Like an Ink Blot on Paper

Testing the Diffusion Hypothesis of Mass Migration, Italy 1876–1920

Yannay Spitzer
yannay.spitzer@huji.ac.il
The Hebrew University of Jerusalem and CEPR

Ariell Zimran
ariell.zimran@vanderbilt.edu
Vanderbilt University and NBER

September 30, 2022

Abstract

Why were the poorer countries of the European periphery latecomers to the Age of Mass Migration? We test the *diffusion hypothesis*, which argues that mass emigration was delayed by an initial lack of well developed migration networks, and that the geographic expansion of these networks in a process of spatial diffusion was the main factor that eventually unleashed mass emigration. We propose a model of migration within a spatial network to formalize this hypothesis and to derive its testable predictions. Focusing on post-unification Italy, we construct a comprehensive commune- and district-level panel of emigration data over four decades, and use it to show that the testable predictions of the diffusion hypothesis are validated by the data. Crucially, we show that Italian mass migration to North America began in a few separate *epicenters* from which it expanded over time in an orderly pattern of spatial expansion, and that the epidemiological characteristics of this expansion match those underlying our model. These findings strongly support the diffusion hypothesis, and call for a revision of our understanding of one of the most important features of the Age of Mass Migration—the *delayed migration puzzle*.

Acknowledgments For helpful comments, we thank Ran Abramitzky, Brian A'Hearn, Michael Amior, Levi Boxell, Carlo Ciccarelli, William Collins, Giuseppe De Arcangelis, Giovanni Facchini, Joseph Ferrie, Michela Giorcelli, Walker Hanlon, Timothy Hatton, Joel Mokyr, Federico Nastasi, Cormac Ó Gráda, John Parman, Hannah Postel, Hillel Rapoport, Marco Tabellini, Gaspare Tortorici; seminar participants at Bar Ilan University, Harvard University, the Hebrew University of Jerusalem (the Department of Economics, the Department of Environmental Economics and Management, and the Migration Workshop), the London School of Economics, Northwestern University, Oxford University, Queen's University Belfast, Stanford University, the University of Michigan, the University of California Davis, the University of Nottingham, the University of Warwick, Vanderbilt University (the Department of Economics and the Robert Penn Warren Institute), and the Economics of Migration Senior Migration Seminar; and conference participants at the 2019 Galatina Summer Meetings, the 2020 Allied Social Sciences Association Meetings, the 2020 H2D2 Research Day at the University of Michigan, the 2020 Economic History Association of Israel Meeting, the 2021 Economic History Society Conference, the 2021 Migration and Development Conference, and the 2022 Economic History Association Conference. Christine Chang, Jenny Liu, Jared Katz, Paul Mitalipov, Elizabeth Nelson, Ilan Pargamin, Sarah Robinson, Stephania Stavropoulos, Claire Whittaker, and Danielle Williamson provided excellent research assistance. We are grateful to Peg Zitko and the Statue of Liberty-Ellis Island Foundation for providing the Ellis Island arrivals data. This project was supported by a Vanderbilt University Research Scholar Grant. Work on this paper was completed while Ariell Zimran was a W. Glenn Campbell and Rita Ricardo-Campbell National Fellow and the William C. Bark National Fellow at the Hoover Institution, Stanford University and while he was a Visiting Scholar at the Opportunity and Inclusive Growth Institute at the Federal Reserve Bank of Minneapolis; funding from both institutions is gratefully acknowledged. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or of the Federal Reserve System. This material is based upon work supported by the National Science Foundation under Grant No. SES-1425598. This research was supported by the Israel Science Foundation (grant No. 3121/21). Additional financial support was provided by the Northwestern University Center for Economic History, The Falk Institute for Economic Research, the Hebrew University of Jerusalem, the German-Israeli Foundation, and Vanderbilt University.

1 Introduction

During the *Age of Mass Migration*, tens of millions of Europeans migrated to the New World and many others relocated within Europe (Abramitzky and Boustan 2017; Hatton and Williamson 1998). This movement is understood to have been primarily driven by large gaps in living standards between the sending and receiving countries (O'Rourke and Williamson 1999; Williamson 1995), which resulted in high returns to migration (e.g., Abramitzky, Boustan, and Eriksson 2012). But a number of fundamental patterns within this broad phenomenon are more difficult to explain, and in many cases are inconsistent with the canonical push-pull paradigm of migration (Sjaastad 1962; Todaro 1969). One such pattern is what we refer to as the *delayed migration puzzle*. Despite having the highest real wages in Europe, western European countries, such as Britain and Germany, were the early leaders in transatlantic mass migration, beginning in the 1840s or earlier (Cohn 2009). Southern and eastern European countries, such as Italy, the Austro-Hungarian Empire, and the Russian Empire, although far poorer and thus facing even larger potential gains from migration, lagged behind for several decades.[1] It was not until the 1890s that these countries suddenly surged to dominance, taking the lead in both the absolute number of migrants and in the rate of migration to the United States (Barde, Carter, and Sutch 2006; Ferenczi and Willcox 1929; Hatton and Williamson 1994, 1998).[2] Why were Europeans from the poorest countries latecomers to mass migration? Why did millions of potential migrants forego for decades the opportunity to earn higher wages abroad before suddenly embracing this technology?

This paper tests the *diffusion hypothesis*, which Gould (1980) originally proposed as an explanation for the delayed migration puzzle. In essence, this hypothesis, as we operationalize it, views emigration as sharing the epidemiological properties of an infectious disease.[3] Just as individuals do not contract an infectious disease unless exposed to someone else who has been infected, this hypothesis holds that, regardless of the strength of the incentive to do so, individuals generally cannot emigrate unless they know someone else who has emigrated.[4] Scaled up to the community level, this implies that, even in places with high emigration potential,[5] emigration is limited in the absence of connections to networks of prior migrants. Further scaled up to the national level, this implies that emigration initially emanates from only a few *epicenters*. As

[1] As Hatton and Williamson (1998, p. 56) put it, "the poorest had the most to gain by a move to higher living standards."

[2] Formally, this implies that the correlation between origin-country real wages and emigration was not negative (as the push-pull model predicts) until the 1890s. Cross-sectional real wage differences were also poor predictors of emigration rates within countries (Baines 1995, ch. 4).

[3] To be clear, we do not intend the use of an epidemiological model or analogy to imply any normative judgements regarding migration. It is instead a useful tool to understand the spread of a phenomenon within a population.

[4] In our formal model, we do permit some pioneers, who can migrate without being connected to a prior migrant. The key is that these pioneers are relatively rare.

[5] By high emigration potential, we mean strong incentives for emigration.

individuals from these places emigrate, their contacts in nearby places became connected to friends and relatives abroad, such that neighboring places sequentially connect one another to networks, leading "the contagion of emigration [to] spread over the map much like an ink blot on paper" (Moya 1998, p. 113).[6] The result of such a process in Italy was that regions closer to the epicenters produced mass emigration early on, whereas regions farther away, many of which faced similar push factors and thus had similar potential for emigration, had to wait, sometimes for several decades, before their potential was abruptly unleashed when they were reached by the expanding networks of migration. In sum, according to the diffusion hypothesis, the main cause for the delay of mass emigration from the European periphery was the initial absence of links to migration networks in otherwise emigration-prone places and the time that it took for these networks to expand over space. The gradual rise in migration to the United States, which appeared to cross the threshold of mass emigration in the 1890s, was, in fact, an accumulation of many sequential rapid surges of local mass emigration that, by the 1890s, had spread to reach a sufficiently large part of Italy.

Some social historians have followed Gould (1980) in considering the diffusion hypothesis to be an important part of the explanation for the evolution of mass migration (e.g., Baines 1995; Lowell 1987; Moya 1998). In economic history, however, Hatton and Williamson's (1998) canonical study of the Age of Mass Migration concluded that the diffusion hypothesis was disproved, and the question of what was responsible for the delayed migration puzzle has largely been dormant since. The dominant explanation, originally due to Thistlethwaite (1960 [1991]) and supported by Hatton and Williamson (1998), is instead the *modernization hypothesis*, according to which a lack of economic development delayed migration, and its eventual onset triggered it by releasing poverty traps, loosening connections to the land, and increasing demographic pressures. The delayed migration puzzle is thus explained by the earlier onset of modernization in northwestern Europe than in the southern and eastern periphery.[7] Networks were important, but their absence was no more than a short-term impediment—where internal conditions were well suited for emigration,[8] some pioneers led the way within a few years, causing migration networks to evolve spontaneously. Returning to the epidemiological analogy introduced above, the modernization hypothesis views migration as sharing the epidemiology of cancer. Places characterized by high risk factors will spontaneously develop high prevalence, and the timing will not depend on proximity to already infected places (though spatial correlation in the

[6] Gould (1980, p. 283) used a similar analogy: "One might describe this process as one of 'diffusion,' at least in the mechanical sense in which a drop of ink on a small piece of blotting paper gradually 'diffuses' over the whole area."

[7] As Hatton and Williamson (1998, p. 46) put it, "mass emigration in Europe had to await the forces of industrialization at home and a glut in the mobile age cohort driven by a demographic transition that industrialization produced."

[8] Internal conditions include both factors incentivizing emigration and local factors that may have been responsible for restraining emigration from places where the incentives for migration were strong. We think of these all collectively as push and pull factors.

underlying conditions might make it appear so). Because it focuses on factors internal to a given location rather than interactions *between* locations, the modernization hypothesis is an *internalist* explanation.

In this paper, we revisit the diffusion hypothesis. Strong indications of the validity of this hypothesis are given by maps documenting the evolution of the geographic origins of the Italian emigration to North America.[9] The maps in Figure 1 report the district-level rates of North America-bound emigration by half decade from the late 1870s.[10] Commune-level maps do the same at a finer level of geography (Figure 2), but with data starting only in 1884. The visual evidence appears to be consistent with the diffusion hypothesis. Migration rates were initially very high in a small number of epicenters. Over time, migration spread in a seemingly orderly, spatially consecutive, manner to the rest of the country. Many regions that had at first produced no migration at all turned out to be enthusiastic participants once this movement reached them; conversely, by the end of the Age of Mass Migration, the initial epicenters no longer stood out in terms of their emigration rates. Our analysis builds on this preliminary evidence by introducing two innovations relative to prior tests of the diffusion hypothesis. This paper is the first to develop an explicit theoretical framework from which we derive testable predictions of the diffusion hypothesis. It is also the first to bring to bear a sufficiently detailed dataset to document diffusion and to test the diffusion hypothesis's predictions. Based on our findings, which largely conform to the predictions of the diffusion hypothesis, our conclusion is that the diffusion hypothesis is the best and most parsimonious explanation for the temporal and spatial patterns of Italian migration, including its delayed beginnings and eventual surge.

The first step in our analysis is to formalize the diffusion hypothesis, proposing a model that combines a simple push-pull model of migration with an underlying epidemiological model of diffusion over a spatial network. This model treats migration as a technology that becomes available to individuals once a person to whom they are linked has migrated. Once they gain the option of emigration, individuals can then choose whether or not to move based on typical push and pull factors, such as income differences or demographic pressures. If they migrate, their connections subsequently gain the option of migration. We explicitly distinguish between intra- and inter-communal diffusion, where the latter is modeled as migration options generated by friends and relatives in neighboring communes. *Pioneers*—individuals who migrate without being linked to a prior migrant—are allowed by this framework, as are personal contacts spanning long distances. When both of these factors are sufficiently rare, our model can generate a persistent delay in emigration from emigration-prone but unlinked regions, and a staggered and spatially ordered entry into mass migration of otherwise similar areas. Conversely, the more frequent are pioneers and long-distance

[9] The construction of these maps is discussed below.
[10] By district, we refer to the Italian *cirdondario* or *distretto*.

social links, the less likely is the evolution of migration to follow a spatial diffusion process; at the limit, the model collapses to a standard push-pull one. Explanations for delayed emigration from emigration-prone regions in this case must rely on internal factors, as the modernization hypothesis does.

When pioneers and long-distance contacts are sufficiently rare, the model makes four testable predictions.

1. *Convergence.* As the technology of emigration spreads across the country, a major source of variation in emigration is removed; over time, the variation diminishes and at the limit approaches the level explained only by variation in internal characteristics. Migration rates across the country thus exhibit σ-convergence. When areas are newly exposed to migration networks, they release a pent-up demand for migration, leading to β-convergence as these areas experience greater growth in emigration than in already-linked areas.

2. *S-shaped local time trends.* At the local level, migration rates should evolve along an S-shaped curve, initially low, then rapidly increasing once networks arrive, and finally reaching a state of saturation in which they plateau around the level determined by push and pull factors.

3. *Correlated destinations.* Since networks are both destination-specific and likely to be shared by neighboring places, the correlation between the destination choices of migrants from any two places in the country of origin should diminish with respect to the distance between them. As networks become more widely diffused over time, the distribution of destinations should become more similar conditional on distance between places.

4. *Spatial expansion and frontier effect.* Mass emigration should be observed initially in only a small number of places and should expand from there in spatial order. This implies the *frontier effect*: in every period, a place's likelihood of entering mass emigration should be negatively related to its distance from the nearest emigration *frontier*—the contour of places that experienced mass emigration in the previous period.

In the second step of our analysis, we test these predictions in the context of Italian emigration during the period 1876–1920, which Gould's (1980) seminal paper cites as a paradigmatic case of mass migration governed by the spatial diffusion of chain migration networks. We focus primarily on Italian migration to North America because its early stages occurred late enough to be captured by our data.[11] We can observe

[11]The other two main streams of emigration were to other countries in Europe and to South America. Emigration to Europe was often characterized by extremely high rates of repeat migration that disproportionately inflate the migration counts, potentially more so in places closer to the border. Emigration to South America started earlier, and so we can only observe the later stages of its development. We also provide analyses in which we study emigration to all destinations, with similar findings to those for emigration to North America.

this movement from its early stages in the 1870s through its surge in the 1900s and peak in the years leading to World War I. Italy constitutes an ideal laboratory in which to test for the existence of a diffusion process. It was sufficiently large and had enough internal variation to enable us to observe diffusion processes evolving gradually within its borders. Moreover, Italy was among the largest migrant source countries, sending 14.3 million migrants during our study period (relative to a 1901 population of about 32.5 million), with nearly 5 million of them headed to the United States. Italian emigration statistics that we digitized for this project document this movement in fine detail, and form the basis of our panel dataset of emigration rates at the district and commune level, covering more than 8,000 communes in over 280 districts with data for 28 consistently defined destinations over 40 years. Compounded by the wealth of commune- and district-level statistics that we collected, we are able, for the first time, to characterize the spatial evolution of Italian emigration and to test the diffusion hypothesis at a time and place in which it is suspected to have operated.

We find that the four testable predictions of our model are borne out in the data. In support of the convergence prediction, we find that the coefficient of variation in commune-level emigration rates to North America fell to less than half its initial value between 1890 and 1914—evidence of σ-convergence. We also find that this pattern was driven by laggards catching up with leaders in a pattern of β-convergence—communes in the bottom quartile of pre-1900 emigration rates to North America experienced a 50-times greater increase in average annual emigration rates after 1900 than communes in the top quartile of pre-1900 emigration rates. Moreover, local time series of emigration followed an S-shaped trend. The average commune, regardless of when it first entered mass migration, experienced an increase in its migration rate from essentially zero to about 25 per thousand over a period of 20 years, after which its emigration rate plateaued. In support of the correlated-destinations prediction, we find that a one-standard deviation increase in distance between two provinces was associated with an increase in the dissimilarity index of their destinations of 0.104, or just under half of a standard deviation, and that this dissimilarity declined by about 0.15, or about 22 percent, from 1886 to 1914. Finally, we show clear visual evidence of spatial expansion of the area of mass emigration. In support of the frontier effect, we find that an increase in distance from the previous half-decade's frontier of mass migration from 0 to 150 km (about one standard deviation) was associated with a decline in the average emigration rate to the United States from just under 5 per thousand to about 1 per thousand. And we find that a one-standard deviation increase in distance from epicenters was associated with a decline in the hazard of entering mass migration of about 40 percent.

The evidence on the spatial expansion of mass migration and the frontier effect is particularly important in distinguishing between the diffusion hypothesis and internalist explanations for the delayed migration puzzle,

such as the modernization hypothesis. According to the diffusion hypothesis, places infect their neighbors in emigration, such that emigration in one place can trigger subsequent emigration in its neighboring places through a *spatial contagion mechanism*. According to internalist views of migration, conditional on all relevant characteristics, emigration from one place does not predict subsequent emigration nearby. The existence of spatial expansion and a frontier effect is thus strong evidence in favor of the diffusion hypothesis. Nonetheless, such patterns can arise even if network connections were not the primary determinants of the timing of mass migration. Specifically, it could be that neighboring places shared unobserved characteristics that affected the timing of the onset of mass migration or the magnitude thereof; or that a time-trend in such characteristics was correlated among neighbors; or even that there did exist a process of spatial diffusion moving across the country, but that the diffusing characteristic was not personal links to previous migrants but some other migration-inducing characteristic, such as local policies (Andrews and Seguin 2015), industrialization (Franck and Galor 2022), or demographic trends (in the spirit of Spolaore and Wacziarg 2009). That is, the question is whether or not the spatial expansion of mass migration was the product of one place's migration *causing* higher migration among its neighbors. Importantly, we find no evidence for the spread of any confounding characteristic in Italy at this time, and it seems *a priori* unlikely that some sort of unobservable evolved in such a sharply cascading spatial way.

Nonetheless, in the third and final step of our analysis, we attempt to build stronger evidence of the existence of the spatial contagion mechanism. We differentiate between the two explanations for spatial expansion and the frontier effect by developing a novel instrumental variables approach by which to estimate a spatial lag model of migration. This approach uses plausibly exogenous variation in the spatial distribution of the population, in conjunction with the (potentially endogenous) strong observed negative correlation between distance from epicenters and emigration. Intuitively, we compare two identical places equidistant from an epicenter of emigration. The only difference between them is that the neighboring population around the first commune is on average closer to the epicenter relative to the second. Under the identifying assumption, the underlying incentives for migration are thus similar. However, because more of the first place's neighbors have been exposed to migration by being closer to the source of the "infection" than in the second, the spatial contagion mechanism predicts higher emigration from the first. This source of variation in exposure to neighbors' migration is exogenous under the assumption that the *orientation* of the spatial distribution of neighbors is independent of its characteristics (i.e., whether they are on average closer or farther from the epicenter), and we use it to identify the causal effect of neighbors' lagged migration on a place's own migration. An analogous analysis based on distance from the frontier of mass migration yields

similar results. Our preferred specifications yield an estimated elasticity of own emigration with respect to the emigration of nearby communes of 0.54–0.60, thus confirming the existence of spatial contagion—a key component of the diffusion hypothesis.

Throughout our analysis, we also identify several patterns that are incongruous with the modernization hypothesis, which is also capable of rationalizing delayed entry into migration by emigration-prone regions and of generating some of the patterns that we document in our analysis. We find no evidence that the areas that led Italian emigration were leaders in modernization; indeed our analysis are all robust to controlling for various measures of local development. Our evidence of the causality of the spatial contagion mechanism is also inconsistent with the modernization hypothesis. We also find that the diffusion processes were specific to the destination, which is inconsistent with the notion that the spatial diffusion of emigration is a result of diffusion of internal factors that cause emigration. Finally, we find evidence of saturation—that time series were S-shaped rather than monotonically increasing, that the initial leaders in emigration no longer stood out by the end of the Age of Mass Migration, and that initial emigration leaders did not experience declining emigration rates. Saturation requires that the factor unleashing emigration not affect the level of emigration, which is implausible if that factor is modernization.

The accumulated evidence from our analysis leads us to the view the diffusion hypothesis as the best available explanation for the set of stylized facts produced by the Italian emigration. Although some of the predictions of our model that we test and verify can plausibly be generated by alternative explanations, the diffusion hypothesis is the most parsimonious theory that can explain them all and is not contradicted by some of them. Our conclusion is therefore that the rising magnitude of emigration from Italy and the evolution over time of its geographic origins was primarily governed by a process of spatial diffusion. Leaving the question of initial conditions aside and taking the 1870s as the starting point, the fact that Italy took two or three more decades to reach mass emigration at the national level can be largely accounted for by the lack of migration networks throughout most of the country in the 1870s.

Most directly, this paper contributes to the literature seeking to explain the delayed migration puzzle. Recent evidence on Jewish emigration from the Pale of Settlement is strongly suggestive of a diffusion process being the primary determinant of its macro trends (Spitzer 2021). We provide the first formalization of the diffusion hypothesis, derive its testable predictions, and carry out its first comprehensive empirical test. Our findings revise our knowledge of the geographic evolution of mass emigration from the European periphery and position the diffusion hypothesis as a strong and plausible rival explanation to the one accepted thus far in the literature.

For several reasons we cannot claim to have fully resolved the delayed migration puzzle. First, we do not explain *why* certain individuals were pioneers and certain places were epicenters. While testing the diffusion hypothesis, we take it as given that some locations and individuals had taken initial leadership and that their selection may not have been random. Moreover, in the absence of a cross-country comparative study we cannot assess whether a similar diffusion process took place earlier in countries that led migration—in which case the ultimate question is as to why the process began later in the periphery—or whether the evolution of emigration in countries that entered mass migration earlier was altogether different—in which case the question is as to why was it so. In particular, we do not test whether there was a continent-wide process of diffusion, in which Italy had to wait until foreign networks percolated through its borders; in fact, some of the Italian epicenters were located far away from its land borders, which suggests that their emergence was either spontaneous or triggered by maritime trade contacts or by rare long-range relations between individuals. In other words, we believe that such events were rare enough such that most Italian regions had to wait for the networks to diffuse, but not so rare as to cause Italy as a whole to wait for cross-border diffusion. What we do believe the evidence tells is that the diffusion hypothesis explains the main patterns of the evolution of Italian mass migration, while leaving aside the question of the timing and the selection of epicenters; in particular, it explains why this movement took several decades to build up before becoming one of the greatest migration flows in modern history. By providing the first evidence supporting the diffusion hypothesis, we show that it is a plausible explanation for the broader delayed migration puzzle.

This paper also offers a more general lesson to the economics of migration. It is well understood that liquidity constraints pose a significant impediment to migration from developing economies, and often the conclusion is that extreme poverty must be alleviated before mass migration is generated (Burchardi, Chaney, and Hassan 2019; Gray, Narciso, and Tortorici 2019; McKenzie and Rapoport 2007). The lesson that we draw from the Italian migration is that *social* liquidity constraints trump financial ones, in the sense that the former are the real bottleneck, and that they can solve the latter. It is possible that the friends and relatives effect may be so strong that it could, by itself, switch a region from little or no migration to extremely high rates of migration within a short period of time and independently of any structural changes, such as poverty alleviation, urbanization, or sectoral shifts.

2 Background

2.1 Italy and the Delayed Migration Puzzle

"[P]ractically all emigration from Italy is primarily due to purely economic causes" (US Congress 1911b, p. 153). This view, expressed in the 1911 report of the Dillingham Commission, reflects the scholarly consensus that Italian mass migration was primarily driven by large and persistent gaps in standards of living between Italy and the destination countries (e.g., Hatton and Williamson 1998). Despite this consensus, no widely accepted theory exists that is capable of explaining either the geographic variation in emigration within Italy or its geographically staggered rollout.

Italy exhibited its own delayed migration puzzle in miniature.[12] Although immigrants to the United States came primarily from the relatively underdeveloped south of Italy, this was not the case for flows to other destinations. More formally, as we show in Figure 3, the correlation between income and emigration to all destinations was positive in the first stages of the Italian migration, only turning to the expected negative sign in the 1890s. Moreover, Italian regions that were seemingly comparable in terms of conditions conducive to mass emigration had widely different timings of its onset.[13] This perplexing fact did not go unnoticed by contemporary observers (US Congress 1911b, p. 164).[14] For example, mass emigration gradually spread south through the western *Mezzogiorno*, from one neighboring province to another: Salerno in Campania in the late 1870s, Cosenza in northern Calabria in the 1880s, Catanzaro and Reggio di Calabria in southern Calabria during the 1890s, and finally Messina, across the strait, around the turn of the century.[15] Real wages were not far apart in these provinces, and at the very least, their ranking was orthogonal to the order in which they entered mass emigration (Federico, Nuvolari, and Vasta 2019). Our general point, applied to this case, is that this 25-year trickle south of emigration across this rather equally poverty-stricken region cannot be explained by underlying economic conditions. This suggests that the *underlying causes* of emigration and the *trigger* that caused the potential for emigration to actually materialize in any given area were separate factors, and that it was a process of spatial diffusion of social networks that sequentially unleashed the latent flow of migration created by these conditions.[16]

[12] For graphs depicting the European delayed migration puzzle, see Online Appendix Figures B.1 and B.2. Online Appendix Figure B.1 shows that US immigration was primarily made up of individuals from Germany, Ireland, and Great Britain until the 1880s, and by 1900 was primarily composed of immigrants from Italy, the Russian Empire and the Austro-Hungarian Empire. Online Appendix Figure B.2 shows that the correlation of emigration and real wages was positive until the 1890s, when emigration from southern and eastern Europe surged.

[13] See Online Appendix Figure B.3 for an example.

[14] Foerster (1919, p. 104) pointed out that "It is significant that emigration should not have originated where misery was greatest."

[15] This geographic progression was described in detail by Foerster (1919, pp. 102–104).

[16] In the words of Foerster (1919, p. 48), "The fact that emigration from Campania was abundant before it became so in

Several explanations for the geographic patterns of emigration in Italy have been proposed. According to MacDonald (1963) and MacDonald and MacDonald (1964), the propensity to emigrate was a result of different constellations of agricultural organization and communal relations.[17] Consistent with the modernization hypothesis, Foerster (1919) suggested that emigration from poorer places was more persistent, but had begun later as a result of liquidity constraints.[18] Italy has also featured as an important case study for validating the modernization hypothesis: while Faini and Venturini (1994) have found statistical evidence consistent with such a role of liquidity constraints in delaying emigration, Hatton and Williamson (1998) found instead that the determinants of migration, both across European countries and across Italian provinces, were real wages, demographic pressures, and the level of employment in agriculture (a negative measure of industrialization). Following Thistlethwaite (1960 [1991]), they concluded that "mass emigration in Europe had to await the forces of industrialization at home and a glut in the mobile age cohort driven by a demographic transition that industrialization produced" (Hatton and Williamson 1998, p. 46).

However, a number of important patterns in Italy and elsewhere in Europe challenge the modernization hypothesis. In general, within-country correlations do not show any systematic correlation, either positive or negative, between economic conditions and emigration in the Age of Mass Migration (Baines 1995). More specifically evidence that mass emigration and economic or demographic modernization emerged in the same places is inconsistent and contested at best. For instance, Jewish emigration from the Pale of Settlement in the Russian Empire began in a few impoverished provinces in the northwest, only later spreading to nearby centers of Polish industrialization, and much later to the relatively well-to-do communities in central and eastern Ukraine (Spitzer 2021). Similarly, Ireland's early leadership in migration predated its industrialization, and there is no evidence of rising demographic pressure there when its emigration first began to surge before the Great Famine (Cohn 2009; Mokyr 1983; Mokyr and Ó Gráda 1982).

Within Italy, the Dillingham Commission noted that industrial and large urban centers tended to produce less emigration (US Congress 1911b, p. 175).[19] The factors highlighted by the modernization hypothesis have also been shown to perform poorly in explaining emigration, both in time-series analysis (Ardeni and Gentili 2014) and when accounting for multiple destinations (Moretti 1999). Moreover, the surge in

Calabria, and that it only as much as ten or fifteen years later assumed large proportions in Sicily, need signify merely that the *occasion which turned a passive into an active cause* arose earlier in one compartment than in another" (emphasis added).

[17] This view, somewhat similar to Hirschman's (1970) *Exit, Voice, and Loyalty*, has gained traction in the socio-historical literature (e.g., Baily 1999; Barton 1975; Silverman 1968; Sturino 1990; Yans-McLaughlin 1977), but came under criticism by Gabaccia (1984a,b, 1988).

[18] "[Emigration] began where there was the chance of saving enough money for passenger fares and has best maintained itself where wages were at a minimum level" (Foerster 1919, p. 104).

[19] "It will be seen that as a rule the heaviest emigration originated in the compartimenti where the proportion of industrial workers was the smallest ... and it is well known that comparatively little Italian emigration originates in the large cities" (US Congress 1911b, p. 175).

emigration from southern Italy around the turn of the century did not correspond to a concurrent wave of industrialization or economic development (Ciccarelli and Fenoaltea 2013; Federico, Nuvolari, and Vasta 2019; Iuzzolino, Pellegrini, and Viesti 2013) in that region. Conversely, the growth of the "industrial triangle" in the northwest, and in particular in the Po Valley (Ciccarelli and Fenoaltea 2013), was not accompanied by a similar emigration surge. Broad demographic trends were also not associated with the evolution of emigration. Northern Italy led the decline in mortality and the increase in life expectancy (Del Panta 1997, p. 10; Vecchi 2011, Table S6), yet some of the early sources of emigration were in the South. Recent attempts to assess the relationship between demographic pressures and emigration have been inconclusive, and lacking statistical power or credible identification (Ardeni and Gentili 2014; Faini and Venturini 1994; Gomellini and Ó Gráda 2013; Hatton and Williamson 1998).

Evidence on the diffusion hypothesis is also scarce and tends to be informal. Gould (1980, Figure 1) highlighted σ-convergence patterns across provinces within regions in Italy, as well as in Hungary and Portugal. He also informally argued that there existed β-convergence and S-shaped sub-national time series in Italy (pp. 282–288). Qualitative evidence similarly indicated diffusion in Scandinavian (Lowell 1987) and Spanish (Moya 1998) emigration. Hatton and Williamson (1998) were the first to provide a formal statistical test of the diffusion hypothesis. They found persistence in the emigration rates of Italian provinces over time (see also Gomellini and Ó Gráda 2013)—evidence that networks were important in determining the size of migratory flows—but failed to find a relationship between literacy—which they viewed as a factor that could have facilitated the spread of information—and emigration rates. As a result, they concluded that diffusion "offers few empirical predictions and says nothing about why emigration rates eventually declined" (p. 15) and that "while such forces [as diffusion] mattered, there is little evidence that persistence or literacy dominated [Italian] provincial emigration rates with anything like the force often assigned to them in the qualitative literature" (p. 121). On the other hand, the case of Jewish emigration from the Russian Empire suggests that diffusion was likely a dominant force in determining emigration rates, as emigration seems to have spread gradually from its single northwestern epicenter towards the east and the south (Spitzer 2021). This movement was also characterized by both σ- and β-convergence.

Prior attempts to test the diffusion hypothesis have been limited by two missing factors—a complete theoretical framework from which to derive the testable predictions of the diffusion hypothesis,[20] and a sufficiently long, rich, and geographically disaggregated panel dataset with which to identify the *inter*-communal transmission of emigration. Therefore, we view the diffusion hypothesis as one that is plausible

[20]This is one of the main contributions that enables us to make an advance over Hatton and Williamson's (1998) operationalization of the diffusion hypothesis as implying persistence and an importance of literacy.

and capable of explaining a fundamental puzzle of the economics of the Age of Mass Migration, but which has yet to be rigorously tested. This paper makes significant advances on both fronts, and thus provides the most comprehensive evidence yet on the validity of the diffusion hypothesis.

2.2 The Role of Networks in Italian Emigration

Is the diffusion hypothesis plausible within the social context of post-unification Italy? Is the existing historical evidence consistent with it? Although the notion that emigration had epidemic-like features was widely recognized by contemporaries,[21] the full implication of the hypothesis—that diffusion was the *primary* determinant of the timing of the onset of mass migration—is a more recent notion, not evident in any contemporary accounts, including Foerster (1919) and the Dillingham Commission Report (US Congress 1911a), arguably the two most comprehensive contemporary inquiries into the causes of Italian emigration.

For such an explanation to be plausible, the social structures that supported emigration must have had certain non-trivial characteristics. They had to be sufficiently strong to support chain migration. They had to be local, yet occasionally crossing community boundaries. And when they did cross community boundaries, they had to reach primarily over short distances, only rarely spanning longer distances. Furthermore, alternative mechanisms that supported migration but did not depend on geographic proximity to previous migration, such as direct recruitment by foreign governments and businesses or poaching by shipping agents, had to be either negligible or themselves dependent on migration networks. Finally, pioneers had to be rare. In what follows, we survey the relevant evidence from the historical literature to evaluate the plausibility of these conditions in Italy during the Age of Mass Migration.

To what extent did Italians engage in chain migration? Much of the debate concerning the sociology of Italian emigration evolved as a reaction to Banfield (1958) and Handlin (1951), who shed doubt on the viability of strong personal and communal relations among south Italian immigrant peasants in the United States, and by implication also on the prospects of strong migrant networks. However, subsequent literature has modified this dismal view of weak social links, showing that both kin- and commune-based ties played an important and constructive role during and after migration (Bell 1979; Briggs 1978; Gabaccia 1984b; Nelli 1967; Vecoli 1964; Yans-McLaughlin 1977). The Italian migration to the United States is portrayed by such studies as being dominated by characteristics of chain migration: early migrants provided funding, information, accommodations, assistance in the labor market, and close examples of successful migration to their friends and kin, who in turn would do the same for theirs (Baily 1999; Cinel 1982; MacDonald and

[21] The usage of metaphors such as "migration fever" prevailed in virtually every sending country (Moya 1998, pp. 95–96).

MacDonald 1964; Sturino 1990). As one immigrant put it, "Immigrants almost always came to join others who had preceded them—a husband, or a father, or an uncle, or a friend" (quoted in Yans-McLaughlin 1977, p. 59). This assertion is supported by recent empirical analysis (Spitzer and Zimran 2018).[22] The importance of chain migration is also made apparent by the ubiquity of town-to-town migration—the specialization of specific towns or small regions in Italy in migration to specific towns in the United States.[23] This pattern was also noticed by contemporary observers, such as the Dillingham Commission, which particularly emphasized the role that letters and the commonality of return migrants played in enabling migration.[24]

While the important role of social networks in the Italian migration is documented beyond doubt in the historical literature, the diffusion hypothesis crucially depends on one particular feature of these networks—that they spread gradually across communes. For this, there had to exist some (though not necessarily many) short-distance contacts across communes, while long distance contacts had to be scarcer or weaker. What historical evidence exists supporting the existence of such contacts? Small-region networks were documented among immigrants in Cleveland (Barton 1975) and among immigrants in Chicago from the Calabrian Rende region (Sturino 1990). Similarly, studies of many smaller US cities found small-region clusters of Italian settlement,[25] and evidence of Italian organizations divided along provincial lines.[26] Weaker evidence to the same effect is the tendency of Italian American communities within the great metropolitan centers, such as New York, Chicago, and Toronto, to cluster by small areas of origin, thus forming "many Little Italies" (Baily 1999; Nelli 1967; Park and Miller 1921; Sturino 1990; Vecoli 1983; Zucchi 1985). In the case study of Antonio Squadrito (Online Appendix C), his followers included residents from four or five different neighboring localities. Outside of the literature on emigration, Lecce, Ogliari, and Orlando (2022) show that social contacts across nearby Italian towns existed in the context of trade, marriage, and linguistic ties.

Were there prevalent alternatives to chain migration that were independent of geographic proximity to previous migrants? Recruited migrant labor was another method on which some Italians relied in their migration to the United States, in particular under the *padrone* system (Iorizzo 1966; Koren 1897; Nelli

[22] In a sample of 31,476 adult Italian passengers arriving at Ellis Island between 1907 and 1925, 33 percent of all males and 72 percent of all females reported joining an immediate family member already present in the United States. Almost all of the rest named other relatives and friends, such that the share of passengers not reporting any contact in the United States was only 5 percent (Spitzer and Zimran 2018, Table A.1). In fiscal years 1908–1910, only 5.9 percent of North Italian and 1.1 percent of South Italian immigrants to the United States did not report joining either a friend or a relative (US Congress 1911c, p. 363, Table 40).

[23] For examples see cases listed by MacDonald and MacDonald (1964, Appendix II) and Cinel (1982, p. 28).

[24] All of these features of chain migration are clearly illustrated in the case study of Antonio Squadrito (Online Appendix C), an early migrant from the Sicilian town of Gualtieri-Sicamino, which, within less than a decade, was followed by "more than one tenth of the population" (Brandenburg 1904, p. 109).

[25] For example, in Buffalo (Yans-McLaughlin 1975, pp. 25–26), St. Louis (Mormino 1986 [2002]), Tampa and Ybor City (Pizzo 1981, pp. 128–130), and Pittsburgh (Bodnar, Simon, and Weber 1982, p. 47).

