.0%	1894	30,740	3.64%	67.3%
1871	3	0.00%	100.0%	1895	42,671	5.05%	91.2%
1872	3	0.00%	100.0%	1896	57,716	6.83%	81.7%
1873	456	0.05%	94.3%	1897	54,334	6.43%	58.0%
1874	335	0.04%	98.5%	1898	60,874	7.20%	13.7%
1875	142	0.02%	98.6%	1899	71,716	8.48%	20.3%
1876	231	0.03%	100.0%	1900	103,474	12.24%	13.5%
1877	76	0.01%	93.4%	Total	845,368	100.0%	60.9%
1878	220	0.03%	98.2%				

Notes: The table reports the percentage of observations per year found in the ITA data set, with missing literacy and missing or invalid place of last residence.

Table A.2. Descriptive statistics. Individuals, surnames and municipalities, sub-samples of 1855-1900 ITA data set.

Sample	Individuals	Surnames	Municipalities
Clean ITA data set	808,066	159,753	5,285
Restricted to individuals with non-missing municipality and surname	346,322	78,784	5,285
Restricted to individuals with non-missing municipality and surnames with common support	277,854	31,621	5,057

Notes: The table reports the number of individuals, surnames and municipalities in each of three sub-samples of the ITA data set.

Table A.3. Descriptive statistics. Distribution of individuals, surnames and municipalities, sub-samples of 1855-1900 ITA data set.

	Mean	Std. Dev.	Percentile 10th	25th	50th	75th	90th	Min	Max
	Clean ITA data set								
Individuals per surname	5.1	27.9	1	1	1	3	7	1	3108
	Restricted to non-missing municipality and surname								
Individuals per surname	4.4	18.4	1	1	1	3	7	1	1504
Municipalities per surname	4.8	16.6	1	1	1	2	8	1	868
Surnames per municipality	72.0	173.3	1	3	18	74	190	1	7093
	Restricted to non-missing municipality and surnames with common support								
Individuals per surname	8.8	28.5	1	1	3	7	17	1	1504
Municipalities per surname	8.8	24.5	1	1	2	6	22	1	868
Surnames per municipality	54.9	122.6	1	3	15	59	143	1	4682

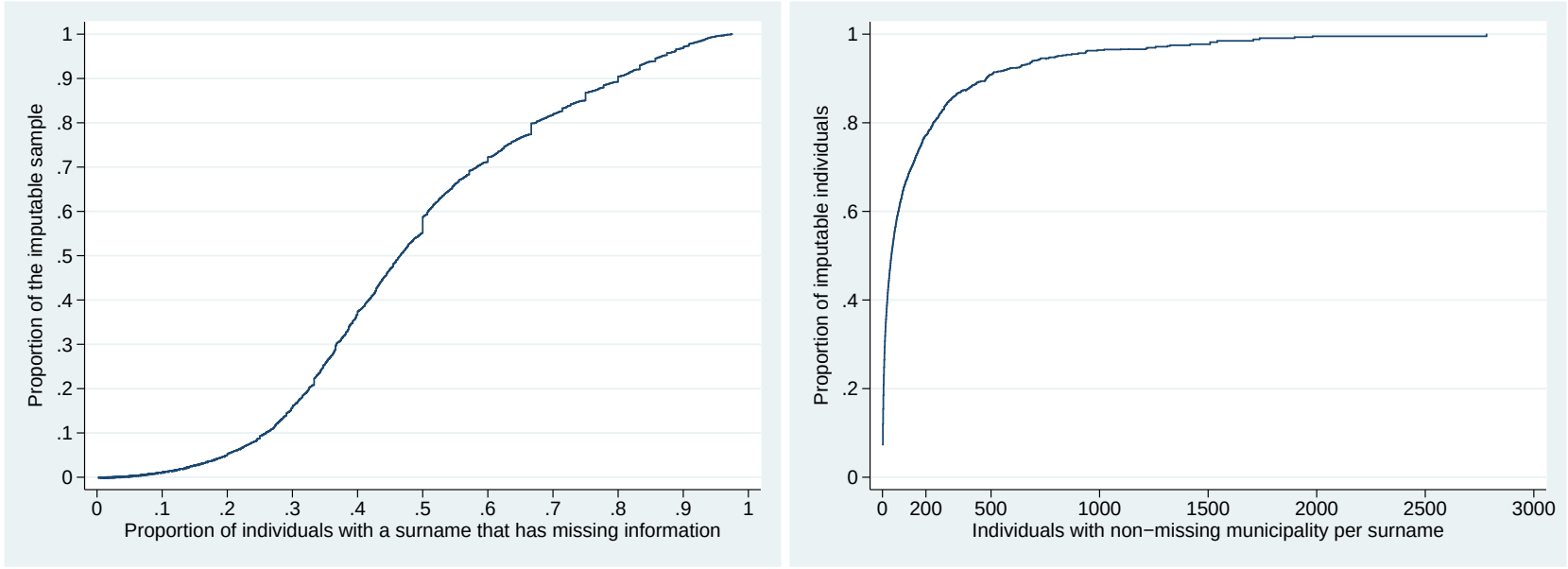
Notes: The table presents descriptive statistics on the distribution of individuals, surnames and municipalities by sub-sample of the 1855-1900 ITA data set.

Table A.4. Data validation results. Comparison of migrant flow distributions across data sources.

	(1) Distribution of migrants by years, 1855-1900 (ITA data set)	(2) Distribution of migrants by years, 1876-1900 (ITA data set)	(3) Distribution of migrants by regions, yearly in 1876-1900 (IT historical data)	(4) Distribution of migrants by years, 1876-1900 (IT historical data)
Distribution of migrants by years, 1855-1900 (US historical data)	1.1772***			
	(0.105)			
Distribution of migrants by years, 1876-1900 (IT historical data)		1.2032***		
		(0.076)		
Distribution of migrants by regions, yearly in 1876-1900 (ITA data set)			1.200**	
			(0.486)	
Distribution of migrants by years, 1876-1900) (US historical data)				1.005***
				(0.066)
Observations	46	25	425	25
R-squared	0.9071	0.9546	0.843	0.9222
Region FE			X	
Year of arrival FE			X	

Notes: Robust standard errors in parenthesis in columns 1, 2 and 4. Standard errors clustered at the regional level and corrected using wild bootstrap in parenthesis in Column 3. The dependent variable is the distribution of Italian migrants by year of arrival as observed in the ITA data set (columns 1 and 2, for the indicated periods) and as observed in Italian historical records (Column 4). In Column 3 the dependent variable is the distribution of Italian migrants by regions of origin for each year of arrival, obtained from historical records. The independent variable is the corresponding distribution, obtained from US historical data (columns 1 and 4), Italian historical data (Column 2) and the ITA data set (Column 3). Significance: *** p<0.01, ** p<0.05, * p<0.1.

Figures for Appendix A



(a) Proportion (b) Number of observations

Figure A.1. Cumulative distribution functions for the imputation of municipality of origin.

Notes: The figure on the left plots the proportion of the imputable sample that receives an imputed municipality of origin, against the proportion of individuals in the clean ITA data set that share the same surname but lack a valid municipality of origin. The figure on the right plots the proportion of the imputable sample that receives an imputed municipality of origin, against the number of observations used to perform the imputation.

Appendix B. ITA data set and US Census: Municipality of origin imputation and validation exercise

The first sub-section of this appendix details the surname matching procedure, the second focuses on the imputation procedure and the third discusses a validation exercise.

B.1 Machine learning surname matching procedure

The surname matching procedure takes advantage of the geographical concentration of Italian surnames, allowing a relatively precise identification of origin municipality by surname.[78] The geographical concentration of Italian surnames is also present in the data: Table C.7 below presents basic descriptive statistics on the surname distribution observed in the ITA data set and the Full US Schedules. For the latter, the table shows that a few very common surnames are shared by many individuals but for a large majority of Italians their surname is unique or shared by a small number of people. Also, most surnames can be found across a limited number of counties. For the former, the table shows, again, that a few very common surnames are shared by many individuals while a large majority of migrants in the sample have a unique surname or one shared by a small number of individuals. Also, most surnames come from a small set of municipalities. Overall, Table C.7 shows that the distributions of surnames across counties, municipalities and individuals are heavily skewed, that is, that surnames are *not* evenly distributed across municipalities, or shared by a large number of Italians across several counties in the US or municipalities in the ITA data set.

Since surnames allow to identify a relatively small set of individuals coming from a small set of municipalities, then they can be used to impute municipality of origin to the stock of Italians observed in the Full US Schedules. As mentioned before, this process faces two challenges. First, the Full US Schedules misreports surnames of Italian wives: 86.4% of married Italian women have the same surname as the male head of the household (their husband), while at least 72.5% of Italian wives got married before migrating (should have a different surname as Italian women do not take their husbands' surname when marrying). Note that 20.4% of all Italians found in the Full US Schedules are women with their husbands' surname, and at least 14.8% of the sample has incorrect surnames, that is, correspond to women that married before migrating. To address this issue I drop women altogether and focus only on men.

[78] See references in Appendix A.

Note that this decision may represent an advantage in terms of the analysis: given the limited presence of women undertaking paid activities in the labour market, it is unclear that the role of income, networks and other factors should have had the same effect on them as on men. Historical studies suggest that migratory flows of women respond more to family reunification decisions, in a context of decisions taken by the (male) household head, affecting the whole household (see for example Hatton & Williamson (1998)). In the extreme case in which all women purely follow a household head, and all women come from the same municipality as the household head, their inclusion or exclusion should not have substantial effects on results.

A second and more important challenge arises from errors in surnames, whether when registered by census officials or shipping companies, or when digitized a century later, or both. Surnames clearly are registered with error as only 37.7% of all Italian men observed in the Full US Schedules can be matched to the ITA data set when using original surnames in both data sets. To address this issue, I adapt a machine-learning matching methodology developed by Feigenbaum (2016), originally used to link individual records in two samples.

The starting point is two sets of clean data. First, the Full US Schedules, featuring all Italian men observed in 1900 with valid surnames (I drop 0.5% of the sample with missing surnames). This data set is comprised of 314,889 individuals with 144,606 different surnames. Second, the ITA data set, featuring Italian migrants who arrived by boat to the US over 1855-1900. This data set stems from the re-imputation exercise explained in Appendix A and is comprised of 682,005 individuals with 78,776 surnames. Surnames in both data sets go through a basic standardization process: setting them in upper case, eliminating multiple blanks, dropping extraneous characters , setting numbers into letters when appropriate, etc.[79] The data sets are then matched by surname: 20,818 surnames from the Full US Schedules can be matched to the ITA data set; this represents 14.4% of the Italian surnames in the US census and 37.7% of the Italian men in the sample.

