Immigrant Diversity and Long-Run Development*

Luigi Minale† Rudi Rocha‡ Bruno Vigna§

March 2025

Abstract

This paper investigates the long-term economic effects of immigrant diversity, focusing on the large immigration wave that Brazil experienced at the turn of the twentieth century, during the Age of Mass Migration. We examine whether municipalities in the State of São Paulo that received a population of immigrants characterized by a more diverse mix of origin countries achieved better long-term economic outcomes. To address potential endogeneity in immigrant diversity, we utilize unique historical, individual-level data on immigrants arriving in São Paulo between 1880 and 1920, coming from passenger arrival records. Our empirical strategy combines time variation in the origin-country composition of immigrants arriving from overseas —driven by external shocks in origin countries—with the timing of the railway network's expansion in São Paulo. Our findings reveal that a one standard deviation increase in immigrant diversity in 1920 is associated with a 7-8% higher income per capita in 2000, a result that is both economically significant and robust to various identification tests. Additionally, we explore the mechanisms driving these long-term effects and find that municipalities with greater immigrant diversity experienced: (i) a shift towards higher employment in manufacturing and services, accompanied by increased occupational diversity within the manufacturing sector; (ii) greater public investment in education, as reflected in higher municipal spending; and (iii) improved educational outcomes over the long term.

Keywords: birthplace diversity, immigration, long-term development.
JEL Classification: C36, N36, O15.

*We are grateful to Martin Fiszbein, Albrecht Glitz, Marco Manacorda, Daniel Oto, Sandra Sequeira, Giuseppe Sorrenti, Jan Stuhler, Felipe Valencia and seminar and conference participants at Tinbergen Institute Amsterdam, University of Pablo de Olavide (Seville), Lisbon Migration Economics Workshop (ISEG), LACEA-EHN Workshop on Historical Development (Santiago de Chile), UCL Economics PhD Alumni Conference (London), European Economic Association Congress, Spanish Economic Association Congress, and CReAM 20th Anniversary Workshop. Luigi Minale gratefully acknowledges financial support from the Spanish Ministry of Science and Innovation and María de Maeztu grant.
†Universidad Carlos III de Madrid, CReAM, and IZA.
‡São Paulo School of Business Administration, Getulio Vargas Foundation (FGV EAESP); São Paulo, Brazil.
§BNDES; Rio de Janeiro, Brazil.

1. INTRODUCTION

The large-scale movement of people over the past centuries has significantly altered the demographic composition and increased the diversity of populations worldwide (Putterman and Weil (2010)). Immigration and the resulting rise in social heterogeneity remain central challenges for modern societies. As Putnam (1997) famously observed, "the most certain prediction we can make about almost any modern society is that it will be more diverse a generation from now than it is today" (p. 137). With this continuous and profound transformation in population composition, understanding how cultural diversity influences economic well-being—and whether it fosters or hinders development—is crucial. This understanding not only sheds light on the development process itself but also informs current debates on migration policies in contemporary societies.

However, the economic effects of diversity remain a topic of significant debate, with both theoretical and empirical studies offering mixed conclusions.[1] On the one hand, diversity can enhance productivity and innovation by increasing the variety of skills in the society, and empirical studies such as Ottaviano and Peri (2006), Alesina et al. (2016), and Bahar et al. (2022) provide indeed evidence that birthplace and cultural diversity can stimulate economic growth. On the other hand, diversity can also have detrimental effects should it trigger mutual mistrust or lack of social cohesion, which can reduce the quantity and quality of public goods (Alesina et al., 1999; Alesina and La Ferrara, 2005). These negative effects, that can be mitigated by intergroup contact (Bazzi et al., 2019; Desmet et al., 2020), can be particularly salient in the short to medium run. In contrast, in the medium to long run, diverse societies might create new forms of social solidarity and new, more encompassing identities (Putnam, 1997). While this conjecture could make diversity ultimately desirable in equilibrium, the empirical literature has mostly focused on cross-country evidence or shorter-term effects.

In this paper, we investigate the relationship between the composition of the immigrant population (birthplace diversity) and the present-day per capita income level of the municipalities in the state of São Paulo. We also examine the mechanisms behind long-term impacts and persistency. For this purpose, we exploit the strong wave of immigration to Brazil in the late nineteenth and early twentieth centuries and sources of exogenous variation in diversity of origin among the state's municipalities. Between 1872 and 1920, roughly 1.8 million immigrants from more than 70 origin countries entered São Paulo. This massive inflow of people corresponded to more than twofold the initial population of the state in 1872, of 837 thousand inhabitants. Diversity spread fast, with external push factors continuously chang-

[1] See Alesina and La Ferrara (2005) for a rich survey of the pros and cons of diversity, with a focus on ethnic diversity.

ing the composition of immigrant arrivals and new railways distributing a varied pool of new inhabitants across the territory. We examine whether municipalities that ended up more diverse by the turn of the 19th century experienced higher output growth, as measured by higher income per capita by the turn of the following century.

Our main empirical strategy compares income per capita in 2000 and other outcomes measured at different moments throughout the 20th century, across municipalities that received a more *vs* less diverse pool of immigrants by the turn of the 19th century – proxied by a birthplace fractionization index computed for 1920. Identification relies on a 2SLS approach to overcome the confounding influence of omitted variables and measurement error. To compute our instrumental variable, we use a unique micro data set containing detailed records from the Immigration Station Houses of São Paulo logbooks. This data set provides individual-level information of approximately 1.6 million immigrants who arrived in São Paulo between 1882 and 1920, including first and last names, gender, civil status, nationality, profession, religion, intrafamily kinship, last country of residence and destination in Brazil (municipality, train station or farm). Based on these logbooks, we constructed a panel of data containing the annual numbers of immigrants by birth country to each destination municipality over the 1882-1920 period. Besides the logbooks of the Immigration Station Houses, we use data from 1872, 1920, 1940, and 2000 Brazilian censuses to characterize our empirical setting and compute dependent and independent variables. We also compute a vast array of auxiliary variables to measure geographic and transport conditions at the municipality level.

More specifically, our instrumental variable strategy exploits two features of immigration toward Brazil in the period under study. First, the total amount and the composition of immigrant flow as long as countries of origin are concerned vary considerably over time, mostly driven by idiosyncratic events occurring in those countries (such as changes in migratory legislation, conflicts, droughts etc). Second, the period under study coincided with the expansion of the railway network in the São Paulo State, which went from east to west, accompanying the occupation of the state's territory. Therefore, we combine variation over time in the composition of immigrants arriving in Brazil with the timing of a municipality's connection to the railway network to obtain exogenous variation in the diversity index of the migrant population that settled in each municipality between 1872 and 1920. The intuition is that a municipality that gets connected to the network when migrant inflow is relatively more heterogeneous will end up with a more diverse pool of migrant population than a municipality connected in a period when migrant inflow occurred to be more homogeneous. Importantly, we only exploit the interaction between the timing of railroads construction and the fluctuation over time of origin-specific immigration flows to construct the instrument.

The resulting variation is likely not correlated with other factors affecting today's per capita income other than the population composition.

We find that immigrants' birthplace diversity positively impacts per capita income in the long run. A one standard deviation increase in the immigrant diversity index in 1920 is associated with a 7-8 percentage points higher income per capita in 2000, an effect that is both statistically robust and economically relevant. We conduct robustness tests to affirm that other confounding factors do not influence the results. First, results remain relatively stable to the inclusion of geographic controls and baseline socioeconomic characteristics (measured in 1872, before the flow of immigrants). Second, we show that the diversity of the immigrant population has an effect over and above its size by controlling for the share of foreigners relative to the local population, measured both in 1872 and in 1920. Third, we show that estimates are robust to including a variable capturing the time elapsed since each municipality has been connected to the railway network.