[26] For example, in San Francisco (Cinel 1982) and Buffalo (Yans-McLaughlin 1975, p. 125).

1964; Peck 2000). However, it was not altogether disconnected from social networks; instead, it depended on them.[27] Even as some agents recruited workers from across Italy, "the emigrant relied on his townspeople to get in touch with the network of agencies and sub-agencies which eventually would lead to a job and cash" (Zucchi 1985, p. 121).[28] Some governments, such as Argentina and Brazil, and later Australia, New Zealand, and certain Canadian provinces, had policies of assistance and subsidies for immigrants (Baines 1995; Kelley and Trebilcock 1998). But migration subsidies were ultimately banned in Italy by the 1902 Prinetti Decree (Baily 1999; Foerster 1919; Gould 1980), and even before that assisted migration was a rarity, particularly among US-bound immigrants. When assisted migration did exist, it was rarely independent of social networks (US Congress 1911b, pp. 61–64). In brief, insofar as overseas emigration was facilitated by such alternatives to chain migration networks, there is little evidence that they were capable of inducing the migration of Italians who were not yet part of these networks. The alternatives were not substitutes but complements to chain migration.

3 Theoretical Model

Our model describes a world in which the diffusion of chain migration networks over space may or may not have been an important determinant of the timing of mass migration. We are then able to determine the model's testable predictions under parameterizations that make network diffusion the primary determinant of the timing of mass migration. These predictions form the basis of our empirical analysis.

As in the traditional push-pull framework (Sjaastad 1962; Todaro 1969), individuals' incentives for migration in our model are determined by push and pull factors, such as real wage gaps between the origin and the destination or factors espoused by the modernization hypothesis. Our main departure from the push-pull paradigm is that we nest the decision of whether or not to migrate within a Susceptible-Infectious-Recovered (SIR)-like epidemiological model, with an underlying geographic network structure and an explicit role for inter-communal transmission.[29]

[27]The recruiting padrone had sub-agents who would travel back to "collect a work force in their home town in Italy" (MacDonald and MacDonald 1964, p. 86), and he "kept his paesani [fellow townsmen] together" (MacDonald and MacDonald 1964, p. 86). The padrone banker was "generally a paesano" (Foerster 1919, p. 391), and the US-based labor boss was "an extension of the informal networks" (Baily 1999, p. 98). The Dillingham Commission agreed with this assessment: "actual and direct contract-labor agreements cannot be considered as the direct or immediate cause of any considerable portion of the European emigration ... immigrants, or at least newly arrived immigrants, are substantially the agencies which keep the American labor market supplied with unskilled laborers from Europe. ... as a rule, each immigrant simply informs his nearest friends that employment can be had and advises them to come. It is these personal appeals which, more than all other agencies, promote and regulate the tide of European emigration to America" (US Congress 1911b, p. 61).

[28]Such was the case of four boys whose departure was assisted by Antonio Squadrito (Online Appendix C).

[29]The SIR model is originally due to Bernoulli (1776) and Kermack and McKendrick (1927), and has been applied in economics by Burnside, Eichenbaum, and Rebelo (2016) and Eichenbaum, Rebelo, and Trabant (2021), among others.

3.1 Basic Setup

Individuals in our model may be in one of three states. They begin as *unlinked*. These individuals are not able to migrate regardless of the incentive to do so. Eventually, these individuals may switch to being *linked*. This switch can occur in two ways. First, an individual switches from unlinked to linked when one of his contacts migrates.[30] Alternatively, the switch from unlinked to linked can occur spontaneously. Linked individuals have access to the migration technology, and being linked is a *necessary* condition for migration. Every period, linked individuals make a choice of whether or not to migrate based on push and pull factors. If they migrate, they become a *migrated* individual and their unlinked contacts become linked. If the migrated individual had become linked spontaneously (rather than through the emigration of one of his contacts), he is a *pioneer*. Individuals have both *intra-communal connections* to other individuals in their same commune and *inter-communal connections* to individuals in other communes. For simplicity, the following discussion will focus on the case of a single destination. When there are multiple destinations, the progress of individuals from unlinked to linked to migrated is separate for each destination and individuals linked to more than one destination decide whether to migrate to one of them or to remain in the origin.

The main state variables of the model for commune i in period t are

$$\mathfrak{S}_{it} = \{U_{it}, L_{it}, M_{it}, N_{it}\},$$

where U_{it}, L_{it}, and M_{it} denote the share of individuals within the commune who belong to each of the three states (unlinked, linked, and migrated). The variable N_{it} is a measure of the exposure of commune i to emigrants in all other communes. It can be thought of as the probability that any out-of-commune contact of an individual in commune i is a migrated person. It takes the form

$$N_{it} = \frac{\sum_{j \neq i} M_{jt} P_j d_{ij}^{\pi}}{\sum_{j \neq i} P_j d_{ij}^{\pi}}, \tag{1}$$

where P_j is the population of commune j, d_{ij} is the distance between communes i and j, and $\pi < 0$ is the rate at which the likelihood that an individual in commune i has a contact in commune j decays with

[30] This switch can capture prior migrants providing material support to potential migrants and the provision of information by prior migrants to potential migrants, among others. All are consistent with the spirit of our model, which requires only that an individual's contacts' migration somehow enable his own. These mechanisms are indistinguishable in our data and we are agnostic as to which one of them carried more weight.

distance. By definition

$$U_{it} + L_{it} + M_{it} = 1$$

$$U_{it}, L_{it}, M_{it}, N_{it} \in [0, 1].$$

The set of main parameters of the model is

$$\Theta = \{\lambda, \delta, \alpha, \pi\},$$

where $\lambda > 0$ is the number of individuals in the same commune to which each individual is connected, which determines the rate of intra-communal transmission; $\delta > 0$ is the number of individuals in other communes to which each individual is linked, which determines the rate of inter-communal transmission; and $\alpha > 0$ is the rate at which individuals spontaneously gain the option to emigrate, which governs the prevalence of potential pioneers.

Let m_{it} denote the probability that a linked individual from commune i chooses to migrate in period t; that is, $e_{it} = m_{it} \times L_{it}$ is the rate of emigration out of the total population. The variation across communes in the probability m_{it} reflects the underlying variation in local incentives for migration, which are distinct from linkage status. The probability m_{it} also reflects any hindrance to emigration from factors internal to commune i, such as (generally speaking) a lack of economic modernization. We assume that in each period, the timeline is as follows. First, individuals who were linked in the previous period decide whether or not to migrate; then, new links are created, caused by individuals who emigrated in the first part of the period (or spontaneous generation). The implied laws of motion for the state variables are then

$$\Delta M_{it} = m_{it} L_{it},$$

which is the change in the fraction of the population that has already migrated. For the fraction linked, the law of motion is

$$\Delta L_{it} = -m_{it} L_{it-1} + [1 - (1 - \alpha)(1 - \lambda \Delta M_{it})(1 - \delta \Delta N_{it})] U_{it};$$

that is, those who migrate are lost from among the linked, and then new linked individuals are created, either spontaneously, from linkages to newly migrating individuals in the commune, or from linkages to newly migrating individuals in other communes. Finally, the fraction of individuals exiting the susceptible

state is the rate of those who become linked:

$$\Delta U_{it} = -[1 - (1 - \alpha)(1 - \lambda \Delta M_{it})(1 - \delta \Delta N_{it})]U_{it}.$$

3.2 Discussion

The diagram in Figure 4 shows a hypothetical chain of events that illustrates the main concepts of the model. There are three communes, A, B, and C. The first individual to migrate was a_1 from commune A. He was a pioneer, in the sense that he migrated after switching spontaneously from susceptible to linked without contact with a prior migrant. He was connected to two other residents of commune A, a_2 and a_3, and his migration converted them from unlinked to linked. This is a case of *intra-communal diffusion* of the migration technology. Eventually, a_2 and a_3 also decided to migrate, converting four more unlinked individuals in commune A to being linked. As the process proceeds, commune A is likely to quickly become *saturated*, in the sense that all individuals would become either linked or will have already migrated, and there would be no more unlinked individuals. At this point, commune A's migration rate is determined solely by push and pull factors and not by the rate at which the migration technology diffuses. Commune A is an *epicenter*, since migration was already common there before arriving in its neighboring communes.

The migration of individual a_3 also linked b_1, an out-of-town contact in neighboring commune B. This is a case of *inter-communal diffusion* of the migration technology, which caused a *spatial contagion* of migration from commune A to commune B. If individual b_1 were eventually to migrate, commune B would likely advance towards saturation with some time lag relative to commune A, transmit the migration technology to its neighboring communes, and so on. Commune C, on the other hand, is further from A, and before receiving the migration technology through an inter-communal linkage, one of its residents, individual c_1 spontaneously gained the option to migrate. If he migrates, he becomes a pioneer and is likely to start a new chain of migration spreading from commune C.

This model is similar to the traditional way that international migration has been modeled (e.g., Hatton and Williamson 1998; McKenzie and Rapoport 2010) in that the fundamental incentives are the same—migrants are driven by push and pull factors—captured by the migration probability m_{it}—such as real wage gaps between origin and destination. In addition, in emphasizing the important role of chain migration in determining the size of migratory flows, our framework shares common ground with the standard analysis of migration, but with two notable differences. First, our conceptualization of the friends and relatives effect at the micro level is different. Our model views a network connection to be a necessary condition for migration,

whereas the standard model views it as simply a cost shifter. But at the aggregate level this difference is largely immaterial, or at most a matter of a different arbitrary choice of functional form: in both models, current aggregate migration from a commune is some function of past migration. More importantly, our model allows the inter-communal diffusion of migrant networks. This feature, which enables the formalization of spatial diffusion, is absent from the standard framework.

Under different parameterizations, our model can capture both diffusionist and internalist explanations for the timing of the onset of mass emigration. Three key parameters distinguish between these two different types of dynamics. To the extent that pioneers are rare (α is small), out-of-commune contacts are frequent (δ is large), and very remote out-of-commune contacts are rare (π is large), migration will be dominated by a process of spatial diffusion, resembling the epidemiology of infectious disease. In most cases, mass emigration will not take off before the migration technology arrives through short-distance inter-communal diffusion. Even regions in which the incentives for emigration are high and no internal characteristic hinders migration may be prevented from producing mass emigration for a long period of time. Moreover, no change in local push factors is necessary for mass emigration to finally be ignited.

Conversely, when the model does not have the above parameterization, it simply collapses in the limit to a standard push-pull model, resembling the epidemiology of cancer.[31] Under this parameterization, if a place does not become a source of mass emigration, it must be because the internal characteristics of that place are not conducive to emigration—either individuals there lack the incentive to emigrate or some characteristic of the place constrains migration (Hatton and Williamson 1998, p. 39). While networks are important, their absence cannot (at least for long) prevent the realization of emigration from places where the incentives are strong: networks supporting migration will spontaneously be generated wherever local factors are conducive to emigration without waiting for them to arrive through the inter-communal transmission process. Thus, conditional on local characteristics, the timing of the onset of mass emigration in a place is independent of whether or not migration was already present among its neighbors.

[31] When the bulk of the country has achieved saturation, the differences between the two parameterizations are largely eliminated. As a result, our model can also capture a phenomenon in which improvements in standards of living eventually reduce migration—when everyone is linked, the migration decision is based solely on push and pull factors, and smaller wage differences reduce the incentive to emigration. Hatton and Williamson (1998) argue that the typical curve of migration has an inverse-U shape, as emigration eventually eliminates real wage gaps, thus reducing the incentive to emigrate. They reject the diffusion hypothesis for failing to predict the downward-sloping side of the curve (p. 15), which is seen in the cases of German and Scandinavian emigration. But this insight regarding the drivers of emigration at saturation shows that, if indeed real wage gaps are eliminated, then a decline following saturation is perfectly consistent with the diffusion hypothesis. The fact of the matter is that no such elimination of real wage gaps occurred between Italy and the United States before World War I.

3.3 Predictions

When the parameters of the model are such that spatial diffusion is dominant, the model makes several predictions that can be evaluated in the data.[32] To be clear, we do not expect evidence supporting any single prediction to individually validate the diffusion hypothesis, as each of them could potentially be explained by alternative hypotheses (some with more difficulty than others). Our goal is to demonstrate how a number of new stylized and striking facts about the Italian emigration can all be parsimoniously explained by the diffusion hypothesis alone.

Prediction 1 (Convergence)**.** The overall cross-commune variation in the rates of emigration caused by underlying variation in push factors is initially augmented by the variation in access to the emigration technology. As a growing number of communes are infected and reach saturation, the latter source of variation is gradually eliminated, such that the overall variation levels off around a lower rate, reflecting only variations in push factors. This leveling is manifested by a pattern of σ-convergence—cross-commune measures of dispersion of migration rates will decline steadily, until they stabilize when the entire country reaches saturation.[33] Second, this process generates a pattern of β-convergence. Communes that are latecomers to migration due to an initial absence of linkage to prior migrants experience rapidly rising migration rates shortly after linkage, whereas communes that are already saturated have higher rates but little or no growth. With convergence in migration rates coming from laggards catching up, the β-convergence prediction is of a strong negative relationship between past migration rates and future growth in migration. To be clear, the diffusion hypothesis does not predict that all communes will converge to a similar rate of emigration. A significant amount of variation may remain even when all communes are exposed to emigration due to variation in local push factors.[34]

Prediction 2 (S-Shaped Local Trends)**.** Before any individual in a commune is linked, the commune's emigration rate will be zero. Once the first individuals in a commune are exposed, intra-communal diffusion will generate a rapid increase in the emigration rate as individuals become linked, emigrate, and link their connections. Eventually, the commune will reach saturation when nearly everyone is linked, and the rate

[32] Online Appendix D presents the results of model simulations, demonstrating that these predictions are implied by the model.

[33] This prediction was first suggested and assessed by Gould (1980), who measured cross-regional Gini coefficients in Italy, Portugal, and Hungary. The prediction of monotone σ-convergence assumes that migration was already well developed in some places. If we began to observe migration at its inception, we would expect an inverse-U-shaped Kuznets-style curve. In effect, we expect in our case to see the downward-sloping portion of this curve.

[34] Gould (1980, p. 314) points out that "The process of diffusion . . . did not guarantee that pioneer migration would be followed by a mass movement increasing in some predetermined mathematical progression. If the conditions were not propitious: if the income gain was insufficiently large, for example, or the conditions of the migrant community unacceptable in some other way, the pioneer movement would prove still-born."

of emigration will stabilize around a level determined by push factors. When combined, these three phases will create an S-shaped local time series of emigration rates. The steadily and gradually rising trend in national emigration rates was in fact an accumulation of many successive and sharply rising local S-curves that also generated the convergence in Prediction 1. This, too, was an observation linked by Gould (1980) to diffusion, and is also a common prediction of the of the technology-adoption literature (e.g., Bass 1969; Comin, Dmitriev, and Rossi-Hansberg 2012; Jovanovic and Lach 1989).[35]

Prediction 3 (Correlated Destinations)**.** Two neighboring places will typically share the same destination-specific networks due to their proximity. Therefore, they should have similar migration options and a similar distribution of destinations. On the other hand, two distant places are more likely to be part of different networks, potentially leading to different destinations. Moreover, as each destination's network spreads across the country, the set of potential destinations of any two places will become increasingly similar. Therefore, the prediction is that the similarity in the distribution of migration destinations of two places should increase with the proximity between places and increase over time.

Prediction 4 (Spatial Expansion, Frontier Effect, and Spatial Contagion)**.** The mechanism that undergirds the diffusion process is *spatial contagion*—the infection of places by neighboring places that had already contracted emigration. The immediate prediction that follows from it is the *frontier effect*: defining the *frontier* in any given period to be the boundary of an area that has already crossed a certain threshold level of emigration, the probability that a place enters mass emigration in the current period is positively related to its proximity to the frontier in the previous period. This ultimately results in *spatial expansion* of mass emigration, such that, starting from the early sources of mass emigration (the epicenters), successive communes will enter mass emigration in spatial order.

To be clear, spatial expansion and the frontier effect can be rationalized by alternative explanations, and Section 6 is dedicated to showing that it is most likely the product of spatial contagion. However, beyond being a prediction that is consistent with the basic mechanism, the frontier effect in itself is a crucial component of the diffusion hypothesis. Places not already experiencing mass migration must only rarely begin to do so unless they are close enough to places where the migration technology has already arrived. If spatial contagion occurs but it is not strong enough to dominate the evolution of emigration and produce the frontier effect, then emigration does not spread primarily by spatial diffusion, and the diffusion hypothesis

[35]The standard SIR model follows the S with a declining portion of the curve coming from immunity due to prior exposure. In this case, we do not predict such a decline for two reasons. First, the rates of emigration, even where they were the highest, were never high enough to completely deplete the population. Second, the continual entry of individuals into the age cohorts associated with emigration would keep the pool of potential emigrants well stocked.

fails. Therefore, the frontier effect is a necessary condition that must be satisfied.

4 Data

4.1 Sources and Construction

Our main data source is the *Statistica della Emigrazione Italiana per l'Estero*. This series of volumes was published approximately every two years from the 1870s to the 1920s by the Italian *Direzione Generale della Statistica*. We digitized three data panels from this source. The first is a panel of annual emigration counts spanning the period 1884–1920 at the level of the commune (*comune*), of which there were more than 8,300 in Italy.[36] The second is a series of annual emigration counts at the district level (*circondario* or *distretto*), of which there were 284, which enables us to extend our temporal coverage to begin 8 years earlier in 1876. The last is an annual panel (1877–1920) of emigration counts for 28 consistently defined destinations (usually countries) at the level of the province (*provincia*), of which there were 69 in Italy. We focus in most of our analysis on three aggregated main destinations—North America, South America, and Europe, which together comprised 96.8 percent of all Italian emigration during the period 1877–1920. These data are based on contemporary jurisdictional boundaries, which experienced some changes during our study period, as well as in the century since. Online Appendix E describes how we harmonized these data to fit consistently defined geographic units.[37]

The *Statistica della Emigrazione per l'Estero* has previously been used to study Italian emigration by a number of studies (e.g., Ardeni and Gentili 2014; Faini and Venturini 1994; Gould 1980; Hatton and Williamson 1998; Moretti 1999), but none have used data at a level finer than the province. The high resolution of the commune-level Italian emigration data that we collected yields perhaps the most detailed data in terms of geographic disaggregation and temporal coverage available on a migration flow as large, as geographically varied in origin and destination, and as historically important as that from Italy in the late nineteenth and early twentieth centuries.[38] These features are essential to our study of the spatio-temporal expansion of migration and thus to our evaluation of the diffusion hypothesis—the spread of emigration over space simply cannot be observed at a sufficiently fine level with data at the level of the province or higher.

[36] We have data for 8,317 communes, though a lack of population counts in some cases limits our sample of communes with known emigration rates to 8,029.

[37] The publications omit tables for 1879 for the district and province-by-destination data and for 1888 for the province-by-destination data. In 1916 and 1917, there was virtually no transatlantic migration because of World War I, and consequently there were no volumes published for these years.

[38] Karadja and Prawitz (2019) and Lowell (1987) use highly detailed data on emigration from Sweden (Karadja and Prawitz 2019; Lowell 1987)—a country with less than one-sixth of Italy's population. Fernández-Sánchez (2021) uses detailed data from a single region of Spain. Work in progress by Fontana et al. (2021) also uses data from this source.

The emigration counts reported in the *Statistica della Emigrazione Italiana per l'Estero* are based on passports issued to emigrants by the Italian government. Although it provides the most comprehensive data available on Italian migration, there are some known issues with this source, such as inaccurate reporting by the mayors (*sindaci*) of the Italian communes.[39] The most concerning issue is a change in Italian law concerning passports in 1901.[40] Prior to this law, passports were helpful, but costly, and not strictly required; after 1901, they became free and compulsory when departing from Italy for trans-Atlantic destinations (Foerster 1919, p. 11). While our analyses will use time fixed effects wherever possible, which should control for this change, it is still potentially concerning that there was a surge in emigration, in particular to the United States, between 1900 and 1901 that may be driven by changes in reporting rather than by true changes in emigration.[41] But US arrival data (Barde, Carter, and Sutch 2006) show growth in Italian arrivals from 1900–1901 that closely matches the increase in our data, reducing this concern.[42] Another issue is that the commune-level data for 1884–1903 aggregate some communes with low but non-zero emigration rates in a single figure for each district. Communes included in this aggregation will appear to have an emigration rate of zero in these years. We address this concern in Online Appendix F, where we repeat our main regressions assigning the aggregate emigration to unlisted communes, with similar results. This concern does not affect the district-level data.

Another concern raised by Foerster (1919, ch. 2) and Hatton and Williamson (1998, ch. 6) is that the distinction in the emigration data between temporary and permanent immigration, when it is made, is unreliable (Foerster 1919, ch. 2; Hatton and Williamson 1998, pp. c. 6). We agree, but we do not view this as a deficiency. Return migration was frequent (Bandiera, Rasul, and Viarengo 2013), but the intended duration of migration upon departure was subject to unpredictable changes (Ward 2017). Our goal is to explain the total movement of labor, permanent or temporary, and therefore we ignore this distinction and count both cases equally. However, the issue of return and repeat migration becomes acute in the northern border regions, where seasonal migration across the border was so frequent that in several communes the total number of leavers throughout the period exceeded the total population. While this is an encouraging indication that even easy overland exits were documented in the data, it leads us to treat border-region emigration in particular, and, more generally, emigration to Europe as a whole, with caution.

Our benchmark specifications use emigration rates based on 1901 population as the denominator, and

[39] See the discussion of the accuracy of the Italian emigration data by Foerster (1919, pp. 10–22).

[40] See Foerster (1919, pp. 11, 21) and Hatton and Williamson (1998, p. 98).

[41] According to Foerster (1919, p. 21), the Italian official statistics were less precise than the American immigration data, and that around 1901 there was a change from under- to over-enumeration of Italian emigrants.

[42] This is shown in Online Appendix Figure B.4. Although there is little difference around 1901, larger differences emerge later in the study period.

we verify robustness to using 1881 population instead (Online Appendix G).[43] Since the smallest geographic unit for which destination data are available is the province, we impute destination-specific emigration rates for each commune and district based on the province-year-specific weights of destinations.[44] As described above, our main focus is on migration to North America (the United States and Canada).[45]

In addition to the emigration data, we draw from a wide range of sources, a battery of commune- and district-level post-unification characteristics that are potentially relevant for determining emigration rates. Their purpose is two-fold. First, they serve as control variables in the various statistical tests for the predictions of the diffusion hypothesis. Second, some of them measure local features that allegedly determined the timing of mass emigration according to the modernization hypothesis or according to other studies of Italian emigration; we thus use them in order to assess the validity of this view. Commune-level data include geographic characteristics (elevation, distance to the coast, and distance to land borders), and distance to the nearest railway line in 1881 (Ciccarelli and Groote 2017). District-level data are digitized from the 1881, 1901, and 1911 censuses and include demographic and labor-force composition—the fraction of the male labor force employed in agriculture or industry, the fraction of the population younger than 15 years old, and the fraction of males aged 15 or older who were literate. Among these, we consider elevation to be one of the most important. We will show below that elevation was one of the strongest predictors of emigration at saturation, regardless of region. We suspect that this regularity, which has not generally been discussed in the literature (c.f., Cinel 1982; Gould 1980; Sturino 1990), can ultimately be explained by a crisis in highland subsistence agriculture. Online Appendix I details all of our sources.

4.2 Summary Statistics

Table 1 summarizes the emigration data by half decades, which will be our main temporal unit of analysis. Column (1) presents data at the district level for all periods. Columns (2)–(8) present data at the commune level, with column (2) presenting data for all commune-periods and columns (3)–(8) by period. The fraction

[43] We use 1901 population because it is the population reported in the 1904–1905 volume of the emigration statistics, which is the first volume in which the emigration of all communes is reported regardless of magnitude. It thus enables us to secure population measures that are most comparable to those of our emigration data. A related issue to the distinction between the legal and actual population is internal migration. While mass rates of internal migration are largely a phenomenon of the Fascist and post-World War II period (A'Hearn and Venables 2013, p. 625), there is some evidence of significant population movements in other settings, such as along the former borders after unification (A'Hearn and Rueda 2022). But it is clear that this was overall a minor phenomenon (and a poorly documented one) relative to international migration. We are grateful to Brian A'Hearn for a helpful discussion on this topic. See also Spitzer, Tortorici, and Zimran (2022) for a discussion of the relative magnitudes of internal and international migration.

[44] A potential consequence of this is that there may be artificial within-province-period correlation in communes' emigration to a particular destination that does not extend past provincial borders. We address this issue by repeating our main results using data for all destinations in Online Appendix H. Our results are qualitatively unaffected in general.

[45] We add Canada because the volumes for early years do not distinguish between the two countries in the provincial counts. Canada comprised 2.4 percent of all migrants to North America in an average year, and never more than 7.5 percent.

of communes with any emigration increased from about 58 percent in 1884–1890 to virtually all by the twentieth century; this includes a surge around 1901, in part due to the diminishing tendency to aggregate migration figures for low-migration communes. Similarly, emigration rates increased over time, in particular for migration to North America and Europe, reaching average rates of 23–27 per thousand during the 1900s. On the other hand, South American migration was already well developed early on, and it did not increase steadily over time (or did so only mildly). Figure 5 presents the time series of emigration by broad region of origin within Italy and broad destination region, illustrating these changes over time in more detail.

The bottom half of Table 1 focuses on whether the emigration rate reached an arbitrary minimal threshold that we consider to indicate mass migration. Similar to the evidence above for the intensive margin, there is a substantial increase around 1900 in the prevalence of mass emigration to North America and Europe, but not to South America. By 1906–1910, 42.6 percent of communes had reached mass migration (an average of at least 5 per thousand per year in a half decade) to North America, and 49.9 percent to Europe, but only 30.5 percent of communes produced mass migration to South America. Similarly, the share of communes within the frontier of mass emigration rose from about 3 percent in the first half decade to 36.6 percent in the last half decade before World War I.

Table 2 presents summary statistics for local characteristics that are potentially relevant for migration incentives or for the timing of the onset of mass migration. In the average district, about half of male labor was employed in agriculture, and just over a fifth in industry (including traditional cottage industries), with somewhat greater shares in the north. Slightly over half of adult males were literate, with much lower literacy in the south. Finally, about a third of the population of the typical district was under age 15. The average commune had an elevation of about 400 meters and in 1881 was about 10 kilometers from the nearest rail line, with a greater distance from rail in the south.

A crucial variable for our study is the distance to the nearest epicenter of emigration. We define an epicenter of emigration to North America as a district that had an average annual emigration rate of at least 1 per thousand to North America during the period 1876–1883, and which did not have a neighboring district with a greater annual average emigration rate to North America in this period. This criterion defines 6 epicenter districts, marked in Figure 6: Sala Consilina in Salerno, Isernia in Campobasso, Corleone in Palermo, Chiavari and Albenga in Genova, and Pozzuoli in Napoli. We compute the distance from each commune to the nearest capital commune of an epicenter district. For South America, we devise a similar definition but with a higher threshold of 5 per thousand because we begin to observe the South America-bound emigration when it was already well developed. The epicenters of South American emigration that

we identify are Lagonegro in Potenza, Chiavari in Genoa, Asiago in Vicenza, and Gemona in Udine, and are also marked in Figure 6. Importantly, while the North and South American sets of epicenters partly intersect (in terms of broader region if not the specific district), each of them encompasses some exclusive regions. Central and southern Italy were most exposed to North American epicenters, whereas northern Italy was much more exposed to South American epicenters (panel B, Table 2). Because the Europe-bound emigration was clearly greatest in the districts sharing a land border with neighboring European countries, we use the distance to this border as the measure of distance to the epicenter.

Finally, we define the *frontier of mass migration* to a destination to be the contour of districts that had ever achieved an emigration rate of at least 5 per thousand by a given half decade. As shown in Figure 7, the evolution of the frontiers diverged meaningfully across the three major destination groups. While there is some similarity between the geographic origins of the three flows—for instance, emigration rates from Tuscany and Latium were low regardless of the destination, and both the North and South American flows had epicenters in Campania, Basilicata, and Calabria—there are also major differences. The north-south divide in the contrasting European and North American flows is clear, but may have been the product of lower migration costs to Europe in the north. Other differences, however, are harder to explain by some destination-specific regional advantages. For example, Veneto in the northeast had extremely high rates of migration to South America, but low rates to North America, and to some extent the same was true in the northwest. Similarly, Sicilian migration to South America came primarily from the southern half of the island, whereas migration to North America spread out from its northwest. The distance to the North American frontier in any half decade is shown at the bottom of Table 1.

Figures 1 and 2 present maps of emigration rates. In Figure 1, these are at the district level beginning in 1876. In Figure 2 they are at the commune level beginning in 1884.[46] Both figures map average annual emigration rates by half decade, dividing communes or districts according to quintiles of the distribution of emigration to North America in the period 1911–1914. At both geographic levels, there is strong visual evidence of a spatial expansion of areas of high emigration rates. Indeed, these maps, together with the results for the other testable predictions, show that inter-communal diffusion not only occurred, but that it was important enough to shape the Italian migration the extent that its spatial nature can easily be seen on a map. In addition to the expansion southward from Salerno along the southwestern coast, discussed above, one can easily observe how emigration spread from Liguria to its neighboring regions of Tuscany and Emilia-Romagna. Expansion is also evident in Sicily from the areas around Palermo towards the east and

[46] Appendix Figures B.5, B.6, B.7, and B.8 present analogous figures for migration to South America and Europe. Appendix Figures B.9 and B.10 present analogous figures for migration to any destination.

south. By the eve of World War I, nearly the entire south of Italy, with the interesting exception of Puglia, had achieved high rates of migration. Moreover, the emigration rates of initial leaders in the last period are not greater than those of other places, undermining claims that their migration was the product of unique local characteristics. Instead, the map of emigration in the last two half decades of our study period (Figure 2, panel f), which we consider to be a period in which the diffusion process had approached saturation, bears a remarkable resemblance to a topographical map of the country (Online Appendix Figure B.11). This is even more strongly evident in comparing emigration to all destinations (Online Appendix Figure B.10, panel f and Online Appendix Figure B.11): the coefficient of correlation between emigration and elevation is 0.37 in the 1911–1914 half decade. This strong relationship between elevation and emigration, previously remarked upon by Gould (1980, pp. 290–291), Cinel (1982, p. 31), and Sturino (1990, p. 14) is interesting in its own right, though investigating its causes is beyond the scope of this paper.

5 Patterns

In this section, we check whether patterns described by each of the four predictions laid out in section 3 are present in the data. Throughout this section, the geographic units of analysis are the district and the commune, and, except where otherwise indicated, the temporal unit of analysis is the half decade, which smooths out short-term fluctuations in emigration without obscuring longer-term trends. The main destination of interest is North America because emigration there is observed from its beginnings (unlike South America) and with less mismeasurement (compared to Europe). We present analogous results for migration to all destinations in Online Appendix H

5.1 Convergence

Gould (1980) and Hatton and Williamson (1998) cite the decline over time in the cross sectional variance of emigration rates across Italian regions as a key piece of evidence in favor of the diffusion hypothesis. Figure 8 examines this trend annually at finer levels of geographic disaggregation, the district and the commune, for emigration to North America. Our preferred measure of this variance is the coefficient of variation because it is both normalized by scale and can account for cases of zero migration. There is clear evidence of σ-convergence over time, with the coefficient of variation decreasing almost monotonically from around 4 in 1880 to just over 1 in 1910.[47]

[47]Due to the tendency to not report the specific migration counts of communes producing little migration in the early years, one might expect to see an exaggeration of the downward trend in cross-commune variation. However, the district totals were

This σ-convergence was not the product of a tendency of all places to converge towards the mean, but of new areas entering migration, while areas that had already experienced migration achieved saturation and stabilizing emigration rates. This can be seen in Figure 9, which tests for β-convergence. In this figure, we compute the average annual emigration rate for each district and commune for the periods before and after 1900. We then plot the relative change between the two periods against the rate in the first period. As the β-convergence prediction implies, there is a strong, negative, and nearly monotonic relationship between these variables. Places with the highest emigration rates in the first period experienced almost no growth. The average commune in the top quartile of pre-1900 migration experienced approximately a quadrupling of emigration rates, while emigration from bottom-quartile communes grew more than 200-fold. Importantly, not a single district and only very few communes experienced a reduction in their emigration rates. This is clear evidence that the β-convergence is not a simple case of mean-reversion or of churning of leading and lagging places due to random shocks. Instead, the rate of migration in the early period was the effective lower bound for the rates in the later period. Considering that real wage gaps relative to destination countries were roughly stable, this is strongly consistent with the notion of saturation—emigration rates plateaued around their full potential when the diffusion of links was completed.[48]

A natural concern is that the β-convergence patterns are spuriously generated by measurement errors or by idiosyncratic random shocks, as the pre-1900 emigration rates enters positively into the right-hand side and negatively into the left-hand side. Table 3 provides a more detailed test of the β-convergence prediction, addressing this concern.[49] We estimate analogous regressions of the form

$$\log\left(\frac{e_{i1}}{e_{i0}}\right) = \alpha + \beta \log(e_{i0}) + \mathbf{x}_i'\gamma + \varepsilon_i, \tag{2}$$

where e_{i0} and e_{i1} are the rates of emigration from commune i before and after 1900, respectively, and $\mathbf{x}_i$ is a vector of controls. Column (1) performs this estimation with no added controls, essentially replicating the findings of Figure 9. Columns (2)–(4) of Table 3 repeat the same estimation with the addition of a variety of control variables and then of province and district fixed effects. These four columns include, under "Falsification," the coefficient from repeating the estimation of equation (2) with emigration in the second

almost aways fully reported, and the trends of the decline in variance based on either commune or district data is very similar. This suggests, first, that the problem of omission of low migration counts is not detrimental; and second, that the decline in variance occurred across larger units rather than within them. This is consistent with the idea that the arriving tide of migration lifts all boats in the same area.

[48] This also means that no place had an inverse-U-shaped emigration curve, regardless of how long its emigration experience was.

[49] Online Appendix Table B.1 performs the same analysis with data at the half-decade level instead of only two observations per place.

period, $\log(e_{i1})$, as the regressor. The fact that this coefficient is positive in columns (3) and (4) indicates that there is some merit to concerns that the relationship may in part be the product of measurement error or idiosyncratic shocks; but since the absolute value of the falsification coefficient is more than 4 times smaller, we can still conclude that, even if the coefficient is biased in this way, the convergence is not fully spurious.[50]

In columns (5)–(8), we take a direct approach to addressing concerns of a spuriously negative relationship. Specifically, we use distance from the nearest epicenter of emigration to North America as an instrument for $\log(e_{i0})$. This strategy exploits the fact, which we explore in more detail below, that emigration expanded spatially from these initial epicenters. To be clear, the object of this analysis is not to identify a causal effect, but to clear the possible source of spurious correlation discussed above. Although the first-stage F-statistics become weak in the most restrictive specifications, the coefficients are statistically significant and closely match those of the OLS regressions. This strongly suggests that the strong β-convergence was not a result of the suspected mechanical bias.

5.2 S-Shaped Local Trends

Prediction 2 is that the typical course of the evolution of emigration at the local level followed an S-shaped curve. Figure 10 plots a smoothed time series of migration rates for district (panels a and b) or communes (panel c and d) with the time measure normalized so that year zero is the first year in which the area had an emigration rate of at least 5 per thousand. Panels (a) and (c) present both the average migration rate for each year-relative-to-threshold bin alongside the median and quartiles of the distribution, smoothed over time. A clear S-shape is evident for both commune- and district-level data for all quartiles and for the mean, the latter with tight confidence intervals.[51] The average commune took about 25 years to make the transition from little migration to saturation.

Panels (b) and (d) of Figure 10 divide the communes and districts by the period in which they first reached the 5-per-thousand emigration rate threshold. Regardless of when the migration surge in a place began, it followed roughly the same path except that in the late 1890s the surge appears to have been somewhat faster. Summing up the evidence on the convergence and the S-shaped time series, the continuous national surge in emigration during the period from unification to World War I was not the product of a rising tide that lifted all boats. Rather, places were sequentially lifted from no migration to their mass

[50] Figure 9 shows the results of a similar robustness check. The negative coefficient in this case for a regression of the change in the emigration rate and the emigration rate in the *later* period (the "Falsification Correlation" in the notes to the figures) implies that the relationship is indeed driven by convergence, even if the slope is downward biased. If the true relationship were zero, then this falsification coefficient would be positive and of the same absolute magnitude. See Spitzer (2021) and Spitzer, Tortorici, and Zimran (2022) for similar analyses.