Then, I generate a space of surname pairs that could potentially be matched (e.g., be the same were they not affected by error) using the Bi-gram algorithm. This algorithm calculates the similarity between every possible (11,391,482,256) pair of surnames. To do so, the algorithm evaluates how similar both surnames are based on how many pairs of characters they share and summarizes match quality in a 0 to 1 index (where 1 corresponds to equal surnames). To reduce the space of potential matches to those with a minimum chance of being good quality matches I drop surname pairs with a Bi-gram

[79] For example turning *DE BRUM* into *DE BRUM*; *BRUM[!* into *BRUM*; *BR0M* into *BROM*.

similarity lower than 0.6, obtaining 6,603,613 surname pairs.[80] I then calculate the Jaro-Winkler similarity index on this reduced set of potential matches. This algorithm measures similarities by focusing on the number of changes (permutation, additions, subtractions, etc.) needed to turn one surname into the other and summarizes match quality in a 0 to 1 index (where 1 corresponds to equal surnames). I then drop pairs for which the Jaro-Winkler index is lower than or equal to 0.5 *and* the Bi-gram index is lower than or equal to 0.7.[81] Then, I add both indexes and drop the bottom 25% of surname pairs. The resulting data set consists of 4,876,401 surname pairs and is called S_{full} from here onward. This represents 0.04% of the starting set of all possible surname pairs, and is the basis of the machine learning process.

The machine learning process requires a set of correct matches as a starting point from which to "learn" (*training data*). I follow a set of rules to code some of the surname pairs in S_{full} as correct matches. First and trivially, the subset of 20,818 surname pairs in which both surnames are identical is classified as a correct match. Then, I fix basic and common typing and spelling mistakes and set as correct any new match (for example, replacing triple consonants with double ones, allowing phonetic transcriptions, among others).[82] I then code pairs as correct matches based on rules regarding differences in surnames' last letters, treatment of duplicate/triplicate letters, and treatment of blanks and apostrophes in compound surnames. Specifically, I code as correct those pairs in which both surnames have more than 4 letters and are identical except for the last letter and this is a vowel.[83] I code as correct the pairs in which both surnames have more than 5 letters and one is equal to the other minus the last letter.[84] I replace duplicate letters with a single one in one data set and set resulting pairs with identical surnames as correct.[85] I code as correct those pairs in which surnames are identical after removal of any blanks, apostrophes or hypens on one of them.[86] Then I code as correct those pairs with high similarity indexes, according to the following rules: i) matches with the highest 1% Jaro-Winkler index and with the highest 10% Bi-gram index if the Jaro-Winkler index is greater than or equal to 0.7 are coded as

[80] As a reference, a similarity of exactly 0.6 corresponds to the surname pair *FAZZIE* and *AGAZZI*. Feigenbaum (2016) uses 0.8.

[81] Jaro-Winkler and Bi-gram similarity indices may differ; the latter may assign higher similarity to strings that heavily feature letter permutations. For example, in my sample *CARAVAGLIO* and *CAVARAGLIO* display a Bi-gram index of virtually 1 but a Jaro-Winkler index of 0.97.

[82] For example, matching *BRRRUM* with *BRRUM*, and *FILIPPO* with *PHILIPPO*.

[83] For example, *BRUMO* and *BRUMU*.

[84] For example, *BRUMO* and *BRUM*.

[85] For example, *BRUMMO* and *BRUMO*, but not *BRRUMO* and *BRUMMO*.

[86] For example, *DE BRUM* and *DEBRUM*, *DE LA BRUM* and *DELABRUM*, *DE'BRUM* and *DEBRUM*, *BRUM'* and *BRUM*, *DE-BRUM* and *DE BRUM*.

correct; ii) matches with the lowest 1% Jaro-Winkler index and with the lowest 1% Bi-gram index if the Jaro-Winkler index is lower than 0.8 are coded as incorrect.

These rules are ad-hoc and simplify the generation of *training data*. Note that the rules overlap, in the sense that many surname pairs are classified as correct through the application of more than one rule.[87] This procedure yields a sample (named S_0) of 351,2312 surname pairs (7.2% of S_{full}) coded into correct (30.4%) or incorrect (69.4%) matches. S_0 is the starting point for the machine learning process, while the remaining 4,525,170 surname pairs (, named S_{ml}) will be used to predict further matches.

The third step is to take S_0 and run a logit regression of the surname match coding against a set of characteristics of the surnames. I include as regressors the length of the US surname, the difference in surname lengths; dummies indicating matches of the Soundex index, the reverse Soundex index, and of the first and last letters of the surnames; the Bi-gram and Jaro-Winkler distances; the number of matches in the ITA dataset for each US surname and its square; and the number of total matches per US and ITA surnames and their squares.[88] I also include interaction terms: the difference in surname lengths and the length of the US surname; first and last letter match; first letter match and the Jaro-Winkler distance; last letter match and the Jaro-Winkler distance; Jaro-Winkler and Bi-gram distances; Soundex match and the Jaro-Winkler distance; and Soundex match and reverse Soundex match.

The fourth step uses regression results to predict whether a match is correct or incorrect; this generates a continuous 0 to 1 variable, interpreted as the probability of a match being correct. Then, a criteria is required to convert this continuous variable into a dummy indicating a correct/incorrect match, and then to compare the prediction with the original manual coding. Feigenbaum (2016) proposes the following: for a given US surname with many potential matches in the ITA dataset, a potential match should be taken as correct when the predicted probability is: i) the highest among all potential matches for that US surname; ii) high enough; iii) higher enough compared to the next best potential match. As S_0 contains both a manual coding

[87] For instance, 21% of the matches coded as correct through the application of spelling rules also display high similarity indexes; 42.3% of the surname pairs coded as correct due to high similarity indexes also comply with other spelling rules.

[88] The Soundex index is a phonetic algorithm that assigns the same code to surnames that are written differently but pronounced similarly in the English language. The Soundex index turns the first letter of the surname and the following three consonants into a four-character string of text. It is thus the same for all surnames with the same initial and following three consonants: M236 corresponds both to *MASTROIANNI* and *MUSTROPIETRO*. Due to some Italian surnames being very long, I also compute the Soundex of the reverse of the surname (*INNAIORTSAM* for *MASTROIANNI*). This captures phonetic similarity of the last part of the surname; as the conventional Soundex approximates phonetic similarities of the first part of it.

and a predicted match probability, thresholds for ii) and iii) above can be chosen to optimize rightly predicted matches (i.e., when the prediction is the same as the manual coding). The author constructs a grid of thresholds and computes, for each point, the ratio of rightly predicted correct matches over total (manually coded) correct matches (abbreviated as PPR), and the ratio of rightly predicted correct matches over total predicted matches (TPR in short). The author shows that there is a trade-off between the ratios (increasing the first relaxes matching criteria and reduces the second), and chooses a point in the grid that maximizes the average of the two ratios and then uses regression results and optimal thresholds to predict matches in the remaining sample. I follow a different approach: I use different optimization criteria, audit the out-of-sample predictions, incorporate new information to the original starting sample and iterate the process.

Empirically, I define a grid of thresholds $0 \leq T_1 \leq 1$ and $0 \leq T_2 \leq 2$, calculate the PPR and TPR for every point and keep T_1 and T_2 with the highest TPR.[89] The algorithm produces more matches of lower quality than under alternative criteria, which is not a concern given that I manually audit a (random) sub-sample of the predictions. Starting from S_0, this procedure yields $T_1 = 0.01$ and $T_2 = 1$, with a PPR of 0.9713 and a TPR of 0.0884: the algorithm classifies as correct matches only 8.8% of the circa 350,000 surname pairs in S_0, but 97.1% of those predictions are right (i.e. the manual coding also classified them as correct matches).

I then use regression results and these optimal thresholds to predict correct and incorrect matches for *all* the surname pairs in S_{full}. I manually analyze a sub-sample of 25,000 new surname pairs and accept 71.3% of the predicted correct matches and 27.8% of the predicted incorrect matches, and recode the rest (overruling the machine-learning prediction). I attach this audited set to the starting sample S_0 and obtain a new starting sample, S_1. I iterate the process eight times in total; I stop iterating once the proportion of suggested correct matches that are manually accepted surpasses 90%. Then, at the end of iteration 7 I choose T_1 and T_2 to maximize the average between the PPR and the TPR; this way the last set of predicted correct matches will be smaller but of higher quality. I opt for conservativeness as I do not audit these predictions.

The resulting data set contains surname pairs classified as matched because: i) surnames are truly identical; ii) of rule application; iii) of high similarity indexes; iv) of the iterated machine-learning process. To remain cautious, I drop surname pairs for which the average between the Bi-gram and Jaro-Winkler similarity indexes is below

[89] I consider 100 evenly distributed points for T_1 and 40 for T_2. Considering more points for either threshold does not significantly affect results.

0.75 and, for each US surname, keep only the Italian surname with the highest average. The final data set allows 68% of the men originally observed in the Census data to be matched to the ITA data set.

This sub-sample of matched Italian men in the Full US Schedules consists of about ten million observations: each one represents an individual and a potential municipality of origin and has an attached weight. These weights are given by the proportion of Italians in the ITA data set that have matching surnames that come from each municipality, and capture the probability that the individuals' surnames come from each municipality. These ten million observations represent 260,199 actual individuals. In the following subsection I use this matched data to impute a municipality of origin to the (matched) Italian migrants observed in the 1900 US Census.

B.2 Municipality of origin imputation

Recall that the US Census captures the stock of Italian migrants and their place of residence in 1900; it has no information on Italian migrants that died or returned to Italy prior to the Census. Moreover, the Census identifies the place of residence in 1900, providing a proxy for the place of first arrival or settlement, which will be imperfect in the case of internal US migration by Italian migrants.

To assign origin municipality to Italians in the Full US Schedules I replicate the imputation procedure explained in Appendix A. I first take the matched data (ten million observations) and aggregate them at the surname and municipality level, effectively computing the number of individuals per surname in the ITA data set. Then I compute the proportion of individuals with surname s that come from municipality m, for all municipalities in the data set ($Share_{s,m} = \frac{M_{s,m}}{\sum_m M_{s,m}}$). I use these proportions as weights, to assign potential municipalities of origin to all Italian men in the Full US Schedules that have a matchable surname (all of them lack information on municipality of origin). Each observation in the matched US-ITA sample represents an Italian man in the census with a matchable surname, a potential municipality of origin and a weight capturing the probability that the his surname comes from that municipality.

The matched data can be aggregated into a Municipalities×Counties matrix, capturing the number of individuals from each municipality residing in each county in the year 1900. Table B.1 shows municipality-county connections arising from this matrix: a link is established if the corresponding cell in the matrix is not zero (there is a positive probability for an Italian from a given municipality to have settled in a given county). On average, a municipality is linked to 180 counties (median of 125). The distribution is skewed: many municipalities are linked to few counties while one, Naples, is linked

to 970. Similarly, each county is linked on average to 704 municipalities out of 5,085 in the ITA data set (median of 257), and a single county, New York, is connected to 4,779 municipalities.

The procedure to impute municipality of origin to Italians observed in the 1900 US Census is affected by several sources and types of measurement error. To ease concerns over this and validate this new dataset, in the next subsection I compare migration flows arising from the matched data with those stemming from other historical sources. Nevertheless, note that Abramitzky et al. (2019) evaluate different methods for record linkage and conclude that coefficient estimates and parameters of interest are similar when using linked samples based on each method.