Why does the diversity of the immigrant stock affect long-run development? To shed light on how immigrant-induced diversity translated into higher income per capita over time, we examine some of the potential channels suggested by the literature. We first analyze the evolution of the structure of employment, which indicates the extent to which the occupational diversity and the variety of goods and services available in the local economy changed over time. We observe that the immigrant diversity is positively associated with a higher share of people working in manufacturing and services, and away from agriculture, both in the short and the long-run. Second, we turn to public investment in schooling and human capital accumulation as potential mechanisms. We find that immigrant diversity is positively associated with investment in public goods, as measured by municipal spending on education and schools per child (both in 1920 and 1940). The larger investment in education materialises into higher enrollment rates (measured in 1940) and average years of schooling in the municipality today.

This paper contributes to three main strands of literature.

First, we contribute to the small but growing group of papers exploring, with historical data, the long-run impact of migration. The most prominent contributions within such literature focus on the Age of Mass Migration experienced by the United States between 1870 and 1920, and document long-run effects of immigration on economic prosperity (Sequeira et al., 2020; Tabellini, 2020) and political views (Giuliano and Tabellini, 2020). However, here we are interested in the effects of the *composition* of the immigrant population, rather than its *size*, which is the focus of most previous research, although with some expectations. For instance, Fulford et al. (2020) show how ancestry composition are associated with changes in local economic development across US counties, while Ager and Bruckner (2013) highlight

opposite effects of fractionalization (positive) and polarization (negative) on local output, while using a supply-push component of immigrant inflows during the age of mass migration as an instrumental variable. Finally, closer to our paper is Rocha et al. (2017), who use variation in immigrants' human capital to examine whether that was conducive to long-run development in São Paulo, and Droller (2017), who documents that Argentinean counties with historically higher shares of European population in 1914 experienced higher per capita GDP about 80 years later. While the author focuses on the impact of European immigrants in general, here we are interested in the impact of the composition of the immigrant population and in within-migrant diversity.

Second, we of course also talk to the broader empirical literature on cultural and birthplace diversity. Some of the papers in this literature are based on cross-country analysis (Alesina et al., 2016; Bove and Elia, 2017; Bahar et al., 2022), while others focus at the U.S. experience and use immigration as a source of variation in birthplace diversity either at the state (Docquier et al., 2020) or at the city level (Ottaviano and Peri, 2006, 2005b) We contribute to this literature by moving away from cross-country analysis, something that can help better identifying causal effects, and by taking a long-run perspective.

Finally, our paper relates to the literature on nation-building and its interaction with diversity. While diversity can pose obvious challenges to nation-building, prior research has highlighted various factors, such as inter-group contact (Bazzi et al., 2019) and compulsory schooling (Bandiera et al., 2013; Murard, 2023), that can mitigate these negative effects and foster the creation of a shared national identity. Our findings align with this literature, suggesting that public investment in education in Brazil may have been used as a tool to homogenize and assimilate an increasingly diverse population.

The remainder of this paper is organized as follows. Section 2 summarizes the historical context. Section 3 contains a description of the data and provides descriptive statistics. We present and discuss our empirical strategy in Section 4. In Section 5 we present the main results and examine the mechanisms behind them. Section 6 concludes.

2. Historical Context

Over 4.1 million immigrants arrived in Brazil between 1872 and 1929, of which nearly 3 million came from Italy, Portugal, Spain, Germany and Japan, and another 600 thousand came from 70 other nations (Bassanezi et al., 2008; Rocha et al., 2017). Regarding European immigration to the Americas in general, estimates are that some 50 million people arrived between the start of the nineteenth century and the outbreak of World War I in 1914. Of these, about 11 million came to South America, of them 38% Italians, 11% Portuguese, 3% French and 3% Germans. This wave of immigration drastically changed the makeup of the

Brazilian population. The Census of 1872 found a total population of 9.930 million people, of whom 383 thousand (3.8%) were foreigners or naturalized citizens. In 1900, the resident population amounted to 17.4 million, of which 1.279 million were foreigners or naturalized citizens, or 7.3% (IBGE, 2007).

Starting in the mid-nineteenth century, the expanding coffee economy created a strong demand for labor and immigration has contributed to the agricultural development as measured by the increase in farm values (Escamilla-Guerrero et al., 2024).[2] The first experiences of immigration other than colonists from Portugal occurred in the late 1810s, with the arrival of Swiss and Germans, who founded the small colonies of Leopoldina and Morro Queimado in today's state of Rio de Janeiro, and in 1824, of Nova Friburgo, also in Rio. Before this experience, a small contingent arrived from the Azores to what is now the state of Santa Catarina, in the mid-seventeenth century. However, the major flow of new immigrants started in the 1850s.

In the then province of São Paulo, where coffee growing was rapidly expanding toward west, the shortage of labor was most critical. The prohibition of the slave trade in 1831 and the Law of the Free Womb (enacted in 1871, by which children born to slave mothers were considered free upon reaching adulthood) made it clear that slavery's days were about to end. Keeping slaves became increasingly risky and expensive.[3] Based on the growing international demand for coffee, foreign immigration to São Paulo occurred with strong governmental intervention. For the purpose of attracting fee laborers, Provincial Law 28 was enacted on March 29, 1884, marking the start of subsidized immigration in São Paulo. It gave priority to the entry of families and adult men, and as an instrument of attraction, established payments to the new immigrants. It created a government apparatus and administrative organization specifically aimed at ordering and promoting immigration. The State Secretariat of Agriculture, Commerce and Public Works (SACOP) was created in 1891, the same year of promulgation of the State Constitution. The Inspectorate of Lands, Colonization and Immigration was created within SACOP. The Secretariat and SACOP were in charge of managing the *Hospedaria dos Imigrantes*, a hostel in the Bom Retiro neighbourhood in the city of São Paulo, later moved to the Brás, which served as a gateway, an inspection station, and marshaling place for the immigrants for their first few days or weeks in Brazil,

[2] Slavery was abolished in Brazil in 1888, but due to the large portion of the population composed of slaves and free blacks and browns (upon manumission), a conscious effort was made in the second half of the nineteenth century by the elites to attract European immigrants.

[3] In 1881, the São Paulo provincial government imposed a tax on the transfer of slaves from other provinces. Besides this, there were "mass escapes and slave resistance, the abolitionist movement and the fear of insurrection (...)" (Goncalves (2009), p.4). In this context, support for subsidized immigration became a pragmatic position of the Paulista coffee growers. "They were neither abolitionists nor slavery supporters; they just needed workers" (ibid, p. 4).

until regularization of their labor contracts. It also worked as a sanitary quarantine site since health conditions were poor among the recently arrived after the transatlantic trip. See a picture of the *Hospedaria dos Imigrantes* from year 1915 in Appendix Figure A1.

Between 1880 and 1900, the province of São Paulo received approximately 940 thousand immigrants (Vasconcellos, 1994). To have an idea of the dimension of this flow, the 1872 Census indicated that the province's total population was 837 thousand. The majority of the immigrants arriving in São Paulo had their travel subsidized and had arranged their final destination in the interior of the province.[4] At the Hostel it was also possible to sign contracts with coffee growers seeking workers (Colistete and Lamounier, 2011). The Hostel's enrollment logs, which were digitized and are available on the Internet for consultation, record the arrival of immigrants of 76 different nationalities.

3. DATA

We use three main sources of data. The first and most novel consists of the records collected at the moment of the immigrants' arrival at the Brás *Hospedaria* in São Paulo. These records reveal the annual numbers of immigrants by birth country and the municipality of destination. Second, we rely on data from the censuses of 1872, 1920, 1940 and 2000 at the municipal level. The third dataset consists of the geographic and socioeconomic variables, tallied by Ipeadata and Embrapa Solos, along with information about transportation infrastructure.