[51] That is, the confidence intervals are such that only an S-shaped curve for the average commune can be drawn within them.

migration potential.

5.3 Correlated Destinations

The data on emigration by destination at the province level enable us to test Prediction 3 regarding correlated destinations. We compute a dissimilarity index V_{ijt} between the destination distributions of the emigration flows of every province pair ij in every half decade t. The dissimilarity index has the convenient feature of being interpretable as the fraction of the emigration flow from province i that would have to have been rerouted to match the destination distribution of province j (or vice versa). Panel (a) of Figure 11 presents a nonparametric regression for each half decade of the dissimilarity indices of each province pair against the distance between the provinces. A positive relationship, is clearly evident: provinces further from one another had more dissimilar destination choices.

Such a relationship, however, is also consistent with an alternative explanation—that the increase in destination dissimilarity with distance between provinces is the result of greater dissimilarity between the characteristics of provinces rather than between their migration networks. What this alternative explanation struggles to explain is why the dissimilarity declined over time over the range of distances, making migration choices more similar across the country. To formalize this pattern, we estimate an equation of the form

$$V_{ijt} = \alpha_t + \beta \log(d_{ij}) + \mathbf{x}'_{ij}\gamma + \varepsilon_{ijt}, \tag{3}$$

where α_t are half-decade fixed effects and d_{ij} is the distance between the capitals of provinces i and j and the controls $\mathbf{x}_{ij}$ are absolute differences between provinces in their individual-level cov1881 agricultural employment share, industrial employment share, literacy rate, and fraction under age 15.[52] Panel (b) of Figure 11 plots the half-decade fixed effects from this regression with the 1876–1880 half decade excluded, as well as the fixed effects from a similar regression limiting attention to province pairs less than 300 kilometers apart. There is a clear decline in the fixed effects over time, and the decline is particularly monotone for province pairs within the 300-kilometer range.

In more intuitive terms, the baseline specification suggests that moving from the 25th to the 75th percentile of distance between provinces entails a 0.17-standard deviation greater dissimilarity. Over the period between 1876–1880 and 1911–1914, the dissimilarity of any pair of provinces diminished by 0.20 standard

[52] We present the estimates of β in Online Appendix Table B.2, using all destinations or different combinations of destinations, with and without control variables. The upward slope of the dissimilarity-distance relationship is robust to the inclusion of controls and to a focus on different time periods or sets of destinations, casting doubt on the notion that increasing differences in local characteristics with distance can explain our results.

deviations of the 1876–1880 distribution. The diffusion hypothesis explains both the spatial and the temporal trends as a result of destination "menus" that are common to neighboring provinces, and become increasingly common to all provinces over time.

5.4 Spatial Expansion and the Frontier Effect

The *frontier effect* is more than simply a prediction that follows from the diffusion hypothesis. It is a necessary condition, in the sense that if it does not hold then the hypothesis as a whole fails. Figure 12 shows the main evidence regarding the frontier effect. In panels (a) and (b), the rates of emigration at the levels of the district and the commune are plotted in a non-parametric regression against the distance from the frontier a half-decade earlier, including in the sample in each period only places that had not yet produced mass migration to North America. When pooling all periods together, the expected trend is sharp and clear (notice that the vertical axis is based on a logarithmic scale). At a distance of 25 kilometers, the rate of emigration is on average 3.2 per thousand at the commune level. It then decays rapidly to 2.4 per thousand at 50 kilometers, 1.5 per thousand at 100 kilometers, and 0.9 per thousand at 150 kilometers; beyond that, the effect of the frontier weakens, as should be expected at distances that are unlikely to allow for personal contacts. The same pattern is apparent, albeit with some volatility, in each half-decade separately as well. These patterns indicate that distance from the frontier is more than a characteristic that correlates with emigration rates; the general absence of high emigration rates beyond 100 kilometers from the frontier suggests that proximity to the expanding frontier of mass migration was, in most cases, necessary for the onset of mass emigration.[53]

A formal test for the frontier effect is presented in panel (c) of Figure 12. We regress emigration rates on period-specific functions of distance from the previous period's frontier. To account for communes with zero recorded emigration, we use the binomial maximum likelihood estimation described in Appendix A. The estimates are universally negative, and are statistically significant until the half decade beginning 1915, consistent with the existence of a frontier effect. Moreover, adding controls for local characteristics does not reduce, and sometimes increases, the estimates. This suggests that the frontier effect is not driven by spatial trends in observed characteristics. Figure 13 shows that the frontier effect was destination specific, formalizing the visual evidence in Figure 7. It repeats the analysis of panel (c) of Figure 12, but in addition to the relationship between emigration to North America and distance to the North American frontier, it

[53]Pooled over all periods, the share of communes entering North American mass migration for the first time was 37.8 percent when the previous half-decade's frontier was less than 50 kilometers away, as opposed to 5.4 percent when it was over 100 kilometers away.

also shows the relationship between migration to European and the South American frontiers. Out of the three main streams, only migration to North America is systematically negatively correlated with distance to the North American frontier.[54]

While the frontier effect is a necessary component of the diffusion hypothesis, for the diffusion hypothesis to hold it must lead to another necessary observable macro pattern: that places situated far from the epicenters contracted emigration much later than places near them. To validate this pattern, panels (a) and (b) of Figure 14 plot non-parametric regressions of the logarithm of half-decade average annual North American emigration rates against distance from the nearest epicenter, which is fixed over time (as opposed to distance from the frontier in Figure 12, which changes each period).[55] Both at the district and the commune level, places closer to the epicenters were emigration leaders throughout the study period, as evidenced by the negative slopes of the non-parametric regressions. There is also a gradual leveling of the curves as more distant areas entered into emigration, as evidenced by the gradual flattening of the slope, to the point that it is nearly flat by the last half decade. Emigration rates from districts under 50 kilometers from the epicenter was 5.8 times greater than from districts 50–100 kilometers away, and 20.8 times greater than those in the range 100–200 kilometers in the period 1876–1880. By 1911-1914 these ratios had shrunk to 0.9 and 2.3, respectively.

We test this pattern formally in panel (c) of Figure 14 by regressing the logarithm of the emigration rate on half-decade-specific functions of distance from the nearest epicenter, using the binomial maximum likelihood method. At both the district and the commune level, the coefficients on distance from epicenter are initially negative and monotonically decline in magnitude over time. As in the frontier regressions, this pattern is robust to controls, and it is notable that after adding them the coefficients of the last periods become indistinguishable from zero. In other words, distance from the epicenters was highly predictive of emigration rates at the beginning of the Italian migration, and ceased to be so by the time it reached saturation.

Finally, to formalize the notion that distance from the epicenters determined the timing of entry into mass emigration, we estimate in Table 4 semiparametric Cox proportional hazard models (Zeng, Mao, and Lin 2016) for the timing of entry into the frontier of mass emigration, focusing on the district as the unit of observation (since that is the unit of analysis that we consider when determining whether a place has

[54] Online Appendix Figure B.14 computes, for every destination pair, a dissimilarity index for each half decade of the sources of migration across Italian provinces. It provides additional evidence that the migration flows to each major destination originated in different areas of Italy.

[55] Results based on the level rather than the logarithm of average annual emigration are presented in Online Appendix Figure B.12.

entered the frontier of mass migration). Column (1) shows that a one-standard deviation increase in distance from the epicenter was associated with a 42 percent lower hazard of achieving mass migration at any time. Columns (2) and (3) show that this pattern is robust to controlling for broad region (i.e., north, south, and center) and to controlling for the various district-level characteristics that we observe. Thus, distance from epicenters was not simply a determinant of rates of emigration, but also of the timing of the onset of mass emigration.

In sum, the findings complement the β-convergence prediction in showing that the laggards who caught up with the leaders were really the more distant places narrowing the gap relative to those closer to the epicenters. In order to enter mass emigration, places had to be situated close to the recent frontier of mass emigration. This, in turn, generated a spatial macro-pattern of expansion from the epicenters outwards, whereby places farther from the epicenters experienced a later onset of mass migration. Moreover, the robustness of these findings to local controls suggests the spatial trends in emigration were not likely a result of systematic spatial trends in underlying characteristics. As a rule, emigration from distant places was delayed, sometimes by decades, for no reason other than their location relative to the epicenters.

5.5 Examining the Modernization Hypothesis

The key advantage of the diffusion hypothesis, and among the strongest evidence of its validity, is that it can parsimoniously explain all of the patterns that we have documented above. Nevertheless, some of these patterns "might be explained in other ways" than by diffusion (Hatton and Williamson 1998, p. 99). As described above, the main alternative to the diffusion hypothesis is the modernization hypothesis, and in this section we assess the power of this hypothesis in explaining the same set of stylized facts.

A key feature of any plausible form of the modernization hypothesis is that more modernized places experienced an earlier onset of mass emigration. At first glance, even this basic feature of the modernization hypothesis is inconsistent with the data: the geographic distribution of early Italian emigration bears little resemblance to that of early Italian industrialization. Full-fledged industrialization in the post-unification era took part in concentrated geographic pockets, primarily in the northwest, which forged ahead in that period (Ciccarelli and Fenoaltea 2013). Some of these industrialization hotspots, such as in Liguria and in the Alpine slopes, were indeed emigration leaders, but the nearby Po Valley never developed mass emigration. Moreover, other emigration epicenters were in the northeast, in the center, and in the south. Some had existing traditional and extractive industries, yet they were generally not a part of the industrialization movement.

To answer this question more systematically, we repeat the hazard regressions on a set of variables representing district-level modernization as of 1881: the share of workers in agriculture (a negative proxy for economic development), the fraction of the population under 15 (a proxy for demographic pressures), literacy, and urbanization measured as the share of the population living in communes with over ten thousand residents. The results are presented in Table 5, where all the explanatory variables are normalized for comparability and the rates of emigration are for all destinations.[56] Column (1) repeats the basic regression from Table 4, but using distance from any emigration epicenter. Columns (2)–(5) report the hazard ratios associated with a one-standard deviation increase in each of the four modernization proxies. Two of these proxies seem to act in a direction opposite to the one predicted by the modernization hypothesis: districts with a more agricultural labor force and with less urban population developed mass emigration earlier. The other two proxies—the fraction of the population under the age of 15 and the rate of literacy—are positively associated with an earlier emigration, as expected, although much more weakly than the distance to the epicenter. When pooling all risk factors together (columns 6–7), literacy no longer has the expected sign and the estimated hazard ratio of the fraction of the population under 15 is reduced to just 1.172, albeit still statistically significant. Distance to the epicenter, however, remains unchanged in its strength as a predictor of a late start, with mass emigration hazard doubling with each standard deviation reduction in distance. We conclude from this that the evidence on the importance of modernization factors in determining the timing of mass emigration is at best mixed and weak, whereas the distance from the epicenter remains first-order predictor.

A naïve version of the modernization hypothesis would argue that modernization factors simply augmented other push factors in increasing the demand for migration.[57] This view, however, fails to offer a simple explanation for the patterns of convergence in emigration rates, since the factors triggering an early emigration would also be associated with higher rates overall. Leaders would simply be situated on a higher emigration path. As shown above, it is hard to make the case that as a rule, early adopters had the highest emigration potential.

A more nuanced version of the modernization hypothesis rationalizes convergence as a result of a process in which places sequentially modernize, experiencing a surge in emigration during this transition, such that the convergence in emigration mirrors convergence in modernization—that is, modernization is a factor that unleashes demand for migration produced by other means. If that were the case, then early modernization

[56] We do not separate emigration by destinations because this separation is not implied by the modernization hypothesis. Destination-specific regressions yield the same qualitative outcomes.

[57] This version is implied in Hatton and Williamson's (1998) analysis.

should predict higher emigration early on, but at a diminishing rate as time goes by, similar to the evidence on the diminishing importance of the distance to the epicenter (Figure 14).

In Figure 15 we examine whether a dynamic in which early modernization diminishes over time as a predictor of emigration rates is borne out in the data. In each period, we regress district-level emigration on distance to epicenters and modernization factors (all normalized, for comparability).[58] In the early periods, the coefficients are qualitatively the same as in the hazard regressions, where only two modernization proxies (literacy and the fraction of the population under 15) have the expected sign and the distance to the epicenter is by far the strongest predictor. Over time, the coefficients indeed converge towards zero, yet the two modernization proxies that affect in the expected direction do so from a relatively low starting point, and in the case of the fraction of the young population the convergence is barely noticeable. We conclude from this that modernization factors are not a likely source for the extraordinarily strong convergence in emigration rates that we document.

Another way in which the diffusion hypothesis outperforms the modernization hypothesis is by providing a straightforward explanation for why migration streams to different destinations were at least partly independent of each other (Figures 7 and 13) and why similarity in destinations increased with proximity and over time (Section 5.3). These patterns are not strictly inconsistent with internalist explanations—it could be argued that the overall level of emigration was not affected by a diffusion process, only the destination was chosen based on the networks that were available in the vicinity. Such an explanation does, however, still emphasize the importance of inter-communal networks and raises questions as to how, given their dominance, emigration can start in their absence. That is, the modernization hypothesis requires a more complex explanation for these patterns, which derive directly from the diffusion hypothesis.

To be clear, we do not argue that internal factors, including different aspects of modernization, did not affect emigration. What we do learn from this cursory glance at the modernization hypothesis is that it does not appear to offer a plausible, simple, and complete explanation for the set of stylized facts of the Italian emigration, neither in its naïve version nor in a more nuanced one. There is no evidence that modernization was systematically associated with the timing of mass emigration, and it is hard to explain the convergence in emigration rates as result of some dynamics of modernization. Most importantly, to the extent that internal factors do play a role in determining the rates of emigration, their impact is small relative to the consistent first-order role of the diffusion process.

[58] Notice the differences in the specification relative to Figure 1: emigration is to all destinations rather than to North America alone and the distance to the epicenter is normalized.

6 The Spatial Contagion Mechanism

The underlying mechanism of the diffusion hypothesis is spatial contagion. Returning to our epidemiological analogy, emigration is a fever—emigration does not arise in a place unless it is infected by its neighbors. In what follows we set out to establish that spatial contagion was indeed the mechanism that caused the diffusion of Italian emigration. The key challenge is to verify that the correlation between a place's rate of emigration and its neighboring place's lagged rates of emigration indeed reflects a causal relationship. The counter-hypothesis which the evidence ought to reject is that the correlation is driven by other spatially correlated and potentially unobserved local factors that cause emigration to arise in neighboring places around the same time. Verifying the causal power and the economic significance of spatial contagion is also important because it is a key differentiating feature that is not inherent to any internalist explanation, including those along the lines of the modernization hypothesis.

6.1 The Rationale of the Instrumental Variables Approach

We develop an instrumental variable approach to identify the causal effect of lagged neighbors' emigration. The baseline estimation equation is a spatial lag model of the form

$$\log(e_{it}) = \alpha_t + \beta \log(e_{\neg it-1}) + \mathbf{x}_i' \delta_t + \varepsilon_{it}, \tag{4}$$

where e_{it} is the rate of emigration from commune i in half decade t, α_t are half-decade fixed effects, $\mathbf{x}_i$ is a vector of controls, and δ_t is a half decade-specific vector of coefficients. We specify equation (4) in logarithmic form rather than in levels in order to make the effects proportional to the level of migration.[59] We focus on data at the level of the commune-half decade in order to smooth out year-to-year shocks. We cluster standard errors at the district level, which permits correlation between any commune-half decade observation within the same district, either over space, over time, or both.[60]

The regressor of interest in equation (4) is $e_{\neg it-1}$, which we refer to as *lagged emigration exposure*. This is a lag of a weighted average of emigration rates of all other communes, with greater weight exerted by nearer and more populous communes. Specifically, we define this object, in a manner analogous to equation

[59] This is consistent, for example, with Mahajan and Yang (2020), who find in linear regressions that the effect of hurricanes on migration increases in the size of the network, which represents the base level of migration. This assumption appears to fit the data much better, at least by the criterion of generating seemingly normal distributions.

[60] Such spatial lag models have recently been used to study diffusion in other settings (e.g., Aidt and Leon-Ablan 2022; Aidt, Leon-Ablan, and Satchell 2022). Ours, however, is the first to pair this model with an IV strategy based on the epidemiological intuition of diffusion.

(1), as

$$e_{\neg it} = \frac{\sum_{j \neq i} e_{jt} N_j d_{ij}^{\theta}}{\sum_{j \neq i} N_j d_{ij}^{\theta}}, \tag{5}$$

where N_j is the population of commune j, d_{ij} is the distance in kilometers between communes i and j, and θ is the rate at which the influence of other communes' j emigration rates on that of commune i decays over distance. The numerator is a measure of effective proximity to emigration, whereas the denominator is a measure of effective proximity to population. The measure $e_{\neg it}$ is thus a population- and distance-weighted average emigration rate in the neighborhood of commune i. The scale of $e_{\neg it}$ is that of an emigration rate, and, like N_{it} in the theoretical model of section 3, it can be thought of as the probability that any of an individual's out-of-commune contacts migrated in the previous period.[61] A value of $\theta < 0$ implies that more distant communes have a smaller impact than do nearer communes. For computational tractability we separate the estimation of θ from that of other parameters. We estimate equation (4) by NLS and arrive at a value of $\theta = -2.83$, which we later use throughout our main analysis.[62] The coefficient of interest in equation (4) is β, which can be interpreted as an elasticity of emigration with respect to lagged emigration exposure.

The obvious deficiency of estimating equation (4) by OLS is that any local determinant of emigration, including unobserved ones, are likely to be spatially correlated, biasing upwards the estimate of β. Identification of the effect of emigration exposure on emigration therefore requires that we find a source of variation in neighbors' emigration that is independent of a place's own internal demand for emigration. The intuition of our instrumental variable approach is to interact the plausibly exogenous variation in the spatial orientation of the neighboring population with the endogenous spatial patterns of emigration described above.

The diagram in Figure 16 illustrates the intuition behind our instrumental variables approach. Consider two communes, A and B, that are observably identical in all their internal characteristics. In particular, they are equidistant from a source of emigration (an epicenter or a frontier) at a distance d_2. The only difference between the two communes is that the neighboring population of commune A is distributed such that on average it is closer to the source than is commune A, whereas the neighboring population of commune B is on average farther from the source than commune B itself. The average neighbor of commune A is therefore

[61] The measure $e_{\neg it}$ satisfies two desirable conditions. The first is that it is robust to splitting communes. The second is that it is robust to uniform changes in population density. If the measure of exposure were a function of the number of emigrants (rather than the local *rate* of emigration), then doubling the population everywhere near a commune (with an accompanying doubling of the number of emigrants) would double an individual's emigration exposure. To reflect the limited number of connections that a person can have, our measure is robust to population density, which would have no impact on the proximity-weighted emigration rate. The implied assumption is that the number of links that any individual has outside of his commune is fixed and independent of the population density in the neighborhood.

[62] As we show in Online Appendix K, our findings are robust to using alternate values of θ selected by estimating alternate forms of equation (4).

likely to be infected by the spreading wave of emigration earlier than is the average neighbor of commune B, not because of any feature that is correlated with any of commune A or B's internal characteristics, but simply due to the different spatial orientation of their neighbors with respect to the source. Therefore, we can construct a valid instrumental variable for emigration exposure based on the weighted distance of a commune's neighbors to the source.

6.2 Implementation of the Instrumental Variables Approach

In practice, the instrumental variable for emigration exposure $\tilde{e}_{\neg it}$ is constructed in a manner analogous to the actual emigration exposure measure $e_{\neg it}$. That is, it is defined as

$$\tilde{e}_{\neg it} = \frac{\sum_{j \neq i} \tilde{e}_{jt} N_j d_{ij}^{\theta}}{\sum_{j \neq i} N_j d_{ij}^{\theta}}, \tag{6}$$

but instead of being a weighted average of actual emigration, $e_j t$, it is a weighted average of estimated emigration, $\tilde{e}_j t$—a measure which is only based on commune j's distance to nearest source, as we detail below. We then estimate equation (4) using $\log(\tilde{e}_{\neg it}$ as an instrument for $\log(e_{\neg it})$.[63] The key identifying assumption underlying this strategy is that, conditional on a commune's distance from the nearest emigration source, the spatial orientation of its neighbors (i.e., whether on average they are closer or farther from the source) is random. Importantly, the distance to the source need not be exogenous; the strategy merely uses the observed fact that a commune's distance to the source is correlated with its emigration rates. Returning to the illustration in Figure 16, since the neighbors of commune A are on average closer to the source, we expect that the actual emigration exposure of commune A, which is a weighted average of the neighbors' actual emigration, will be greater. For the same reason, the predicted emigration of the neighbors of commune A is on average greater, and as a result the weighted average of these predicted measures will also be greater. The latter is then used as an instrument for the former. Notice that this source of variation, stemming from random differences that are likely to be small in the geographic orientation of the neighboring population, is at risk of suffering from low statistical power; as we discuss below, this poses difficulties in some, but not all, of our specifications.

To construct $\tilde{e}_{jt}$, we estimate a non-parametric regression of the form

$$\log(e_{jt} + \varepsilon) = f_t(z_j) + u_{it}; \tag{7}$$

[63] It is not necessary to adjust standard errors for the use of this generated instrument (Wooldridge 2002, pp. 116–117).

that is, a period-specific non-parametric regression of log emigration on distance from the nearest emigration epicenter.[64] We then set $\tilde{e}_{jt} = \exp[\hat{f}_t(z_j)]$.

In all of our specifications, we control for a commune-half decade's own value of predicted emigration $\tilde{e}_{it}$, computed as in equation (7), allowing the coefficient to vary by half decade. This is the straightforward way of controlling for the commune's own location with respect to the source, as required by the IV strategy; the identifying variation is in predicted emigration exposure conditional on own predicted emigration, and using $\tilde{e}_{it}$ ensures that both expected values are calculated in the same way. Our baseline specifications restrict the sample to communes situated between 50 and 250 kilometers from the sources of emigration.[65]

6.3 Hidden Threats to the Validity of the Instrument

The baseline specifications also include controls for the geometry of national borders and for population density. Each of these are crucial in addressing potential risks to the validity of the instrument which are not plainly visible. First, consider a commune that is located on the contours of Italy—either on the coast or on the land border. For obvious geometric reasons (and because we do not observe communes directly on the other side of the land border), any source of emigration within the country will tend to be closer to the commune's average neighbors than to the commune itself.[66] This regularity will cause a positive correlation between predicted emigration exposure and position along the coast or the land border. If such locations have systematically different emigration potentials, this would amount to an endogeneity problem stemming from the violation of the assumption that the spatial orientation of the commune's neighbors is random. We address this threat by controlling for half decade-specific functions of distance to the coast and distance to the land border.

The second hidden threat comes from variation in population density around a commune. Consider a case in which the predicted emigration of a commune is a downward sloping convex function of distance to the source (which in our case, it is). Then a reduction of the population density around the commune would be associated with a greater average predicted emigration of its neighbor, while keeping own predicted emigration unchanged.[67] If population density is correlated with unobserved determinants of emigration,

[64] The addition of $\varepsilon = 0.0001$ on the left-hand side of equation (7) is needed to ensure that commune-half decades with no emigration are included in the construction of emigration exposure.

[65] Removing the range restriction strengthens the results substantially.

[66] For a simple example, consider a commune located at the corner of a grid. When the source of emigration is located strictly inside the grid, there exists a positive radius around the corner commune within which every other commune is closer to the source than itself. Exceptions to this rule are conceivable in real-life data, but are unlikely to occur systematically.

[67] It is easy to see this in the simple example in which the population is positioned along a line; a proportional expansion of the neighboring population away from the commune would leave the weight of each commune unchanged, but due to Jensen's inequality, the increase in predicted emigration of the communes closer to the source will be greater than the reduction in predicted emigration of communes farther from the source.

this would create endogeneity in the instrument. The straightforward way to address this issue is to control for a measure of population density, for which we use the denominator of the ratio in the right-hand side of equation (5).

Another hidden threat comes from the fact that the predicted values of commune i's neighbors $\tilde{e}_{jt}$ are in small part based on commune i's realized rate of emigration, which introduces a very small room for endogeneity in predicted emigration exposure.[68] This is also accounted for by controlling for the commune's own lagged predicted emigration $\tilde{e}_{it-1}$. An alternative way to remove this source of endogeneity altogether is to estimate equation (7) while excluding the source catchment into which a commune falls. For example, the emigration exposure for communes whose nearest emigration epicenter is the district of Corleone is constructed by estimating equation (7) for all communes except those for which the nearest epicenter is Corleone and then applying the predicted emigration to all communes in the Corleone catchment. For distance from the mass emigration frontier, a commune's estimated emigration is constructed by estimating equation (7) excluding a commune's own province.[69] We discuss the results of this alternative approach below.

6.4 Results

Table 6 presents the results of our estimation.[70] Column (1) includes no controls beyond those described above, and only half-decade fixed effects. The first-stage F-statistic clearly passes the Staiger and Stock (1997) threshold. The estimated elasticity of a commune's own emigration with respect to the portion of lagged neighbors' emigration driven by variation in distance from epicenters is approximately 1 and statistically significant, as is to be expected in the presence of a spatial contagion mechanism. Column

[68] Consider commune i at a distance d from the source. The predicted rate of emigration for communes $j \neq i$ that are at a distance of just above or below d is a weighted average of actual emigration rates that includes that of commune i. Hence, the emigration rate e_{it} feeds in to the estimated emigration exposure $\tilde{e}_{\neg it}$.

[69] The solution is different because the frontier is not a small number of places, each with a well-defined catchment that could be excluded. Each commune has its own unique closest point along the frontier, and thus there is no natural definition of catchments.

[70] Analogous OLS results are in Online Appendix Table B.3. Online Appendix Table B.4 presents OLS results without the limitation to communes within 50 and 250 kilometers of emigration epicenters. Online Appendix Table B.5 presents results of using a more standard approach of estimating a spatial lag model of the form

$$\log(e_{it}) = \alpha_t + \beta \log(e_{\neg it-1}) + \gamma \log(e_{it-1}) + \mathbf{x}_i' \delta_t + \varepsilon_{it}$$

instrumenting for lagged emigration exposure using half decade-specific functions of neighbors' characteristics (e.g., Aidt and Leon-Ablan 2022). This is operationalized by using the binomial maximum likelihood estimator of Appendix A to estimate neighbors' emigration as a function of all observables (including distance from epicenter) and then constructing an instrument as in equation (6). We do not restrict communes on the basis of distance from the epicenter. The results are qualitatively similar to the main results. We prefer those presented in the main text because these alternate results require the less-plausible assumption that the observable characteristics of neighboring communes do not directly affect a commune's own emigration.

(2) adds controls for year-specific functions of latitude, longitude, elevation, exposure to cities,[71] district shares of employment in industry and agriculture in 1881, 1881 distance to rail, and 1881 district-level literacy and fraction aged 15 or less. Doing this only slightly attenuates the coefficient on lagged emigration exposure. Columns (3)–(8) add increasingly fine fixed effects at the region, region-half decade, province, province-half decade, district, and district half-decade, respectively. The magnitude of the coefficient is further attenuated, and the precision of the estimate and the strength of the instrument (as measured by the first-stage F-statistic) diminish, particularly when controlling for period-specific regional fixed effects. As mentioned above, the identifying variation is likely very small, and thus as the control becomes very tight statistical power is eroded. Nevertheless, with the exception of the specifications that control for district fixed effects, in which the standard errors are too large to identify elasticities smaller than 1, the estimated elasticities are in the range 0.5–1 and are at least marginally statistically significant.[72]

Table 7 repeats the same estimation, but uses the distance to the dynamic frontier rather than the static epicenter to construct the instrument. Since, as shown above, the predictive power of the distance to the epicenter diminishes over time whereas that of the distance to the frontier does not, this is likely to increase the strength of the instrument without violating its exogeneity.[73] As in the regressions that estimate the frontier effect (Section 5.4), the sample changes for each half decade to include only communes that had not yet reached the frontier. The results are qualitatively the same as in Table 6, but with a meaningful improvement in the strength of the instrument and in statistical power.[74]

In Online Appendix Tables B.8 and B.9, we repeat the analyses of Tables 6 and 7, but, as described above, we construct the instrument based on excluding a commune's own catchment. Both the estimates and the statistical power are somewhat sensitive to the exclusion of own catchment and own province from the prediction of emigration rates, but the qualitative results remain unchanged: estimates that have sufficient statistical power point at an elasticity in the range of 0.5–1.

To be sure, the results of this section are neither precise nor stable, likely owing to the fact that they derive from the limited variation in the spatial orientation of communes' neighboring population, leaving little statistical power under a large enough set of fixed effects. The identification also relies on the inclusion

[71] That is, we construct a measure analogous to equation (5), but use city population instead of number of emigrants. The cities in question are Bologna, Catania, Firenze, Genova, Messina, Milano, Napoli, Palermo, Roma, Torino, and Venezia. We include this measure because urban areas are likely to be differentially responsive when they are reached by the expanding tide of emigration. Returning to Figure 16, if commune A is neighbored by urban areas and commune B is not, we expect differences in the degree to which they are exposed to emigration by their neighbors.

[72] Online Appendix Table B.6 repeats the same estimation, but drops any observation in the top and bottom one percent of residuals from a regression of the instrument on the controls used in all columns. Doing so strengthens the results.

[73] Using the distance to the frontier requires taking a stand on how to predict emigration for communes that are within the frontier. We assign them a distance of zero.

[74] Online Appendix Table B.7 shows results dropping communes with extreme residuals and they are again stronger.

of a number of controls, most notably the commune's own predicted emigration. In part for this reason, this evidence on the spatial contagion mechanism is not meant to stand alone as proof of the diffusion hypothesis. We view it instead as part of a body of evidence together with the results of Section 5, which altogether is most parsimoniously explained by the diffusion hypothesis and is difficult to rationalize under alternate hypotheses. Nevertheless, despite the demanding constraints, the outcomes of this exercise broadly point at an economically significant contagion effect that is consistent with the diffusion hypothesis: communes that were exposed to greater emigration in their vicinity for reasons other than a shared ecology produced a greater emigration flow, most likely at an elasticity of 0.5 or more. While we cannot pinpoint the precise reasons behind this effect, it appears that it had to do with physical proximity between populations. Based on our reading of the sociology of the Italian migration (Section 2.2) and of other similar movements, we argue that the most plausible explanation is that there existed inter-communal diffusion of migration networks. Friends and relatives enabled both the emigration of individuals in their home towns, as well as that of their contacts in neighboring communes, such that the networks expanded from one commune to the next.[75]

7 Summary of Robustness Checks

The Online Appendix presents a variety of robustness checks for the results presented in sections 5 and 6. Online Appendix F repeats the main results incorporating communes that are not listed in the earlier emigration statistics volumes, but which may have been included in the data for "Other Communes" provided for each district, by allocating this extra emigration equally to all of these unlisted communes. Online Appendix G uses 1881 population as the basis for computing emigration rates instead of 1901 population. Online Appendix H repeats the results using data on emigration to all destinations instead of only to North America. Among other issues, this addresses the concern that the correlation of emigration from communes in the same province may have been inflated by the fact that the emigration-by-destination data (which are not used in this case) are available only at the province level. Online Appendix J repeats the results including communes with no emigration in a particular half decade. which are otherwise excluded due to the use of the logarithm of emigration as the main outcome in many analyses. Online Appendix K repeats the results of section 6 using different values of θ to compute lagged emigration exposure and the instrument.

[75] An alternative explanation could be that emigration in one commune caused some change—such as in the economy or the culture of the place—and that this change spilled over to neighboring communes and caused emigration there. The local effects of emigration in that period is indeed a subject that lacks quantitative evidence. While we suspect that such effects did exist, we find it hard to believe that their spillovers to neighboring communities were sufficiently strong to be a major cause for emigration. Yet even if they were, such spillover effects would have been part of a slightly different version of the diffusion hypothesis, which would have kept all of the important implication of our preferred version of this hypothesis that stipulates that personal contacts were the key substrate over which inter-communal diffusion occurred.

Finally, Online Appendix L repeats the estimates of section 6 using increasingly fine geographic fixed effects.

8 Conclusion

The question of why emigration from the European periphery was delayed is one of the fundamental puzzles of the Age of Mass Migration. In this paper, we test the diffusion hypothesis—an influential yet heretofore not rigorously tested explanation for this puzzle that attributes the delay to an absence of chain migration networks and the subsequent onset of mass migration to the spatial expansion of these networks. We formalize this hypothesis by developing a model of migration within a spatial network and deriving the hypothesis's testable predictions. We then use newly constructed data on one of the most important episodes of emigration in the Age of Mass Migration—Italian emigration to North America in the period 1876–1920—to develop new evidence supporting the diffusion hypothesis. We show that the four testable predictions of the diffusion hypothesis that we derive from our epidemiological framework are confirmed by the data. The most novel and important result in this regard is the spatial expansion of emigration, beginning in a handful of epicenters, and expanding at a rate of less than 100 kilometers per half decade. Then, we develop a novel instrumental variables strategy that exploits the epidemiological differences between the two main explanations for the delayed migration puzzle. On the basis of this strategy, we find strong evidence that the spatial expansion of migration reflected diffusion along social channels, as envisioned by the diffusion hypothesis, rather than through some other mechanism. Besides showing that a diffusion process operated in Italy, our results strongly suggest that diffusion was not simply one factor that affected migration but a first-order determinant, one that can alone explain a number of important macro-patterns of the Italian emigration. The delay of mass emigration from Italy was thus the product of the slow pace by which the initial seeds of emigration generated networks that eventually spread across the country. Although our results constitute the strongest empirical basis to date for the diffusion hypothesis, they do not definitively disprove the modernization hypothesis. Doing so would require taking a structural approach, which would enable the evaluation of counterfactual scenarios for the evolution of Italian emigration—a task that we leave for future research. Nonetheless, our results weaken substantially the footing of the modernization hypothesis by showing that it is difficult to parsimoniously rationalize the main empirical patterns of the Italian migration using it or any other internalist explanation.

More broadly, the literature on migration in economic history has struggled for decades to explain or come to terms with the ubiquity of evidence that the push-pull paradigm simply does provide a simple explanation

for too many important features of the Age of Mass Migration. Hatton and Williamson's (1998) pathbreaking work was a heroic attempt in that respect. Their view regarding the role of industrialization, and of economic modernization more broadly, rationalized some of the contradictions to the push-pull paradigm without deviating from it, and while ruling out an important role for spatial diffusion. Given the evidence from our new high-resolution data of the Italian migration, we have shown that the role of diffusion in shaping the course of mass migration was considerably more important than previously realized, and that the current paradigm may have an insufficient explanatory power. Instead, we have suggested a new synthesis between the push-pull paradigm and an epidemiological diffusionist view of migration. This new synthesis seems to fit a number of refreshed and new stylized facts, and to offer a simple, parsimonious, and comprehensive explanation for the delayed migration puzzle, at least for Italy.

References

A'Hearn, Brian and Valeria Rueda (2022). "Internal Borders and Population Geography in the Unification of Italy." *Journal of Economic History* Forthcoming.

A'Hearn, Brian and Anthony J. Venables (2013). "Regional Disparities: Internal Geography and External Trade." In *The Oxford Handbook of the Italian Economy Since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 21, pp. 599–630.

Abramitzky, Ran and Leah Platt Boustan (2017). "Immigration in American Economic History." *Journal of Economic Literature* 55:4, pp. 1311–1345.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102:5, pp. 1832–1856.

Aidt, Toke and Gabirel Leon-Ablan (2022). "The Interaction of Structural Factors and Diffusion in Social Unrest: Evidence from the Swing Riots." *British Journal of Political Science* 52, pp. 869–885.

Aidt, Toke, Gabirel Leon-Ablan, and Max Satchell (2022). "The Social Dynamics of Collective Action: Evidence from the Diffusion of the Swing Riots, 1830–1831." *Journal of Politics* 84:1, pp. 209–225.

Andrews, Kenneth T. and Charles Seguin (2015). "Group Threat and Policy Change: The Spatial Dynamics of Prohibition Politics, 1890–1919." *American Journal of Sociology* 121:2, pp. 475–510.

Ardeni, Pier Giorgio and Andrea Gentili (2014). "Revisiting Italian Emigration before the Great War: A Test of the Standard Economic Model." *European Review of Economic History* 18, pp. 452–471.

Baily, Samuel L. (1999). *Immigrants in the Lands of Promise: Italians in Buenos Aires and New York City, 1870–1914*. Cornell University Press.

Baines, Dudley (1995). *Emigration from Europe 1815–1930*. Cambridge: Cambridge University Press.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo (2013). "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics* 102, pp. 23–47.