B.3 Validation exercise

First, I compare the migration trends at the national level observed in the US 1900 Census data set with those recorded in historical sources of the United States and Italy. First, I compute the proportion of all Italians observed in the Census that arrived in each different year: $Share_{cUS,y} = \frac{M_{cUS,y}}{\sum_y M_{cUS,y}}$; M indicates the stock of migrants, cUS denotes the source of the data, y indicates year, and $\sum_y M_{US,y}$ represents the stock of migrants in a given period (e.g., 1820 to 1900). That is, all the Italians observed in the US Census of 1900 that declared having arrived in year y divided by the stock that declared having arrived in the whole period covered by the data set. I compute the same measure ($Share_{hUS,y} = \frac{M_{hUS,y}}{\sum_y M_{hUS,y}}$) using US historical records (Carter et al., 2006) and Italian ($Share_{hIT,y} = \frac{M_{hIT,y}}{\sum_y M_{hIT,y}}$) ones (Commissariato Generale Dell'Emigrazione, 1927). In the first case $1820 \leq y \leq 1900$ while in the second $1876 \leq y \leq 1900$, to allow comparability of the series. I then estimate the following equations, through OLS and using robust standard errors:

$$Share_{cUS,y} = \alpha + \beta Share_{hUS,y} + \varepsilon_y \tag{B.1}$$

$$Share_{cUS,y} = \alpha + \beta Share_{hIT,y} + \varepsilon_y \tag{B.2}$$

Note that equation B.1 comprises US Census data and US historical records, and hence uses data for the period 1820 to 1900 (full coverage of the Census data set). Unfortunately, Italian historical records cover only the period 1876 to 1900, then the estimation of equation B.2 uses only data for that sub-period.

Second, I compare the proportion of yearly migrants by Italian region to the US captured in the (matched) US Census with that observed in the pure unmatched ITA data set. I aggregate the data for matched Italian men observed in the US Census at the

region of origin (inferred from imputed municipality of origin) and year of arrival levels, computing $Share_{cUS,r,y} = \frac{M_{cUS,r,y}}{M_{cUS,y}}$; r indicates each Italian region of origin. $Share_{cUS,r,y}$ measures the proportion of all (observed and matched) Italian (men) arriving in a given year, that came from each region. I construct the same measure for all Italian migrants (men and women) observed in the ITA data as: $Share_{ITA,r,y} = \frac{M_{ITA,r,y}}{M_{ITA,y}}$. Then I estimate the following equation, through OLS and clustering standard errors at the regional level:[90]

$$Share_{cUS,r,y} = \alpha + \beta Share_{ITA,r,y} + d_r + d_y + \varepsilon_{r,y} \tag{B.3}$$

Here d_r and d_y are region and year fixed effects, respectively. Note that equation B.3 uses data for 1855-1900, due to availability of ITA data. Results are presented in Table B.2 at the end; columns 1 to 3 correspond to the estimation of equations 1 to 3 above. A good fit between the two variables under study should yield a high R-squared and a coefficient of around one.

Results in columns 1 and 2 show that at the national level the matched US-ITA data set quite accurately tracks both the US and Italian historical records. This suggests that the matched data adequately captures the true pattern of yearly arrivals, reducing concerns over measurement error. Results in Column 3 show that the flows of Italian migrants *by region* inferred from the matched US-ITA data set are very similar to those found in the unmatched ITA data set. This suggests that the matched sample is similar to the raw flow of Italians observed in the manifest data, which itself tracks closely Italian historial records (as shown in Appendix A). These regional yearly flows are affected by measurement error in surnames in both the US census and the ITA dataset, by return and internal migration of Italians in the US, and by migrant mortality, among other issues. Then, this result reduces concerns over all these sources of measurement error. In particular, these results suggest that the sub-sample of matched men in the US 1900 Census is representative of the yearly regional Italy-US flows during the period.[91] See Section 6.2 for a detailed discussion of measurement error.

[90] Because of the small number of clusters, I correct the standard errors using 'wild bootstrap' as developed by Cameron et al. (2008).

[91] Imputation errors concerning municipalities from the same region do not affect the observed regional distribution, only mis-imputations that cross over regional borders do. Also, different errors could be cancelling each other out.

Tables for Appendix B

Table B.1. Match results. Municipality-county connections.

	Mean	Std. dev.	Median	9^{th} decile	Min	Max
Counties per municipality	180.5	172.9	125	433	1	970
Municipalities per county	704.8	909.7	257	2118	1	4779

Notes: The table presents descriptive statistics on the degree of connectedness between municipalities and counties. Statistics are computed from the sample of matched men.

Table B.2. Match data validation. Comparison of migrant flow distributions by data sources.

Variables	(1) distribution of migrants by year, 1820-1900 Full US Schedules 1900	(2) distribution of migrants by year, 1876-1900 Full US Schedules 1900	(3) distribution of migrants by regions, yearly in 1855-1900 Matched men in Full US Schedules 1900
distribution of migrants by year, 1820-1900 (US historical data)	0.926***		
	(0.043)		
distribution of migrants by year, 1876-1900 (IT historical data)		0.758***	
		(0.097)	
distribution of migrants by regions, yearly in 1855-1900 (ITA data set data)			0.956**
			(0.457)
Observations	83	25	736
R-squared	0.8527	0.7279	0.821
Region FE			X
Year of arrival			X

Notes: Robust standard errors in parenthesis in columns 1 and 2. Standard errors clustered at the regional level and corrected using wild bootstrap in parenthesis in Column 3. The dependent variable is the distribution of Italian migrants by year of arrival as observed in the Full US Schedules (columns 1 and 2, for the indicated periods) and as observed only for matched men in that data set (Column 3). The independent variable is the corresponding distribution, obtained from US historical data (Column 1) and Italian historical data (columns 2 and 3). Significance: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Appendix C. Additional Tables and Figures in the text

Tables

Table C.1. Descriptive statistics. Italians, ITA data set.

Variables	% in sample
Gender	
Men	74.4%
Departure Port	
Naples	42.9%
Genoa	19.2%
Havre	15.2%
Palermo	6.2%
Occupation	
Labourer	33.6%
Farmer	17.2%
Peasant	6.9%
Countryman	5.3%
Shoemaker	1.8%
Tailor	1.4%
Other	33.8%
	Mean
Age	26.7
Year of arrival	1893.1
Literacy rate	38.3%
$Network_{m,1897}$	-
Observations	845,368

Period	% Missing last residence
1855-1892	80.2%
1893-1897	68.3%
1898-1900	15.6%
1855-1900	60.9%

% that can be imputed back	83.7%

Notes: The table reports descriptive statistics for the full ITA sample (1855-1900). 'Other' captures 355 occupations that represent less than 1.5% of the sample each. The bottom panel reports the proportion of the full sample with missing information on place of last residence for three subperiods, and the share of migrants with missing or invalid last residence (in the 1855-1900 full sample) for which municipality can be imputed back.

Table C.2. Descriptive statistics. Variables and municipality coverage, 1881 Italian census summary.

Variables	Coverage	
	Municipalities	% of total pop.
Population	8259	100.0%
Literacy rate (by gender)	69	17.6%
Employment rate (by gender)	205	24.8%
% of working-age population employed in agriculture (by gender)	205	24.8%
% of population in 14 age groups (by gender)	69	17.6%
% of population single, married or widowed (by gender)	69	17.6%

Notes: The table reports the number of municipalities with data, for each group of variables taken from the 1881 Italian census summary. For the municipalities covered the table also reports the sum of their population as a share of the total Italian population.

Table C.3. Descriptive statistics. Municipality characteristics, 1881 Italian Census.

Variables	Municipalities	Descriptive statistics	
		Mean	Std. dev.
Population	8,259	3,676.0	11437.6
Male literacy rate	69	0.5065	0.1530
Female literacy rate	69	0.3933	0.1707
Male employment rate	205	0.8125	0.0479
Female employment rate	205	0.4261	0.1861
% of working-age men employed in agriculture	139	0.2795	0.1750
% of working-age women employed in agriculture	139	0.1129	0.1527
% of men that are single	69	0.5796	0.0209
% of women that are single	69	0.5838	0.0231
% of men aged 10 to 20	69	0.3022	0.0301
% of men aged 21 to 35	69	0.2456	0.0290
% of men aged 36 to 55	69	0.2329	0.0181
% of men aged 56+	69	0.1281	0.0248
% of women aged 10 to 20	69	0.3086	0.0288
% of women aged 21 to 35	69	0.2269	0.0156
% of women aged 36 to 55	69	0.2404	0.0178
% of women aged 56+	69	0.1365	0.0232
Women/men ratio, ages 10 to 20	69	0.9889	0.0549
Women/men ratio, ages 21 to 35	69	0.9070	0.1323
Women/men ratio, ages 36 to 55	69	1.0039	0.1013
Women/men ratio, ages 55+	69	1.0415	0.1115
Distance to main port	8,259	128.5	82.7

Notes: The table reports descriptive statistics for municipality characteristics, taken from the Italian census summary of 1881. The column Municipalities list the number of municipalities of origin for which the data is available. I use latitude and longitude information from the Italian statistical agency (ISTAT) to calculate the distance to the nearest of the three main Italian ports in the period (Genoa, Naples and Palermo)

Table C.4. Descriptive statistics. Literacy rate by province, 1881 Italian Census.

Province	Literacy rate			Province	Literacy rate		
	Men	Women	All		Men	Women	All
Alessandria	70.6%	52.6%	61.8%	Massa e Carrara	46.1%	22.7%	33.9%
Ancona	39.6%	23.4%	31.3%	Messina	24.2%	10.6%	17.3%
Aquila degli Abruzzi	38.5%	13.6%	25.0%	Milano	68.0%	63.6%	65.8%
Arezzo	35.2%	19.9%	27.8%	Modena	45.8%	32.3%	39.1%
Ascoli Piceno	29.3%	13.8%	21.3%	Napoli	39.4%	27.8%	33.6%
Avellino	27.6%	8.7%	18.0%	Novara	76.5%	61.9%	68.9%
Bari delle Puglie	23.8%	13.5%	18.6%	Padova	47.8%	29.3%	38.7%
Belluno	70.0%	45.5%	57.0%	Palermo	32.4%	19.3%	25.9%
Benevento	28.0%	8.5%	18.2%	Parma	39.4%	28.3%	33.9%
Bergamo	70.9%	65.8%	68.4%	Pavia	62.4%	52.1%	57.3%
Bologna	49.7%	37.6%	43.8%	Perugia	33.8%	18.4%	26.3%
Brescia	66.5%	59.5%	63.1%	Pesaro Urbino	31.3%	20.3%	25.8%
Cagliari	25.1%	10.2%	17.9%	Piacenza	39.5%	33.0%	36.4%
Caltanissetta	23.0%	8.5%	15.9%	Pisa	46.8%	27.8%	37.7%
Campobasso	28.8%	8.0%	17.9%	Porto Maurizio	72.0%	50.7%	61.3%
Caserta	30.7%	13.4%	22.0%	Potenza	23.0%	7.5%	14.8%
Catania	22.9%	10.6%	16.7%	Ravenna	35.5%	29.7%	32.6%
Catanzaro	25.5%	7.0%	16.2%	Reggio di Calabria	22.7%	8.2%	15.3%
Chieti	26.6%	9.4%	17.8%	Reggio nell'Emilia	47.2%	28.5%	37.9%
Como	76.5%	64.7%	70.6%	Roma	48.3%	34.4%	41.8%
Cosenza	22.6%	5.8%	13.6%	Rovigo	47.0%	25.3%	36.2%
Cremona	59.4%	52.5%	56.0%	Salerno	28.8%	11.8%	20.0%
Cuneo	69.1%	53.5%	61.4%	Sassari	31.0%	16.3%	23.9%
Ferrara	40.4%	25.9%	33.3%	Siena	35.5%	25.0%	30.6%
Firenze	46.4%	35.3%	40.9%	Siracusa	22.5%	9.6%	16.1%
Foggia	30.9%	15.1%	23.0%	Sondrio	75.9%	64.4%	70.0%
Forli	32.0%	23.7%	27.9%	Teramo	23.3%	8.7%	16.0%
Genova	61.5%	47.7%	54.5%	Torino	81.5%	68.6%	74.9%
Girgenti	22.6%	8.7%	15.6%	Trapani	23.7%	11.6%	17.6%
Grosseto	38.2%	27.7%	33.7%	Treviso	56.7%	35.4%	46.3%
Lecce	26.5%	12.6%	19.5%	Udine	62.1%	26.9%	44.2%
Livorno	60.9%	50.5%	55.7%	Venezia	51.4%	37.6%	44.5%
Lucca	52.3%	31.2%	41.1%	Verona	61.7%	42.5%	52.4%
Macerata	32.9%	16.3%	24.3%	Vicenza	61.9%	38.1%	50.1%
Mantova	52.9%	38.2%	45.7%				

Notes: literacy rates are reported for individuals aged 6 and older. Source: 1881 Italian census summary (Ministero di Agricoltura, Industria e Commercio, 1884).