Economic growth and population densification over the years increased the administrative division in the state of São Paulo. While in 1872 it was divided into 88 municipalities, in 1920 there were 202 and in 2000 the number had reached 645. Our analysis relies on the division into 202 municipalities of 1920: the 645 municipalities of 2000 are aggregated according to the borders of 1920, and the original data of the 88 municipalities of 1872 are connected to the corresponding territories of 1920. Therefore, a single municipality in 1872 can correspond to several in 1920, but the territories are comparable in time between 1920 and 2000.

[4] According to data from the Museum of Immigration, 60% of the immigrants who arrived at the Hostel between 1889 and 1915 had their travel subsidized by the a Brazilian government (national or state), with the rest being classified as spontaneous or self-financed immigrants.

3.1. BIRTHPLACE DIVERSITY INDEX

To derive a measure of birthplace diversity among the immigrant population, we start from the more general index of fractionalization used in the literature:[5]

$$Fractionalization_{pop} = \sum_{i=1}^{I} s_i \times (1 - s_i) = 1 - \sum_{i=1}^{I} (s_i)^2$$

Where s_i is the proportion on the total population of individuals born in country 'i'. In particular, $s_{i=1}$ denotes the share of natives. Inspired by the Herfindahl–Hirschman index of concentration (HHI), the fractionalization index measures the probability that two individuals in each given municipality, chosen at random, have two different countries of birth. Zero fractionalization means no diversity, when all the residents of the municipality were born in the same country, while fractionalization equal to one occurs when each inhabitant of the municipality was born in a different country. The above fractionalization index, however, has two dimensions and it is important to separate them (Alesina et al., 2016). The first is the proportion of foreigners $(1 - s_1)$,[6] which is independent of the country of origin and captures the scale of the immigrant flows. The second is the intragroup variety of immigrants. In other words, the proportion of foreigners measures the diversity in the intergroup dimension (between natives and non-natives) while the variety of foreigners measures the intragroup dimension (only among the foreigners). We are interested in the impact of this latter component.[7] We therefore refer to the following measure of diversity within the immigrant group:

$$Div_{migrants} = 1 - \sum_{j}^{J} (s_j)^2 \tag{1}$$

Where s_j indicates the proportion of immigrants with origin j among the total number of non-natives: $s_j = (s_i)/(1 - s_1)$. Intuitively, the index is calculated analogously to the fractionalization one but only captures diversity in the immigrant population without depending on the size of the immigrant population. We will call this measure Immigrant Diversity Index.

---

[5] For example, see Alesina and La Ferrara (2005), Beach and Jones (2017), Ottaviano and Peri (2005a), and Ager and Bruckner (2013).

[6] Here s_1 is the proportion of natives.

[7] For groups formed all of foreigners $(1 - s_1)$, independent of country of origin, the index boils down to the sum of only two terms, s_1 and $(1 - s_1)$ and can be written as:

$$Frac = 1 - \sum_{i} (s_i)^2 = s_1 * (1 - s_1) + (1 - s_1) * s_1 = 2 * s_1 * (1 - s_1)$$

Thus, the index is a function of $(1 - s_1)$ and leads to confusion between the size of the scale of the flow with the intragroup variety (only among foreigners).

3.2. The Logbooks of the Immigrant's *Hospedaria*

Most of the logbooks of the Brás Immigrant's Hostel in São Paulo were digitized by the State Culture Secretariat in the 1990s. The Hostel building now houses the Immigration Museum of São Paulo, which makes the digitized database available at its website.[8] This database contains information on approximately 1.6 million immigrants who arrived in Brazil after 1882.[9] The records give the first and last names of the new arrivals, as well as their gender, civil status, nationality, profession, religion, intrafamily kinship, last country of residence, destination in Brazil (municipality, train station or farm) and information about the financing of their travel (if subsidized by the government – central or state – by immigration promotion societies, or spontaneous or self-sponsored). Based on these logbooks, we constructed a database that describes the annual numbers of immigrants by origin country and destination municipality. When the immigrant's destination was a train station or farm, we researched the location to attribute the destination municipality correctly. The records where the station or farm could not be located were discarded.[10]

3.3. Outcomes and Control Variables

Our main dependent variable is the natural logarithm of the per capita income at the municipal level, calculated based on the microdata from the 2000 census, but adapted to the administrative arrangement of 202 municipalities in 1920. In the discussion of the mechanisms by which diversity affects the long-term development of localities, we use census data from 1872, 1920, and 2000.

Among the control variable used are (i) geographic characteristics of the municipalities; (ii) 1872 census data, to examine the socioeconomic conditions before the immigration shock; (iii) a dummy for the presence of a railroad serving the municipality, which takes the value of one for years when a railroad line crossed through a municipality, and zero otherwise. Below we detail the three blocks of variables.

The controls for geographic characteristics of the municipalities come from the Ipeadata and Embrapa Solos databases. We use dummies that indicate the predominant soil type in each municipality (among the four most common in the state: argisol, latosol, cambisol and spodisol). They are included in the regressions as controls to deal with the potential het-

[8] http://www.museudaimigracao.org.br/.

[9] Although estimates from the historical literature indicate the arrival of approximately 2.5 million immigrants at the Hostel (Holloway, 1984; Vangelista, 1991; Mauch and Vasconcellos, 1994; Goncalves, 2009), only 65% of the records have been digitized. According to the Center for Preservation, Research and Reference (CPPR) of the Immigration Museum, not all the logbooks have been digitized for preservation.

[10] The Immigrant Hostel's logbooks contain 1,264,204 records up to 1920, of which 689,161 identify a destination. Of these, 2,863 (0.4%) give locations outside the state of São Paulo and 1,303 (0.2%) give destinations that could not be assigned to a municipality from what was described in the digital database.

erogeneity of the soil quality or agricultural productivity. Among the geographic variables are latitude, longitude, distance from the capital (city of São Paulo) and elevation. Before the wave of immigration, the regions more to the northeast of the state had smaller population density. Therefore, it is important to consider the relative geographic isolation of a municipality, because this can affect the timing of the economic expansion.

The census data for 1872 describe the socioeconomic characteristics in the period before the influx of immigrants to the state of São Paulo. The set of variables includes the proportion of slaves, children enrolled in schools and literacy rate of the total population, as well as the ratio of people occupied by sector (agriculture, industry and services/retail) and population density.

The third dataset refers to railroad infrastructure. The movement of the agricultural frontier and development of urbanization of the state were closely related to the means of transport necessary to haul the exportable output (Dean, 1977). The expansion of the railway network was often financed by coffee growers, such that they were usually among the stockholders of the railroad companies (Goncalves (2008)). We use historical data on tailway expansion and line construction to assign to each municipality the year in which it has been connected to the railway network. The data was collected from the website *Estações Ferroviárias*.[11]

3.4. Descriptive Statistics

Table 1 presents a summary of the descriptive statistics based on the administrative division of 1920, with 202 municipalities. Panel A brings the data on diversity, proportion of foreigners in the total population and presence of railroads in the municipality. The diversity index presented refers to the intragroup diversity, analogous to fractionalization, commonly used in the literature, but independent of the native's shares. It only reflects the diversity among foreigners and takes values between zero and one, where one means that every foreigner in the municipality had different origin. As far as the proportion of foreigners is concerned, it rose from 1% to 13% between 1872 and 1920. The proportion of connected municipalities was 2% in 1882, approximately 20% (40 municipalities) in 1880, and reached 74% (149 municipalities) in 1920.