Banfield, Edward C. (1958). *The Moral Basis of a Backward Society*. Free Press.

Barde, Robert, Susan B. Carter, and Richard Sutch (2006). "Table Ad106–120: Immigrants, by country of last residence—Europe, 1820–1997." In *Historical Statistics of the United States*. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright (ed.). Cambridge: Cambridge University Press, pp. 1.560–1.563.

Barton, Josef J. (1975). *Peasants and Strangers: Italians, Rumanians, and Slovaks in an American City, 1890–1950*. Cambridge: Harvard University Press.

Bass, Frank M. (1969). "A New Product Growth for Model Consumer Durables." *Management Science* 15:5, pp. 215–227.

Bell, Rudolph M. (1979). *Fate and Honor, Family and Village: Demographic and Cultural Change in Rural Italy since 1800*. Chicago: University of Chicago Press.

Bernoulli, Daniel (1776). "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir." In *Mémoires de mathematique et de physique*. Academie Royale des Sciences.

Bodnar, John, Roger Simon, and Michael P. Weber (1982). *Lives of Their Own: Blacks, Italians, and Poles in Pittsburgh, 1900–1960*. Urbana and Chicago: University of Illinois Pres.

Brandenburg, Broughton (1904). *Imported Americans: The Story of the Experiences of a Disguised American and His Wife Studying the Immigration Question*. New York: Frederick A. Stokes Company.

Briggs, John Walker (1978). *An Italian Passage: Immigrants to Three American Cities, 1890–1930*. New Haven: Yale University Press.

Burchardi, Konrad B., Thomas Chaney, and Tarek A. Hassan (2019). "Migrants, Ancestors, and Foreign Investments." *Review of Economic Studies* 86:4, pp. 1448–1486.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo (2016). "Understanding Booms and Busts in Housing Markets." *Journal of Political Economy* 124:4, pp. 1088–1147.

Ciccarelli, Carlo and Stefano Fenoaltea (2013). "Through the Magnifying Glass: Provincial Aspects of Industrial Growth in Post-Unification Italy." *Economic History Review* 66:1, pp. 57–85.

Ciccarelli, Carlo and Peter Groote (2017). "Railway Endowment in Italy's Provinces, 1839–1913." *Rivista di Storia Economica* 33, pp. 45–88.

Cinel, Dino (1982). *From Italy to San Francisco: The Immigrant Experience*. Stanford: Stanford University Press.

Cohn, Raymond L. (2009). *Mass Migration Under Sail: European Immigration to the Antebellum United States*. New York: Cambridge University Press.

Comin, Diego A., Mikhail Dmitriev, and Esteban Rossi-Hansberg (2012). "The Spatial Diffusion of Technology." NBER Working Paper 18534.

Del Panta, Lorenzo (1997). "Infant and Child Mortality in Italy, Eighteenth to Twentieth Century: Long-term Trends and Territorial Differences." In *Infant and Child Mortality in the Past*. Alain Bideau, Bertrand Desjardins, and Héctor Pérez Brignoli (ed.). New York: Oxford University Press. Chap. 1, pp. 7–21.

Eichenbaum, Martin, Sergio Rebelo, and Mathias Trabant (2021). "The Macroeconomics of Epidemics." *Review of Financial Studies* 34:11, pp. 5149–5187.

Faini, Riccardo and Alessandra Venturini (1994). "Italian Emigration in the Pre-War Period." In *Migration and the International Labor Market 1850–1939*. Timothy J. Hatton and Jeffrey G. Williamson (ed.). London: Routledge, pp. 72–90.

Federico, Giovanni, Alessandro Nuvolari, and Michelangelo Vasta (2019). "The Origins of the Italian Regional Divide: Evidence from Real Wages, 1861–1913." *Journal of Economic History* 79:1, pp. 63–98.

Ferenczi, Imre and Walter F. Willcox (1929). *International Migrations*. New York: National Bureau of Economic Research.

Fernández-Sánchez, Martín (2021). "Mass Emigration and Human Capital over a Century: Evidence from the Galician Diaspora." Mimeo., LISER.

Foerster, Robert F. (1919). *The Italian Emigration of Our Times*. New York: Russell & Russell.

Fontana, Nicola, Marco Manacorda, Gianluca Russo, and Marco Tabllini (2021). "Emigration and Long-Run Economic Development: The Effects of the Italian Mass Migration." Mimeo., HBS.

Franck, Raphaël and Oded Galor (2022). "Technology-Skill Complementarity in Early Phases of Industrialization." *Economic Journal* 132:642, pp. 618–643.

Gabaccia, Donna (1984a). *From Sicily to Elizabeth Street: Housing and Social Change among Italian Immigrants, 1880–1930*. SUNY Press.

——— (1984b). "Kinship, Culture, and Migration: A Sicilian Example." *Journal of American Ethnic History* 3:2, pp. 39–53.

——— (1988). "The Transplanted: Women and Family in Immigrant America." *Social Science History* 12:3, pp. 243–253.

Gomellini, Matteo and Cormac Ó Gráda (2013). "Migrations." In *The Oxford Handbook of the Italian Economy Since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 10, pp. 271–302.

Gould, John D. (1980). "European Inter-Continental Emigration: The Role of 'Diffusion' and 'Feedback'." *Journal of European Economic History* 9:2, pp. 267–315.

Gray, Rowena, Gaia Narciso, and Gaspare Tortorici (2019). "Globalization, Agricultural Markets and Mass Migration: Italy, 1881–1912." *Explorations in Economic History* 74, 101276.

Handlin, Oscar (1951). *The Uprooted: The Epic Story of the Great Migrations that Made the American People*. Boston: Little Brown.

Hatton, Timothy J. and Jeffrey G. Williamson (1994). "Latecomers to Mass Emigration." In *Migration and the International Labor Market 1850–1939*. Timothy J. Hatton and Jeffrey G. Williamson (ed.). London: Routledge, pp. 55–71.

——— (1998). *The Age of Mass Migration: Causes and Economic Impact*. New York: Oxford University Press.

Hirschman, Albert O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations and States*. Cambridge: Harvard University Press.

Iorizzo, Luciano John (1966). *Italian Immigration and the Impact of the Padrone System*. Syracuse: Syracuse University Press.

Iuzzolino, Giovanni, Guido Pellegrini, and Gianfranco Viesti (2013). "Regional Convergence." In *The Oxford Handbook of the Italian Economy Since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 20, pp. 571–598.

Jovanovic, Boyan and Saul Lach (1989). "Entry, Exit, and Diffusion with Learning by Doing." *American Economic Review* 79:4, pp. 690–699.

Karadja, Mounir and Erik Prawitz (2019). "Exit, Voice and Political Change: Evidence from Swedish Mass Migration to the United States." *Journal of Political Economy* 127:4, pp. 1864–1925.

Kelley, Ninette and Michael Trebilcock (1998). *The Making of the Mosaic: A History of Canadian Immigration Policy*. Toronto: University of Toronto Press.

Kermack, William O. and Anderson G. McKendrick (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character* 115:772, pp. 700–721.

Koren, John (1897). *The Padrone System and Padrone Banks*. Washington: US Government Printing Office.

Lecce, Giampaolo, Laura Ogliari, and Tommaso Orlando (2022). "Resistance to Institutions and Cultural Distance: Brigandage in Post-Unification Italy." *Journal of Economic Growth* Forthcoming.
Lowell, Briant L. (1987). *Scandinavian Exodus: Demography and Social Development of 19th Century Rural Communities*. Boulder: Westview.
MacDonald, John S. (1963). "Agricultural Organization, Migration and Labour Militancy in Rural Italy." *Economic History Review* 16:1, pp. 61–75.
MacDonald, John S. and Leatrice MacDonald (1964). "Institutional Economics and Rural Development: Two Italian Types." *Human Organization* 23:2, pp. 113–118.
Mahajan, Parag and Dean Yang (2020). "Taken By Storm: Hurricanes, Migrant Networks, and US Immigration." *American Economic Journal: Applied Economics* 12:2, pp. 250–277.
McKenzie, David and Hillel Rapoport (2007). "Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico." *Journal of Development Economics* 84, pp. 1–24.
——— (2010). "Self-Selection Patterns in Mexico-US Migration: The Role of Migration Networks." *Review of Economics and Statistics* 92:4, pp. 811–821.
Mokyr, Joel (1983). *Why Ireland Starved: An Analytical and Quantitative Study of Irish Poverty, 1800–1851*. Boston: George Allen and Unwin.
Mokyr, Joel and Cormac Ó Gráda (1982). "Emigration and Poverty in Prefamine Ireland." *Explorations in Economic History* 19, pp. 360–384.
Moretti, Enrico (1999). "Social Networks and Migrations: Italy 1876–1913." *International Migration Review* 33:3, pp. 640–657.
Mormino, Gray Ross (1986 [2002]). *Immigrants on the Hill: Italian-Americans in St. Louis, 1882–1982*. Columbia: University of Missouri Press.
Moya, Jose C. (1998). *Cousins and Strangers: Spanish Immigrants in Buenos Aires, 1850–1930*. Berkeley: University of California Press.
Nelli, Humbert S. (1964). "The Italian Padrone System in the United States." *Labor History* 5:2, pp. 153–167.
——— (1967). "Italians in Urban America: A Study in Ethnic Adjustment." *International Migration Digest* 1:3, pp. 38–55.
O'Rourke, Kevin Hjortshøj and Jeffrey G. Williamson (1999). *Globalization and History: The Evolution of a Nineteenth-Century Atlantic Economy*. Cambridge: MIT Press.
Park, Robert E. and Herbert A. Miller (1921). *Old World Traits Transplanted*. New York: Harper & Brothers.
Peck, Gunther (2000). *Reinventing Free Labor: Padrones and Immigrant Workers in the North American West, 1880–1930*. Cambridge: Cambridge University Press.
Pizzo, Anthony P. (1981). "The Italian Heritage in Tampa." In *Little Italies in North America*. Robert F. Harney and Vincenza Scarpaci (ed.). Multicultural History Society of Ontario.
Silverman, Sydel F. (1968). "Agricultural Organization, Social Structure, and Values in Italy: Amoral Familism Reconsidered." *American Anthropologist* 70:1, pp. 1–20.
Sjaastad, Larry A. (1962). "The Costs and Returns of Human Migration." *Journal of Political Economy* 70:5, pp. 80–93.
Spitzer, Yannay (2021). "Pogroms, Networks, and Migration: The Jewish Migration from the Russian Empire to the United States 1881–1914." Mimeo., Hebrew University of Jerusalem.
Spitzer, Yannay, Gaspare Tortorici, and Ariell Zimran (2022). "International Migration Responses to Natural Disasters: Evidence from Modern Europe's Most Destructive Earthquake." NBER Working Paper 27506.
Spitzer, Yannay and Ariell Zimran (2018). "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925." *Journal of Development Economics* 134, pp. 226–247.
Spolaore, Enrico and Romain Wacziarg (2009). "The Diffusion of Development." *Quarterly Journal of Economics* 124:2, pp. 469–529.
Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression With Weak Instruments." *Econometrica* 65:3, pp. 557–586.
Sturino, Franc (1990). *Forging the Chain: A Case Study of Italian Migration to North America, 1880–1930*. Toronto: Multicultural History Society of Ontario.
Thistlethwaite, Frank (1960 [1991]). "Migration from Europe Overseas in the Nineteenth and Twentieth Centuries." In *A Century of European Migrations, 1830–1930*. R. J. Vecoli and S. M. Sinke (ed.). Urbana: University of Illinois Press. Chap. 1, pp. 17–57.
Todaro, Michael P. (1969). "A Model of Labor Migration and Urban Unemployment in Less Developed Countries." *American Economic Review* 59:1, pp. 138–148.

US Congress (1911a). *Reports of the Immigration Commission: Emigration Conditions in Europe*. Washington, DC, 61st Congress, 3rd Session, Senate Document No. 748: Government Printing Office.

——— (1911b). *Reports of the Immigration Commission: Emigration Conditions in Europe*. Vol. 4. Washington, DC, 61st Congress, 3rd Session, Senate Document No. 748: Government Printing Office.

——— (1911c). *Reports of the Immigration Commission: Statistical Review of Immigration 1820–1910—Distribution of Immigrants 1850–1900*. Vol. 3. Washington, DC, 61st Congress, 3rd Session, Senate Document No. 756: Government Printing Office.

Vecchi, Giovanni (2011). *In ricchezza e in povertà. Il benessere degli italiani dall'Unità ad oggi*. Bologna: Il Mulino.

Vecoli, Rudolph J. (1964). "Contadini in Chicago: A Critique of the Uprooted." *Journal of American History* 51:3, pp. 404–417.

——— (1983). "The Formation of Chicago's 'Little Italies'." *Journal of American Ethnic History* 2:2, pp. 5–20.

Ward, Zachary (2017). "Birds of Passage: Return Migration, Self-Selection and Immigration Quotas." *Explorations in Economic History* 64, pp. 37–52.

Williamson, Jeffrey G. (1995). "The Evolution of Global Labor Markets since 1830: Background Evidence and Hypotheses." *Explorations in Economic History* 32, pp. 141–196.

Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Yans-McLaughlin, Virginia (1975). *Italian Women and Work: Experience and Percception*. Center for the Historical Study of Societies, State University of New York at Binghamton.

——— (1977). *Family and Community: Italian Immigrants in Buffalo, 1880–1930*. Urbana and Chicago: University of Illinois Press.

Zeng, Donglin, Lu Mao, and D. Y. Lin (2016). "Maximum Likelihood Estimation for Semiparametric Transformation Models with Interval-Censored Data." *Biometrika* 103:2, pp. 253–271.

Zucchi, John E. (1985). "Italian Hometown Settlements and the Development of an Italian Community in Toronto, 1875–1935." In *Gathering Place: Peoples and Neighbourhoods of Toronto, 1834–1945*. Robert F. Harney (ed.). Toronto: Multicultural History Society of Ontario.

Tables

Table 1: Summary statistics for time-varying variables

	District	Commune						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	All	All	1884–1890	1891–1895	1896–1900	1901–1905	1906–1910	1911–1914
Any Emigration	0.990	0.772	0.582	0.566	0.674	0.969	1.000	0.995
	(0.100)	(0.420)	(0.493)	(0.496)	(0.469)	(0.173)	(0.000)	(0.070)
Emigration Rates (per k)								
All Destinations	15.564	15.674	9.509	10.039	11.518	22.896	26.373	27.480
	(21.615)	(22.899)	(20.080)	(21.942)	(24.335)	(26.095)	(20.669)	(24.590)
North America	3.334	4.805	1.446	1.502	2.009	7.184	9.035	8.935
	(6.339)	(8.843)	(4.382)	(4.114)	(5.042)	(11.952)	(11.166)	(10.706)
South America	2.658	3.249	3.151	3.660	3.387	3.903	4.472	3.478
	(3.882)	(5.068)	(5.530)	(6.852)	(5.284)	(5.173)	(4.864)	(3.717)
Europe	9.173	8.448	4.313	4.669	5.919	11.273	12.450	14.622
	(20.045)	(19.038)	(15.955)	(18.314)	(21.940)	(21.908)	(17.203)	(21.770)
Mass Emigration								
All destinations (>10 per thousand)	0.464	0.446	0.272	0.276	0.310	0.613	0.772	0.770
	(0.499)	(0.497)	(0.445)	(0.447)	(0.463)	(0.487)	(0.420)	(0.421)
North America (> 5 per thousand)	0.184	0.246	0.084	0.093	0.125	0.312	0.426	0.452
	(0.388)	(0.431)	(0.278)	(0.290)	(0.331)	(0.463)	(0.494)	(0.498)
South America (> 5 per thousand)	0.164	0.217	0.216	0.237	0.241	0.277	0.305	0.228
	(0.370)	(0.412)	(0.412)	(0.425)	(0.428)	(0.448)	(0.460)	(0.420)
Europe (> 5 per thousand)	0.327	0.307	0.134	0.132	0.163	0.417	0.499	0.521
	(0.469)	(0.461)	(0.341)	(0.338)	(0.369)	(0.493)	(0.500)	(0.500)
Within North American Frontier	0.173	0.194	0.031	0.069	0.095	0.132	0.281	0.366
	(0.379)	(0.395)	(0.173)	(0.253)	(0.293)	(0.339)	(0.450)	(0.482)
Distance to North American Frontier (km)	288.062	246.347	410.926	392.245	391.055	349.372	74.415	53.385
	(241.148)	(243.451)	(221.203)	(238.929)	(240.525)	(226.968)	(73.068)	(70.924)
Observations	2,545	56,083	7,909	8,029	8,029	8,029	8,029	8,029
Units	284	8,029	7,909	8,029	8,029	8,029	8,029	8,029

Notes: Observations are at the district-half decade level in column (1) and at the commune-half decade level in columns (2)–(8). Standard deviations in parentheses.

Table 2: Summary statistics for time-invariant variables

Variable	(1) All	(2) North	(3) Center	(4) South
Panel A: District-level Data				
District Share of Male Labor in Agriculture	0.547	0.555	0.547	0.534
	(0.109)	(0.101)	(0.135)	(0.109)
District Share of Male Labor in Industry	0.216	0.237	0.191	0.192
	(0.070)	(0.074)	(0.058)	(0.056)
District Adult Male Literacy Rate	0.468	0.582	0.453	0.279
	(0.183)	(0.149)	(0.127)	(0.055)
District Population Fraction Under Age 15	0.328	0.334	0.309	0.327
	(0.024)	(0.024)	(0.026)	(0.018)
Observations	284	154	41	89
Panel B: Commune-level Data				
Distance to Railroad (1881, km)	9.853	8.871	7.992	12.486
	(12.437)	(12.178)	(9.209)	(13.758)
Mean Elevation (m)	451.418	471.039	390.918	445.771
	(425.914)	(503.853)	(260.907)	(324.431)
Distance to North America Epicenter (km)	149.796	162.926	138.351	132.067
	(90.280)	(68.928)	(78.142)	(120.294)
Distance to South America Epicenter (km)	168.408	124.097	219.139	222.421
	(101.971)	(49.260)	(107.815)	(127.637)
Distance to European Border (km)	240.432	45.426	238.222	586.443
	(264.104)	(37.066)	(122.753)	(173.120)
Distance to Coast (km)	68.738	106.680	30.388	20.045
	(57.649)	(50.310)	(26.493)	(18.982)
Observations	8,028	4,371	1,187	2,470

Notes: Observations are at the district level in Panel A and at the commune level in Panel B. Standard deviations in parentheses. Observation numbers are the minimum with observations for all variables, excluding places whose emigration rates cannot be calculated due to a lack of population data. Distance from railroad is 0 for any commune with a rail line passing through it in 1881, and the distance from the nearest commune border to the rail line for all other communes. Distance to the epicenters of North America- and South America-bound emigration are from the commune centroid to the centroid of the epicenter district's capital city. Distance to the European border is from the commune centroid.

Table 3: β-convergence

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.521^a	-0.738^a	-0.833^a	-0.856^a	-0.514^a	-0.793^a	-0.828^a	-0.909^a
	(0.021)	(0.017)	(0.011)	(0.012)	(0.043)	(0.063)	(0.158)	(0.210)
Observations	5,856	5,855	5,855	5,855	5,856	5,855	5,855	5,849
R-squared	0.592	0.768	0.881	0.914	0.592	0.765	0.799	0.822
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	55.107	90.635	11.503	7.393
FE	None	None	P	D	None	None	P	D
Falsification	−0.150	−0.010	0.192	0.187				
	(0.054)	(0.072)	(0.064)	(0.056)				

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune. Dependent variable is the change in the log of the emigration rate to North America from the pre-1900 period to the period 1900 and later. Controls include latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Instrument is the distance to the nearest epicenter of emigration to North America. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the post-1900 period; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

Table 4: Survival time regressions

Variables	(1)	(2)	(3)
Distance to North American epicenter	0.583^a	0.655^a	0.647^a
	(0.054)	(0.052)	(0.061)
Observations	284	284	284
Broad region FE	No	Yes	Yes
Controls	No	No	Yes

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: This table presents estimated hazard ratios for semiparametric Cox proportional hazard models for the timing of entry into the frontier of mass migration to North America. Hypothesis testing is relative to a null hypothesis of a hazard ratio of 1. All variables are standardized to have mean 0 and standard deviation 1. The unit of observation is a district. The data run from the 1876–1880 half decade to the 1916–1920 half decade. The date of entering the mass migration frontier is intervaled by half decade. Broad region fixed effects are for the center and south (with the north as the excluded category). Controls are a district's share of agricultural employment, share of industrial employment, literacy rate, and fraction under age 15, all in 1881, as well as its mean elevation.

Table 5: Survival time regressions

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Distance to epicenter	0.524^a					0.501^a	0.505^a
	(0.039)					(0.058)	(0.060)
Ag. labor share		1.238^a				1.089	1.089
		(0.075)				(0.078)	(0.078)
Fraction under 15			1.381^a			1.172^a	1.160^b
			(0.080)			(0.066)	(0.069)
Literacy rate				1.355^a		0.839^b	0.796^b
				(0.076)		(0.071)	(0.087)
Share Urban 1881					0.711^a	0.756^a	0.767^a
					(0.055)	(0.071)	(0.076)
Observations	284	284	284	284	284	284	284
Broad region FE	No	No	No	No	No	No	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: This table presents estimated hazard ratios for semiparametric Cox proportional hazard models for the timing of entry into the frontier of mass migration to any destination. Hypothesis testing is relative to a null hypothesis of a hazard ratio of 1. All variables are standardized to have mean 0 and standard deviation 1. The unit of observation is a district. The data run from the 1876–1880 half decade to the 1916–1920 half decade. The date of entering the mass migration frontier is intervaled by half decade. Broad region fixed effects are for the center and south (with the north as the excluded category).

Table 6: Spatial contagion results, epicenter-based IV

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.956^a	0.888^a	0.670^a	0.657^a	0.602^a	0.474^c	0.498	0.166
	(0.104)	(0.122)	(0.177)	(0.214)	(0.153)	(0.243)	(0.472)	(0.453)
Observations	31,463	31,463	31,463	31,463	31,463	31,462	31,463	31,427
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	30.71	42.06	18.83	12	21.50	8.649	22.73	6.589

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table 7: Spatial contagion results, frontier-based IV

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.942^a	0.887^a	0.642^a	0.759^a	0.667^a	0.636^a	0.613	0.412^c
	(0.136)	(0.160)	(0.159)	(0.166)	(0.175)	(0.162)	(0.583)	(0.232)
Observations	11,207	11,206	11,206	11,205	11,206	11,196	11,206	11,170
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	64.51	53.08	59.57	45.29	48.95	44.71	19.26	27.80

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Figures

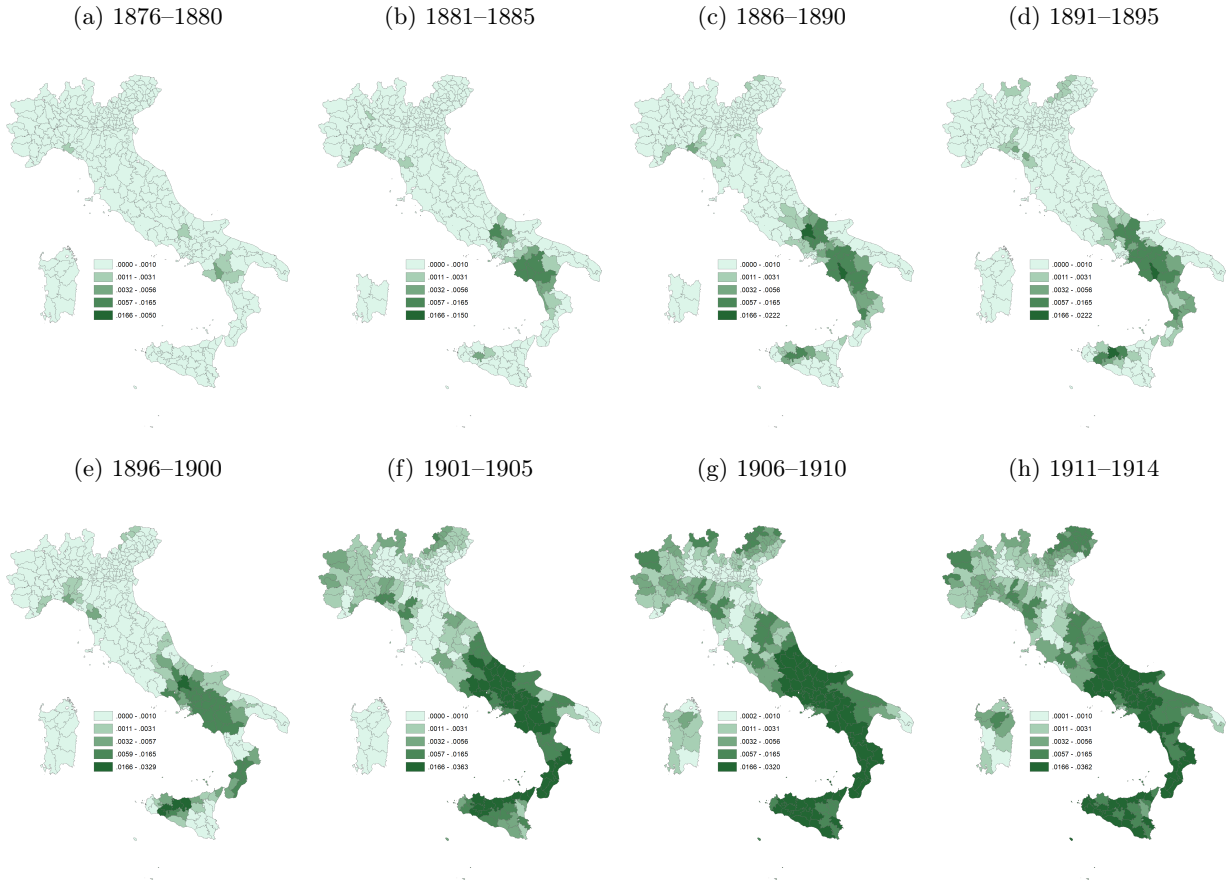


Figure 1: District-level emigration rates to North America

Note: Each panel presents a district's average annual emigration rate to North America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have greater emigration rates.

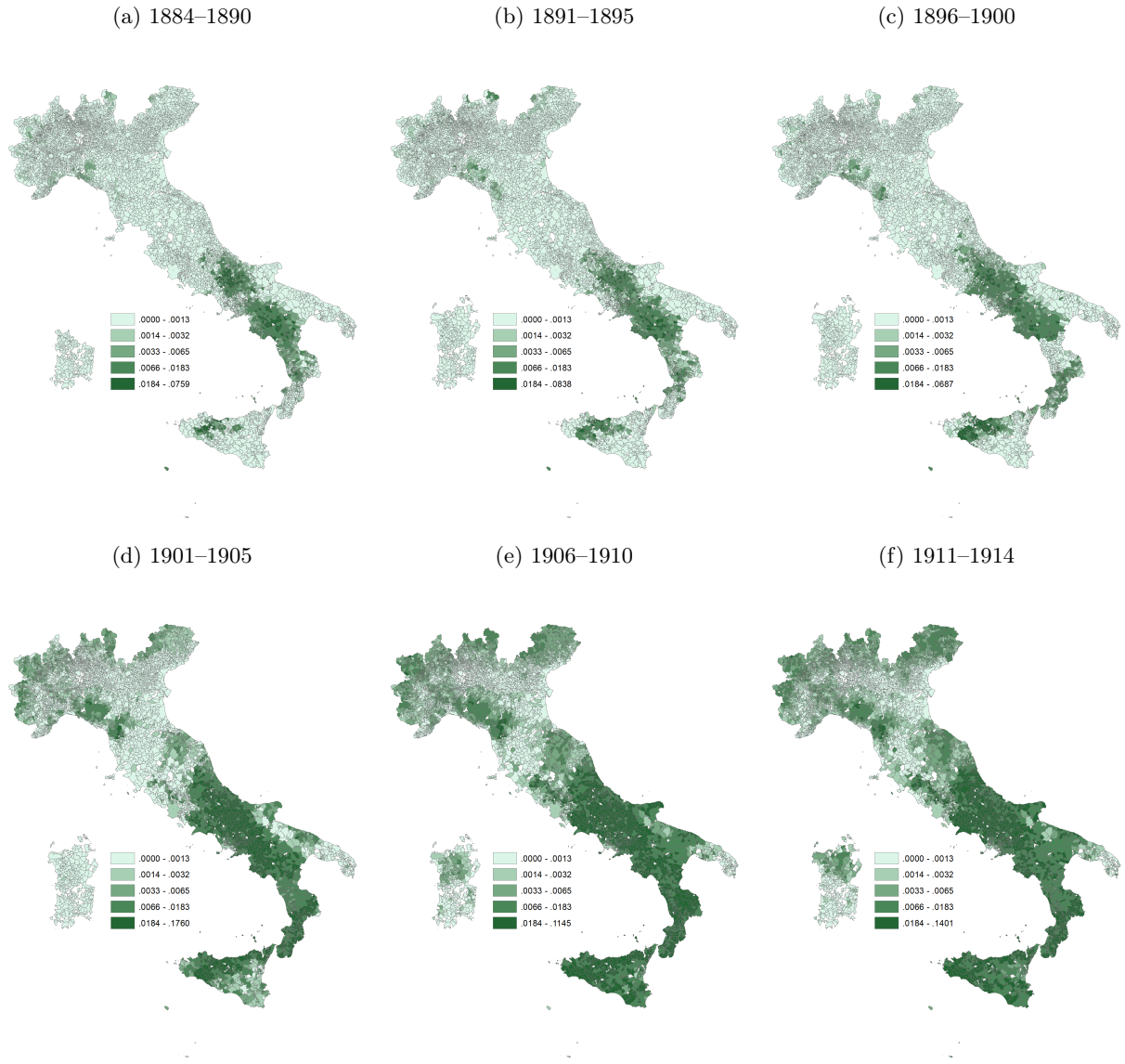


Figure 2: Commune-level emigration rates to North America

Note: Each panel presents a commune's average annual emigration rate to North America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have greater emigration rates.

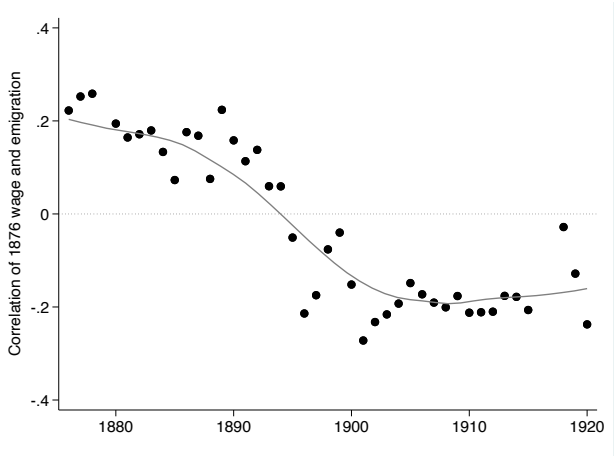


Figure 3: Correlation of province-level emigration and 1876 wages

Source: Wage data are from Federico, Nuvolari, and Vasta (2019).

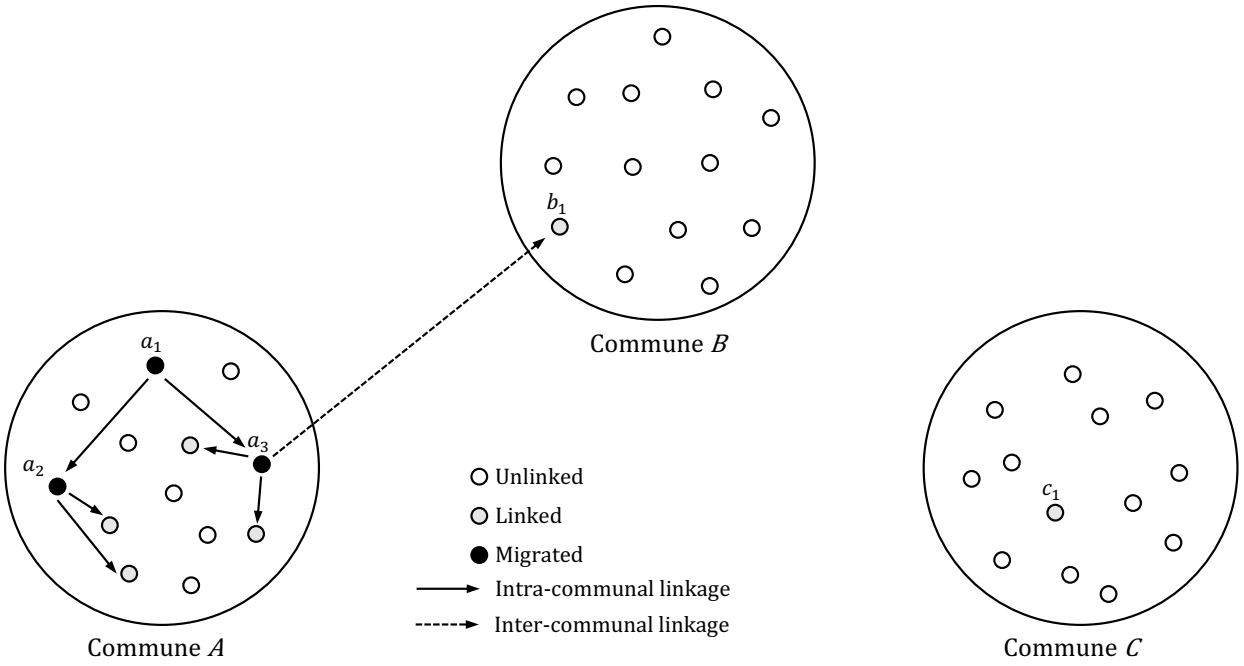
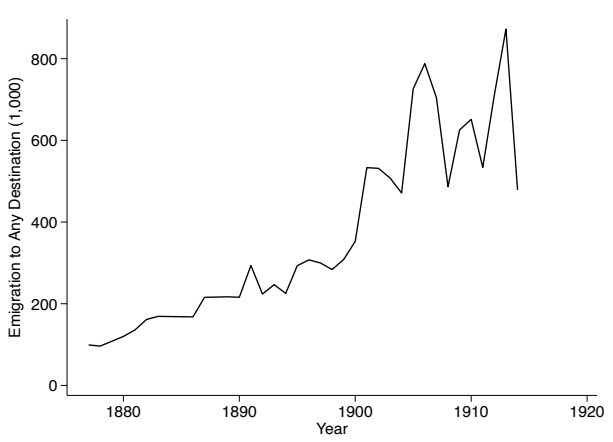


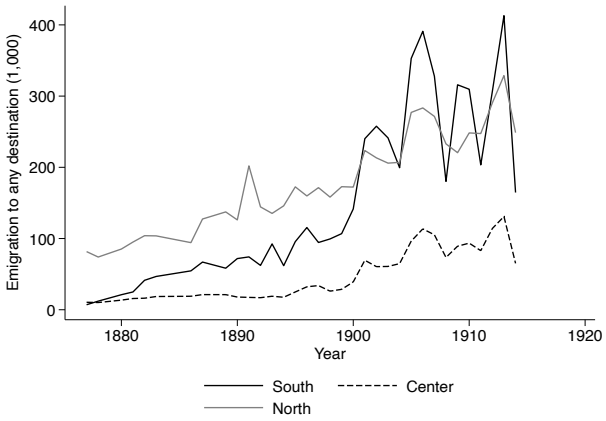
Figure 4: Intra- and inter-communal transmission of the migration technology

Note: See explanation in section 3. Arrows indicate the direction of the diffusion of the migration "technology," not the direction of linkage.