Table C.5. Descriptive statistics. Distribution of arrivals by state and year, Full US Schedules.

	Year of arrival (in intervals)										
State	1800-54	1855-59	1860-64	1865-69	1870-74	1875-79	1880-84	1885-89	1890-94	1895-1900	1800-1900
Alabama	0.89%	0.55%	0.42%	0.28%	0.18%	0.18%	0.18%	0.21%	0.13%	0.19%	0.2%
Arizona	0.49%	0.28%	0.18%	0.16%	0.22%	0.32%	0.15%	0.16%	0.15%	0.09%	0.1%
Arkansas	0.73%	0.55%	0.30%	0.28%	0.12%	0.08%	0.14%	0.05%	0.03%	0.14%	0.1%
California	21.77%	21.75%	28.03%	21.56%	16.52%	13.84%	7.15%	5.24%	3.91%	2.64%	4.7%
Colorado	1.38%	0.46%	0.85%	0.87%	1.33%	1.56%	1.58%	1.61%	1.43%	1.25%	1.4%
Connecticut	0.73%	0.92%	1.27%	1.55%	1.91%	2.42%	3.16%	3.80%	4.11%	4.74%	4.0%
Delaware	0.08%			0.08%	0.09%	0.08%	0.29%	0.25%	0.23%	0.25%	0.2%
Dist. of Columbia	1.30%	1.29%	1.39%	0.32%	0.68%	0.23%	0.25%	0.21%	0.18%	0.13%	0.2%
Florida	0.97%	0.83%	0.61%	0.20%	0.18%	0.13%	0.15%	0.16%	0.18%	0.18%	0.4%
Georgia	0.41%	0.65%	0.18%	0.32%	0.14%	0.10%	0.06%	0.06%	0.04%	0.01%	0.0%
Idaho	0.24%	0.09%	0.06%	0.20%	0.18%	0.21%	0.14%	0.17%	0.17%	0.14%	0.2%
Illinois	4.22%	6.73%	4.54%	4.64%	4.82%	4.95%	6.22%	5.75%	5.62%	3.42%	4.9%
Indiana	1.54%	1.11%	0.48%	0.36%	0.36%	0.31%	0.27%	0.25%	0.25%	0.24%	0.3%
Iowa	0.24%	0.46%	0.30%	0.52%	0.26%	0.33%	0.22%	0.25%	0.24%	0.24%	0.2%
Kansas	0.41%	0.55%	0.36%	0.20%	0.20%	0.31%	0.24%	0.23%	0.25%	0.13%	0.2%
Kentucky	1.79%	1.94%	1.51%	1.67%	0.60%	0.30%	0.26%	0.15%	0.09%	0.04%	0.1%
Louisiana	9.34%	9.12%	10.77%	10.28%	5.27%	4.13%	3.22%	2.93%	3.75%	3.18%	3.6%
Maine	0.32%	0.74%	0.54%	0.44%	0.31%	0.22%	0.14%	0.16%	0.19%	0.40%	0.3%
Maryland	1.38%	1.66%	1.15%	0.75%	0.91%	0.83%	0.59%	0.45%	0.48%	0.47%	0.5%
Massachusetts	4.63%	5.81%	4.72%	5.88%	6.42%	4.25%	4.74%	5.20%	5.93%	6.64%	6.0%
Michigan	1.30%	1.01%	0.67%	0.56%	1.19%	0.94%	0.97%	1.19%	1.25%	1.46%	1.3%
Minnesota	0.97%	0.46%	0.18%	0.75%	0.47%	0.59%	0.62%	0.52%	0.46%	0.41%	0.5%
Mississippi	0.97%	1.66%	0.42%	0.64%	0.52%	0.42%	0.26%	0.11%	0.11%	0.15%	0.2%
Missouri	4.63%	5.81%	2.60%	2.66%	2.09%	1.39%	1.35%	0.99%	0.90%	0.61%	0.9%
Montana	0.49%	0.09%	0.42%	0.40%	0.42%	0.49%	0.42%	0.51%	0.45%	0.45%	0.5%
Nebraska	0.16%	0.28%	0.30%	0.36%	0.22%	0.32%	0.27%	0.24%	0.17%	0.08%	0.2%
Nevada	0.57%	0.37%	1.09%	0.79%	1.14%	1.21%	0.36%	0.25%	0.18%	0.21%	0.3%
New Hampshire	0.08%	0.09%	0.30%	0.04%	0.30%	0.12%	0.13%	0.11%	0.15%	0.26%	0.2%
New Jersey	5.04%	3.41%	5.39%	5.88%	7.06%	8.27%	8.88%	9.40%	8.75%	8.61%	8.6%
New Mexico	0.41%	0.18%	0.06%	0.12%	0.20%	0.26%	0.23%	0.12%	0.14%	0.08%	0.1%
New York	13.81%	12.26%	16.95%	20.01%	30.46%	36.19%	38.43%	38.54%	38.47%	39.44%	37.7%
North Carolina	0.16%	0.09%	0.06%	0.08%	0.14%	0.04%	0.02%	0.01%	0.10%	0.01%	0.0%
North Dakota			0.06%		0.02%	0.03%	0.06%	0.08%	0.16%	0.18%	0.1%
Ohio	4.31%	3.69%	2.60%	3.02%	2.26%	1.88%	2.15%	2.24%	2.35%	2.37%	2.3%
Oklahoma				0.12%	0.07%	0.17%	0.14%	0.16%	0.11%	0.11%	0.1%
Oregon	0.73%	0.18%	0.42%	0.60%	0.49%	0.54%	0.33%	0.26%	0.22%	0.12%	0.2%
Pennsylvania	5.61%	5.62%	5.39%	7.66%	7.00%	7.56%	11.12%	12.82%	13.31%	15.36%	13.6%
Rhode Island	0.32%	0.55%	0.30%	0.36%	0.91%	1.25%	1.16%	1.81%	1.93%	2.20%	1.8%
South Carolina	0.32%	0.18%	0.18%	0.08%	0.14%	0.05%	0.04%	0.04%	0.04%	0.01%	0.0%
South Dakota	0.24%	0.18%	0.06%		0.11%	0.13%	0.11%	0.09%	0.08%	0.05%	0.1%
Tennessee	2.03%	1.47%	1.09%	1.03%	0.55%	0.46%	0.26%	0.15%	0.16%	0.14%	0.3%
Texas	1.62%	2.58%	1.03%	1.79%	1.46%	1.12%	1.48%	0.88%	0.84%	0.43%	0.8%
Utah	0.81%	1.20%	0.54%	0.48%	0.18%	0.23%	0.24%	0.20%	0.24%	0.17%	0.2%
Vermont		0.09%	0.06%	0.12%	0.15%	0.08%	0.27%	0.26%	0.46%	0.56%	0.4%
Virginia	1.22%	0.83%	0.42%	0.36%	0.52%	0.36%	0.27%	0.15%	0.11%	0.07%	0.2%
Washington	0.24%	0.46%	0.97%	0.60%	0.60%	0.49%	0.58%	0.59%	0.46%	0.32%	0.4%
West Virginia	0.16%	0.28%	0.12%	0.32%	0.18%	0.22%	0.24%	0.28%	0.43%	0.95%	0.6%
Wisconsin	0.65%	1.38%	0.48%	0.67%	0.42%	0.31%	0.36%	0.54%	0.46%	0.45%	0.4%
Wyoming	0.24%	0.09%	0.12%		0.04%	0.08%	0.09%	0.15%	0.16%	0.20%	0.2%
All	0.25%	0.22%	0.34%	0.52%	1.66%	2.27%	8.48%	14.54%	25.74%	39.03%	

Notes: The table reports the share of Italian arrivals by state for year or arrival intervals, computed from the Full US Schedules.

Table C.6. Descriptive statistics. Distribution of Italians by region of origin and arrival year, ITA data set.

	Year of arrival (in intervals)									
Region of origin	1855-59	1860-64	1865-69	1870-74	1875-79	1880-84	1885-89	1890-94	1895-1900	1855-1900
Campania	17.6%	16.9%	21.2%	25.6%	18.4%	28.2%	31.1%	30.2%	28.2%	29.01%
Sicilia	9.1%	10.1%	11.3%	9.5%	9.1%	13.7%	12.6%	14.7%	19.7%	16.73%
Calabria	5.4%	9.2%	6.8%	6.7%	6.4%	9.8%	10.1%	11.1%	11.4%	10.94%
Basilicata	6.2%	3.0%	3.6%	8.8%	5.4%	8.7%	8.8%	7.3%	6.2%	7.02%
Molise	3.1%	3.7%	8.0%	3.9%	3.5%	6.9%	7.5%	7.3%	5.8%	6.50%
Abruzzi	3.3%	3.8%	3.0%	3.9%	4.4%	4.7%	4.9%	5.8%	5.9%	5.62%
Piemonte	9.7%	6.7%	10.0%	7.6%	5.6%	4.5%	4.6%	4.0%	4.2%	4.20%
Lombardia	10.7%	11.6%	6.9%	7.0%	7.9%	5.4%	4.2%	4.0%	3.6%	3.94%
Puglia	2.4%	1.3%	1.1%	1.9%	1.8%	2.3%	2.7%	2.7%	3.0%	2.79%
Lazio	2.2%	3.9%	3.1%	2.2%	2.6%	2.2%	2.3%	2.4%	2.9%	2.63%
Liguria	14.5%	15.8%	11.3%	10.3%	18.8%	5.4%	3.1%	2.5%	1.9%	2.50%
Emilia-Romagna	5.3%	3.6%	2.7%	4.1%	5.7%	2.7%	2.5%	2.3%	2.2%	2.33%
Toscana	5.1%	5.1%	2.8%	3.8%	5.4%	2.4%	2.3%	2.4%	1.8%	2.09%
Veneto	2.2%	2.1%	6.6%	1.9%	2.3%	1.4%	1.4%	1.5%	1.5%	1.45%
Marche	0.7%	1.5%	0.7%	1.2%	1.3%	0.7%	0.7%	0.8%	0.9%	0.83%
Umbria	0.5%	0.4%	0.8%	0.8%	0.3%	0.5%	0.3%	0.4%	0.4%	0.38%
Sardegna	1.1%	0.7%	0.1%	0.4%	0.5%	0.3%	0.3%	0.3%	0.3%	0.29%
Friuli	0.3%	0.5%	0.0%	0.3%	0.4%	0.2%	0.2%	0.2%	0.3%	0.24%
Valle d'aosta	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.03%
All	0.14%	0.10%	0.06%	0.10%	0.16%	7.81%	15.48%	29.95%	46.21%	

Notes: The table reports the share of Italian arrivals by region of origin for year intervals, taken from the ITA data set.