Panel B brings a descriptive summary of the geographical variables used as controls. Given that all the municipalities in the sample are in the same state, there is very small variance of latitude, longitude and elevation. Regarding soil type, latosol is predominant in 55% of the municipalities, while argisol is in 37%. This is not surprising, since these are the two leading soil types in the state of São Paulo and both are suitable for growing coffee,

[11] www.estacoesferroviarias.com.br.

the main crop in the region at the end of the nineteenth century. Panels C, D and E, in turn, report the socioeconomic characteristics of the sample at three times: 1872, 1920 and 2000. Of particular note is that the average municipal population doubled between 1872 and 1920 while the proportion of literate people only grew by 30%. In contrast, in 2000 universal basic education held sway, reflected in a literacy rate and school enrollment rate near 100%. The structural transformation of the economy between 1920 and 2000 also deserves special attention, since the preponderant sector in terms of occupation shifted from the primary to the tertiary sector and industry employed 25% of the workforce (versus 10% in 1920).

TABLE 1: DESCRIPTIVE STATISTICS

VARIABLES	(1) Avg.	(2) St.Dev.	(3) Min	(4) Max	(5) N
A. Diversity e Share of Foreigners					
1920 Birthplace Diversity Index	0.431	0.284	0.000	0.821	202
1872 Share of Foreigners	0.012	0.016	0.000	0.081	202
1920 Share of Foreigners	0.125	0.096	0.000	0.357	202
1872 Dummy railway	0.024	0.156	0.000	1.000	202
1920 Dummy railway	0.738	0.441	0.000	1.000	202
B. Geographical Variables					
Distance to the capital (log km)	5.139	0.693	2.078	6.243	202
Latitude	-22.594	0.989	-25.015	-20.131	202
Longitude	-47.607	1.404	-51.448	-44.386	202
Elevation (x 100m)	6.023	1.866	0.010	11.980	202
Latosols Dummy	0.546	0.455	0.000	1.000	202
Argisols Dummy	0.374	0.435	0.000	1.000	202
Cambisols Dummy	0.114	0.300	0.000	1.000	202
Spondosols Dummy	0.012	0.095	0.000	1.000	202
C. 1872 Census Variables					
Proportion of slaves	0.154	0.086	0.039	0.531	202
Proportion of literate (>6 y.o.)	0.202	0.110	0.048	0.452	202
Proportion of children enrolment	0.145	0.101	0.027	0.764	202
Population (x 1,000)	11.147	7.198	1.566	41.751	202
Proportion in agriculture	0.594	0.099	0.351	0.908	202
Proportion in industry	0.109	0.044	0.020	0.244	202
Proportion in services/retail	0.296	0.091	0.058	0.569	202
D. 1920 Census Variables					
Proportion of literate (>6 y.o.)	0.296	0.099	0.103	0.699	202
Schools/school-age child (x 1,000)	0.387	0.377	0.000	1.863	202
Professors/school-age child (x 1,000)	9.562	6.938	0.826	45.955	202
Population (x 1,000)	22.635	43.101	2.917	577.621	202
Proportion in agriculture	0.784	0.133	0.071	0.957	202
Proportion in industry	0.091	0.071	0.005	0.495	202
Proportion in services/retail	0.125	0.072	0.035	0.549	202
Proportion of catholics	0.849	0.182	0.098	1.000	202
Proportion of protestants (1940)	0.019	0.015	0.000	0.076	202
E. 2000 Census Variables					
Proportion of literate (>6 y.o.)	0.952	0.016	0.854	0.989	202
Proportion of children enrolment	0.962	0.018	0.863	0.991	202
Schools/school-age child (x 1,000)	0.387	0.377	0.000	1.863	202
Professors/school-age child (x 1,000)	82.833	21.228	19.670	152.683	202
Years of schooling (>5 y.o.)	5.538	0.660	3.532	7.109	202
Population (x 1,000)	183.329	818.392	2.867	11,086.8	202
Per capita income (log)	5.693	0.272	4.791	6.392	202
Proportion in agriculture	0.197	0.134	0.003	0.595	202
Proportion in industry	0.252	0.087	0.097	0.545	202
Proportion in services/retail	0.533	0.098	0.254	0.784	202

Note: all panels use the 202 municipalities sample considering the 1920 administrative boundaries. The geographical data source is Ipeadata (distance to the State's capital, latitude, longitude e elevation) and Embrapa Solos (kinds of soil). The railway presence indicator for every municipality was constructed using data from www.estacoesferroviarias.com.br. The remaining of the data correspond to the municipality's socioeconomic characteristics available in the Census (1872, 1920 and 2000).

4. EMPIRICAL STRATEGY

In this paper we investigate the relationship between immigrant's birthplace diversity and the present-day per capita income level of the municipalities in the state of São Paulo. For this purpose, we focus on the large wave of immigration to Brazil occurred in the late nineteenth and early twentieth centuries. At the end of the immigration period, around 1920, different municipalities in the São Paulo state ended up with different composition of the immigrant population. We start exploring our research question by estimating the following OLS equation:

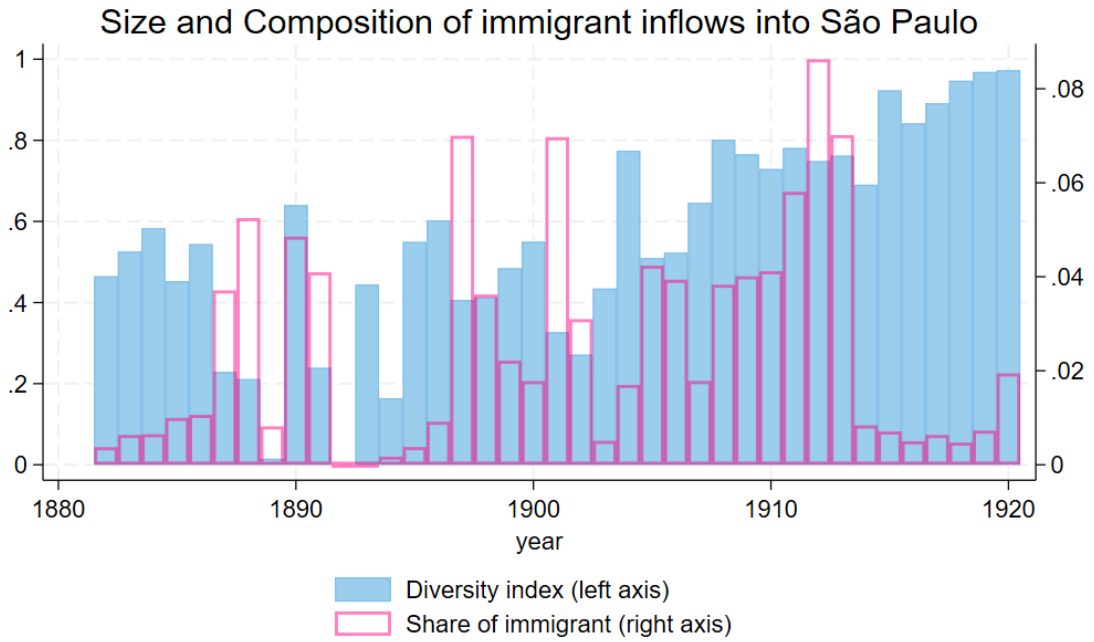
$$y_{m,2000} = \beta_0 + \beta_1 Div_{m,1920} + X'_m\gamma + e_m \tag{2}$$

Where $y_{m,2000}$ is the per capita income in 2000 of municipality m; the term $Div_{m,1920}$ is the birthplace diversity index of immigrants defined previously; and the term X_m is a vector of control variables that includes geographic characteristics, such as latitude, longitude, distance from the capital and elevation, along with soil type, as well as socioeconomic characteristics before the large influx of immigrants, extracted from the 1872 census, such as the proportion of slaves, foreigners, children enrolled in school, literacy rate of the total population, the proportion of workers occupied by sector (agriculture, industry and services/commerce) and population density. Lastly, in the more complete specifications, we also include controls for the proportions of foreigners in 1872 and 1920 to assure that the effect of diversity is not affected by the scale of the flow of immigrants.