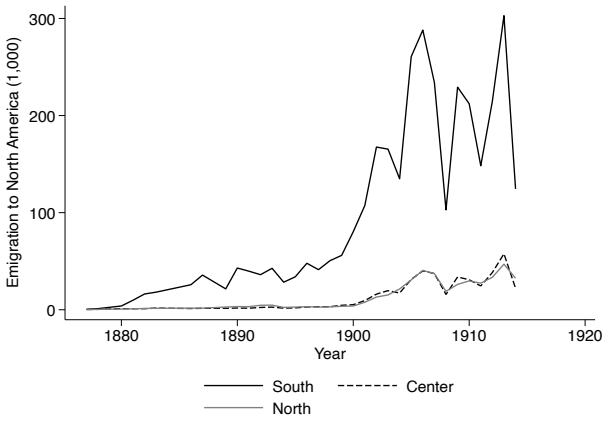
(a) All Destinations



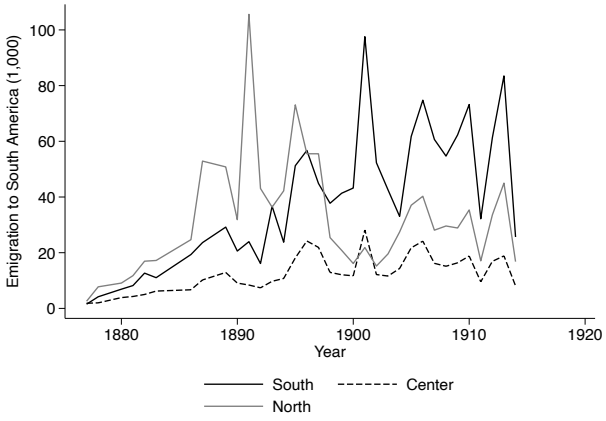
(b) All Destinations



(c) North America



(d) South America



(e) Europe

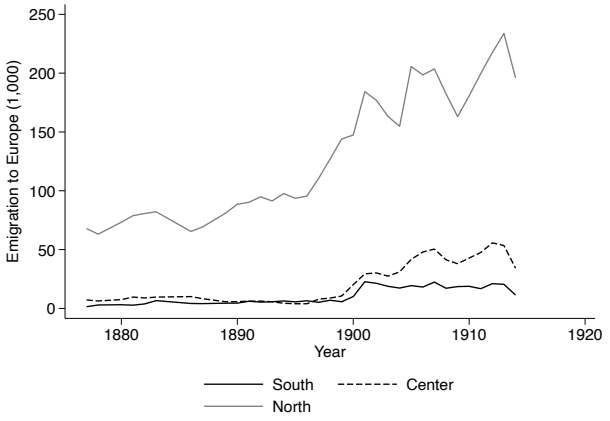


Figure 5: Emigration by origin and destination, 1876–1914

Note: These figures are based on our province-by-destination data. South includes the regions of Abruzzo, Campania, Puglia, Basilicata, Calabria, Sicilia, and Sardinia. Center includes the regions of Liguria, Toscana, Marche, Umbria, and Latium. North includes the regions of Piemonte, Lombardia, Veneto, and Emilia Romagna.

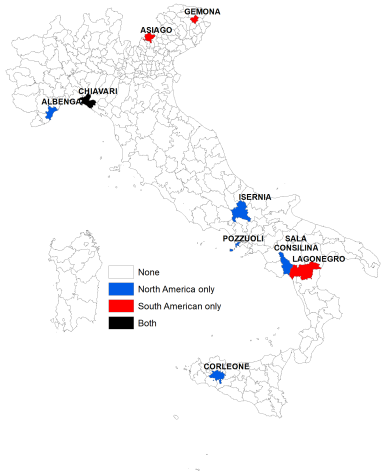


Figure 6: Epicenters of mass migration to North and South America

Note: The districts in blue (Albenga, Isernia, Pozzuoli, Sala Consilina, and Corleone) are epicenters for North American migration only. The districts in red (Asiago, Gemona, and Lagonegro) are epicenters for South American migration only. Chiavari (in black) is an epicenter for migration to both North America and South America. Epicenters are defined as described in section 4.2.

(a) North America (b) South America (c) Europe

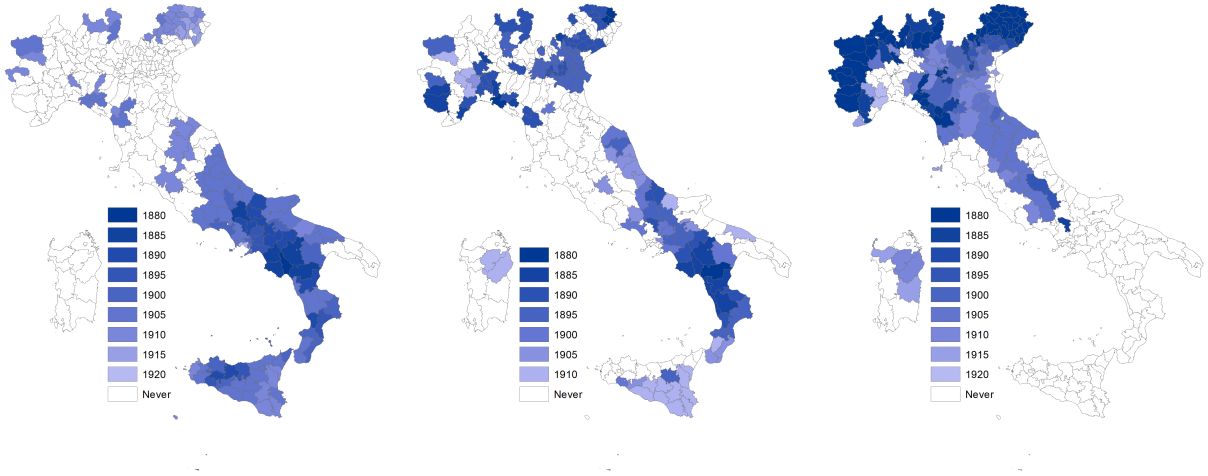


Figure 7: Frontiers of mass migration by destination

Note: Districts are shaded according to the half decade in which they first achieved an average annual emigration rate to the listed destination of at least 5 per thousand. Darker districts entered the frontier earlier.

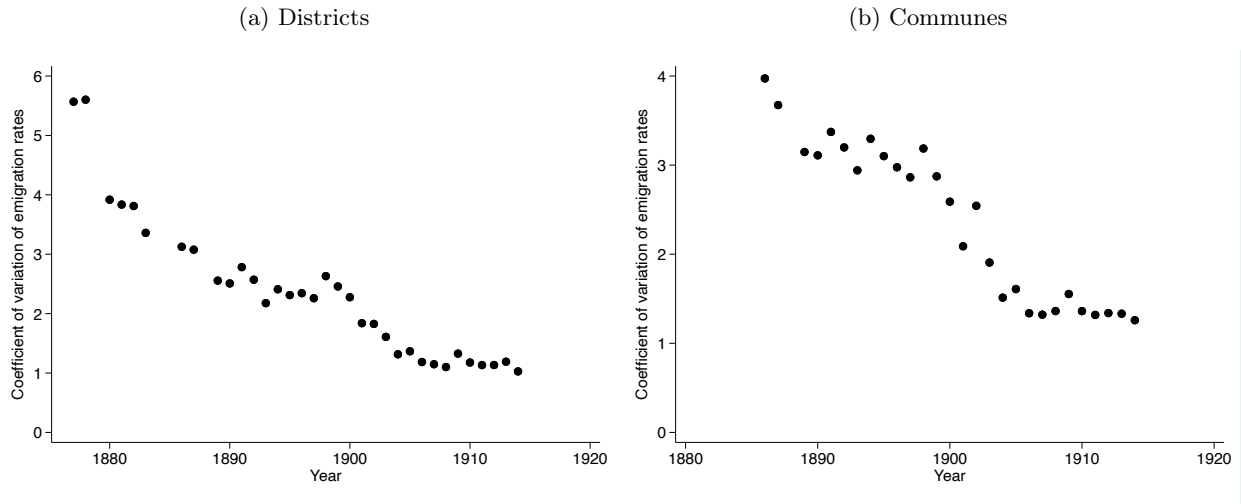


Figure 8: σ-convergence in emigration rates to North America

Note: Each point represents the coefficient of variation in emigration rates to North America in a particular year.

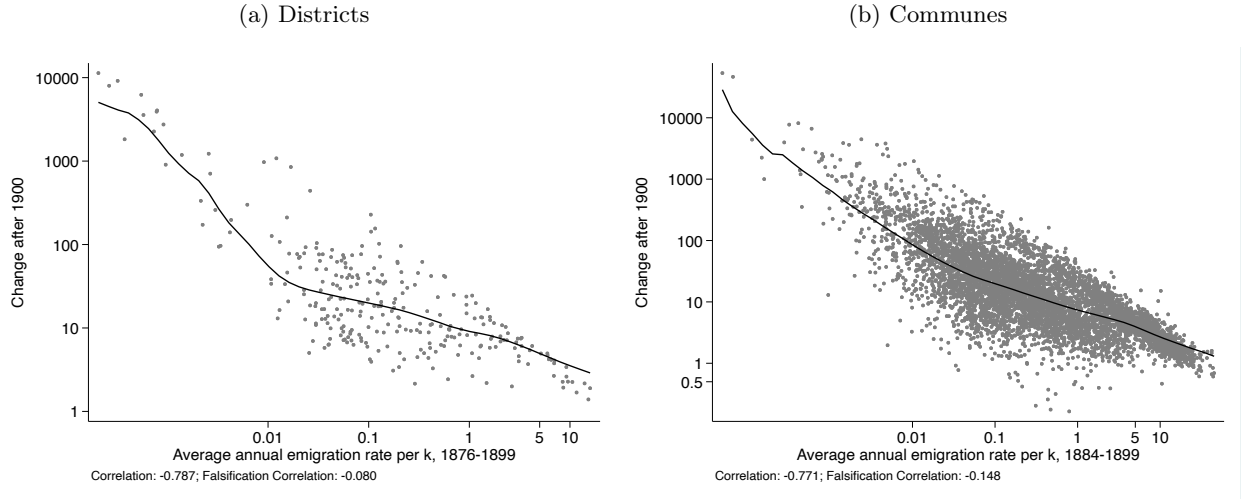


Figure 9: β-convergence in emigration rates to North America

Note: Each point represents a commune or district. The x-axis is the average annual emigration rate for a district for 1876–1899 or a commune for 1884–1899 on a log scale. The y-axis is the ratio of the average emigration rate before and after 1900, also on a log scale. The falsification correlation is the correlation of the change in emigration and emigration after 1900; that it is not positive indicates that the negative relationship shown in the graphs is unlikely to be spurious, as explained in section 5.1.

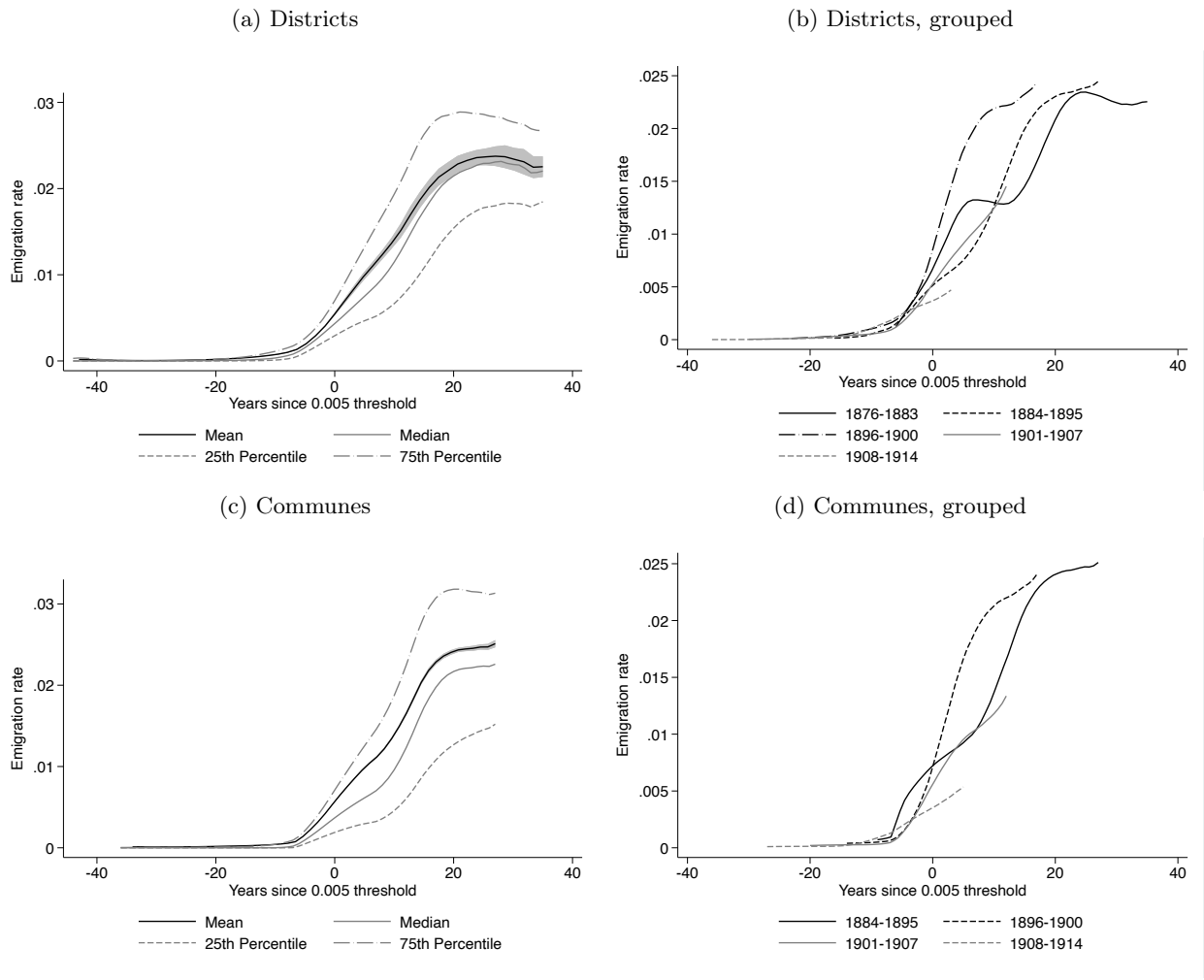
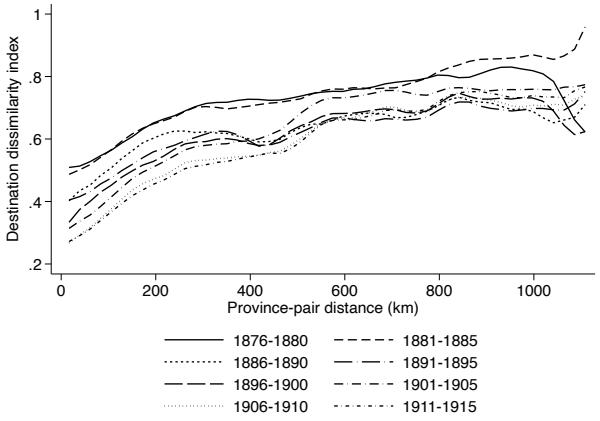


Figure 10: S-shaped time series of migration to North America

Note: Panels (a) and (c) plot a non-parametric regression (the mean), as well as quartiles of emigration rates to North America against time, normalized so that year 0 is the first year in which a place had an emigration rate of at least 5 per thousand. Shaded areas are 95-percent confidence intervals for the mean. Panels (b) and (d) are the same as (a) and (c) but divide areas according to the half decade in which they crossed the threshold.

(a) Nonparametrics



(b) OLS with Half-Decade FE

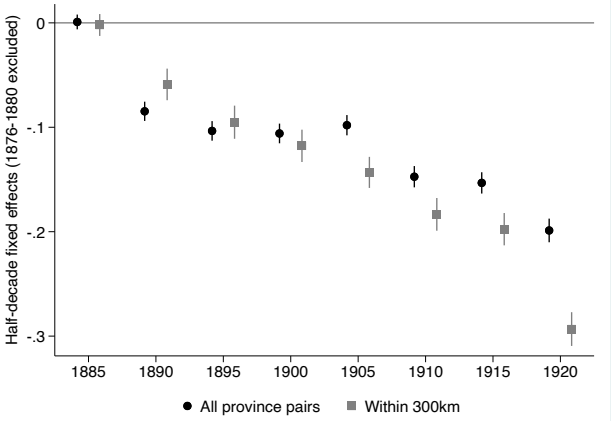


Figure 11: Destination dissimilarity by distance and half decade

Note: Panel (a) plots non-parametric regressions for each half decade of the dissimilarity index between two provinces' emigration and the distance between them. Panel (b) plots half-decade fixed effects from a regression of dissimilarity on province-pair distance and these fixed effects, excluding the 1876–1880 half decade.

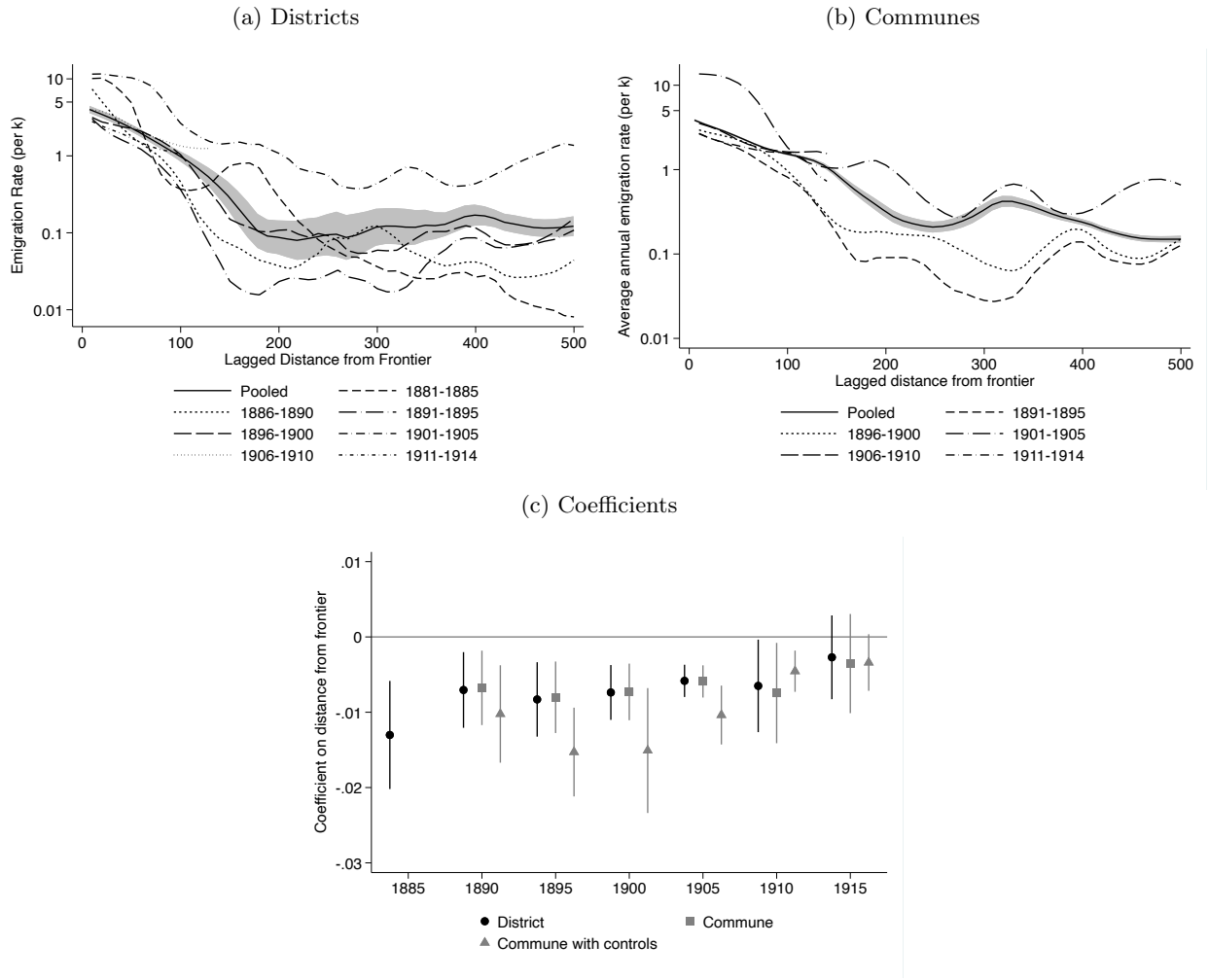


Figure 12: Emigration rates to North America by distance to the mass migration frontier (km)

Note: Panels (a) and (b) present non-parametric regressions of the log of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to districts that had not yet achieved this threshold. Shaded areas are 95-percent confidence intervals. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of lagged distance from the frontier of mass migration to North America and plots the coefficients on lagged distance from the frontier. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

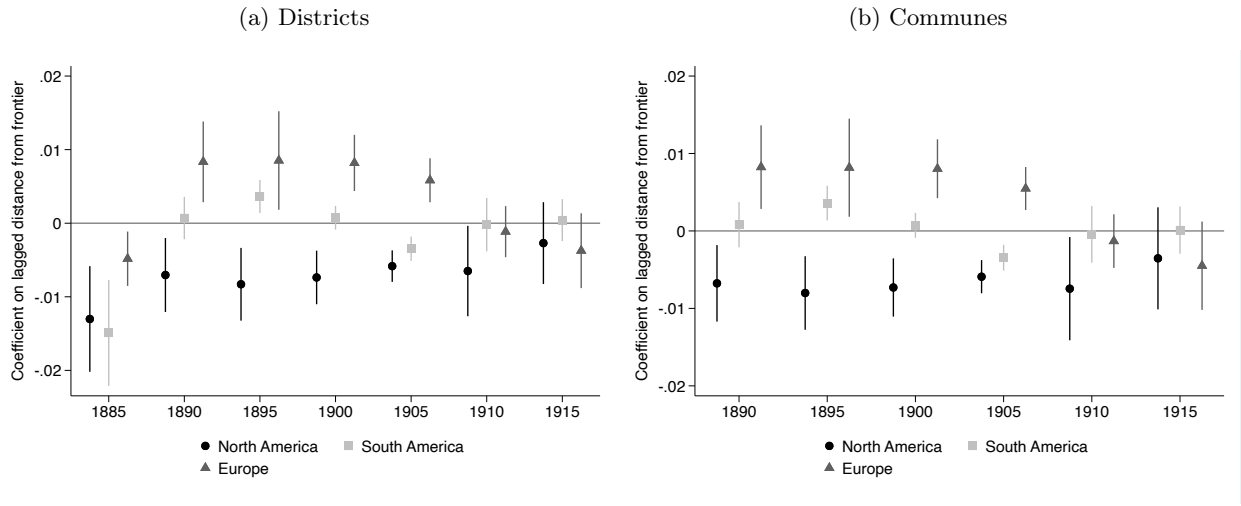


Figure 13: Emigration to various destinations by distance to the mass migration frontier for North America (km)

Note: This figure repeats the binomial maximum likelihood regressions of panel (c) of Figure 12, but includes results for migration to South America and Europe in addition to those for migration to North America.

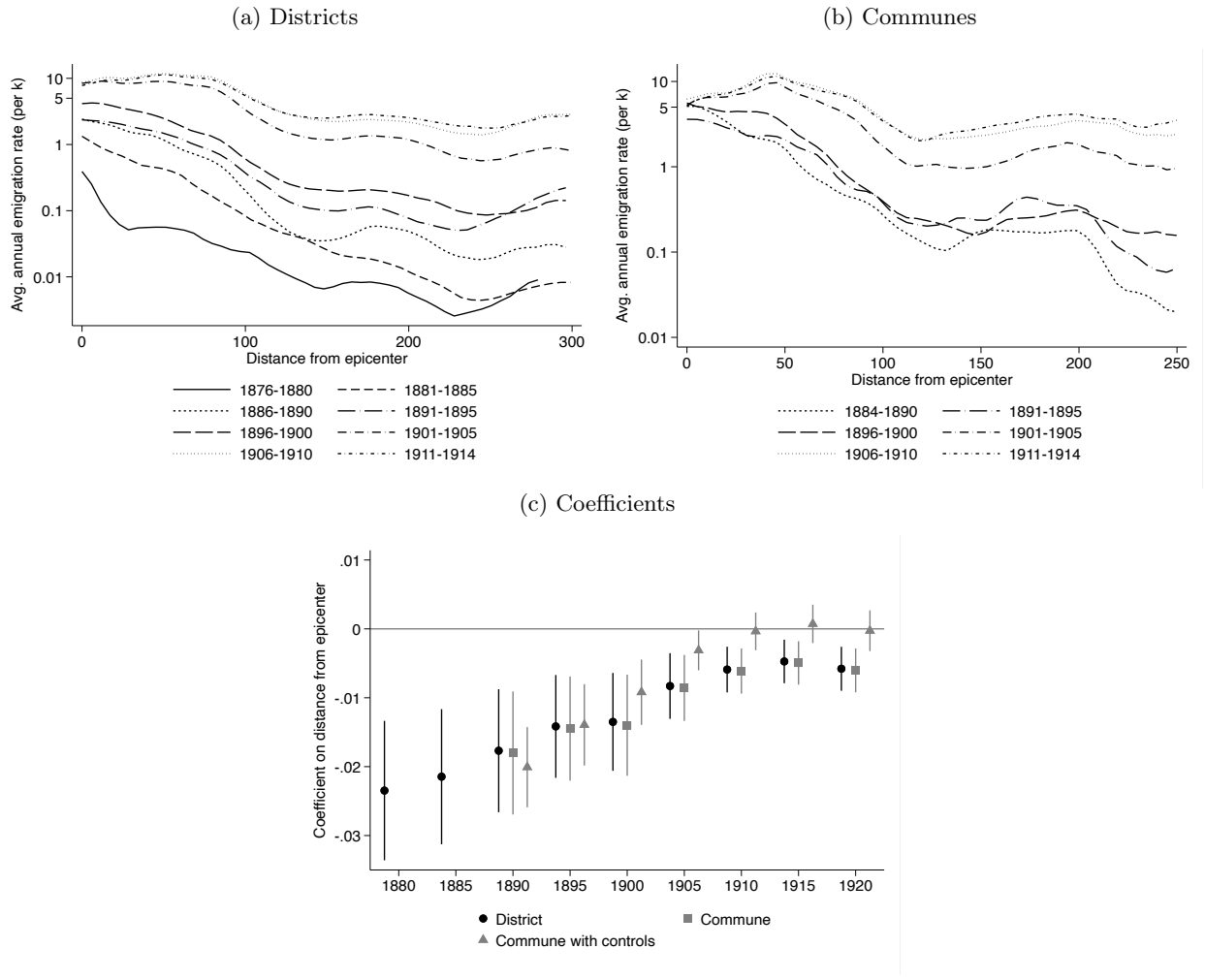


Figure 14: Emigration rates to North America by distance to epicenter (km)

Note: Panels (a) and (b) plot non-parametric regressions of the log of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration to North America. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of distance from the nearest epicenter of emigration to North America and plots the coefficients on distance from epicenter. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

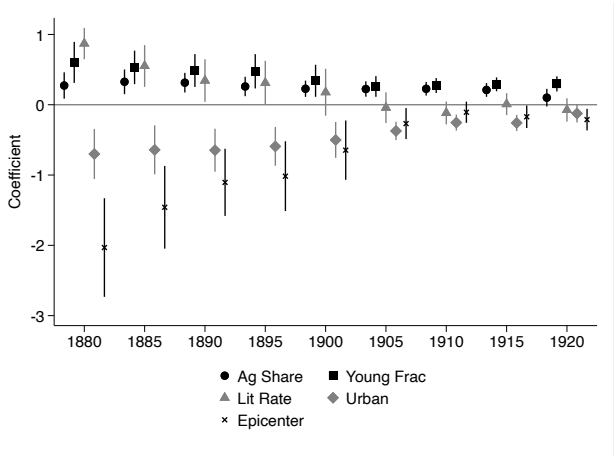


Figure 15: Relationship of migration to various local characteristics

Note: This figure presents the results of a regression of emigration to any destination on year-specific functions of various district characteristics. All explanatory variables are standardized to have mean zero and standard deviation one.

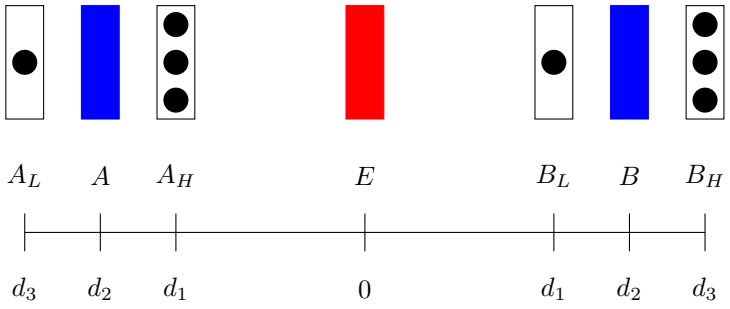


Figure 16: Illustration of the identification strategy

Note: Rectangles indicate communes and circles indicate population. The population of communes A and B and of the epicenter commune E are unimportant to the example and are not specified. Communes A_H and B_H are more populous than communes A_L and B_L, respectively, as indicated by the greater number of circles within the former than the latter. The number line indicates each commune's distance from the epicenter commune E; for instance, communes A and B are both at distance d_2 from the epicenter.

A Binomial Maximum Likelihood Regression

In Figures 14, 12, and 13, we estimate a regression in which commune or district i's emigration rate in period t, p_{it}, is expressed in the logit form

$$p_{it} = \frac{\exp(\nu_t + \eta_t z_i + \mathbf{x}_i'\gamma_t)}{1 + \exp(\nu_t + \eta_t z_i + \mathbf{x}_i'\gamma_t)},$$

where z_i is commune i's distance from the nearest epicenter of emigration to North America, ν_t is a period-specific intercept, η_t and γ_t are period-specific coefficients, and $\mathbf{x}_i$ are controls. This method is intended to address observations of zero migration by treating these as cases in which all individuals have a strictly positive migration probability, p_{it}, but the realization of every resident of the commune is to stay. After determining this logit migration demand, we use the binomial distribution to determine the probability that a given number of emigrants are observed from commune i in time period t given p_{it} and the commune's baseline population N_i, which enables us to estimate ν_t, η_t, and γ_t by maximum likelihood. The log-likelihood function after removing constants is

$$\mathfrak{L} = \sum_i \sum_t e_{it} N_i \log(p_{it}) + (1 - e_{it}) N_i \log(1 - p_{it}),$$

and the model is estimated by maximum likelihood.

Online Appendix for

Like an Ink Blot on Paper
Testing the Diffusion Hypothesis of Mass Migration, Italy 1876–1920

Yannay Spitzer
The Hebrew University of Jerusalem & CEPR

Ariell Zimran
Vanderbilt University & NBER

September 30, 2022

B. Additional Tables and Figures 67
C. The Migration of Antonio Squadrito's Group in October 1903 83
D. Model Simulations 86
E. Preparing Official Statistics for Analysis 87
F. Results Including Data on "Other Communes" 89
G. Results with 1881 Population 94
H. Results for Migration to All Destinations 101
I. List of Archival Sources 107
J. Results Including Observations with No Migration 108
K Robustness to Choice of θ 111
L Results With Grid Fixed Effects 114

B Additional Tables and Figures

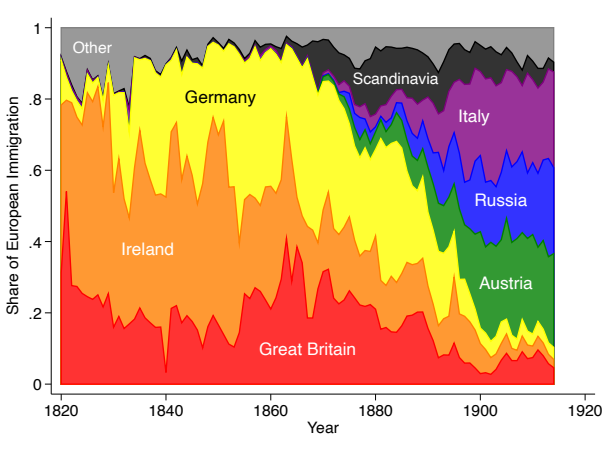


Figure B.1: Distribution of origin countries for US immigration from Europe

Source: Barde, Carter, and Sutch (2006)

Note: This graph shows the share of European immigration to the United States coming from each source country. The "Austria" data are from Barde, Carter, and Sutch's (2006) data for "Other Central Europe," which cover Central Europe other than Germany and Poland.

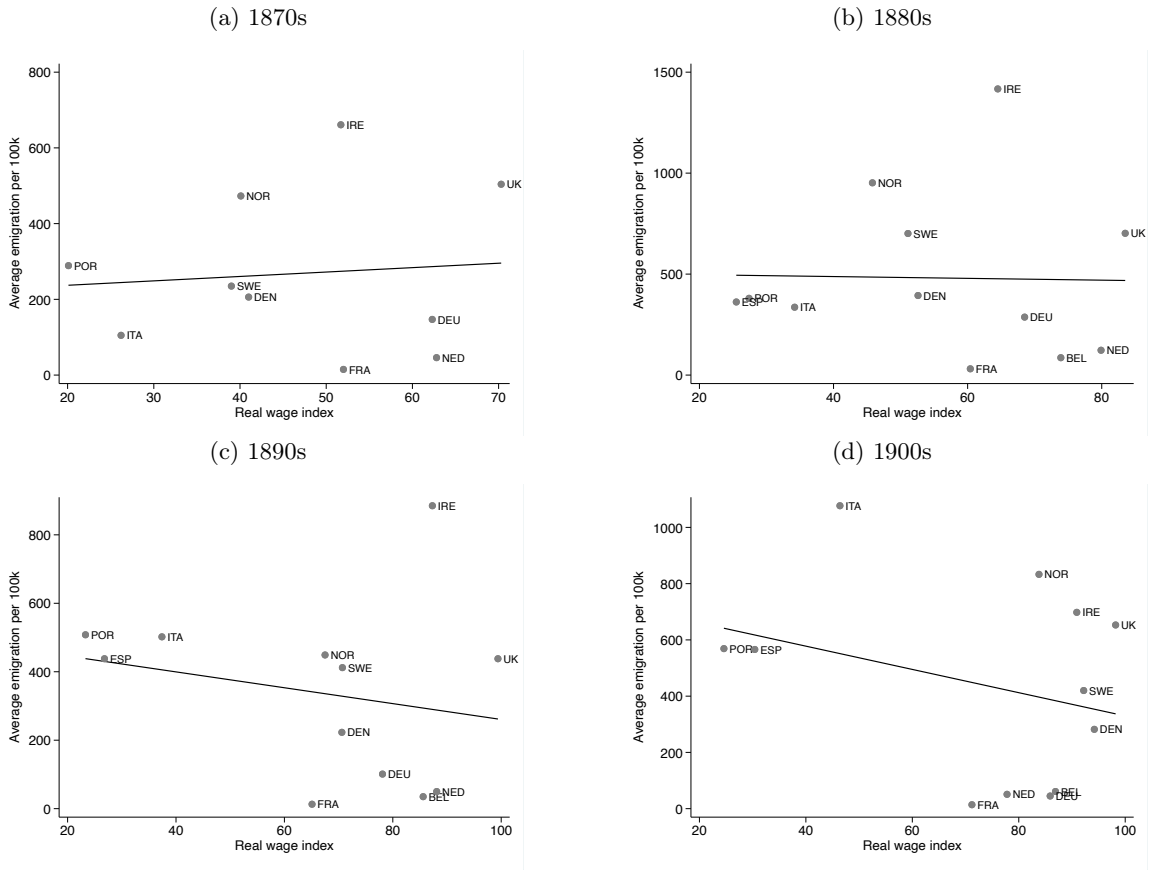


Figure B.2: Emigration and real wages

Source: Emigration data are from Ferenczi and Willcox (1929). Wage data are from Hatton and Williamson (1998).

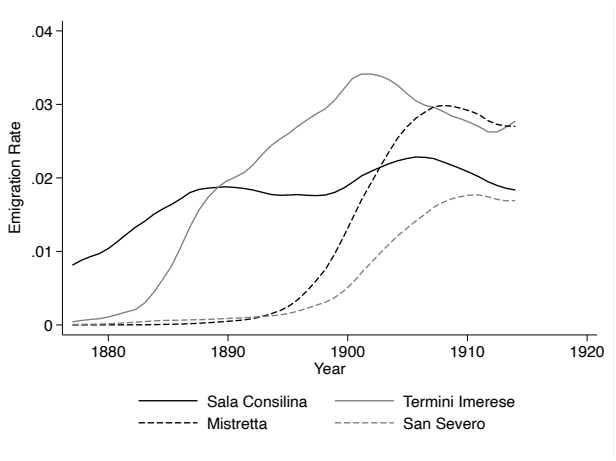


Figure B.3: Time series of emigration to North America from selected districts

Note: The four districts in this figure are Sala Consilina in Salerno, Termini Imerese in Palermo, Mistretta in Messina, and San Severo in Foggia. The epicenter district of Sala Consilina was selected because it had the highest emigration rate to North America in the period 1876–1883. The remaining three districts were selected because their estimated pre-1884 emigration rates as implied by their observables were among the most similar to that predicted for Sala Consilina. Some discretion was exercised in the choice of these example districts for purposes of exposition. The time series are smoothed using a local linear regression. The main takeaway in this figure is that, even though the four districts were observationally very similar, they experienced very different time series of emigration, surging into S-shapes in order of their distance from the nearest epicenter (not necessarily Sala Consilina).