Table C.7. Descriptive statistics. Surname distributions by data source.

	Mean	Std. dev.	Median	9^{st}th decile	Min	Max
Full US Schedules						
Individuals per surname	3.2	10.6	2	5	1	1165
Counties per surname	1.4	2.4	1	2	1	185
Surnames per county	146.2	1106.1	6	240	1	36923
ITA data set						
Individuals per surname	8.6	39.4	2	15	1	3108
Municipalities per surname	4.8	16.6	1	8	1	868
Surnames per municipality	72.0	173.3	17	190	1	7093

Notes: The table presents descriptive statistics on the distribution of surnames across individuals and counties, from the US census data (top panel), and across individuals and municipalities, from the ITA data set.

Table C.8. Descriptive statistics. Distribution of Italian men by province of origin and year of arrival, matched Full US Schedules data.

Province of origin	Year of arrival			Province of origin	Year of arrival		
	Median	Mean	Std. dev.		Median	Mean	Std. dev.
Alessandria	1892	1890.43	8.58618	Massa e Carrara	1892	1890.23	8.74241
Ancona	1893	1891.53	7.27958	Messina	1893	1891.44	7.83667
Aquila Degli Abruzzi	1893	1891.7	7.58273	Milano	1892	1890.94	7.98648
Arezzo	1893	1890.71	8.27467	Modena	1893	1891.05	7.82843
Ascoli Piceno	1893	1891.27	7.87864	Napoli	1893	1891.22	7.58555
Avellino	1893	1891.61	7.18528	Novara	1892	1890.8	8.06966
Bari Delle Puglie	1894	1891.99	7.56608	Padova	1892	1890.05	8.65981
Belluno	1892	1891.28	7.45095	Palermo	1893	1891.63	7.315
Benevento	1893	1891.83	7.13381	Parma	1892	1890.56	8.41421
Bergamo	1893	1891.27	7.80096	Pavia	1892	1889.68	9.14459
Bologna	1892	1890.9	7.81186	Perugia	1893	1891.64	7.30939
Brescia	1892	1890.91	7.88918	Pesaro e Urbino	1893	1891.51	7.80629
Cagliari	1893	1891.44	7.32359	Piacenza	1892	1889.95	8.89034
Caltanissetta	1893	1891.94	6.95282	Pisa	1892	1891.07	7.57242
Campobasso	1893	1891.58	7.21853	Porto Maurizio	1893	1891.66	7.60978
Caserta	1893	1891.84	7.17091	Potenza	1892	1890.76	7.63783
Catania	1893	1891.64	7.39405	Ravenna	1893	1891.54	7.45271
Catanzaro	1893	1891.88	7.25715	Reggio di Calabria	1894	1892.08	7.13671
Chieti	1893	1891.74	7.39359	Reggio Nell'Emilia	1892	1890.85	7.86977
Como	1892	1890.25	8.31273	Roma	1893	1891.47	7.6457
Cosenza	1893	1891.52	7.27636	Rovigo	1893	1891.91	6.83589
Cremona	1893	1891.38	7.74459	Salerno	1893	1891.2	7.35119
Cuneo	1893	1891.19	7.9215	Sassari	1893	1890.67	8.8185
Ferrara	1893	1891.5	7.64445	Siena	1893	1891.12	7.3733
Firenze	1892	1890.55	8.47694	Siracusa	1894	1891.86	7.41564
Foggia	1893	1891.56	7.03427	Sondrio	1893	1891.48	7.37247
Forli	1892	1890.85	7.89873	Teramo	1893	1891.66	7.25884
Genova	1890	1886.41	11.5068	Torino	1892	1890.92	7.83941
Girgenti	1893	1891.83	7.35855	Trapani	1893	1891.49	7.23248
Grosseto	1893	1890.91	8.29698	Treviso	1894	1892.49	7.311
Lecce	1893	1891.39	7.44326	Udine	1892	1890.69	8.012
Livorno	1892	1890.81	7.9782	Venezia	1893	1890.73	7.66356
Lucca	1892	1889.7	9.24859	Verona	1892	1890.31	8.88142
Macerata	1893	1891.74	7.38285	Vicenza	1893	1891.4	7.46522
Mantova	1893	1891.63	7.50346				

Notes: The table reports the standard deviation, mean and median year of arrival by province of origin, computed from the sample of matched Italian men in the Full US Schedules.

Table C.9. First stage results, Municipality-County model.

Instruments	(1) First-wave municipality-county outmigration rate	(2) Average county income × % of first-wave migrants in county	(3) % of first-wave municipality-US migrants in county
	Specification 1		
Instrument: First-wave municipality-county outmigration rate	0.0050*** (0.0001)	0.0185*** (0.0020)	0.0022*** (0.0002)
Instrument: average county income × % of first-wave migrants in county	-0.0002*** (0.0001)	0.2294*** (0.0059)	0.0256*** (0.0007)
Instrument: % of first-wave municipality-US migrants in county	0.0016*** (0.0005)	-1.2516*** (0.0427)	-0.1352*** (0.0050)
Observations	46727679	46727679	46727679
F-statistic (Sanderson-Windmeijer)	4450.7	2345.0	2343.8
Time FE	X	X	X
County controls			
Time FE × Municipality controls			
County and Municipality FE			
Municipality × County FE			
Municipality × Time FE			
Time × County FE			
	Specification 2		
Instrument: First-wave municipality-county outmigration rate	0.0050*** (0.0001)	0.0185*** (0.0020)	0.0022*** (0.0002)
Instrument: average county income × % of first-wave migrants in county	-0.0003*** (0.0001)	0.2250*** (0.0061)	0.0250*** (0.0007)
Instrument: % of first-wave municipality-US migrants in county	0.0023*** (0.0005)	-1.2166*** (0.0445)	-0.1304*** (0.0053)
Observations	46727679	46727679	46727679
F-statistic (Sanderson-Windmeijer)	4385.0	2239.2	2237.7
Time FE	X	X	X
County controls	X	X	X
Time FE × Municipality controls	X	X	X
County and Municipality FE			
Municipality × County FE			
Municipality × Time FE			
Time × County FE			

Continued on next page

Continued from previous page

Instruments	(1) First-wave municipality-county outmigration rate	(2) Average county income × % of first-wave migrants in county	(3) % of first-wave municipality-US migrants in county
	Specification 3		
Instrument: First-wave municipality-county outmigration rate	0.0050***	0.0159***	0.0019***
	(0.0001)	(0.0019)	(0.0002)
Instrument: average county income × % of first-wave migrants in county	0.0001	0.2274***	0.0254***
	(0.0001)	(0.0068)	(0.0008)
Instrument: % of first-wave municipality-US migrants in county	-0.0009	-1.2690***	-0.1378***
	(0.0007)	(0.0522)	(0.0064)
Observations	46727679	46727679	46727679
F-statistic (Sanderson-Windmeijer)	3342.0	600.7	594.5
Time FE	X	X	X
County controls	X	X	X
Time FE × Municipality controls	X	X	X
County and Municipality FE	X	X	X
Municipality × County FE			
Municipality × Time FE			
Time × County FE			
	Specification 4		
Instrument: First-wave municipality-county outmigration rate			
Instrument: average county income × % of first-wave migrants in county		0.0265***	
		(0.0009)	
Instrument: % of first-wave municipality-US migrants in county			
Observations	46727679	46727679	46727679
F-statistic (Sanderson-Windmeijer)		875.4	
Time FE			
County controls			
Time FE × Municipality controls			
County and Municipality FE			
Municipality × County FE	X	X	X
Municipality × Time FE	X	X	X
Time × County FE	X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variables are the first-wave migration rate from each municipality to each county (in thousands, Column 1), the interaction between average county income and the proportion of first-wave migrants to the US that settled in each county (Column 2), and the proportion of first-wave migrants to the US that settled in each county (Column 3). The reported independent variables are the counterfactual versions of the three dependent variables just mentioned. These counterfactuals are based on the yearly distribution of first-wave non-Italian (and non-German) continental European migrants across counties. Each panel refers to a different specification. The first panel includes time fixed effects, panel 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Panel 3 adds county and municipality fixed effects; Panel 4 adds two-way interactions between time, county and municipality fixed effects. F-statistics are based on the Sanderson-Windmeijer statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.10. First stage results, Municipality-US model.

Instruments	(1) First-wave municipality-US outmigration rate	(2) Mean income of average first-wave migrant	(3) First-wave municipality-US outmigration rate	(4) Mean income of average first-wave migrant	(5) Mean income of average first-wave migrant
	Specification 1		Specification 2		Specification 3
Instrument: First-wave municipality-US outmigration rate	1.7331***	0.0293***	1.7262***	0.0265***	
	(0.0262)	(0.0056)	(0.0240)	(0.0054)	
Instrument: Mean income of average first-wave migrant	-0.0107	0.1362***	-0.0874***	0.1210***	0.2914***
	(0.0151)	(0.0310)	(0.0122)	(0.0315)	(0.0049)
Observations	23556	23556	23556	23556	23556
F-statistic (Sanderson-Windmeijer)	1061.5	20.3	898.0	16.0	3498.3
Time FE	X	X	X	X	X
Time FE × Muni. characteristics			X	X	X
Municipality FE					X
Province × Time FE					
	Specification 4		Specification 5		Specification 6
Instrument: First-wave municipality-US outmigration rate	1.7298***	0.0092**	1.7256***	0.0089**	
	(0.0285)	(0.0044)	(0.0266)	(0.0045)	
Instrument: Mean income of average first-wave migrant	-0.0119	0.1362***	-0.0861***	0.1251***	0.2870***
	(0.0123)	(0.0296)	(0.0117)	(0.0309)	(0.0050)
Observations	23556	23556	23556	23556	23556
F-statistic (Sanderson-Windmeijer)	7221.5	22.8	8287.6	17.6	3323.5
Time FE					
Time FE × Muni. characteristics			X	X	X
Municipality FE					X
Province × Time FE	X	X	X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variables are: the first-wave migration rate from each municipality to the US (in thousands, columns 1 and 3) and the weighted average of mean county income (weights given by the proportion of first-wave migrants residing in each county, columns 2, 4 and 5). In all columns, the main independent variables are the counterfactual versions of the two dependent variables in the table. These counterfactuals are based on the yearly distribution of first-wave non-Italian non-German continental European migrants across counties. Specifications 1 controls for time fixed effects, specification 2 adds municipality characteristics (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time, and specification 3 adds municipality fixed effects. Specifications 4, 5 and 6 repeat this pattern but always including time-province fixed effects. 1st stage F-statistics are based on the Sanderson-Windmeijer statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.11. Bartik income instrument results, Municipality - County model. Later-wave municipality-county outmigration rates.