Employing this specification, however, it is not possible to guarantee the identification of causal impacts. Migrants might non-randomly select into destination areas according to their economic prospects and other factors, generating a spurious correlation between birthplace diversity in 1920 and long-run development. Besides migrant non-random selection, during the more than 100 years, other variables (omitted or unobservable) correlated with immigration and the composition of the workforce might have affected the long-run economic performance. There is also a potential error in measuring the independent variable due to incompleteness of historical records. All these factors contribute to making OLS estimations of equation 2 potentially biased.

To identify causal effect, we develop an instrumental variable strategy that exploits two main features of immigration toward Brazil in the period under study. First, the total amount and the composition of immigrant flow in terms of countries of origin varied considerably over time, mostly driven by idiosyncratic events occurring in those countries, such as changes in migratory legislation, conflicts, and droughts. As Figure 1 shows, we observe independent variation in the size and the composition of the immigrant inflows over time. The periods

FIGURE 1: SIZE AND COMPOSITION OF IMMIGRANT INFLOWS INTO SÃO PAULO: 1882-1920



Note: The figure reports the annual numbers of immigrants arriving at *Hospedaria* and the annual diversity indexes of those immigrants, according to the data in its records.

of more diverse flows (in terms of countries of origin) are not systematically those with the largest arrival of immigrants, as also shown in Appendix Figure A2.

Second, the period under study coincided with the expansion of the railway network in the São Paulo State, which went from east to west, accompanying the occupation of the state's territory (see Appendix Figure A3). Therefore, we combine variation over time in the composition of immigrants arriving in Brazil with the timing of a municipality's connection to the railway network to obtain exogenous variation in the diversity index of the migrant population that settled in each municipality between 1872 and 1920. The intuition is that a municipality that got connected to the network in a period when the migrant inflow happened to be relatively more heterogeneous ended up with a more diverse pool of migrants.

4.1. ZERO-STAGE AND SYNTHETIC DIVERSITY INDEX CONSTRUCTION

To construct our instrument, we proceed in various steps. First, we use yearly panel data at the municipality-country of origin level to estimate a set of *zero-stage* equations, one for each origin country. We work with the five main countries of origin (Italy, Spain, Portugal, Germany and Japan), which accounted for about 85% of all the immigrants arrived during

the period. More specifically, we estimate:

$$Imm^{c}_{m,t} = \alpha_0 + \alpha_1 ImmFlow^{c}_{t} xRailAccess_{m,t} + \alpha_2 RailAccess_{m,t} + \kappa_m + \mu_t + u^{c}_{m,t} \quad (3)$$

For each country of origin c we regress the number of immigrants arriving into the destination municipality m in year $t \in \{1882, ..., 1920\}$, on a set of municipality dummies, year dummies, an indicator for the municipality being connected to the railway network at time t and an interaction term between the latter and the number of immigrants arriving into the state of São Paulo from country c in year t. The parameter α_1 captures the differential effect that connection to the railway has on, for instance, Italian immigrant settlement during periods of high aggregate immigration from Italy relative to periods of low aggregate immigration from Italy. Notice that by estimating a different equation for each country of origin, we allow the effect of access to the rail network and its interaction with the flow of immigrants to be origin-country specific.

Table 2 presents estimates of the zero-stage equation. The estimated coefficients for the parameter of interest, the interaction term between the total inflows from a specific country and the municipality's access to the railway network, are positive and highly significant in all the five specifications. The positive coefficients suggest that municipalities receive more immigrants when connected to the railway network. Notice that the value of origin-specific immigrant flows per year is captured by year fixed effects. While the coefficients on the rail-access indicator, all non statistically different from zero except from one, indicate that access to railway does not affect the number of migrants received when the aggregate amount of immigrants arriving in the state is zero.

As a next step in constructing our instrument we predict the number of immigrants who have settled by municipality and year from 1882 to 1920 for each origin country. Importantly, in predicting the number of immigrants we only rely on the interaction term in the equation above, as follows:

$$\widehat{Imm}^{c}_{m,t} = \widehat{\alpha_1} ImmFlow^{c}_{t} \times RailAccess_{m,t} \quad (4)$$

Therefore, we only exploit the interaction between the timing of railroads construction and the fluctuation over time of origin-specific immigration flows. The resulting variation is likely not correlated with any other factors affecting today per capita income other than the population composition. Based on the predicted values for $\widehat{Imig}^{c}_{m,t}$, we sum the predicted yearly number of immigrants from origin c that arrived in municipality m over the period 1882 to 1920. With the predicted origin-specific total number of immigrants who have settled during the 1882 to 1992 period in each municipality, we can finally construct our instrument:

TABLE 2: ZERO-STAGE ESTIMATES

Dependent variable	Number of immigrants arriving to the municipality				
	ITA	ESP	POR	GER	JAP
	(1)	(2)	(3)	(4)	(5)
Imm Flow x Rail Access	0.011	0.008	0.007	0.006	0.008
	(0.002)***	(0.001)***	(0.003)**	(0.002)***	(0.003)**
Rail Access	-18.826	-12.081	-9.484	-0.487	-0.312
	(7.549)**	(10.786)	(11.703)	(0.611)	(2.217)
Observations	7,878	7,878	7,878	7,878	7,878
R-squared	0.456	0.381	0.210	0.274	0.113
Year FE	YES	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES	YES

Note: All columns report OLS results, where the dependent variable is the number of immigrants from country 'c' arriving at municipality 'm' in year 't'. Robust standard errors are clustered by municipality according to the 1920 administrative division. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

the predicted diversity index in 1920 for municipality m, which is defined as our Synthetic Diversity Index ($SDI_{m,1920}$ or SDI), and is calculated according to the formula in equation 1.

4.2. TWO STAGES LEAST SQUARES AND IDENTIFICATION

We are now ready to estimate the long-run impact of immigrant birthplace diversity on economic prosperity using the SDI as an instrument for the 1920 Diversity Index. In particular, we estimate a 2SLS model, where equations (5) and (6) below are the first and second stage, respectively.

$$Diversity_{m,1920} = \delta_0 + \delta_1 SDI_{m,1920} + \delta_2 ShareForeigners_m + \delta_3 RailTiming_m + X_m'\theta + v_m \tag{5}$$

$$y_{m,2000} = \gamma_0 + \gamma_1 \widehat{Diversity}_{m,1920} + \gamma_2 ShareForeigners_m + \gamma_3 RailTiming_m + X_m'\eta + \varepsilon_m \tag{6}$$

In the first stage (equation 5), we regress the birthplace diversity index on the municipality 'm' in 1920 on our instrument, the SDI, and controls. The analysis is performed on the cross-section of municipalities. In the second stage (equation 6), 'y' represents some economic outcome of interest, such as per capita income, employment composition or government spending in education, generally measured in 2000 or at an intermediate time between

1920 and 2000.

Because we are interested in the composition of the immigrant stock, rather than in its size, one important control variable we include to distinguish the effect of birthplace diversity from that of the size of the immigrant population is the share of immigrants on the total population, measured either in 1872 or in 1920 ($ShareForeigners_m$). If the Diversity Index and the share of immigrants in the population were correlated, omitting the latter would lead to biased estimates of δ_1 and γ_1, as long as the share of immigrants directly affects the outcomes we are interested in.

We leverage on the timing when municipalities are connected to the railway network to construct the instrument. A potential threat to exclusion restriction is that early/late connection to the network might independently affect long-run development through, for instance, facilitating trade. Further, it might be the case that municipalities that connected relatively earlier to the railway network tended to receive a more diverse flow of immigrants. To control for early/late connection to the railway network, we include the number of years elapsed between the railway network connection and year 1920.