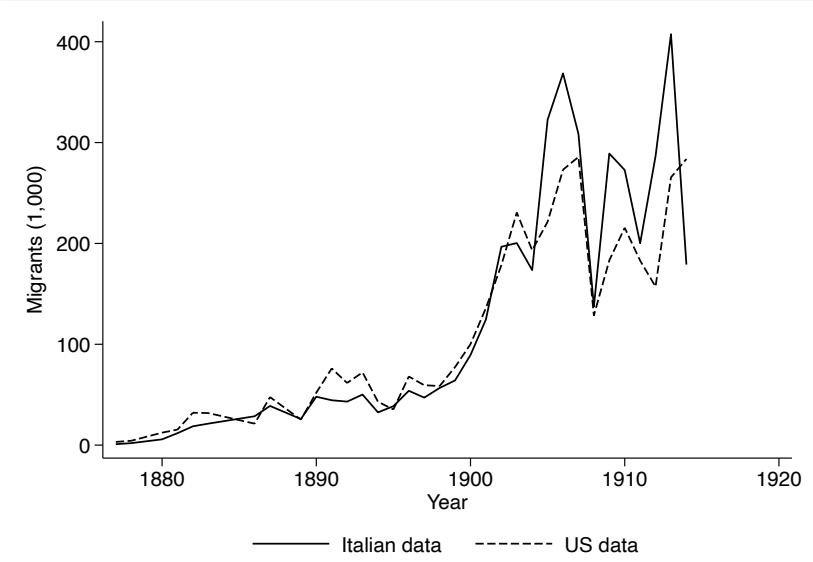


Figure B.4: Comparison of Italian emigration data and US immigration data

Note: The Italian data are for North America-bound emigrants from our transcriptions of the *Statistica della Emigrazione Italiana per l'Estero* and are based on calendar years. The US data are for immigrants arriving from Italy from Barde, Carter, and Sutch (2006) and are based on fiscal years.

Table B.1: β-convergence, half decades

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.226^a	-0.376^a	-0.542^a	-0.598^a	-0.214^a	-0.398^a	-0.534^a	-0.381^a
	(0.011)	(0.014)	(0.013)	(0.013)	(0.022)	(0.043)	(0.046)	(0.049)
Observations	36,329	36,326	36,326	36,326	36,329	36,326	36,326	36,326
R-squared	0.520	0.621	0.678	0.698	0.162	0.338	0.405	0.370
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	25.907	20.563	21.883	31.579
FE	None	None	P	D	None	None	P	D
Falsification	0.101	0.225	0.414	0.478				
	(0.015)	(0.017)	(0.017)	(0.017)				

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune-half decade. Dependent variable is the change in the log of the emigration rate to North America from one half deade to the next. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. All specifications include half-decade fixed effects. Instrument is a time-specific function of the logarithm of distance to the nearest epicenter of emigration to North America. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the current (rather than last period's) emigration; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

Table B.2: Destination dissimilarity and distance between provinces

Variables	(1) All	(2) All	(3) Major	(4) Minor
log(Distance)	0.134^a	0.121^a	0.114^a	0.072^a
	(0.003)	(0.004)	(0.004)	(0.004)
Observations	21,114	21,114	21,114	21,114
R-squared	0.303	0.310	0.217	0.122
Controls	No	Yes	Yes	Yes

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Dependent variable is the dissimilarity index in the emigration destination distribution of the two provinces making up a province pair in a given half decade. Unit of observation is a province pair-half decade. Standard errors clustered by province pair. All regressions include half-decade fixed effects. Major destinations are US, Canada, France, Argentina, Uruguay, Switzerland, Austria-Hungary, Germany, and Brazil. Controls are absolute differences in agricultural and industrial employment shares, literacy rates, and fraction under age 15.

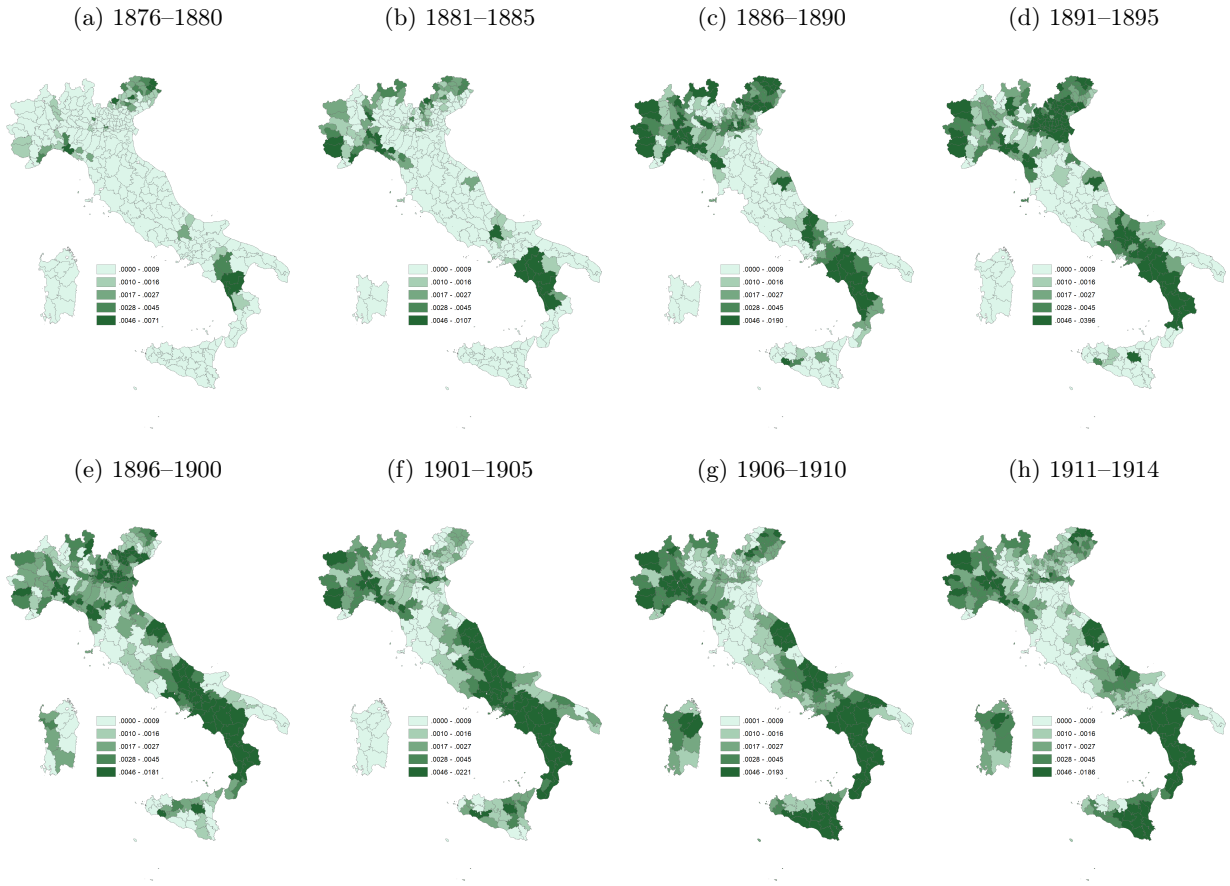


Figure B.5: District-level emigration rates to South America

Note: Each panel presents a district's average annual emigration rate to South America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

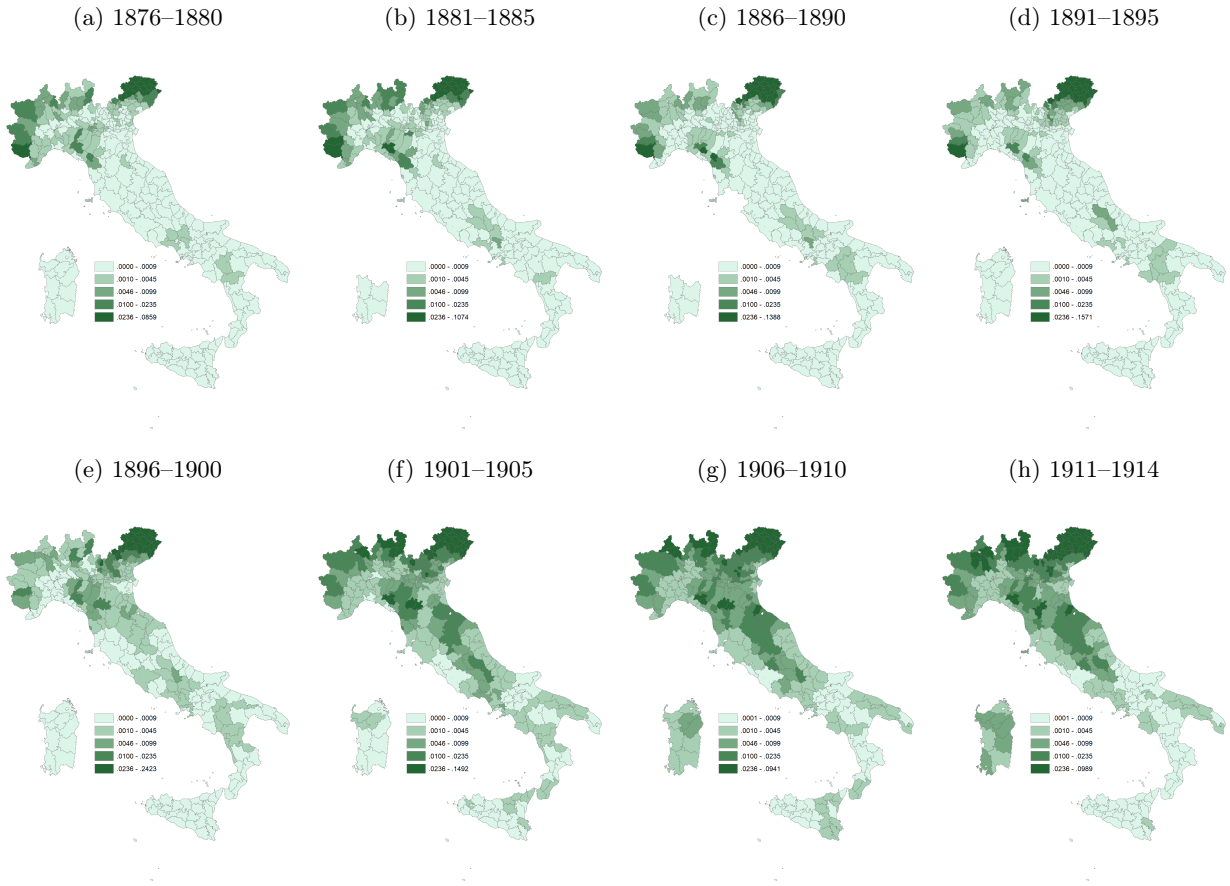


Figure B.6: District-level emigration rates to Europe

Note: Each panel presents a district's average annual emigration rate to Europe in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

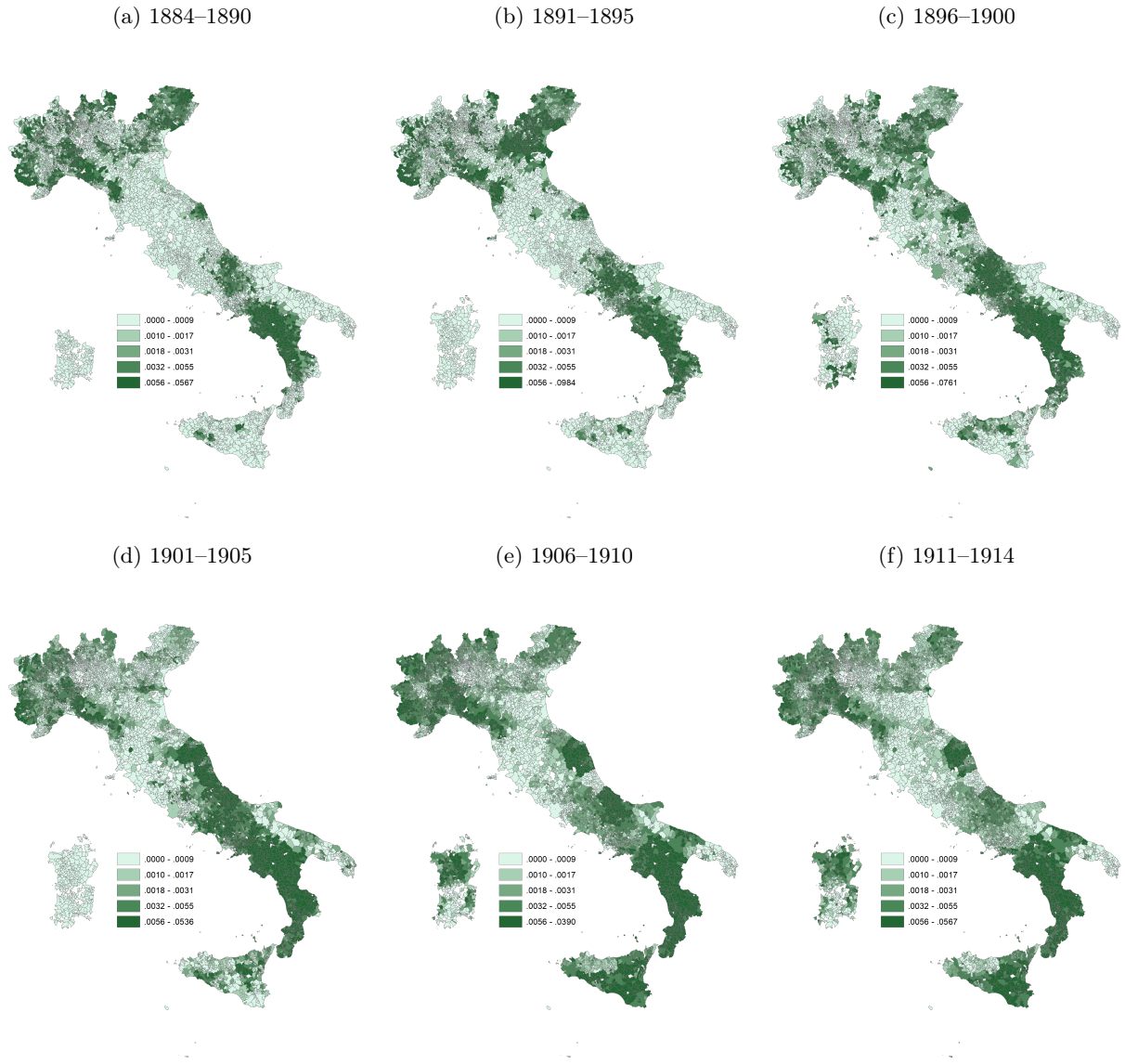
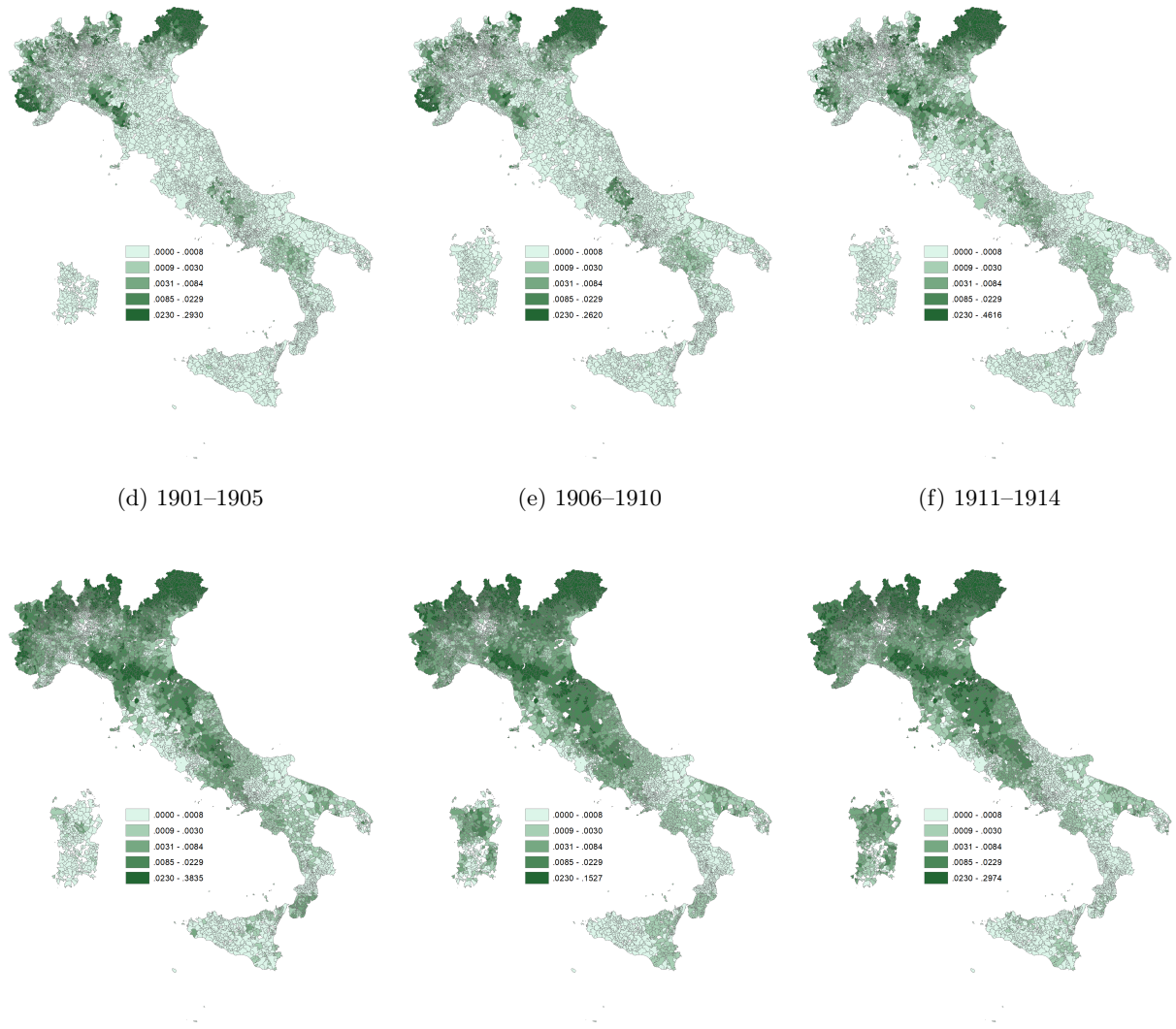


Figure B.7: Commune-level emigration rates to South America

Note: Each panel presents a commune's average annual emigration rate to South America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

(a) 1884–1890 (b) 1891–1895 (c) 1896–1900



(d) 1901–1905 (e) 1906–1910 (f) 1911–1914

Figure B.8: Commune-level emigration rates to Europe

Note: Each panel presents a commune's average annual emigration rate to Europe in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

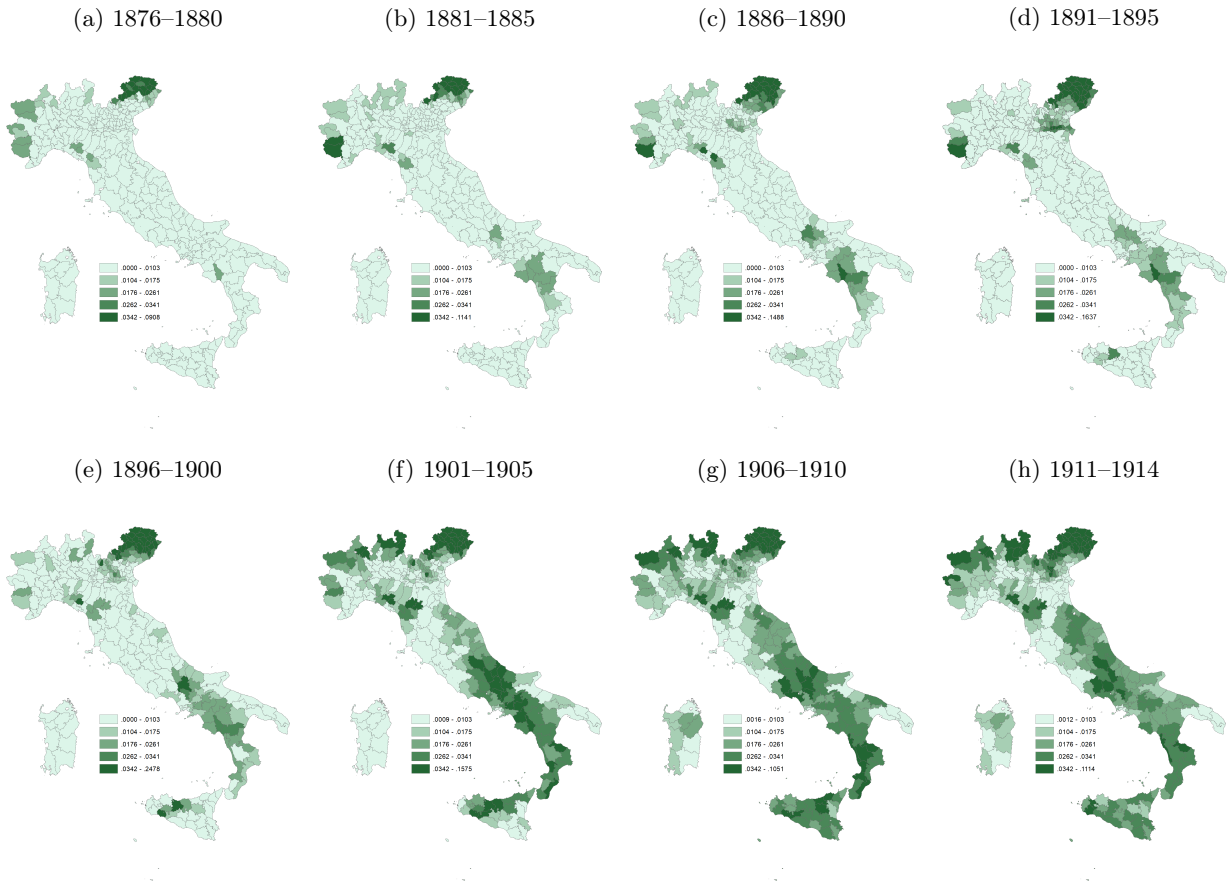
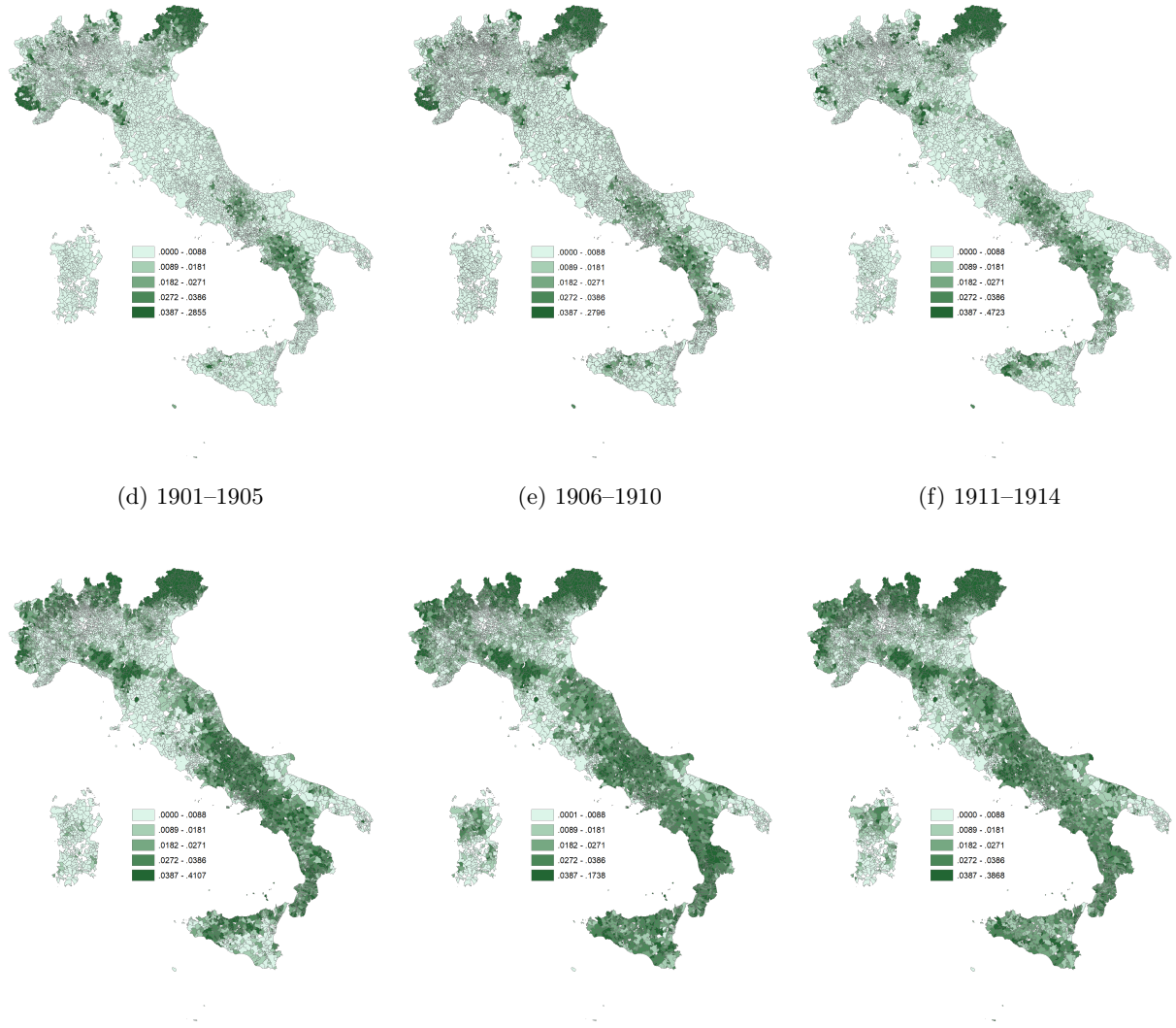


Figure B.9: District-level emigration rates to any destination

Note: Each panel presents a district's average annual emigration rate to any destination in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

(a) 1884–1890 (b) 1891–1895 (c) 1896–1900



(d) 1901–1905 (e) 1906–1910 (f) 1911–1914

Figure B.10: Commune-level emigration rates to any destination

Note: Each panel presents a commune's average annual emigration rate to any destination in the period in question. Scale is based on quintiles of emigration rates in 1911–1914. Darker areas have higher emigration rates.

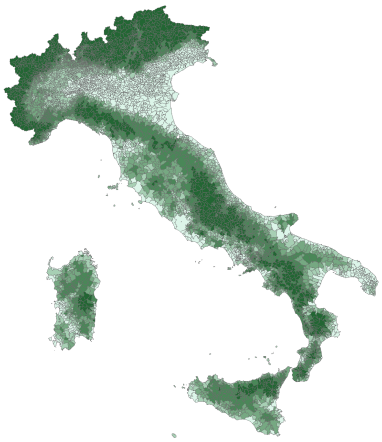
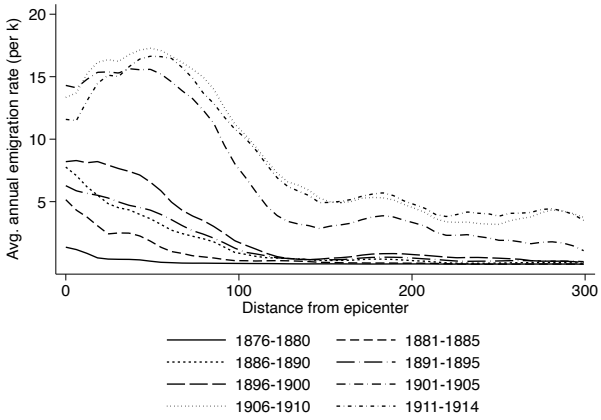


Figure B.11: Elevation

Source: Shuttle Radar Topography Mission (Jet Propulsion Laboratory 2014)

Note: Darker shading indicates greater elevation.

(a) Districts



(b) Communes

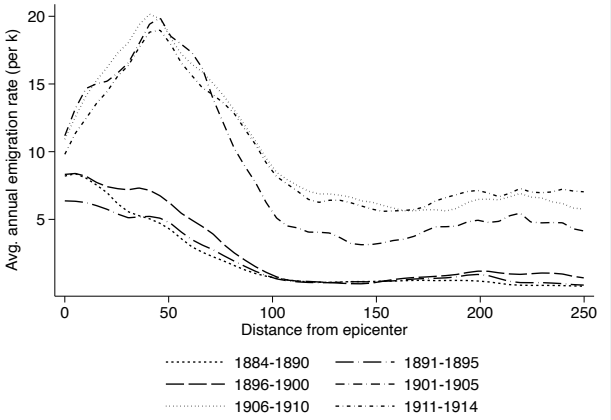


Figure B.12: Emigration rates to North America by distance to epicenter (km)

Note: These figures plot non-parametric regressions of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration to North America.

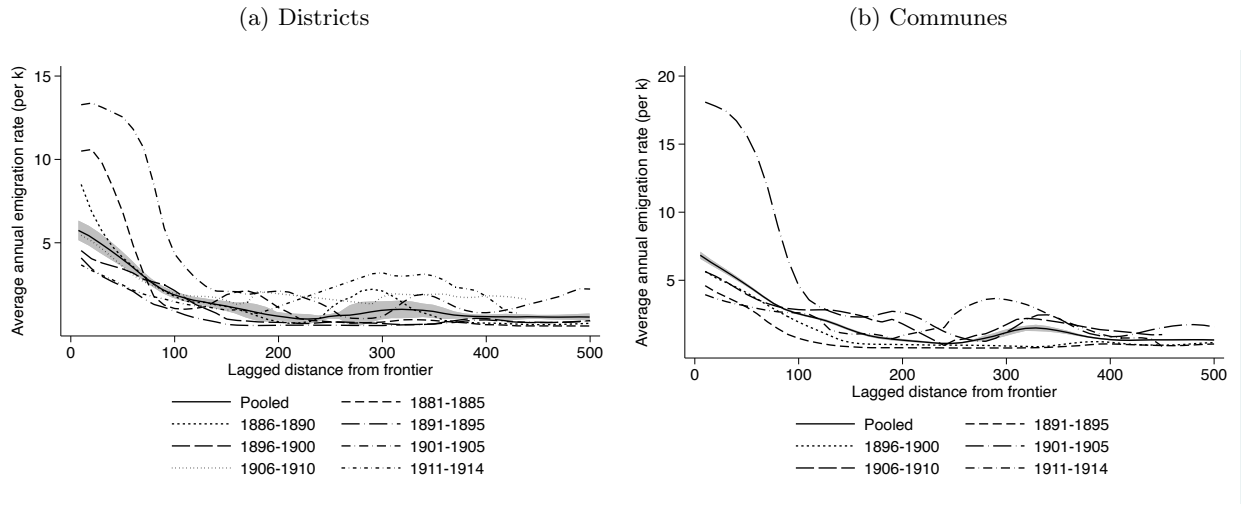


Figure B.13: Emigration rates to North America by distance to the mass migration frontier (km)

Note: These figures present non-parametric regressions of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to communes in districts that had not yet achieved this threshold. Shaded areas are 95-percent confidence intervals.

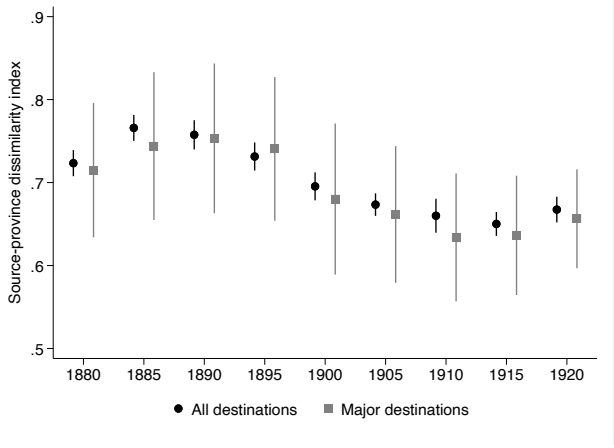


Figure B.14: Source dissimilarity over time

Note: This figure plots coefficients from a regression of source-province dissimilarity for each pair of destinations on half-decade fixed effects. The source-province dissimilarity for destination pair ij is the fraction of the migratory flow to destination i that would have to have its place of origin changed to match the origin distribution of the migratory flow to destination j (or vice versa). Standard errors clustered at the destination-pair level. Major destinations are the United States, Canada, Brazil, Argentina, Uruguay, Austria-Hungary, France, Germany, and Switzerland.

Table B.3: Spatial contagion results, OLS, 50-250km from epicenters

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.957[a]	0.947[a]	0.810[a]	0.843[a]	0.659[a]	0.722[a]	0.530[a]	0.601[a]
	(0.031)	(0.031)	(0.032)	(0.033)	(0.032)	(0.030)	(0.032)	(0.035)
Observations	31,463	31,463	31,463	31,463	31,463	31,463	31,463	31,463
R-squared	0.717	0.750	0.762	0.772	0.781	0.812	0.792	0.839
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.4: Spatial contagion results, OLS

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.979[a]	0.935[a]	0.823[a]	0.847[a]	0.669[a]	0.718[a]	0.557[a]	0.594[a]
	(0.026)	(0.027)	(0.027)	(0.026)	(0.026)	(0.025)	(0.027)	(0.031)
Observations	41,169	41,165	41,165	41,165	41,165	41,165	41,165	41,165
R-squared	0.732	0.762	0.773	0.784	0.791	0.822	0.803	0.849
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.5: Spatial contagion results, standard instrumentation approach

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.524[a]	0.568[a]	0.871[a]	0.886[a]	0.598[a]	0.664[a]	0.535[a]	0.568[a]
	(0.056)	(0.072)	(0.201)	(0.195)	(0.136)	(0.125)	(0.103)	(0.072)
Lagged Own Emigration	0.368[a]	0.349[a]	0.219[a]	0.234[a]	0.281[a]	0.329[a]	0.256[a]	0.349[a]
	(0.018)	(0.018)	(0.075)	(0.070)	(0.040)	(0.030)	(0.030)	(0.018)
Observations	35,287	35,284	35,329	35,329	35,329	35,327	35,329	35,284
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	153.1	119.3	28.74	29.94	76.30	57.47	117.6	119.3

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1
Notes: Standard errors clustered at the district level. All specifications include at least half-decade fixed effects. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, distance to coast, distance to the european land border, population, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.6: Spatial contagion results, epicenter-based IV, dropping extreme residuals

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.949[a]	0.891[a]	0.670[a]	0.663[a]	0.590[a]	0.476[b]	0.526	0.183
	(0.111)	(0.137)	(0.178)	(0.205)	(0.152)	(0.205)	(0.512)	(0.508)
Observations	31,011	31,011	31,011	31,011	31,011	31,007	31,011	30,977
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	30.24	32.50	17.93	11.93	23.08	11.40	18.61	5.024

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1
Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Observations in the top and bottom 1 percent of residuals in a regression the instrument on all controls are excluded. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.7: Spatial contagion results, frontier-based IV, dropping extreme residuals

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	1.019[a]	1.022[a]	0.795[a]	0.919[a]	0.824[a]	0.789[a]	0.744	0.612[b]
	(0.146)	(0.175)	(0.181)	(0.184)	(0.226)	(0.190)	(1.026)	(0.284)
Observations	11,056	11,056	11,056	11,054	11,056	11,047	11,056	11,020
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	56.07	56.64	54.52	48.38	40.21	41.68	15	22.13

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Observations in the top and bottom 1 percent of residuals in a regression the instrument on all controls are excluded. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.8: Spatial contagion results, epicenter-based IV, other catchments

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.745[a]	0.762[a]	0.561[b]	0.492	0.674[a]	0.503	1.020[b]	−0.613
	(0.199)	(0.143)	(0.265)	(0.443)	(0.155)	(0.356)	(0.481)	(1.500)
Observations	31,463	31,463	31,463	31,463	31,463	31,462	31,463	31,427
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	11.75	16.35	6.131	2.342	17.51	3.729	9.849	1.016

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: The instrument in this table is constructed on the basis of the relationship between emigration and epicenter distance in catchment areas other than that of the commune in question. Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table B.9: Spatial contagion results, frontier-based IV, other catchments

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.362	0.913[a]	0.799[a]	0.669[b]	0.717	0.621[c]	0.094	0.659
	(0.237)	(0.227)	(0.268)	(0.307)	(0.655)	(0.377)	(0.787)	(0.458)
Observations	11,195	11,194	11,194	11,193	11,194	11,184	11,194	11,158
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	38.90	28.33	24.73	19.61	25.90	13.45	13.82	10.88

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: The instrument in this table is constructed on the basis of the relationship between emigration and distance from the mass migration frontier in provinces other than that of the commune in question. Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

C The Migration of Antonio Squadrito's Group in October 1903

Antonio Squadrito was born around 1877 in the small Sicilian town of Gualtieri-Sicamino, near Messina. In 1898 he decided to migrate to the United States, among the first in his commune to do so. In New York he had a "distant relative from a northern province," and his passenger manifest listed his American contact as an uncle living on 21st Street, but he paid for his travel by borrowing money from his father, Giovanni. He arrived at the Battery on July 7, 1898, and his first job was in a quarry in Rhode Island. Soon thereafter, several opportunities arose. He befriended another Italian who owned a barbershop in Stonington, CT, and joined him as an employee, gradually paying off his loan. Shortly after, his boss had to leave the business and forced Antonio to take a loan to purchase the shop. The shop prospered, and Antonio had his older married brother Giuseppe come in to help him. Giuseppe was followed by their father and two younger brothers. In June 1903 Antonio married Harriet H. Burtch-Gardiner, who, at 66, was 41 years his senior.[76] That the same summer, he travelled back to his hometown with the purpose of helping the migration of a large number of friends and relatives. By that time, five years after Antonio had first left Sicily, emigration was already widespread in Gualtieri-Sicamino.