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	Bartik instrument, by occupation			
First-wave municipality-county outmigration rate	1.1598***	1.1591***	1.1618***	
	(0.0435)	(0.0435)	(0.0436)	
% of first-wave municipality-US migrants in county	-0.4789***	-0.3888***	-1.3931***	
	(0.0347)	(0.0331)	(0.0877)	
Average county income	0.0002***	0.0006***	0.0072***	
	(0.0000)	(0.0000)	(0.0004)	
Interaction: average county income × % of first-wave migrants in county	0.0580***	0.0473***	0.1654***	0.4666***
	(0.0040)	(0.0039)	(0.0103)	(0.0273)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	595.8	578.4	163.0	2288.2
	Bartik instrument, by industry			
First-wave municipality-county outmigration rate	1.1593***	1.1586***	1.1605***	
	(0.0435)	(0.0434)	(0.0435)	
% of first-wave municipality-US migrants in county	-0.2724***	-0.1733***	-0.8161***	
	(0.0301)	(0.0291)	(0.0586)	
Average county income	0.0002***	0.0007***	-0.1533***	
	(0.0000)	(0.0000)	(0.0197)	
Interaction: average county income × % of first-wave migrants in county	0.0336***	0.0219***	0.0973***	0.3895***
	(0.0035)	(0.0034)	(0.0069)	(0.0222)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	531.8	511.7	19.1	2214.8
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the municipality-county per-thousand migration rate in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. Only IV estimations are reported: key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties, and with Bartik versions of average county income. In the top panel counterfactual county income is constructed applying national changes in the number of individuals by occupation to county baseline occupation figures; in the bottom panel the exercise is based on changes at the industry level. 1st stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.12. Bartik income instrument results, Municipality - County model. County concentration of later-wave municipality-US migrants.

Variables	(1) % of later-wave municipality-US migrants in county	(2) % of later-wave municipality-US migrants in county	(3) % of later-wave municipality-US migrants in county	(4) % of later-wave municipality-US migrants in county
	Bartik instrument, by occupation			
First-wave municipality-county outmigration rate	0.0007	0.0006	0.0013	
	(0.0021)	(0.0021)	(0.0020)	
% of first-wave municipality-US migrants in county	0.0735***	0.1057***	0.1604***	
	(0.0133)	(0.0138)	(0.0260)	
Average county income	0.0000***	0.0002***	-0.0003**	
	(0.0000)	(0.0000)	(0.0001)	
Interaction: average county income × % of first-wave migrants in county	-0.0034**	-0.0072***	-0.0141***	-0.0719***
	(0.0016)	(0.0016)	(0.0031)	(0.0137)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	595.8	578.4	163.0	2288.2
	Bartik instrument, by industry			
First-wave municipality-county outmigration rate	0.0008	0.0006	0.0015	
	(0.0021)	(0.0021)	(0.0020)	
% of first-wave municipality-US migrants in county	0.0390***	0.0720***	0.0623***	
	(0.0148)	(0.0155)	(0.0220)	
Average county income	0.0001***	0.0002***	-0.0318***	
	(0.0000)	(0.0000)	(0.0042)	
Interaction: average county income × % of first-wave migrants in county	0.0007	-0.0032*	-0.0025	-0.0783***
	(0.0017)	(0.0018)	(0.0025)	(0.0108)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	531.8	511.7	19.1	2214.8
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the proportion of later-wave migrants from a municipality that settled in a county, in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. Only IV estimations are reported: key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties, and with Bartik versions of average county income. In the top panel counterfactual county income is constructed applying national changes in the number of individuals by occupation to county baseline occupation figures; in the bottom panel the exercise is based on changes at the industry level. 1st stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.13. Bartik income instrument results, Municipality - US model. Later-wave municipality-US outmigration rates.

Variables	(1) Later-wave municipality-US outmigration rate	(2) Later-wave municipality-US outmigration rate	(3) Later-wave municipality-US outmigration rate	(4) Later-wave municipality-US outmigration rate	(5) Later-wave municipality-US outmigration rate	(6) Later-wave municipality-US outmigration rate
	Bartik instrument, by occupation					
First-wave municipality-US outmigration rate	1.2653***	1.2625***		1.2236***	1.2233***	
	(0.0488)	(0.0496)		(0.0573)	(0.0569)	
Mean income of average first-wave migrant	1.1371***	0.8571**	4.4907***	0.0842	0.2248	2.4227***
	(0.3451)	(0.3536)	(0.2040)	(0.2246)	(0.2631)	(0.1398)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1^{st} stage)	11.2	8.8	6366.3	12.3	9.5	5323.3
	Bartik instrument, by industry					
First-wave municipality-US outmigration rate	1.2674***	1.2648***		1.2236***	1.2234***	
	(0.0487)	(0.0495)		(0.0572)	(0.0569)	
Mean income of average first-wave migrant	0.9467***	0.6324*	4.5753***	-0.0166	0.1085	2.4893***
	(0.3359)	(0.3442)	(0.2058)	(0.2346)	(0.2726)	(0.1412)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1^{st} stage)	10.0	7.8	6938.6	11.1	8.6	5603.0
Time FE	X	X	X			
Time FE × Munic. baseline controls		X	X		X	X
Municipality FE			X			X
Province × Time FE				X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the municipality-US per-thousand later-wave migration rate, in three five-year intervals after 1885. The main independent variables are: the first-wave migration rate from each municipality to the United States, and a weighted average of county income, with weights equal to the proportion of first-wave migrants in each county. Column 1 includes time fixed effects, Column 2 adds municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds municipality fixed effects. Columns 4-6 repeat the estimations in 1-3 incluiding province-year fixed effects. Only IV estimations are reported: key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties, and with Bartik versions of average county income. In the top panel counterfactual county income is constructed applying national changes in the number of individuals by occupation to county baseline occupation figures; in the bottom panel the exercise is based on changes at the industry level. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.14. Alternative instruments results, Municipality - County model. Later-wave municipality-county outmigration rates.

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	IV: non-Italian Europeans			
First-wave municipality-county outmigration rate	1.1829***	1.1824***	1.1832***	
	(0.0432)	(0.0432)	(0.0432)	
% of first-wave migrants in county	0.0238	0.0730***	-0.4117***	
	(0.0263)	(0.0259)	(0.0404)	
Average county income	0.0001***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0012	-0.0070**	0.0497***	0.3681***
	(0.0030)	(0.0030)	(0.0047)	(0.0193)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	1966.3	2050.4	859.4	3408.0
	IV: all non-Italians			
First-wave municipality-county outmigration rate	1.1817***	1.1812***	1.1819***	
	(0.0435)	(0.0435)	(0.0435)	
% of first-wave migrants in county	0.0241	0.0756***	-0.3900***	
	(0.0243)	(0.0242)	(0.0380)	
Average county income	0.0001***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0013	-0.0073***	0.0472***	0.3552***
	(0.0028)	(0.0028)	(0.0044)	(0.0187)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	1943.7	1995.0	583.7	3740.6
	IV: non-Italian Continental Europeans			
First-wave municipality-county outmigration rate	1.1592***	1.1585***	1.1592***	
	(0.0445)	(0.0445)	(0.0445)	
% of first-wave municipality-US migrants in county	-0.1446***	-0.0657**	-0.5645***	
	(0.0298)	(0.0286)	(0.0454)	
Average county income	0.0002***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0185***	0.0092***	0.0677***	0.3025***
	(0.0034)	(0.0033)	(0.0053)	(0.0173)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	1171.2	1240.3	417.6	2195.0
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Continued on next page

Continued from previous page

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	IV: non-Italian Continental Europeans (excl. Germany, Sweden and Norway)			
First-wave municipality-county outmigration rate	1.1381***	1.1370***	1.1386***	
	(0.0456)	(0.0456)	(0.0456)	
% of first-wave municipality-US migrants in county	-0.0753***	0.0534*	-0.4861***	
	(0.0280)	(0.0275)	(0.0497)	
Average county income	0.0002***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0103***	-0.0048	0.0583***	0.2208***
	(0.0032)	(0.0031)	(0.0058)	(0.0145)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	476.7	494.6	158.9	1245.7
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the municipality-county per-thousand migration rate in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. Only IV results are reported. Key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from across US counties. The first panel considers all non-Italian Europeans (incluiding the UK); the second considers all non-Italians (adding African, Asian, Latinamerica, etc.). The third uses non-Italian Continental Europeans, and the fourth excludes also Germany, Sweden and Norway. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.15. Alternative instruments results, Municipality - County model. County concentration of later-wave municipality-US migrants.

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	IV: non-Italian Europeans			
First-wave municipality-county outmigration rate	1.1829***	1.1824***	1.1832***	
	(0.0432)	(0.0432)	(0.0432)	
% of first-wave migrants in county	0.0238	0.0730***	-0.4117***	
	(0.0263)	(0.0259)	(0.0404)	
Average county income	0.0001***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0012	-0.0070**	0.0497***	0.3681***
	(0.0030)	(0.0030)	(0.0047)	(0.0193)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	1966.3	2050.4	859.4	3408.0
	IV: all non-Italians			
First-wave municipality-county outmigration rate	1.1817***	1.1812***	1.1819***	
	(0.0435)	(0.0435)	(0.0435)	
% of first-wave migrants in county	0.0241	0.0756***	-0.3900***	
	(0.0243)	(0.0242)	(0.0380)	
Average county income	0.0001***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0013	-0.0073***	0.0472***	0.3552***
	(0.0028)	(0.0028)	(0.0044)	(0.0187)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	1943.7	1995.0	583.7	3740.6
	IV: non-Italian Continental Europeans			
First-wave municipality-county outmigration rate	1.1592***	1.1585***	1.1592***	
	(0.0445)	(0.0445)	(0.0445)	
% of first-wave municipality-US migrants in county	-0.1446***	-0.0657**	-0.5645***	
	(0.0298)	(0.0286)	(0.0454)	
Average county income	0.0002***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0185***	0.0092***	0.0677***	0.3025***
	(0.0034)	(0.0033)	(0.0053)	(0.0173)
Observations	46727679	46727679	46727679	46727679
F-stat. (1st stage)	1171.2	1240.3	417.6	2195.0
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Continued on next page

Continued from previous page

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	IV: non-Italian Continental Europeans (excl. Germany, Sweden and Norway)			
First-wave municipality-county outmigration rate	1.1381***	1.1370***	1.1386***	
	(0.0456)	(0.0456)	(0.0456)	
% of first-wave municipality-US migrants in county	-0.0753***	0.0534*	-0.4861***	
	(0.0280)	(0.0275)	(0.0497)	
Average county income	0.0002***	0.0004***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0103***	-0.0048	0.0583***	0.2208***
	(0.0032)	(0.0031)	(0.0058)	(0.0145)
Observations	46727679	46727679	46727679	46727679
F-stat. (1^{st} stage)	476.7	494.6	158.9	1245.7
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the proportion of later-wave migrants from a municipality that settled in a county, in three five-year intervals after 1885. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. Only IV results are reported. Key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from across US counties. The first panel considers all non-Italian Europeans (incluiding the UK); the second considers all non-Italians (adding African, Asian, Latinamerica, etc.). The third uses non-Italian Continental Europeans, and the fourth excludes also Germany, Sweden and Norway. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table C.16. Alternative instruments results, Municipality - US model. Later-wave municipality-US outmigration rates.