Finally, the term X_m includes several municipality level covariates. First, we control for time-invariant geographical characteristics such as distance to the state's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Second, we include some socioeconomic municipality characteristics measured at baseline (1872) such as literacy ratio, the share of slaves, of workers in agriculture, industry and services/retail.

Table 3 reports estimates from the first-stage regression. The coefficient associated to the Synthetic Diversity Index is precisely estimated across the board (statistically significant at 1% level in all specifications). It suggests that our instrument is a strong predictor of the actual birthplace Diversity Index in 1920. The first column reports estimates from the baseline specification, which includes only geographical and socioeconomic municipality controls. Adding the share of foreigners measured either in 1920 (column 2) or 1872 (column 3) leaves the coefficients almost unchanged. This result is reassuring as it provides evidence that the instrument loads on variation uncorrelated with the immigrant stock's size and helps dispel doubts about our 2SLS estimates factoring in the direct effect of the size of immigrant stock. In columns 4 to 7, we further add the time elapsed since the connection to the railway network (Railway Connection Timing), first on its own (column 4) and then in addition to the share of foreigners (columns 5 and 6). The estimated coefficient on Railway Connection Timing is positive and statistically significant, the estimates of our IV get smaller, but the first stage remains strong. Even in our most complete specifications (columns 5 and 6), the F-test value is around 40-41, well above the conventional thresholds for weak instruments.

TABLE 3: FIRST STAGE

Dependent Variable	1920 Observed Diversity Index					
	(1)	(2)	(3)	(4)	(5)	(6)
1920 Synthetic Diversity Index	0.764	0.753	0.703	0.487	0.491	0.486
	(0.061)***	(0.061)***	(0.068)***	(0.076)***	(0.076)***	(0.076)***
Share of Foreigners (1872)		0.007			0.003	
		(0.004)*			(0.003)	
Share of Foreigners (1920)			0.575			0.026
			(0.238)**			(0.239)
Railway Connection Timing				0.009	0.008	0.008
				(0.001)***	(0.001)***	(0.001)***
Observations	202	202	202	202	202	202
Partial-F (Synthetic Diversity Index)	156.07	149.96	106.47	40.84	41.25	40.87
R-squared	0.548	0.560	0.560	0.650	0.651	0.650
Geo Controls	YES	YES	YES	YES	YES	YES
Socioecon Controls (1872)	YES	YES	YES	YES	YES	YES

Note: The Observed Diversity Index, the Synthetic Diversity Index and the Share of Foreigners (both in 1872 and 1920) are standardised to have mean 0 and SD 1. Railway Connection Timing indicates the time (in years) elapsed, for each mnicipality, between the connection to the railway network and 1920. Geographical controls: distance to the State's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Socioeconomic characteristics taken from the Census (1872): share of slaves, literacy ratio, share of workers in agriculture, industry and services/retail. All regressions use the 202 municipalities sample considering the 1920 administrative boundaries. Robust standard errors are reported in parenthesis. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

5. THE IMPACT OF BIRTHPLACE DIVERSITY ON LONG-RUN DEVELOPMENT

5.1. MAIN RESULTS

We start the description of results by discussing OLS estimates of the impact of birthplace diversity on long-term per capita income. All specifications reported in Table 4 include two sets of controls, geographic municipality characteristics and time-varying socioeconomic characteristics measured at baseline (in year 1872). In column 1 we find a positive point estimate of 0.097. In columns 2 and 3 we include the proportion of foreigners measured respectively in year 1872 and 1920. The share of immigrants, measured in 1920, is positively correlated with per capita income in year 2000, but the our coefficient of interest remains robust and slightly decreases to 0.085. Column 4 adds the time (in years) elapsed between the connection of the municipality to the railway network and year 1920. Municipalities that have been connected to the railway network earlier appear to have slightly higher income per capita in year 2000, however, the inclusion of such control makes decrease the coefficient on the diversity index marginally to 0.074. Columns 5 and 6, finally, bring together the share of foreigners as well as the railway connection timing variable. When doing so, the coefficient on the immigrant diversity varies between 0.074 and 0.071, and is both highly statistically significant and economically sizable. To interpret its magnitude: a one standard

TABLE 4: DIVERSITY AND LONG-RUN DEVELOPMENT - 2SLS

Dependent Variable	Ln per capita Income in 2000					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: OLS					
1920 Immigrant Diversity Index	0.097	0.096	0.085	0.074	0.074	0.071
	(0.014)***	(0.014)***	(0.015)***	(0.021)***	(0.021)***	(0.021)***
Share of Foreigners (1872)		0.006			0.002	
		(0.020)			(0.020)	
Share of Foreigners (1920)			0.049			0.040
			(0.025)*			(0.027)
Railway Connection Timing				0.002	0.002	0.002
				(0.001)*	(0.002)	(0.002)
	Panel B: Second stage					
1920 Immigrant Diversity Index	0.107	0.106	0.087	0.081	0.081	0.068
	(0.018)***	(0.019)***	(0.023)***	(0.039)**	(0.038)**	(0.039)*
Share of Foreigners (1872)		0.004			0.002	
		(0.018)			(0.018)	
Share of Foreigners (1920)			0.048			0.040
			(0.028)*			(0.026)
Railway Connection Timing				0.002	0.002	0.002
				(0.002)	(0.002)	(0.002)
	Panel C: Reduced form					
1920 Synthetic Diversity Index	0.078	0.076	0.058	0.037	0.038	0.031
	(0.014)***	(0.014)***	(0.016)***	(0.018)**	(0.018)**	(0.018)*
Share of Foreigners (1872)		0.017			0.005	
		(0.024)			(0.023)	
Share of Foreigners (1920)			0.065			0.040
			(0.027)**			(0.028)
Railway Connection Timing				0.005	0.004	0.004
				(0.001)***	(0.001)***	(0.001)***
Observations	202	202	202	202	202	202
R-squared Reduced Form	0.517	0.520	0.535	0.549	0.549	0.555
R-squared OLS	0.566	0.566	0.576	0.573	0.573	0.579
Partial-F first stage	156.07	149.96	106.47	40.84	41.25	40.87
Geographical Controls	YES	YES	YES	YES	YES	YES
1872 Socioecon Controls	YES	YES	YES	YES	YES	YES

Note: The dependent variable is the natural log of the per capita income in 2000 at the municipal level. The Observed Diversity Index, the Synthetic Diversity Index and the Share of Foreigners (both in 1872 and 1920) are standardised to have mean 0 and SD 1. Railway Connection Timing indicates the time (in years) elapsed, for each mnicipality, between the connection to the railway network and 1920. Geographical controls: distance to the State's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Socioeconomic characteristics taken from the Census (1872): share of slaves, literacy ratio, share of workers in agriculture, industry and services/retail. All regressions use the 202 municipalities sample considering the 1920 administrative boundaries. Robust standard errors are reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

deviation increase in immigrant diversity in year 1920 is associated with a 7.1-7.4 percent higher income per capita in year 2000.

Estimation by OLS, as discussed above, can lead to biased estimates. We therefore now discuss results from 2SLS estimates (Panel b of Table 4) based on our instrumental variable strategy. Similarly to the OLS case, the 2SLS estimates indicate a positive impact of immigrant diversity on long-term per capita income. Coefficients are fairly stable across specifications and remain stable also when controlling for the railway connection timing (columns 4-6) although standard errors get larger. The 2SLS estimates are slightly larger than OLS ones, except for those in column 6, possibly due also to idiosyncratic measurement error in the independent variable (observed diversity), generating attenuation bias of the coefficients estimated by OLS. In our most complete specifications (columns 5 and 6) the coefficients indicate that a one standard deviation increase in immigrant diversity is associated with a 6.8-8.1 percent higher income per capita in year 2000, an economically relevant effect. Finally, Panel C of Table 4 presents reduced-form estimates. Results show positive impacts of our instrument on current per capita income.