While in Sicily, Antonio collected a large group of individuals whose migration he facilitated (listed in Table C.1), mainly close and more distant relatives from Gualitieri-Sicamino and from other neighboring places. Among them were his sister-in-law, her four-year-old niece, her brother, and her nephew, all from Gualtieri-Sicamino and destined for Stonington. The others had other destinations in the United States, where they reported having relatives. A sixteen-year-old girl—a cousin from the neighboring commune of San Filipo—and five young men—all neighbors and family friends from Gualtieri-Sicamino—were traveling to Boston and to New York. Four farmer boys from Soccorso,[77] a detachment (*frazioni*) of Gualtieri-Sicamino, were on their way to the mines in Pennsylvania. They reported relatives in Philadelphia, but in reality they were illegally contracted laborers, and the uncle of one of them was the middleman (perhaps a *padrone* of sorts) who helped to recruit them. The entire group left for Messina en route to Napoli. From Napoli they embarked on the steamship *Prinzess Irene* on October 2, 1903, and arrived at Ellis Island on October 14. Broughton Brandenburg, the journalist and self-proclaimed immigration specialist who followed Squadrito's entourage and documented their migration, noted that this sort of group migration organized by a friend or relative was so common, that "The most notable feature was the ease with which one could detect that every seventh or eighth person had been to America before, and now had gathered around him a group of

[76] Her wealth, estimated at $60,000, was inherited from her deceased husband, a whaling ship captain (Brandenburg 1904, p. 44).

[77] Brandenburg (1904) mistakenly referred to it as "Socosa."

from two to thirty friends, relatives, and neighbors, going over in his care, just as our party was going in the care of Antonio Squadrito and myself" (Brandenburg 1904, p. 172).

In fact, the group was planned to be larger, as they had expected passengers from other communes to join them in Messina. These were Giuseppe Cardillo, accompanied by a few other people, and the Papalia family from Monforte San Giorgio, a small town situated about ten kilometers west of Gualtieri-Sicamino. Cardillo's hometown is unknown and so is the specific relation between the two families and the Squadritos. Eventually, according to Brandenburg (1904, p. 133), Cardillo's group decided to postpone their travel and the Papalias ended up taking the next steamer. Indeed, two weeks later, on October 28, Michele and Maria Papalia, originally from Monforte San Giorgio, and their five-year-old daughter Rosina were recorded arriving at Ellis Island on board the steamship *Lahn*, where they were listed as American citizens returning home to New York. All in all, the extended group that planned their joint voyage comprised of neighbors, friends, relatives, and other acquaintance from five different localities, at least four of which were within a short distance from one another.

How does this case fit the theoretical framework proposed in section 3? Clearly, it shows that the reality was more complex than the stylized story about a linear chain in which one individual links others in his geographic environment who depend on him, and leads them to the same destination. It is not clear, for example, how crucial the role played by Antonio Squadrito's relative was in enabling his own migration in 1898, and therefore it is impossible to tell whether or not he was a real pioneer. Even if he were linked by his relative, it is hard to tell whether this linkage conformed to our assumption that social contacts were largely local, because although he was a relative, according to Brandenburg he was from a "northern province" (Brandenburg 1904, p. 43). Furthermore, many in the group relied on additional contacts in the United States. They were supported by Antonio, but he was not their sole sponsor, and it is probable that they would have migrated even without his help. Indeed, only a few were destined to join him in Stonington. Networks merged and diverged to different destinations, and it is unknown whether the emigration from Gualtieri-Sicamino could be traced back to a single local founding father or to several ancestors separately linked from other communes, and whether any of them were virtual pioneers. Nevertheless, those going to other destinations were still relying on other personal links, usually family members. Even those who were in reality contracted laborers were recruited through a relative. If the case of Squadrito's group is indicative, then in a broad sense, the Italian transatlantic movement occurred within local networks based both on intra-communal and on short-distance inter-communal links. This is precisely the core insight that the theoretical framework that we propose is meant to capture.

Table C.1: Antonio Squadrito's group, on board *Prinzess Irene*, arriving October 14, 1903

First Name	Last Name	Sex	Age	Relation to Antonio Squadrito	Place of Origin	Joining	Destination
Antonio	Squadrito	M	26		Gualtieri-Sicamino	Brothers, Giuseppe, Carmelo, and Gaetano	Stonington, CT
Carmela	Squadrito	F	32	Sister in law	Gualtieri-Sicamino	Husband, Giuseppe Squadrito (Antonio's brother)	Stonington, CT
Caterina	Squadrito	F	4	Niece	Gualtieri-Sicamino	Father, Giuseppe Squadrito	Stonington, CT
Giovanni	Pulejo	M	49	Brother in law, probably also a cousin	Gualtieri-Sicamino	Brother, Nicola	Boston, MA
Felice	Pulejo	M	16	Nephew	Gualtieri-Sicamino	Uncle, Nicola	Boston, MA
Concetta	Fomica	F	15	Cousin	San Filipo	Uncle, Stefano Senedile, Boston	Boston, MA
Antonio	Nastasia	M	16	Neighbor	Gualtieri-Sicamino	Uncle, Tommaso Trovato, Boston	Boston, MA
Gaetano	Mullura	M	16	Neighbor	Gualtieri-Sicamino	Uncle, Nicolo Puleo, Boston	Boston, MA
Nicola	Curro	M	27	Family friend	Gualtieri-Sicamino	Cousin, Angelo Ragusa, New York	New York, NY
Nunzio	Giunta	M	23	Fellow townsman	Gualtieri-Sicamino	Cousin, New York	New York, NY
Antonio	Genino	M	21	Fellow townsman	Gualtieri-Sicamino	Uncle, Giuseppe Maucino, Philadelphia	Philadelphia, PA
Salvatore	Niceta	M	20	Farm boy from detached village	Soccorso	Brother, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Benedetto	Runzio	M	21	Farm boy from detached village	Soccorso	Cousin, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Luciano	Sofia	M	17	Farm boy from detached village	Soccorso	Cousin, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Salvatore	Damico	M	23	Farm boy from detached village	Soccorso	Brother in law, Antonio Salvatore, Philadelphia	Philadelphia, PA

Sources: Brandenburg (1904) and the Statue of Liberty-Ellis Island Foundation

D Model Simulations

E Preparing Official Statistics for Analysis

The data that we collected from the *Statistica della Emigrazione Italiana per l'Estero* volumes and from the 1881 Italian census required considerable preparation before they could be used for analysis. At the commune level, the main difficulties are the changing of commune names over time, and the combination or division of communes to form other communes. A key source for this effort was the *Comuni e Loro Popolazione ai Censimenti dal 1861 al 1951*, published by ISTAT (the Italian statistical bureau) in 1960. This publication describes the changing borders of communes, allowing us to create consistently defined communes over the entire sample period, based on borders in 1904. Another difficulty arose from the existence of two sometimes conflicting records for the same commune-year in cases when two different volumes presented data for the same year. In this case, we used data from the later-published volume.

Our analysis also requires knowing the geographic location of each commune. For communes that still exist (the vast majority), we were able to simply match the list of commune names to a GIS file of modern communes whose historical provinces could be determined using a shapefile of historic province boundaries provided by ISTAT. This was more difficult in the case of historic communes that were consistently defined throughout our study period but have since ceased to exist. For instance, the commune of Santo Stefano di Briga existed throughout our study period, but has since been incorporated into the commune of Messina. The best guess of geographic location that we are able to derive is thus to place Santo Stefano di Briga in the same place as Messina. This simplification is a possible source of error, but because most communes are quite small, the resulting error is likely to be small.

Another issue was the mapping of districts. To our knowledge, no shapefiles of Italian districts exist. We constructed the shapefile that we use by merging the polygons of all communes assigned to a particular district. For communes that were created after our study period, we determined the commune of which they were once a part, and assign the modern commune to the district of the historic commune from which it was split. Comparison of our resultant shapefile to a map that we were able to locate of historical districts shows that our generated shapefile is quite accurate.

The precise number of districts varied over the study period due to the consolidation of the smaller *distretti* in the territories annexed from Austria in 1866 (Veneto) into larger *circondari*; but we were able to identify 284 consistently defined units with emigration data, corresponding to the *circondari* and *distretti* existing at the beginning of our study period.

Another issue arose from the fact that northern provinces that were previously part of the Austro-Hungarian Empire (the region of Veneto) had *distretti* instead of *circondari*. We treat both of these as

districts, but the *distretti* were smaller and were eventually eliminated, creating provinces with a single *circondario*. For the emigration data, we can reconstruct the *distretti* totals from the commune-level data. For the census data, we must use province-level data on literacy and employment for these northern provinces.

F Results Including Data on "Other Communes"

This appendix addresses the fact that for years 1903 and earlier, the emigration of some communes was not listed in the *Statistica della Emigrazione Italiana per l'Estero*, but was instead included in an aggregate report for each district under the header of "Other Communes in this District." To ensure that this is not responsible for driving results, we allocate this unassigned emigration equally to the excluded communes and repeat the main results. Since this does not affect the district-level data, those results are not repeated here.

Figure F.1: σ-convergence in emigration rates to North America

Note: Each point represents the coefficient of variation in emigration rates to North America in a particular year.

Table F.1: β-convergence

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.566^{a}	-0.704^{a}	-0.784^{a}	-0.825^{a}	-0.609^{a}	-0.784^{a}	-0.589^{b}	-0.788^{a}
	(0.026)	(0.023)	(0.019)	(0.016)	(0.047)	(0.051)	(0.249)	(0.211)
Observations	7,930	7,929	7,929	7,929	7,930	7,929	7,929	7,928
R-squared	0.594	0.764	0.877	0.911	0.591	0.756	0.634	0.687
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	97.846	95.140	5.961	6.719
FE	None	None	P	D	None	None	P	D
Falsification	−0.065	−0.036	0.390	0.486				
	(0.068)	(0.084)	(0.054)	(0.048)				

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune. Dependent variable is the change in the log of the emigration rate to North America from the pre-1900 period to the period 1900 and later. Controls include latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Instrument is the distance to the nearest epicenter of emigration to North America. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the post-1900 period; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

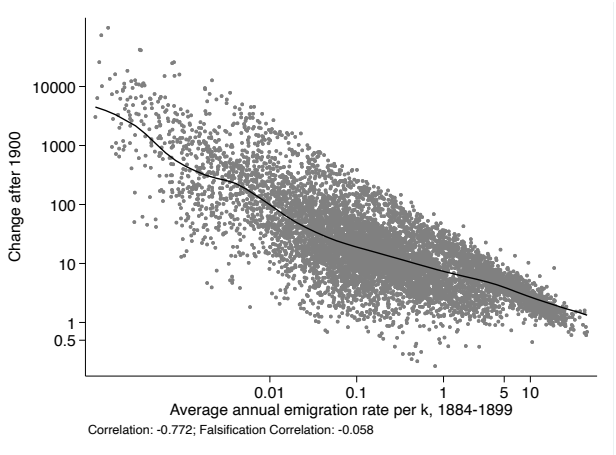


Figure F.2: β-convergence in emigration rates to North America

Note: Each point represents a commune. The x-axis is the average annual emigration rate for 1884–1899 on a log scale. The y-axis is the ratio of the average emigration rate before and after 1900, also on a log scale. The falsification correlation is the correlation of the change in emigration and emigration after 1900; that it is not positive indicates that the negative relationship shown in the graphs is unlikely to be spurious, as explained in section 5.1.

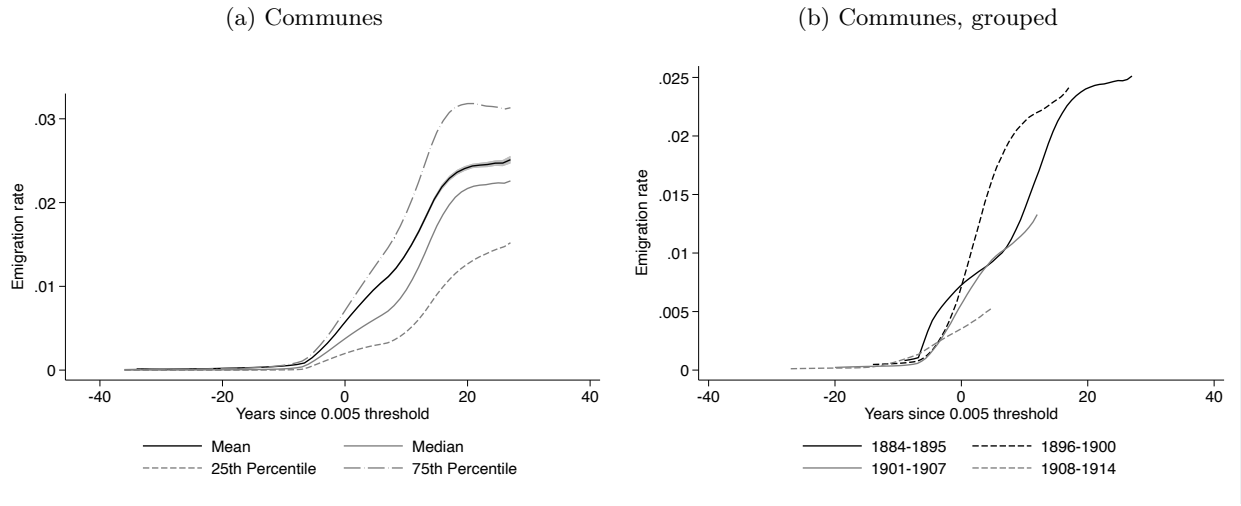


Figure F.3: S-shaped time series of migration to North America

Note: Panels (a) plots a non-parametric regression of emigration rates to North America against time, normalized so that year 0 is the first year in which a place had an emigration rate of at least 5 per thousand. The shaded area is a 95-percent confidence interval. Panel (b) divides communes according to the half decade in which they crossed the threshold.

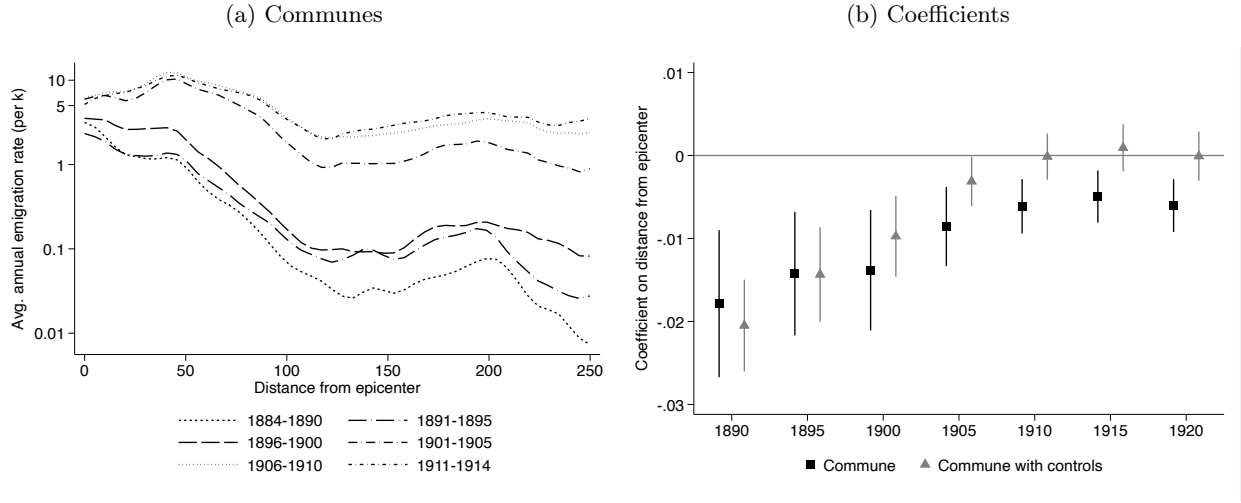


Figure F.4: Emigration rates to North America by distance to epicenter (km)

Note: Panel (a) plots non-parametric regressions of the log of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration to North America. Panel (b) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of distance from the nearest epicenter of emigration to North America and plots the coefficients on distance from epicenter. Panel (b) also includes a regression controlling for half decade-specific functions of various controls.

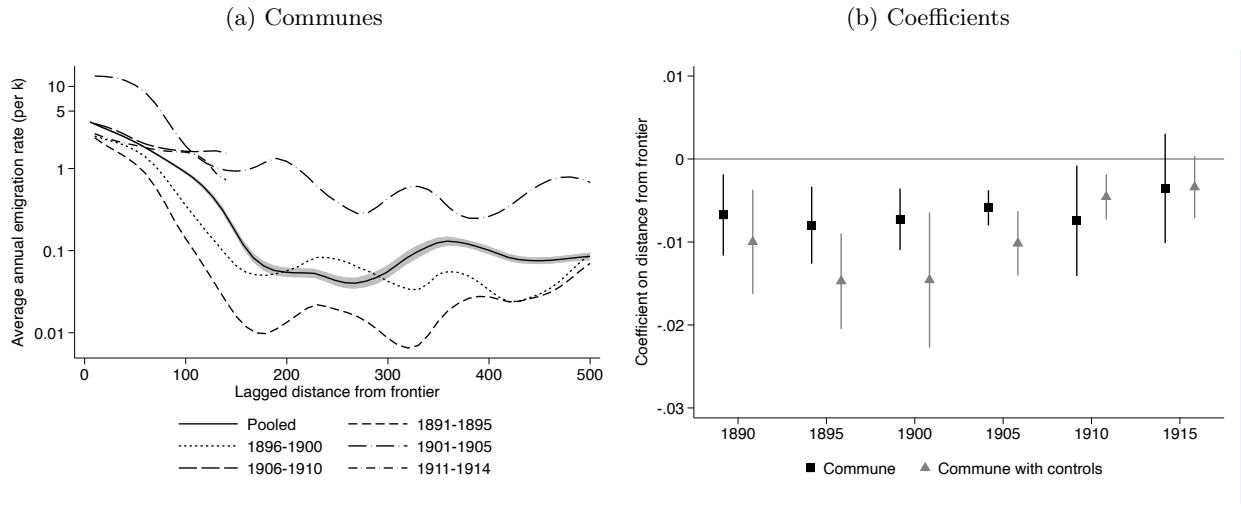


Figure F.5: Emigration rates to North America by distance to the mass migration frontier (km)

Note: Panel (a) presents non-parametric regressions of the log of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to districts that had not yet achieved this threshold. The shaded areas is a 95-percent confidence interval. Panel (b) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of lagged distance from the frontier of mass migration to North America and plots the coefficients on lagged distance from the frontier. Panel (b) also includes a regression controlling for half decade-specific functions of various controls.

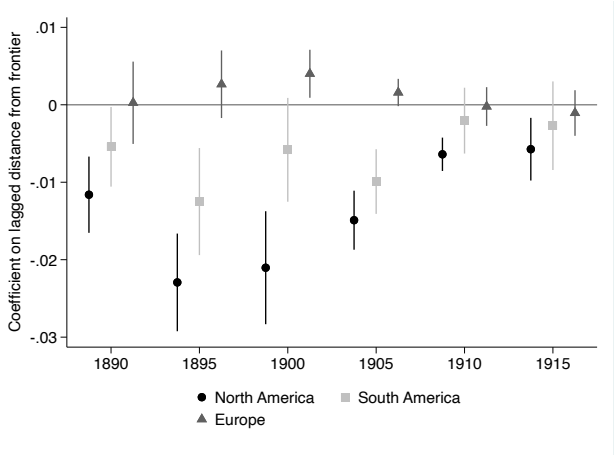


Figure F.6: Emigration to various destinations by distance to the mass migration frontier for North America (km)

Note: This figure repeats the binomial maximum likelihood regressions of panel (b) of Figure F.5, but includes results for migration to South America and Europe in addition to those for migration to North America.

Table F.2: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.936[a]	1.014[a]	0.769[a]	0.682[a]	0.748[a]	0.497[b]	1.170[b]	0.244
	(0.110)	(0.108)	(0.179)	(0.217)	(0.160)	(0.225)	(0.457)	(0.283)
Observations	36,697	36,697	36,697	36,697	36,697	36,697	36,697	36,668
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	37.86	54.59	20.25	15.27	26.81	12.92	34.18	12.72

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table F.3: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.834[a]	0.889[a]	0.543[b]	0.627[a]	0.617[a]	0.490[a]	0.365	0.355[c]
	(0.192)	(0.187)	(0.217)	(0.208)	(0.181)	(0.170)	(0.424)	(0.186)
Observations	12,670	12,668	12,668	12,667	12,668	12,658	12,668	12,639
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	77.19	83.85	78.39	70.18	80.99	79.84	60.60	54.37

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

G Results with 1881 Population

The main results use 1901 population as the denominator in calculating emigration rates. This appendix repeats the main results using 1881 population as the denominator.

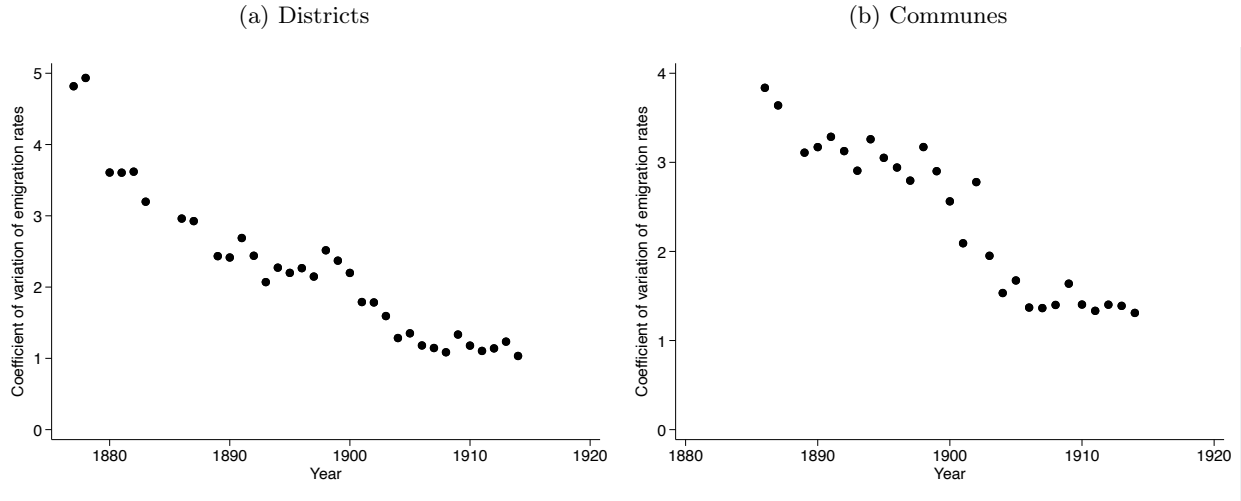


Figure G.1: σ-convergence in emigration rates to North America

Note: Each point represents the coefficient of variation in emigration rates to North America in a particular year.

Table G.1: β-convergence

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.525^a	-0.751^a	-0.854^a	-0.873^a	-0.533^a	-0.830^a	-0.896^a	-0.999^a
	(0.022)	(0.018)	(0.011)	(0.011)	(0.046)	(0.068)	(0.171)	(0.232)
Observations	5,837	5,836	5,836	5,836	5,837	5,836	5,836	5,831
R-squared	0.570	0.751	0.874	0.908	0.570	0.743	0.785	0.796
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	53.203	88.865	11.978	7.805
FE	None	None	P	D	None	None	P	D
Falsification	−0.096	0.110	0.347	0.343				
	(0.055)	(0.069)	(0.061)	(0.057)				

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune. Dependent variable is the change in the log of the emigration rate to North America from the pre-1900 period to the period 1900 and later. Controls include latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Instrument is the distance to the nearest epicenter of emigration to North America. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the post-1900 period; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

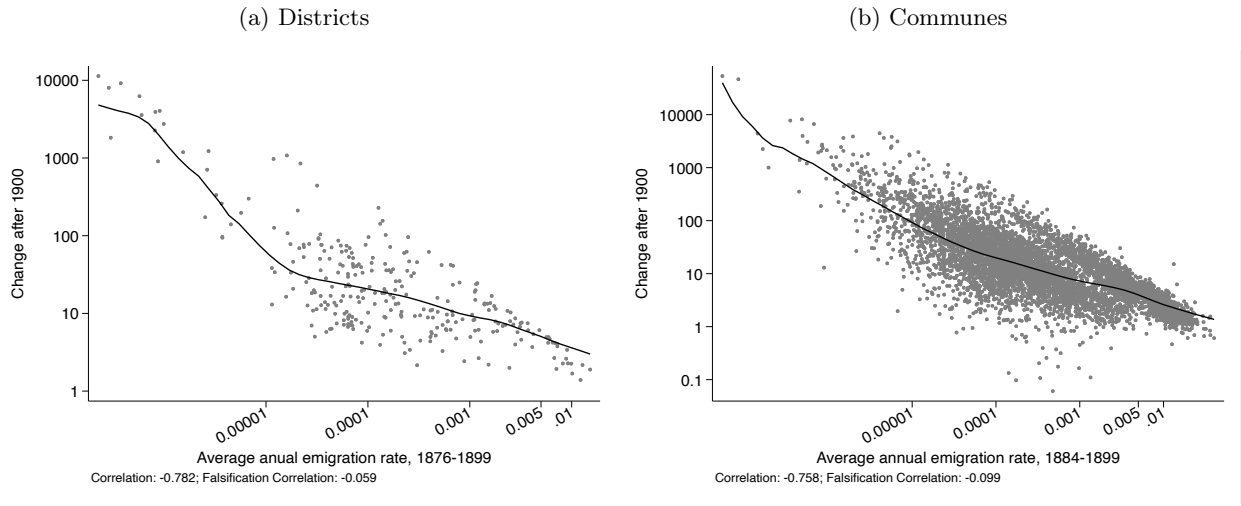


Figure G.2: β-convergence in emigration rates to North America

Note: Each point represents a commune or district. The x-axis is the average annual emigration rate for a district for 1876–1899 or a commune for 1884–1899 on a log scale. The y-axis is the ratio of the average emigration rate before and after 1900, also on a log scale. The falsification correlation is the correlation of the change in emigration and emigration after 1900; that it is not positive indicates that the negative relationship shown in the graphs is unlikely to be spurious, as explained in section 5.1.

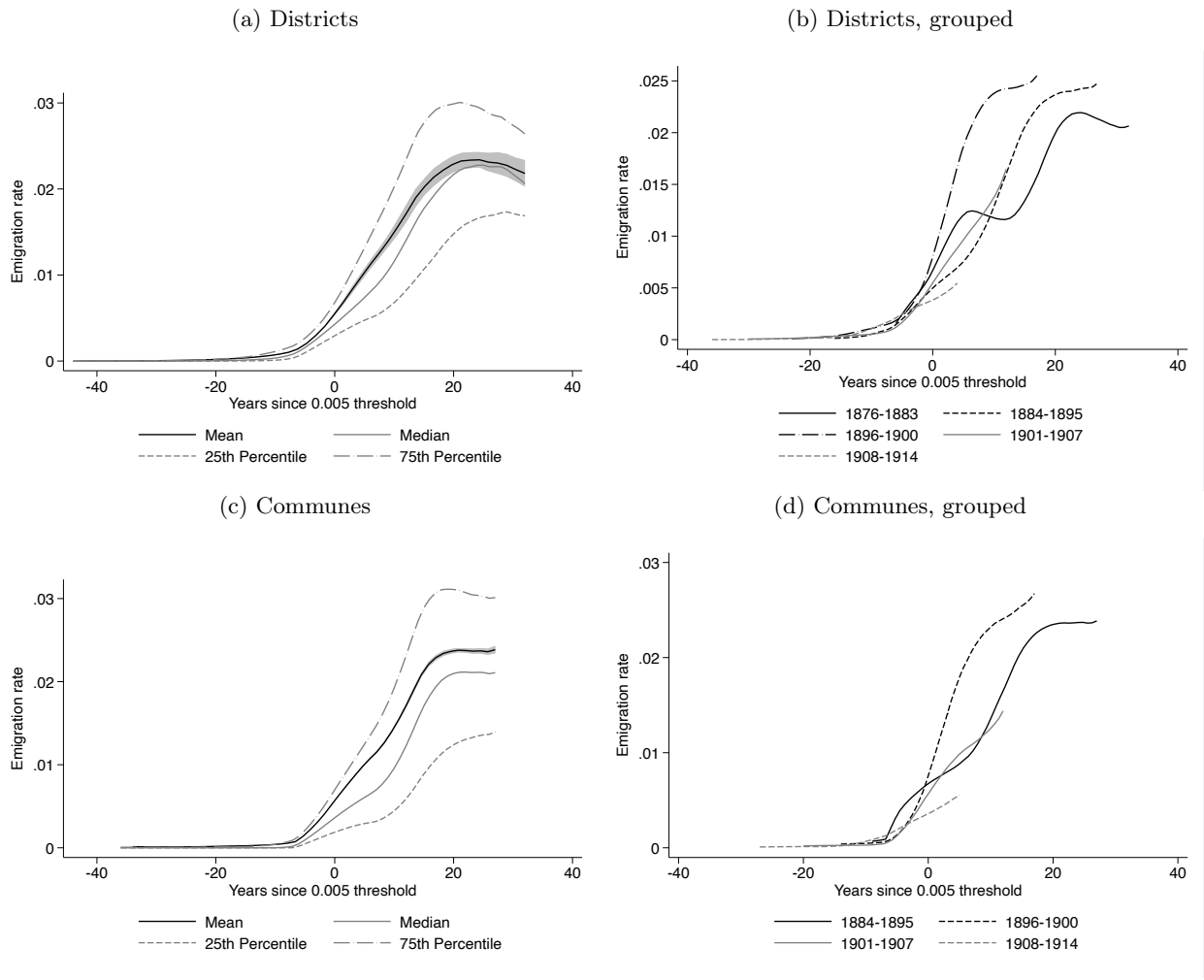


Figure G.3: S-shaped time series of migration to North America

Note: Panels (a) and (c) plot a non-parametric regression (the mean), as well as quartiles of emigration rates to North America against time, normalized so that year 0 is the first year in which a place had an emigration rate of at least 5 per thousand. Shaded areas are 95-percent confidence intervals for the mean. Panels (b) and (d) are the same as (a) and (c) but divide areas according to the half decade in which they crossed the threshold.