Variables	(1) Later-wave municipality-US outmigration rate	(2) Later-wave municipality-US outmigration rate	(3) Later-wave municipality-US outmigration rate	(4) Later-wave municipality-US outmigration rate	(5) Later-wave municipality-US outmigration rate	(6) Later-wave municipality-US outmigration rate
	IV: non-Italian Europeans					
First-wave municipality-US outmigration rate	1.2563***	1.2532***		1.2206***	1.2201***	
	(0.0492)	(0.0499)		(0.0576)	(0.0572)	
Mean income of average first-wave migrant	1.7001***	1.4928***	4.5485***	0.3859	0.6111*	2.4638***
	(0.4697)	(0.5155)	(0.2066)	(0.2568)	(0.3282)	(0.1407)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1st stage)	9.7	7.0	5893.6	10.4	7.5	5119.0
	IV: all non-Italians					
First-wave municipality-US outmigration rate	1.2567***	1.2535***		1.2208***	1.2203***	
	(0.0493)	(0.0499)		(0.0579)	(0.0575)	
Mean income of average first-wave migrant	1.6759***	1.4667***	4.5722***	0.3565	0.5842*	2.4845***
	(0.4614)	(0.5089)	(0.2074)	(0.2599)	(0.3317)	(0.1410)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1st stage)	10.2	7.5	6017.0	11.1	8.0	5140.5
	IV: non-Italian Continental Europeans					
First-wave municipality-US outmigration rate	1.2601***	1.2572***		1.2227***	1.2223***	
	(0.0495)	(0.0503)		(0.0580)	(0.0576)	
Mean income of average first-wave migrant	1.5308***	1.2930***	4.5316***	0.3533	0.5211*	2.4483***
	(0.4251)	(0.4393)	(0.2070)	(0.2265)	(0.2760)	(0.1403)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1st stage)	9.7	7.4	5501.6	10.3	7.9	4840.4
Time FE	X	X	X			
Time FE × Munic. baseline controls		X	X		X	X
Municipality FE			X			X
Province × Time FE				X	X	X

Continued on next page

Continued from previous page

Variables	(1) Later-wave municipality-US outmigration rate	(2) Later-wave municipality-US outmigration rate	(3) Later-wave municipality-US outmigration rate	(4) Later-wave municipality-US outmigration rate	(5) Later-wave municipality-US outmigration rate	(6) Later-wave municipality-US outmigration rate
	IV: non-Italian Continental Europeans (excl. Germany, Sweden and Norway)					
First-wave municipality-US outmigration rate	1.2716***	1.2693***		1.2294***	1.2293***	
	(0.0491)	(0.0499)		(0.0581)	(0.0577)	
Mean income of average first-wave migrant	1.0330***	0.7340**	4.4112***	0.0587	0.1818	2.3405***
	(0.3535)	(0.3627)	(0.2073)	(0.2283)	(0.2714)	(0.1387)
Observations	23556	23556	23556	23556	23556	23556
F-stat. (1^{st} stage)	9.1	6.9	4137.0	10.0	7.8	3826.3
Time FE	X	X	X			
Time FE × Munic. baseline controls		X	X		X	X
Municipality FE			X			X
Province × Time FE				X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1885. The dependent variable is the municipality-US per-thousand later-wave migration rate, in three five-year intervals after 1885. The main independent variables are: the first-wave migration rate from each municipality to the United States, and a weighted average of county income, with weights equal to the proportion of first-wave migrants in each county. Column 1 includes time fixed effects, Column 2 adds municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds municipality fixed effects. Columns 4-6 repeat the estimations in 1-3 incluiding province-year fixed effects. Only IV results are reported. Key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from across US counties. The first panel considers all non-Italian Europeans (incluiding the UK); the second considers all non-Italians (adding African, Asian, Latinamerica, etc.). The third uses non-Italian Continental Europeans, and the fourth excludes also Germany, Sweden and Norway. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.17. 1890 threshold results, Municipality - County model. Later-wave municipality-county outmigration rates.

Variables	(1) Later-wave municipality-county outmigration rate	(2) Later-wave municipality-county outmigration rate	(3) Later-wave municipality-county outmigration rate	(4) Later-wave municipality-county outmigration rate
	OLS			
First-wave municipality-county outmigration rate	0.618***	0.6179***	0.6048***	
	(0.0132)	(0.0131)	(0.0132)	
% of first-wave municipality-US migrants in county	-0.0717***	-0.0705***	-0.0413***	
	(0.0072)	(0.0071)	(0.0062)	
Average county income	0.0004***	0.0006***	0.0004***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0082***	0.0080***	0.0037***	0.0143***
	(0.0009)	(0.0009)	(0.0007)	(0.0026)
Observations	31151786	31151786	31151786	31151786
R-squared	0.608	0.609	0.619	0.890
	IV			
First-wave municipality-county outmigration rate	0.6856***	0.6854***	0.6857***	
	(0.0153)	(0.0153)	(0.0153)	
% of first-wave municipality-US migrants in county	0.1477*	0.2010***	-0.1448***	
	(0.0188)	(0.0187)	(0.0292)	
Average county income	0.0001***	0.0003***	0.0003***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0169***	-0.0232***	0.0174***	0.2957***
	(0.0022)	(0.0021)	(0.0034)	(0.0222)
Observations	31151786	31151786	31151786	31151786
F-stat. (1^{st} stage)	1059.8	980.0	196.2	612.6
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1890. The dependent variable is the municipality-county per-thousand migration rate in two five-year intervals after 1890. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.18. 1890 threshold results, Municipality - County model. County concentration of later-wave municipality-US migrants.

Variables	(1) % of later-wave municipality-US migrants in county	(2) % of later-wave municipality-US migrants in county	(3) % of later-wave municipality-US migrants in county	(4) % of later-wave municipality-US migrants in county
	OLS			
First-wave municipality-county outmigration rate	0.0223***	0.0220***	0.0116***	
	(0.0014)	(0.0013)	(0.0008)	
% of first-wave municipality-US migrants in county	-0.0438***	-0.0429***	-0.0173***	
	(0.0039)	(0.0039)	(0.0032)	
Average county income	0.0003***	0.0003***	0.0000	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	0.0070***	0.0069***	0.0031***	-0.0118***
	(0.0005)	(0.0005)	(0.0004)	(0.0037)
Observations	31151786	31151786	31151786	31151786
R-squared	0.176	0.178	0.252	0.701
	IV			
First-wave municipality-county outmigration rate	-0.0016	-0.0017	-0.0014	
	(0.0011)	(0.0011)	(0.0010)	
% of first-wave municipality-US migrants in county	0.1493***	0.1652***	0.1611***	
	(0.0142)	(0.0143)	(0.0219)	
Average county income	-0.0000	0.0001***	0.0001***	
	(0.0000)	(0.0000)	(0.0000)	
Interaction: average county income × % of first-wave migrants in county	-0.0134***	-0.0153***	-0.0150***	-0.0816***
	(0.0017)	(0.0017)	(0.0025)	(0.0154)
Observations	31151786	31151786	31151786	31151786
F-stat. (1^{st} stage)	1059.8	980.0	196.2	612.6
Time FE	X	X	X	
County controls		X	X	
Time FE × Municipality baseline controls		X	X	
County and Municipality FE			X	
Municipality × County FE				X
Municipality × Time FE				X
Time × County FE				X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1890. The dependent variable is the proportion of later-wave migrants from a municipality that settled in a county, in two five-year intervals after 1890. The independent variables are: the municipality-county first-wave migration rate, average county income, the share of first-wave migrants from a municipality to the US that settled in each county, and their interaction. Column 1 includes time fixed effects, Column 2 adds time-varying county characteristics (distance to New York, proportion of men, average age, proportion of black, school attendance rate (ages 6 to 15), literacy rate (ages 10 and older), and proportion of rural population) and municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds county and municipality fixed effects; Column 4 adds two-way interactions between time, county and municipality fixed effects. In the bottom panel, key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Table C.19. 1890 threshold results, Municipality - US model. Later-wave municipality-US outmigration rates.

Variables	(1) Later-wave municipality-US outmigration rate	(2) Later-wave municipality-US outmigration rate	(3) Later-wave municipality-US outmigration rate	(4) Later-wave municipality-US outmigration rate	(5) Later-wave municipality-US outmigration rate	(6) Later-wave municipality-US outmigration rate
			OLS			
First-wave municipality-US outmigration rate	0.7506***	0.7481***		0.7320***	0.7319***	
	(0.0129)	(0.0132)		(0.0148)	(0.0147)	
Mean income of average first-wave migrant	0.0324**	0.0183	2.4354***	-0.0081	-0.0060	1.3371***
	(0.0131)	(0.0121)	(0.2409)	(0.0124)	(0.0120)	(0.1369)
Observations	15704	15704	15704	15704	15704	15704
R-squared	0.892	0.895	0.964	0.913	0.913	0.972
			IV			
First-wave municipality-US outmigration rate	0.7511***	0.7513***		0.7410***	0.7413***	
	(0.0130)	(0.0130)		(0.0150)	(0.0149)	
Mean income of average first-wave migrant	1.2892***	1.1008**	4.5584***	0.2846	0.4540	2.2186***
	(0.4740)	(0.5451)	(0.2407)	(0.2603)	(0.3892)	(0.1581)
Observations	15704	15704	15704	15704	15704	15704
F-stat. (1^{st} stage)	5.5	3.3	3052.1	4.1	2.6	2727.5
Time FE	X	X	X			
Time FE × Munic. baseline controls		X	X		X	X
Municipality FE			X			X
Province × Time FE				X	X	X

Notes: Standard errors clustered at the municipality level in parenthesis. All regressions are weighted by baseline municipality population. The threshold to separate first-wave and later-wave variables is set to 1890. The dependent variable is the municipality-US per-thousand later-wave migration rate, in two five-year intervals after 1890. The main independent variables are: the first-wave migration rate from each municipality to the United States, and a weighted average of county income, with weights equal to the proportion of first-wave migrants in each county. Column 1 includes time fixed effects, Column 2 adds municipality characteristics at baseline (distance to main ports, shares of male and female workers employed in agriculture, male and female employment and literacy rates, and share of male and female population in five age groups) interacted with time fixed effects. Column 3 adds municipality fixed effects. Columns 4-6 repeat the estimations in 1-3 incluiding province-year fixed effects. In the bottom panel, the key independent variables are instrumented with counterfactuals based on the distribution of first-wave migrants from continental Europe (excluding Italy and Germany) across US counties. 1^{st} stage F-statistics are based on the Kleinbergen-Paap Wald statistic. Significance: *** p<0.01, ** p<0.05, * p<0.1.