Finally, while we observe positive impacts of birthplace diversity among immigrants on long-term income per capita, a related question is whether such diversity had any effects on income inequality as well. As shown in Appendix Table A1, we find a positive relationship between the immigrant diversity in 1920 and the Gini coefficient at the municipality level in year 2000. Such relationship is statistically significant when estimating OLS but looses significance when implementing our 2SLS estimation, although the coefficients' size remain similar.

5.2. Structural Transformation and Occupational Diversity

Tables 4 shows that the diversity of origin of the immigrants who arrived between 1882 and 1920 induced higher per capita income in 2000 in São Paulo's municipalities. In this section we explore the mechanisms through which this shock propagated over time allowing immigrant diversity to affect long-term income. In particular, we consider two main mechanisms, the impact on structural transformation and occupational diversity, and human capital accumulation.

First, we observe that higher immigrant diversity induced a structural transformation of the economic activity in the following decades. The places with immigrants from more diverse origins in 1920 achieved a faster transfer of jobs from agriculture to manufacturing and to the service sector. A potential explanation for the long-term economic benefits of diversity is that during the first stages of development, the immigration provided a relatively more educated supply of labor, as documented by Rocha et al. (2017), and possibly with

TABLE 5: STRUCTURAL TRANSFORMATION AND OCCUPATIONAL DIVERSITY

	1920 OLS (1)	1940 OLS (2)	2000 OLS (3)	1920 2SLS (4)	1940 2SLS (5)	2000 2SLS (6)
	Panel A: Share Agricultural Occupation					
1920 Imm. Diversity Index	-0.023	-0.025	-0.036	-0.030	-0.033	-0.062
	(0.009)***	(0.012)**	(0.011)***	(0.019)	(0.028)	(0.024)***
	Panel B: Share Industrial Occupation					
1920 Imm. Diversity Index	0.012	0.004	0.000	0.009	0.027	0.034
	(0.004)***	(0.006)	(0.007)	(0.010)	(0.020)	(0.018)*
	Panel C: Share Services/Retail Occupation					
1920 Imm. Diversity Index	0.011	0.022	0.034	0.020	0.007	0.024
	(0.005)**	(0.008)***	(0.009)***	(0.011)*	(0.014)	(0.017)
	Panel D: Industrial Occupational Diversity					
1920 Imm. Diversity Index	0.006		0.010	0.058		0.052
	(0.009)		(0.008)	(0.019)***		(0.020)***
Observations	202	202	202	202	202	202

Note: Diversity Index is standardised to have mean 0 and SD 1. All specifications include Immigrant Share (measured in 1920), Railway Connection Timing, Geographical and socioeconomic controls (measured in 1872). Geographical controls: distance to the State's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Socioeconomic characteristics taken from the Census (1872): share of slaves, literacy ratio. All regressions use the 202 municipalities sample considering the 1920 administrative boundaries. Robust standard errors are reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, *$p<0.1$.

diversity of specific knowledge, skills and ideas necessary for more diverse industries to take off. This benefit may have been particularly captured in the context of the regional and urban economies, since cities act as engines of economic growth precisely by permitting people with different ideas to interact.

Table 5 shows that immigrant diversity prompted the structural transformation of the local economy by shifting occupation to more complex segments (manufacturing and services). Results reveal that municipalities that received a more diverse set of immigrants between 1880 and 1920 displayed a smaller proportion of labor employed in agriculture (Panel A) and larger proportions in manufacturing (Panel B), services and retail (Panel C). While IV estimates for the share of workers in the agricultural sector in 2000 are negative and statistically significant at the 1% level, coefficients for the manufacturing and services sectors are positive, although imprecisely estimated. Yet, overall the results suggest that these sectors absorbed workers from the agricultural sector overtime. Together with OLS estimates, results also suggest that this move may have begun as soon as the immigration flow was completed, as the share of workers in the service sector was already greater circa 1920.

Finally, in Panel D of Table 4 we investigate the effect of immigrant diversity on the occupational diversity within the manufacturing sector in 1920 and 2000. Industries used to calculate the this occupational diversity index in 1920 are the extractive, construction, wearing apparel, wood and furniture, metallurgy, chemicals, and foods. For the index in 2000, data available are broken down as in 1920 but with the addition of machinery, automotive and plastics. The 2SLS estimates (columns 4 and 6) indicate that municipalities that received a more diverse set of immigrants tend to have an higher manufacturing occupational diversity, both in 1920 and in 2000. We conjecture that a more diverse flow of immigrants increased the potential for complementarity among skills and professions brought from the places of origin, as suggested by Murard (2018) and Menyhert (2018) in other contexts.

5.3. Schooling Investment and Human Capital Accumulation

We now examine whether a significant difference exists between municipalities that were more *versus* less diverse on government investment in schooling and educational outcomes. Panel A of Table 6 reports OLS estimates, while Panel B reports 2SLS ones. Column 1 and 2 show that municipalities with higher diversity had more schools per school-age child in 1920 and in 1940, as well as in per capita public expenditure in education measured in 1940. Finally, columns 4 and 5 show higher levels of enrollments rates among children 7-14 years old in 1940 (column 4) and more years of schooling in year 2000 (column 5). The 2SLS coefficient in column 5 indicates that municipalities that experienced a one standard

TABLE 6: HUMAN CAPITAL AND SCHOOLING

	Schools per capita 1920	Schools per capita 1940	Education expenditure per capita 1940	Enrollment rate 7-14 y.o 1940	Years of education 2000
	(1)	(2)	(3)	(4)	(5)
	Panel A: OLS				
1920 Imm. Diversity Index	0.027	0.287	0.196	0.027	0.138
	(0.040)	(0.162)*	(0.068)***	(0.009)***	(0.049)***
	Panel B: 2SLS				
1920 Imm. Diversity Index	0.249	0.761	0.373	0.040	0.259
	(0.077)***	(0.392)*	(0.124)***	(0.024)*	(0.111)**
Observations	202	202	202	202	202

Note: Diversity Index is standardised to have mean 0 and SD 1. All specifications include Immigrant Share (measured in 1920), Railway Connection Timing, Geographical and socioeconomic controls (measured in 1872). Geographical controls: distance to the State's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Socioeconomic characteristics taken from the Census (1872): share of slaves, literacy ratio, share of workers in agriculture, industry and services/retail. All regressions use the 202 municipalities sample considering the 1920 administrative boundaries. Robust standard errors are reported in parenthesis. Significance levels: *** p<0.01, ** p<0.05, *p<0.1.

deviation increase in diversity have on average 0.26 more years of schooling in year 2000.

The results in Table 6 not only suggest that long-term human capital improved in more diverse municipalities, which is directly instrumental for economic development (Rocha et al., 2017), but also that diversity may not have triggered mistrust or lack of social cohesion, as investments in public goods were also greater. In that sense, cohesion, new forms of solidarity and more encompassing identities may have been forged over time, as conjectured by Putnam (1997).

6. CONCLUSIONS

In the four decades that followed 1880, Brazil, and particularly the state of São Paulo, received a large wave of foreign immigration from various origin countries. This article explored this episode and measured the impact of the diversity brought about by immigration on long-term development at the municipal level. To identify causal effects, we used unique historical individual data on immigrants arriving in São Paulo between 1880 and 1920, and developed an instrumental variable strategy that combined time variation in the composition of immigrants arriving from overseas with the timing of the railway network expansion in the state. We fuond that municipalities that received a more diverse mix of origin countries

ended up having greater long-term economic outcomes.