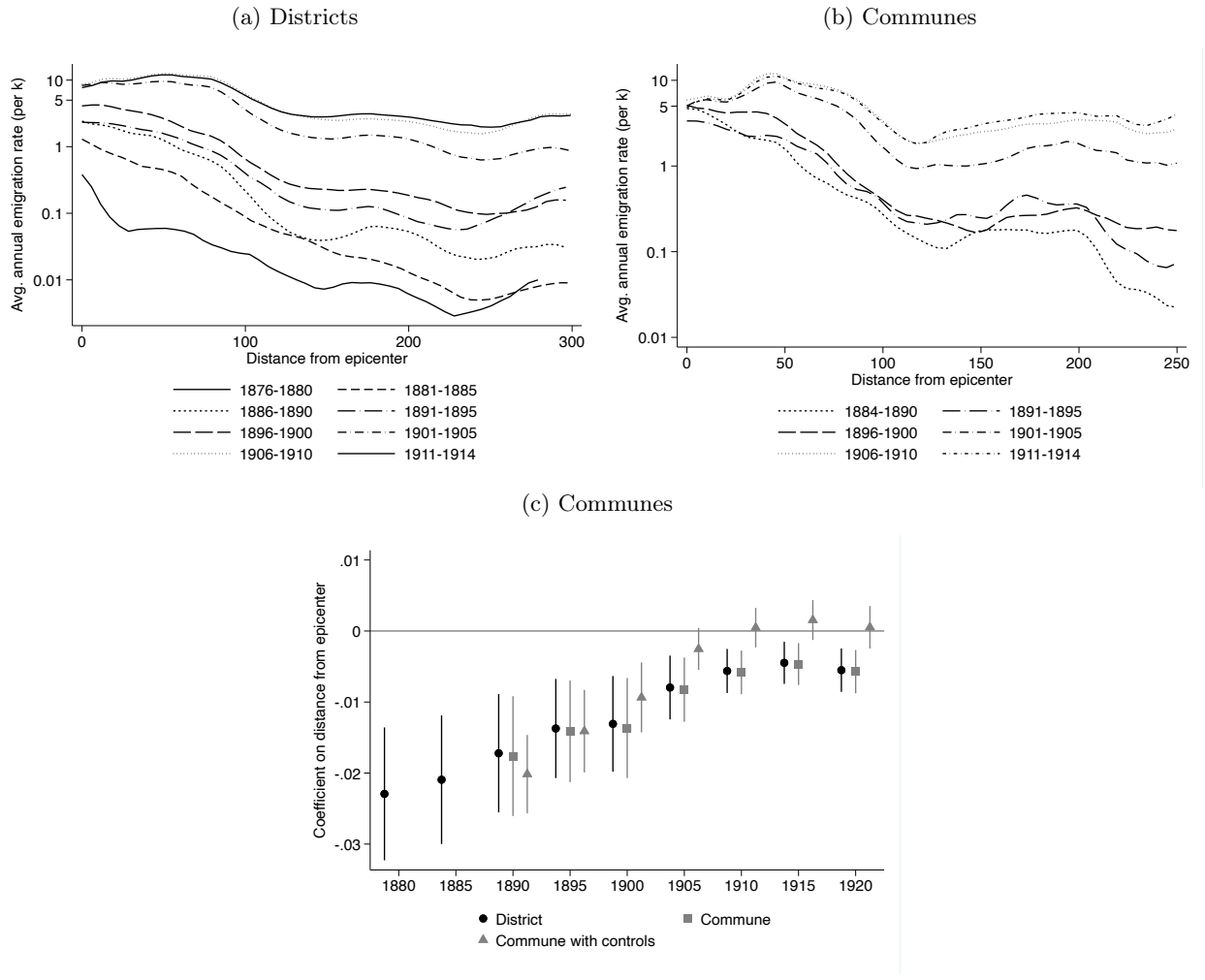


Figure G.4: Emigration rates to North America by distance to epicenter (km)

Note: Panels (a) and (b) plot non-parametric regressions of the log of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration to North America. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of distance from the nearest epicenter of emigration to North America and plots the coefficients on distance from epicenter. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

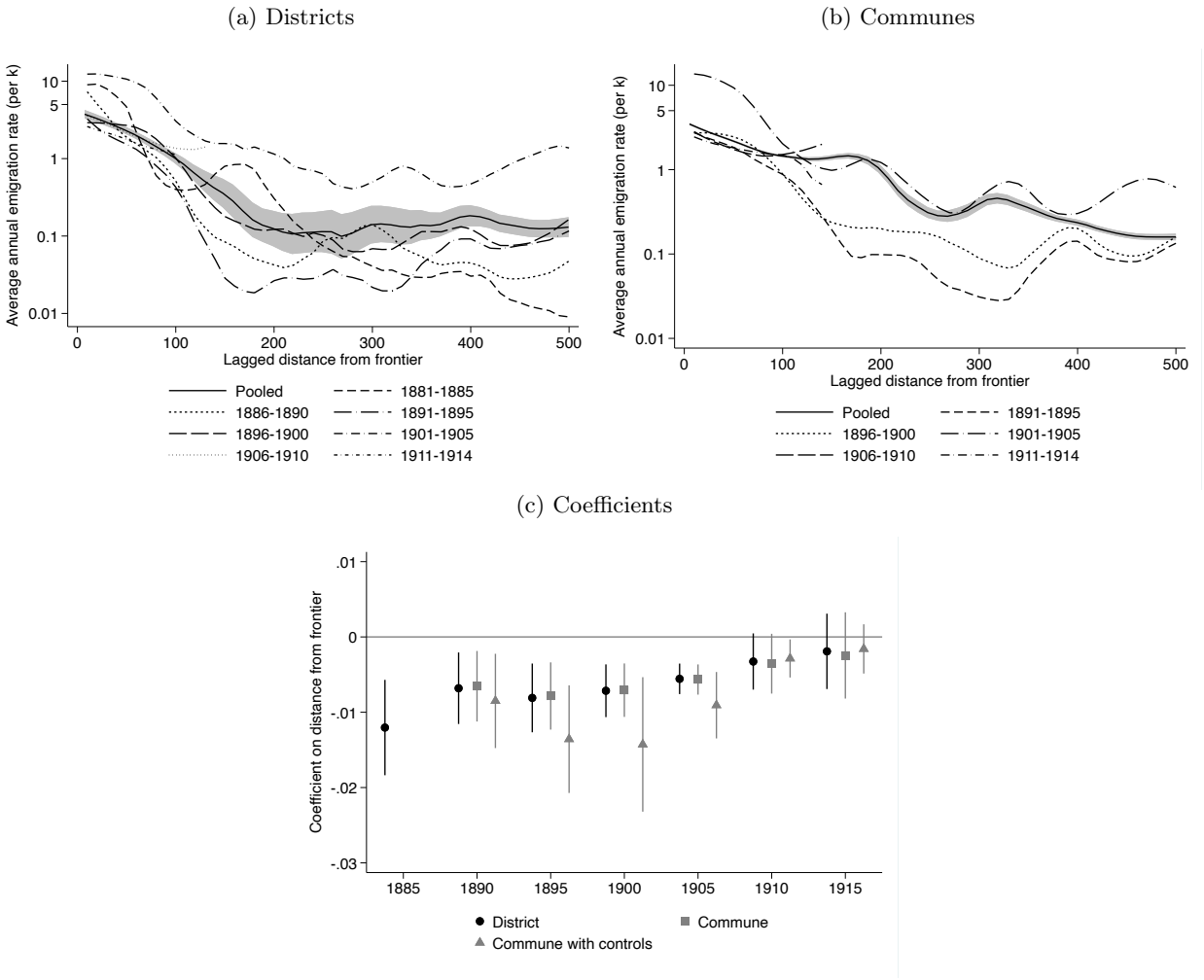


Figure G.5: Emigration rates to North America by distance to the mass migration frontier (km)

Note: Panels (a) and (b) present non-parametric regressions of the log of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to districts that had not yet achieved this threshold. Shaded areas are 95-percent confidence intervals. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of lagged distance from the frontier of mass migration to North America and plots the coefficients on lagged distance from the frontier. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

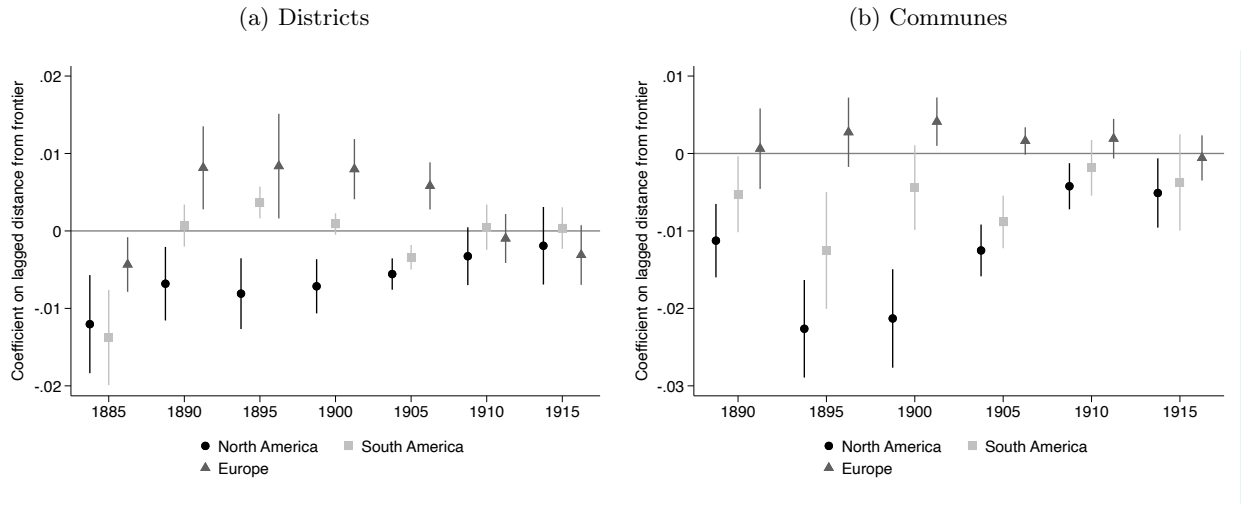


Figure G.6: Emigration to various destinations by distance to the mass migration frontier for North America (km)

Note: This figure repeats the binomial maximum likelihood regressions of panel (c) of Figure G.5, but includes results for migration to South America and Europe in addition to those for migration to North America.

Table G.2: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.957[a]	0.899[a]	0.654[a]	0.663[a]	0.652[a]	0.631[a]	0.512	0.389
	(0.095)	(0.124)	(0.196)	(0.236)	(0.141)	(0.211)	(0.436)	(0.399)
Observations	31,573	31,573	31,573	31,573	31,573	31,572	31,573	31,537
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	35.88	40.98	15.98	10.08	24.35	10.47	24.66	6.687

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table G.3: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.925[a]	0.922[a]	0.715[a]	0.847[a]	0.721[a]	0.695[a]	0.697	0.538[b]
	(0.126)	(0.150)	(0.150)	(0.159)	(0.170)	(0.156)	(0.545)	(0.219)
Observations	11,372	11,371	11,371	11,370	11,371	11,361	11,371	11,335
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	78.87	75.57	72.66	52.69	55	51.58	19.18	30.33

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

H Results for Migration to All Destinations

This appendix repeats the main results of the paper, but focuses on migration to all destinations rather than on migration to North America alone. This addresses the concern that some of the local correlation in emigration rates could be the product of the fact that the emigration-by-destination data are available only at the province level. It focuses on distance to all epicenters rather than only distance to epicenters of emigration to North America and on the frontier of mass emigration to any destination rather than only on the frontier of mass migration to North America. The results are, for the most part, qualitatively unchanged, with the exception of the S-shaped time series. As expected, these are not S-shaped, but continuously increasing after places cross the mass migration threshold.

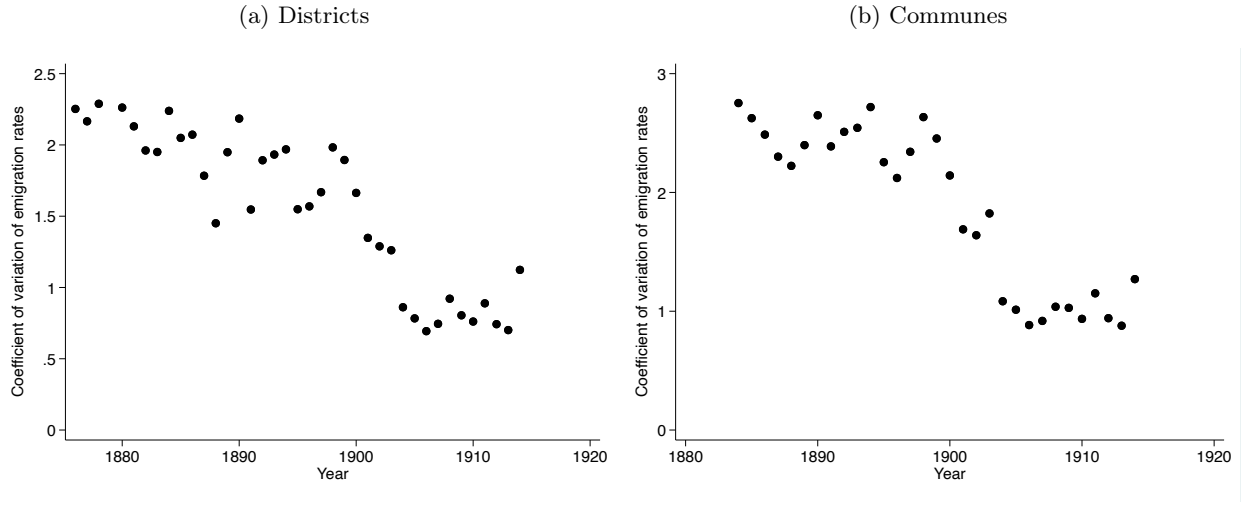


Figure H.1: σ-convergence in emigration rates to all destinations

Note: Each point represents the coefficient of variation in emigration rates to all destinations in a particular year.

Table H.1: β-convergence

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.691^a	-0.790^a	-0.775^a	-0.799^a	-0.615^a	-0.806^a	-0.578^a	-0.741^a
	(0.019)	(0.017)	(0.013)	(0.013)	(0.079)	(0.050)	(0.141)	(0.166)
Observations	6,104	6,103	6,103	6,103	6,104	6,103	6,103	6,100
R-squared	0.641	0.768	0.833	0.880	0.633	0.767	0.673	0.743
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	70.939	81.112	19.551	9.543
FE	None	None	P	D	None	None	P	D
Falsification	0.149	0.221	0.238	0.223				
	(0.055)	(0.063)	(0.050)	(0.051)				

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune. Dependent variable is the change in the log of the emigration rate to any destination from the pre-1900 period to the period 1900 and later. Controls include latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Instrument is the distance to the nearest epicenter of emigration. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the post-1900 period; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

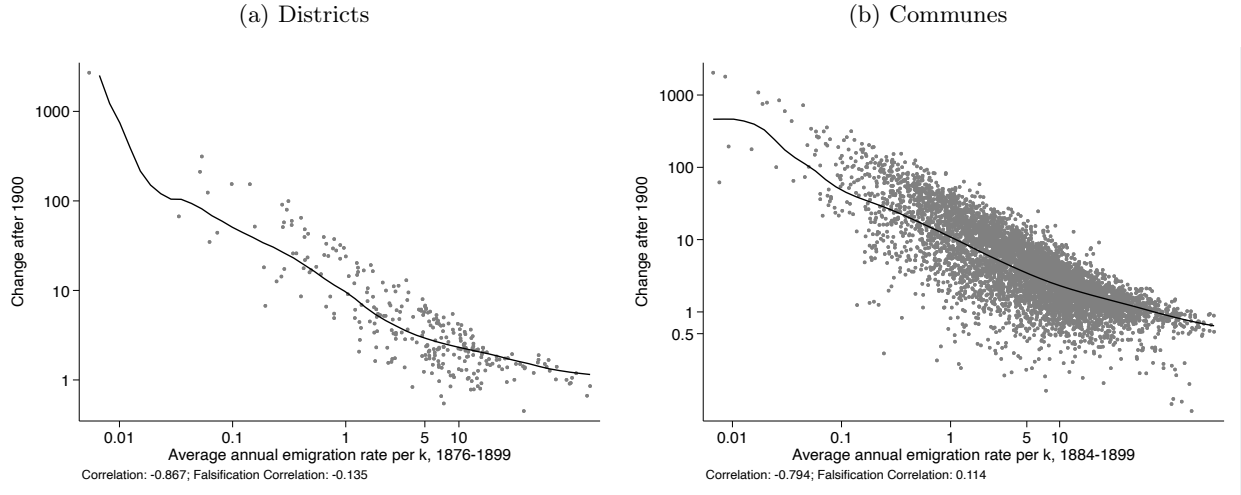


Figure H.2: β-convergence in emigration rates to all destinations

Note: Each point represents a commune or district. The x-axis is the average annual emigration rate for a district for 1876–1899 or a commune for 1884–1899 on a log scale. The y-axis is the ratio of the average emigration rate before and after 1900, also on a log scale. The falsification correlation is the correlation of the change in emigration and emigration after 1900; that it is not positive (or if it is, that its magnitude is considerably less than the plotted negative correlation) indicates that the negative relationship shown in the graphs is unlikely to be spurious, as explained in section 5.1.

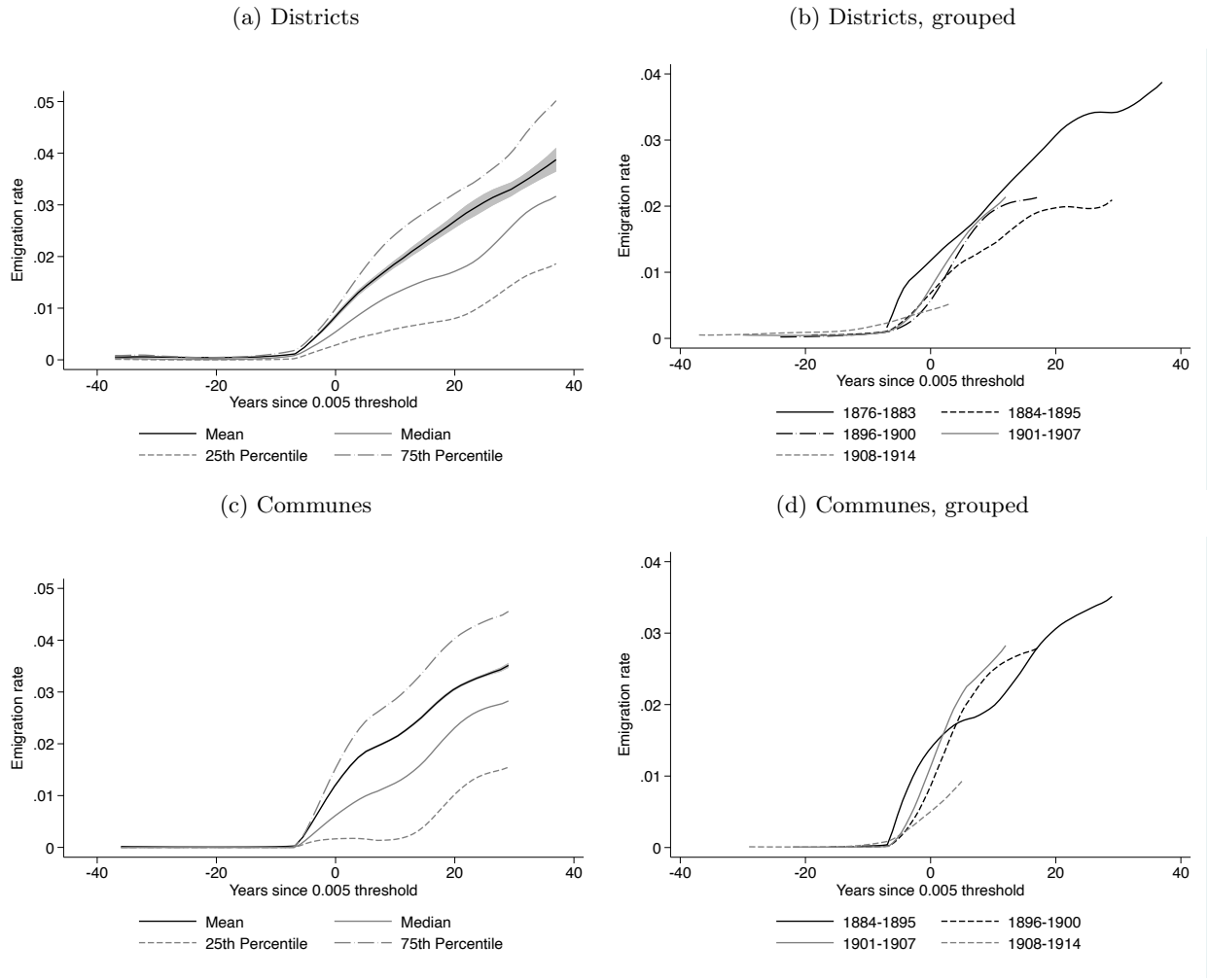


Figure H.3: S-shaped time series of migration to all destinations

Note: Panels (a) and (c) plot a non-parametric regression (the mean), as well as quartiles of emigration rates to any destination against time, normalized so that year 0 is the first year in which a place had an emigration rate of at least 5 per thousand. Shaded areas are 95-percent confidence intervals for the mean. Panels (b) and (d) are the same as (a) and (c) but divide areas according to the half decade in which they crossed the threshold.

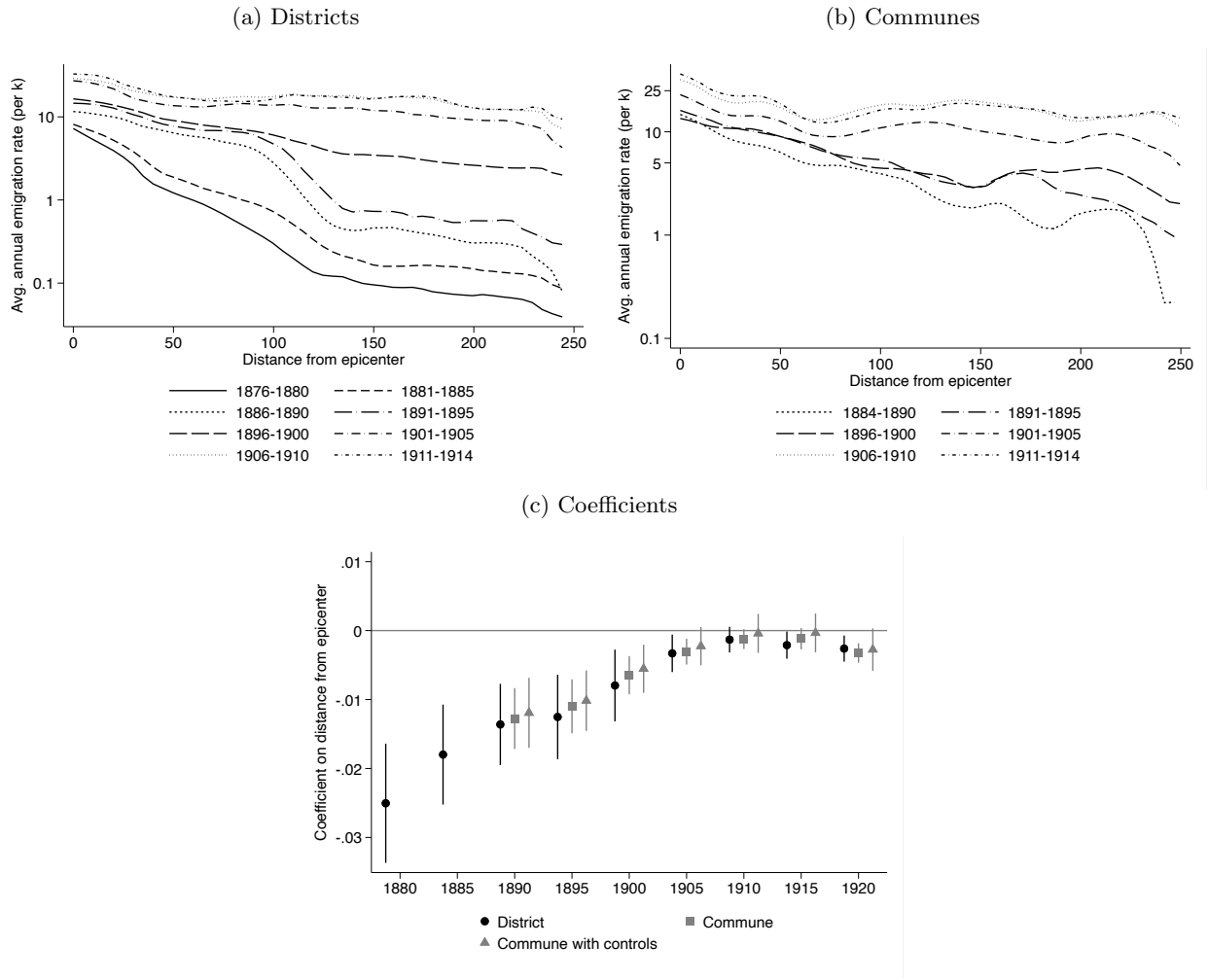


Figure H.4: Emigration rates to all destinations by distance to epicenter (km)

Note: Panels (a) and (b) plot non-parametric regressions of the log of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of distance from the nearest epicenter of emigration and plots the coefficients on distance from epicenter. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

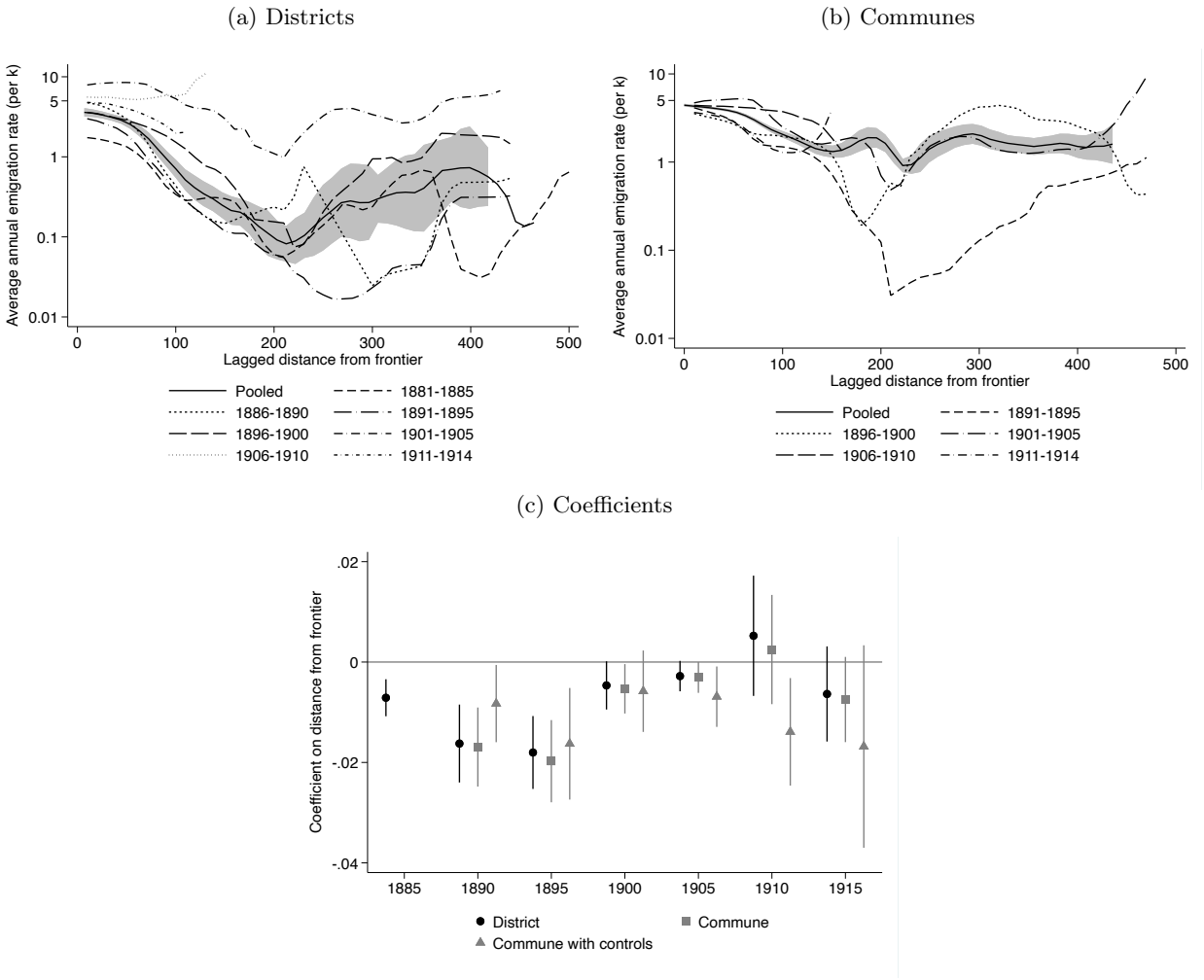


Figure H.5: Emigration rates to all destinations by distance to the mass migration frontier (km)

Note: Panels (a) and (b) present non-parametric regressions of the log of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to districts that had not yet achieved this threshold. Shaded areas are 95-percent confidence intervals. Panel (c) estimates a binomial maximum likelihood regression of emigration rates on half decade-specific functions of lagged distance from the frontier of mass migration and plots the coefficients on lagged distance from the frontier. Panel (c) also includes a regression controlling for half decade-specific functions of various controls.

Table H.2: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	−0.222	0.779[a]	0.654[a]	0.607[a]	0.711[a]	0.689[b]	0.455	0.621
	(1.268)	(0.143)	(0.227)	(0.224)	(0.234)	(0.330)	(0.331)	(0.454)
Observations	20,238	20,238	20,238	20,238	20,238	20,238	20,238	20,186
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	1.259	22.94	17.15	13.83	18.73	11.53	7.485	7.136

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to any destination. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table H.3: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.315	0.300	0.289	0.464[c]	0.392	0.643[b]	0.266	0.125
	(0.283)	(0.266)	(0.328)	(0.267)	(0.322)	(0.289)	(0.485)	(0.366)
Observations	1,689	1,688	1,688	1,684	1,688	1,682	1,685	1,668
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	27.43	34.73	26.66	31.22	45.01	55.30	33.01	37.16

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

I List of Archival Sources

J Results Including Observations with No Migration

Whenever the object of interest is the logarithm of emigration, communes with no emigration in a particular half decade must be excluded. In this appendix, we repeat the main results using $\log(e_{it}+\varepsilon)$, where $\varepsilon = 10^{-5}$ instead of $\log(e_{it}$ in order to incorporate these commune-half decades into the analysis. This is not necessary when the binomial maximum likelihood regression is used (since that is designed to account for cases of zero migration), and so this appendix only repeats the results where the change is necessary. The results are qualitatively unchanged.

Table J.1: β-convergence

Variables	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV	(7) IV	(8) IV
Lagged Own Emigration	-0.589^a	-0.766^a	-0.852^a	-0.887^a	-0.502^a	-0.752^a	-0.552^b	-0.768^a
	(0.024)	(0.022)	(0.015)	(0.014)	(0.062)	(0.062)	(0.276)	(0.202)
Observations	8,029	8,028	8,028	8,028	8,029	8,028	8,028	8,027
R-squared	0.578	0.747	0.879	0.912	0.565	0.747	0.691	0.806
Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
1st Stage F	.	.	.	.	69.557	73.957	6.681	4.627
FE	None	None	P	D	None	None	P	D
Falsification	0.029	0.124	0.332	0.393				
	(0.060)	(0.069)	(0.061)	(0.069)				

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the district level. Unit of observation is a commune. Dependent variable is the change in the log of the emigration rate to North America from the pre-1900 period to the period 1900 and later. Controls include latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Instrument is the distance to the nearest epicenter of emigration to North America. P denotes province-level fixed effects included. D denotes district-level fixed effects included. The falsification coefficient is the coefficient from regressing the change in emigration on emigration in the post-1900 period; if it is either negative or positive but of a smaller magnitude than the main coefficient of interest, this is evidence that the relationship is not spurious.

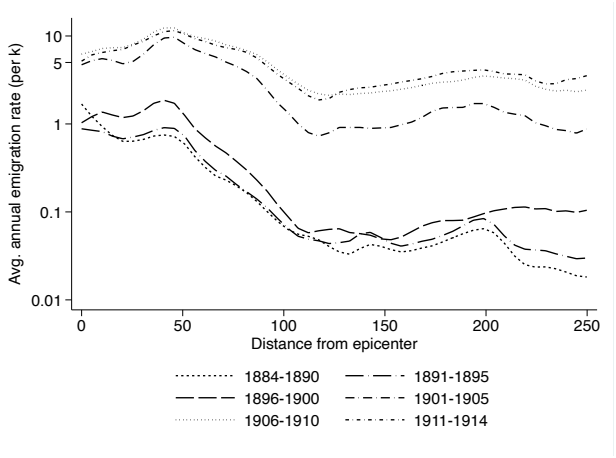


Figure J.1: Emigration rates to North America by distance to epicenter (km)

Note: This figure plots a non-parametric regressions of the log of the average annual emigration rate for each half decade against distance to the nearest epicenter of emigration to North America.

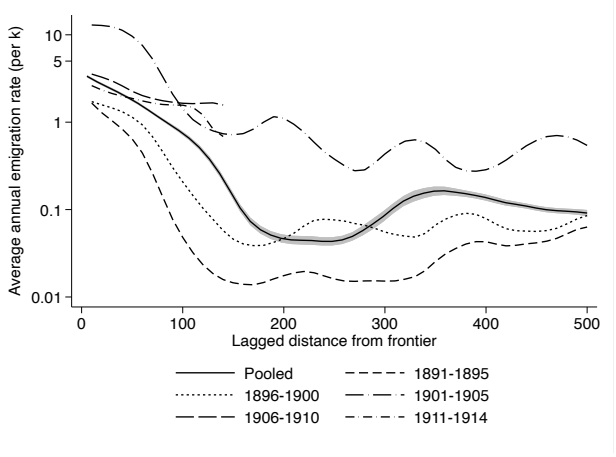


Figure J.2: Emigration rates to North America by distance to the mass migration frontier (km)

Note: This figure presents non-parametric regressions of the log of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 5 per thousand by the previous half decade, limiting the sample to districts that had not yet achieved this threshold. Shaded area is a 95-percent confidence interval.

Table J.2: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.725[a]	0.763[a]	0.436[c]	0.447[c]	0.402[c]	0.425	0.175	0.127
	(0.134)	(0.131)	(0.247)	(0.266)	(0.226)	(0.291)	(0.499)	(0.443)
Observations	37,128	37,128	37,128	37,128	37,128	37,128	37,128	37,104
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	29.04	46.22	18.28	13.05	24.78	10.55	27.01	8.264

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table J.3: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.723[a]	0.949[a]	0.685[a]	0.789[a]	0.706[a]	0.617[a]	0.185	0.480[c]
	(0.185)	(0.225)	(0.212)	(0.201)	(0.228)	(0.211)	(0.642)	(0.272)
Observations	12,877	12,875	12,875	12,873	12,875	12,864	12,875	12,850
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	60.61	45.27	45.27	42.01	42.23	41.23	29.92	23.63

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

K Robustness to Choice of θ

This appendix verifies the robustness of the results in section 6 to alternative choices of the parameter θ, which governs the rate at which the influence of other communes on the emigration exposure of a commune declines with the distance between them. In particular, two alternate values of θ are considered. Whereas that in the main text was chosen on the basis of estimating equation (4) by non-linear least squares without controls, the alternate values in this appendix were chosen after performing this estimation with controls and with geographic fixed effects. The results are qualitatively unaffected.

K.1 θ=-2.97

Table K.1: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.718^a	0.834^a	0.601^a	0.588^b	0.563^a	0.450^c	0.422	−0.120
	(0.182)	(0.128)	(0.188)	(0.236)	(0.150)	(0.248)	(0.483)	(0.588)
Observations	31,463	31,463	31,463	31,463	31,463	31,462	31,463	31,427
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	11.92	32.02	14.45	8.617	19.71	7.520	18.23	4.240

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table K.2: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.769[a]	0.835[a]	0.621[a]	0.734[a]	0.657[a]	0.599[a]	0.611	0.331
	(0.179)	(0.167)	(0.163)	(0.171)	(0.183)	(0.163)	(0.639)	(0.243)
Observations	11,207	11,206	11,206	11,205	11,206	11,196	11,206	11,170
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	28.90	40.05	46.05	35.77	35.20	32.58	13.70	19.94

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

K.2 θ=-2.71

Table K.3: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	1.109[a]	0.939[a]	0.729[a]	0.713[a]	0.629[a]	0.482[c]	0.563	0.376
	(0.081)	(0.118)	(0.175)	(0.206)	(0.164)	(0.255)	(0.468)	(0.397)
Observations	31,463	31,463	31,463	31,463	31,463	31,462	31,463	31,427
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	52.15	51.70	22.77	15.54	22.22	9.515	25.54	9.090

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

Table K.4: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	1.079^a	0.941^a	0.664^a	0.784^a	0.680^a	0.672^a	0.628	0.486^b
	(0.119)	(0.157)	(0.159)	(0.166)	(0.173)	(0.166)	(0.537)	(0.230)
Observations	11,207	11,206	11,206	11,205	11,206	11,196	11,206	11,170
Additional FE	None	None	C	CT	P	PT	D	DT
Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	96.83	62.79	68.92	52.57	60.42	54.20	25.20	35.53

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include at least half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. C denotes region (compartimento)-level fixed effects. P denotes province-level fixed effects. D denotes district-level fixed effects. CT, PT, and DT denote region-time, province-time, and district-time fixed effects.

L Results With Grid Fixed Effects

This appendix repeats the estimation of section 6, but instead of using fixed effects based on actual geographic divisions (i.e., region, province, and district), we use fixed effects for grids of various sizes, ranging from grid cells of 90-by-90 kilometers to 15-by-15 kilometers. This method, based on that used by Barsbai et al. (2017), is intended to show that the estimates are not the product of bias caused by unobservables by making a coefficient stability argument—if the coefficients are largely unchanged in the face of fixed effects for finer and finer grid cells, it is unlikely that local characteristics are responsible for the relationship. Although the results for the epicenter-based instrument are in many cases rendered statistically insignificant by these very fine controls, the results for the frontier-based instrument are robust, and moreover are largely stable across specifications.

Table L.1: Spatial contagion results, epicenter-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.508^{c}	0.345	0.583^{a}	0.425	0.621^{a}	0.675	0.709^{a}	1.767^{c}
	(0.303)	(0.792)	(0.220)	(0.617)	(0.179)	(0.636)	(0.160)	(0.928)
Observations	31,463	31,435	31,463	31,409	31,463	31,280	31,463	30,341
Additional FE	G	GT	G	GT	G	GT	G	GT
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grid Size	90	90	60	60	30	30	15	15
F-statistic	12.01	2.053	23.97	3.563	37.16	3.045	47.88	3.924

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Sample limited to communes between 50 and 250km of an epicenter of emigration to North America. Standard errors clustered at the district level. All specifications include half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the epicenter, local population, distance to coast, and distance to the European frontier, half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. G denotes the inclusion of grid fixed effects for grid cells of the specified size (e.g., 90-by-90km). GT denotes the inclusion of grid-half decade fixed effects.

Table L.2: Spatial contagion results, frontier-based instrument

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Emigration Exposure	0.660[a]	0.778[a]	0.791[a]	0.821[a]	0.602[c]	0.691[b]	0.979[c]	0.601
	(0.200)	(0.186)	(0.235)	(0.191)	(0.332)	(0.294)	(0.570)	(0.465)
Observations	11,205	11,182	11,202	11,166	11,199	11,093	11,159	10,674
Additional FE	G	GT	G	GT	G	GT	G	GT
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grid Size	90	90	60	60	30	30	15	15
F-statistic	48.58	53.78	34.03	59.92	25.05	37.11	14.44	26.99

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: Sample limited to commune-half decades between 50 and 250km of the frontier of mass migration to North America. Standard errors clustered at the district level. All specifications include half-decade fixed effects and control for half decade-specific functions of own predicted lagged emigration based on distance from the mass migration frontier, local population, distance to coast, and distance to the European frontier, half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. G denotes the inclusion of grid fixed effects for grid cells of the specified size (e.g., 90-by-90km). GT denotes the inclusion of grid-half decade fixed effects.

References

Barde, Robert, Susan B. Carter, and Richard Sutch (2006). "Table Ad106–120: Immigrants, by country of last residence—Europe, 1820–1997." In *Historical Statistics of the United States*. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright (ed.). Cambridge: Cambridge University Press, pp. 1.560–1.563.

Barsbai, Toman, Hillel Rapoport, Andreas Steinmayr, and Christoph Trebesch (2017). "The Effect of Labor Migration on the Diffusion of Democracy: Evidence from a Former Soviet Republic." *American Economic Journal: Applied Economics* 9:3, pp. 36–69.

Brandenburg, Broughton (1904). *Imported Americans: The Story of the Experiences of a Disguised American and His Wife Studying the Immigration Question*. New York: Frederick A. Stokes Company.

Ferenczi, Imre and Walter F. Willcox (1929). *International Migrations*. New York: National Bureau of Economic Research.

Hatton, Timothy J. and Jeffrey G. Williamson (1998). *The Age of Mass Migration: Causes and Economic Impact*. New York: Oxford University Press.

Jet Propulsion Laboratory (2014). *Shuttle Radar Topography Mission* [machine-readable database]. Pasadena: California Institute of Technology.