Figures

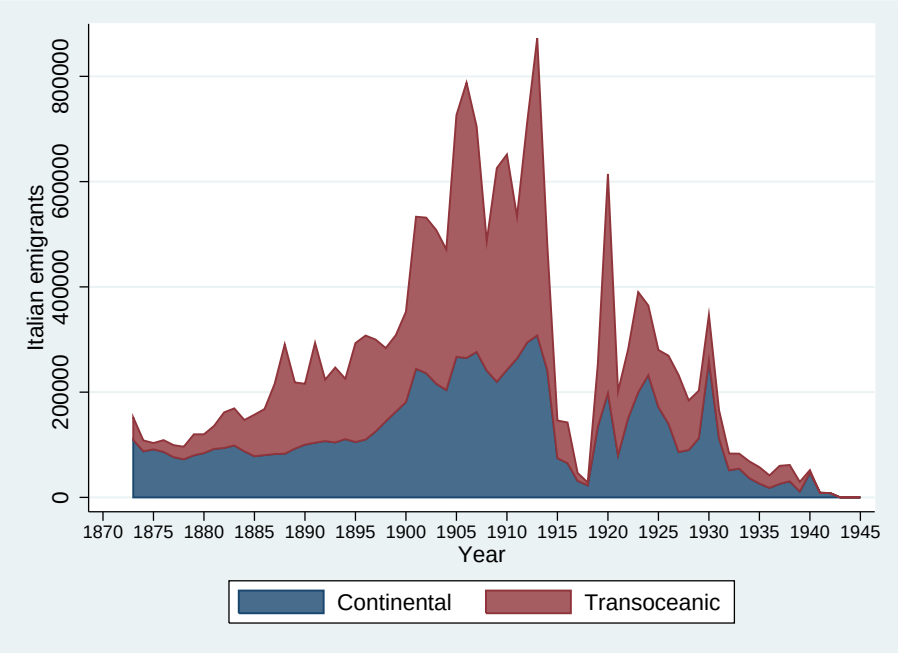


Figure C.1. Italian emigrants by year (1870-1945).

Notes: The figure plots outflows from Italy towards other European countries and to Transoceanic destinations (the Americas, Australia, Asia, etc.). Source: Istituto Nazionale di Statistica (2011).

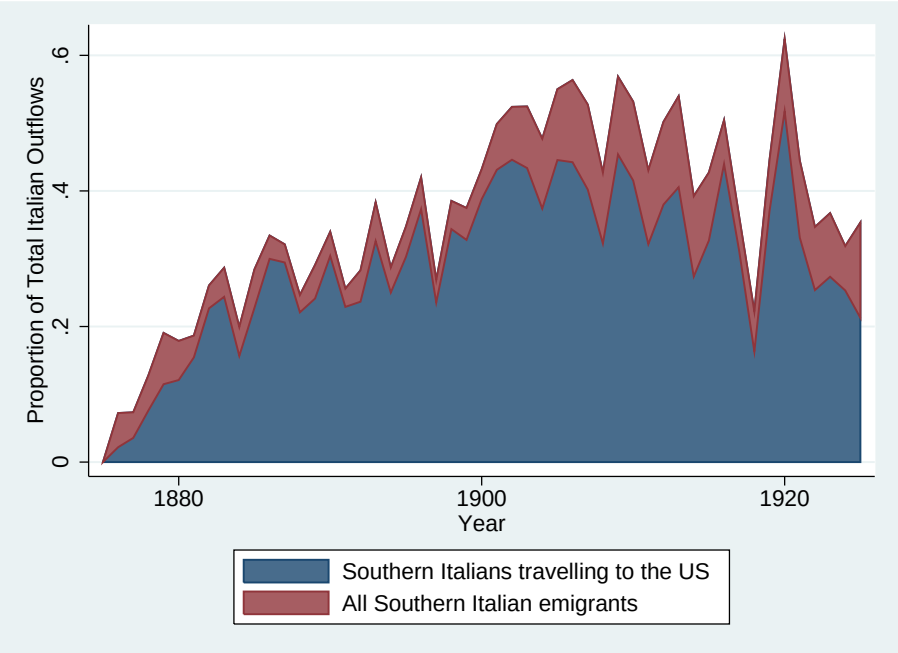


Figure C.2. Total Italian outflows. Participation of southern Italians and southern Italians to the US.

Notes: The figure plots the proportion of total Italian outflows in 1876-1925 that departed from south Italy, and the proportion that departed from south Italy exclusively toward the United States. Source: Commissariato Generale Dell'Emigrazione (1927).

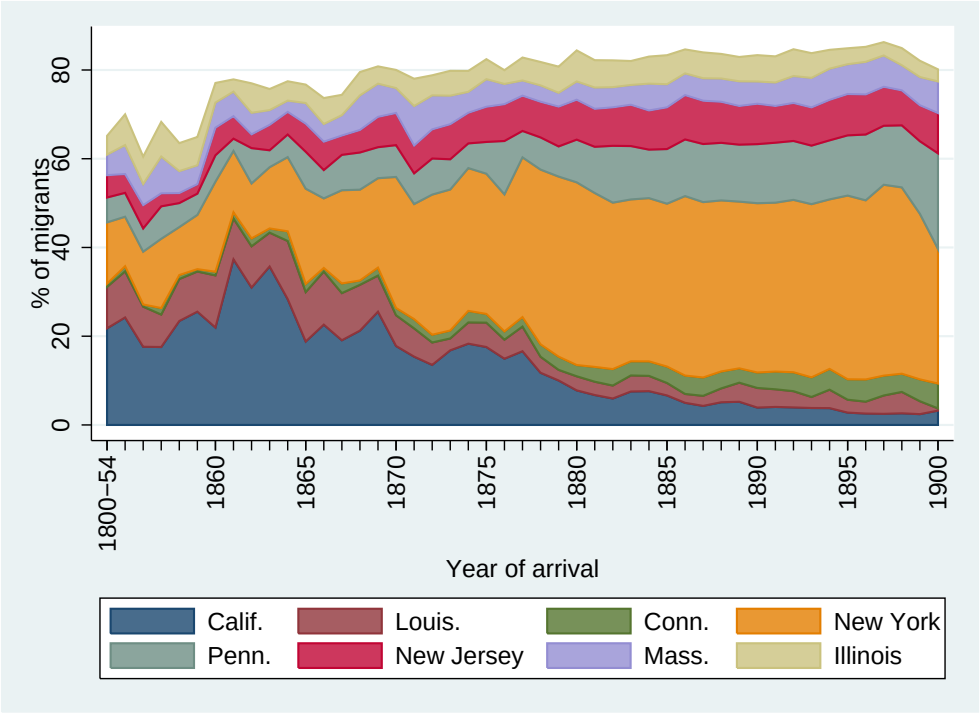


Figure C.3. Distribution of Italian migrants by State of residence and Year of arrival.

Notes: The figure plots the share of Italian arrivals to the United States computed from the Full US Schedules, by year of arrival and State of residence. Arrivals prior to 1854 are presented together for ease of presentation. The figure includes only the main eight receiving states, together accounting for more than 80% of Italian arrivals in the period.

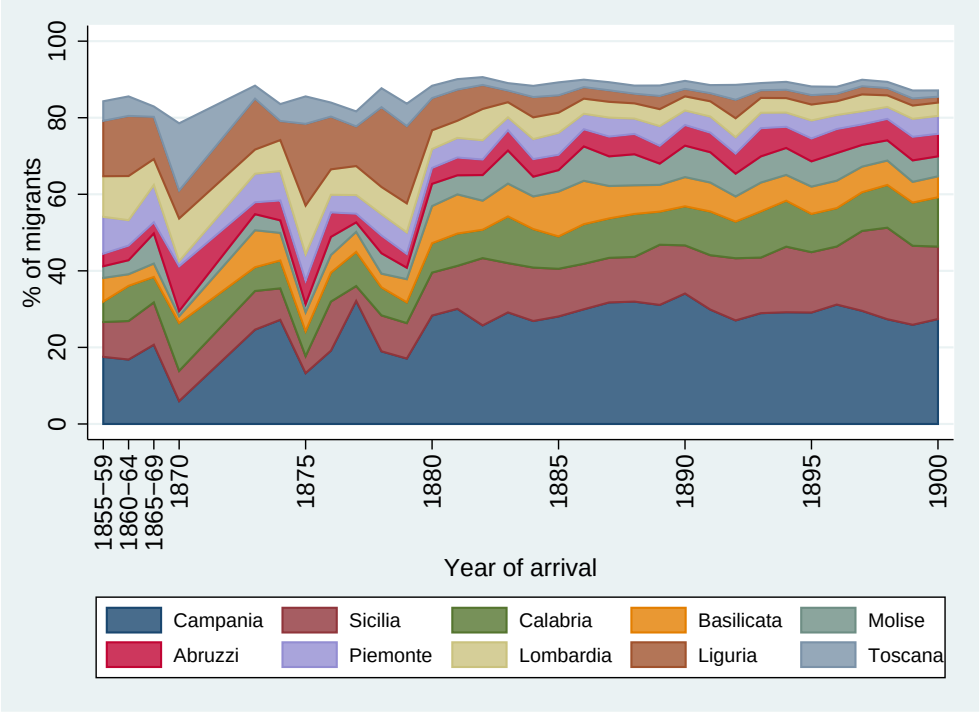


Figure C.4. Distribution of migrants by year of departure and region of origin.

Notes: The figure plots the share of Italian arrivals to the United States computed from ITA data, by year of arrival and Region of origin. Arrivals prior to 1870 are grouped in intervals for ease of presentation. The figure includes only the main ten sending regions, together accounting for more than 85% of Italian emigrants in the period.

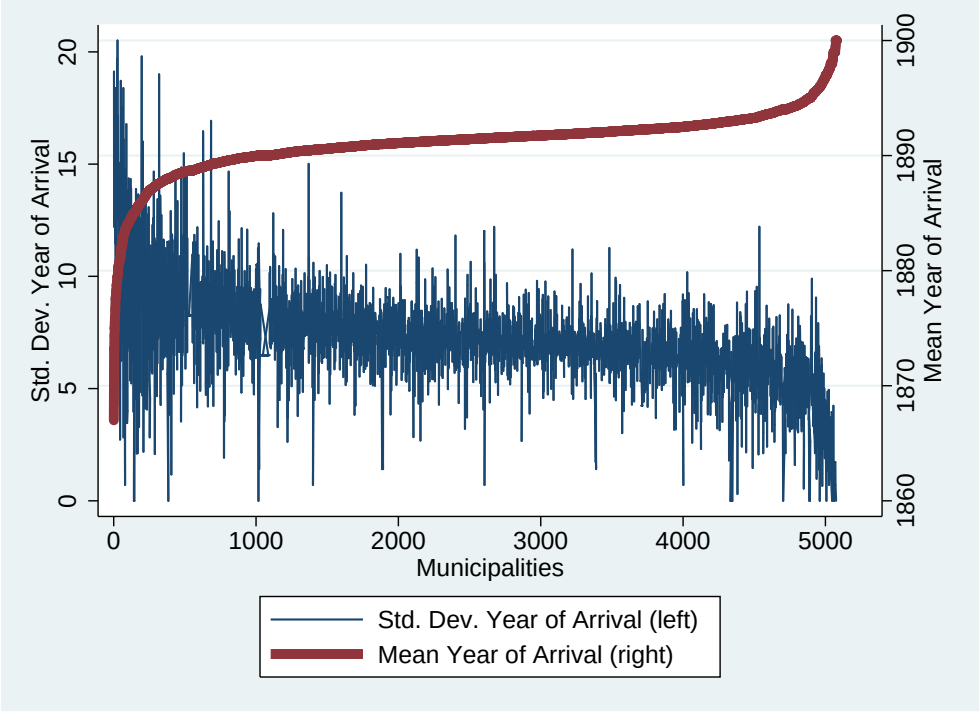


Figure C.5. Average and Standard Deviation of year of arrival by municipality of origin.

Notes: The figure plots average Year of Arrival and Standard Deviation by Municipality of Origin, computed from the sample of matched men in the Full US Schedules.