We estimate that a one standard deviation increase in immigrant diversity in 1920 is associated with a 7-8% higher income per capita in 2000. This effect is economically relevant and robust to robustness checks. Among the mechanisms underpinning these long-term effects, we examined the role of structural transformation and human capital formation, as well as investments in public goods, such as educational inputs. These results are relevant as long as history has shown that immigration and diversity have continuously reshaped the composition of populations around the world.

REFERENCES

Ager, P. and Bruckner, M. (2013). Cultural diversity and economic growth: Evidence from the united states during the age of mass migration. *European Economic Review*, 64:76–97.

Alesina, A., Baqir, R., and W., E. (1999). Public Goods and Ethnic Divisions. *Quarterly Journal of Economics*, 114(4):1243–1284.

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, 21(2):101–138.

Alesina, A. and La Ferrara, E. (2005). Ethnic diversity and economic performance. *Journal of economic literature*, 43(3):762–800.

Bahar, D., Rapoport, H., and Turati, R. (2022). Birthplace diversity and economic complexity: Cross-country evidence. *Research Policy*, 51(8):103991.

Bandiera, O., Rasul, I., and Viarengo, M. (2013). The making of modern america: migratory flows in the age of mass migration. *Journal of Development Economics*, 102:23–47.

Bassanezi, M. S. C. B., Scott, A. S. V., Bacellar, C. d. A. P., and Truzzi, O. M. S. (2008). *Atlas da Imigração Internacional em São Paulo, 1850–1950*. Editora UNESP, São Paulo.

Bazzi, S., Gaduh, A., Rothenberg, A. D., and Wong, M. (2019). Unity in diversity? how intergroup contact can foster nation building. *American Economic Review*, 109(11):3978–4025.

Beach, B. and Jones, D. B. (2017). Gridlock: Ethnic diversity in government and public good provision. *American Economic Journal: Economic Policy*, 9(1):112–136.

Bove, V. and Elia, L. (2017). Migration, diversity, and economic growth. *World Development*, 89:227–239.

Colistete, R. P. and Lamounier, M. L. (2011). The end of plantation? coffee and land inequality in early twentieth century. *Annals of Statistics*, 7:1–30.

Dean, W. (1977). *Rio Claro. Um sistema brasileiro de grande lavoura, 1820-1920*. Paz e Terra, Rio de Janeiro, 1 edition.

Desmet, K., Gomes, J. F., and Ortuño-Ortín, I. (2020). The geography of linguistic diversity and the provision of public goods. *Journal of Development Economics*, 143:102384.

Docquier, F., Turati, R., Valette, J., and Vasilakis, C. (2020). Birthplace diversity and economic growth: evidence from the us states in the post-world war ii period. *Journal of Economic Geography*, 20(2):321–354.

Droller, F. (2017). Migration, population composition and long-run economic development: Evidence from settlements in the pampas. *The Economic Journal*, pages 2321–2352.

Escamilla-Guerrero, D., Papadia, A., and Zimran, A. (2024). The effects of immigration in a developing country: Brazil in the age of mass migration. Technical report, National Bureau of Economic Research.

Fulford, S. L., Petkov, I., and Schiantarelli, F. (2020). Does it matter where you came from? ancestry composition and economic performance of us counties, 1850–2010. *Journal of Economic Growth*, 25(3):341–380.

Giuliano, P. and Tabellini, M. (2020). The seeds of ideology: Historical immigration and political preferences in the united states. Technical report, National Bureau of Economic Research.

Goncalves, P. C. (2008). Mercadores de braços: riqueza e acumulação na organização da emigração européia para o novo mundo.

Goncalves, P. C. (2009). A cidade de são paulo: um entreposto de braços para a lavoura cafeeira. *Revista Cordis de História Social da Cidade*, 2.

Holloway, T. H. (1984). *Imigrantes para o café: café e sociedade em São Paulo, 1886-1934*. Paz e Terra, Rio de Janeiro.

Mauch, C. and Vasconcellos, N. (1994). *Os Alemães no Sul do Brasil: Cultura, Etnicidade e História*. Editora Ulbra, Canoas.

Menyhert, B. (2018). Economic growth spurred by diversity: Evidence from the austro-hungarian monarchy. pages 1–81.

Murard, E. (2023). Long-term effects of the 1923 mass refugee inflow on social cohesion in greece. *World Development*, 170:106311.

Murard, E.; Sakalli, S. O. (2018). Mass refugee inflow and long-run prosperity: Lessons from the greek population ressetlement. *IZA Discussion Paper n.11613*.

Ottaviano, C. and Peri, G. (2005a). Cities and cultures. *Journal of Urban Economic*, 58(1):304–337.

Ottaviano, G. I. and Peri, G. (2005b). Cities and cultures. *Journal of Urban Economics*, 58(2):304–337.

Ottaviano, G. I. and Peri, G. (2006). The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography*, 6(1):9–44.

Putnam, R. D. (1997). E Pluribus Unum: Diversity and Community in the Twenty-first Century. *Scandinavian Political Studies*, (30):137–174.

Putterman, L. and Weil, D. (2010). Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality. *Quarterly Journal of Economics*, 125(4):1627–1682.

Rocha, R., Ferraz, C., and Soares, R. S. (2017). Human capital persistence and and development. *American Economic Journal: Applied Economics*, 9(4):105–136.

Sequeira, S., Nunn, N., and Qian, N. (2020). Immigrants and the making of america. *The Review of Economic Studies*, 87(1):382–419.

Tabellini, M. (2020). Gifts of the immigrants, woes of the natives: Lessons from the age of mass migration. *The Review of Economic Studies*, 87(1):454–486.

Vangelista, C. (1991). *Os Braços da Lavoura: imigrantes e caipiras na formação do mercado de trabalho paulista (1850-1930)*. Hucitec, São Paulo.

Vasconcellos, H. D. (1994). Oscilações do movimento imigratório no brasil. *Revista de Imigração e Colonização*, 1(2):23–47.

7. Online Appendix

For online publication only

TABLE A1: IMMIGRANT DIVERSITY AND INEQUALITY

Dependent Variable:	Gini coefficient in 2000					
	OLS	OLS	OLS	2SLS	2SLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
1920 Immigrant Diversity Index	0.007	0.007	0.009	0.004	0.004	0.011
	(0.004)*	(0.004)*	(0.004)**	(0.010)	(0.010)	(0.010)
Observations	202	202	202	202	202	202
Geo Controls	YES	YES	YES	YES	YES	YES
Socioecon Controls (1872)	YES	YES	YES	YES	YES	YES
Railway Connection Timing	YES	YES	YES	YES	YES	YES
Share of Foreigners (1872)		YES			YES	
Share of Foreigners (1920)			YES			YES

Note: The dependent variable is Gini coefficient in 2000 at the municipal level. The Observed Diversity Index, the Synthetic Diversity Index and the Share of Foreigners (both in 1872 and 1920) are standardised to have mean 0 and SD 1. Railway Connection Timing indicates the time (in years) elapsed, for each mnicipality, between the connection to the railway network and 1920. Geographical controls: distance to the State's capital, latitude, longitude and elevation and dummies for each existing kind of soil (latosoil, argisoil, cambisoil and spodosoil). Socioeconomic characteristics taken from the Census (1872): share of slaves, literacy ratio, share of workers in agriculture, industry and services/retail. All regressions use the 202 municipalities sample considering the 1920 administrative boundaries. Robust standard errors are reported in parenthesis. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

FIGURE A1: IMMIGRANT'S HOSPEDARIA



Note: Immigrants in the inner courtyard of the Hospedaria de São Paulo, in 1915. Collection of the Memorial do Imigrante, São Paulo.

FIGURE A2: SIZE AND COMPOSITION OF ANNUAL IMMIGRANT INFLOWS

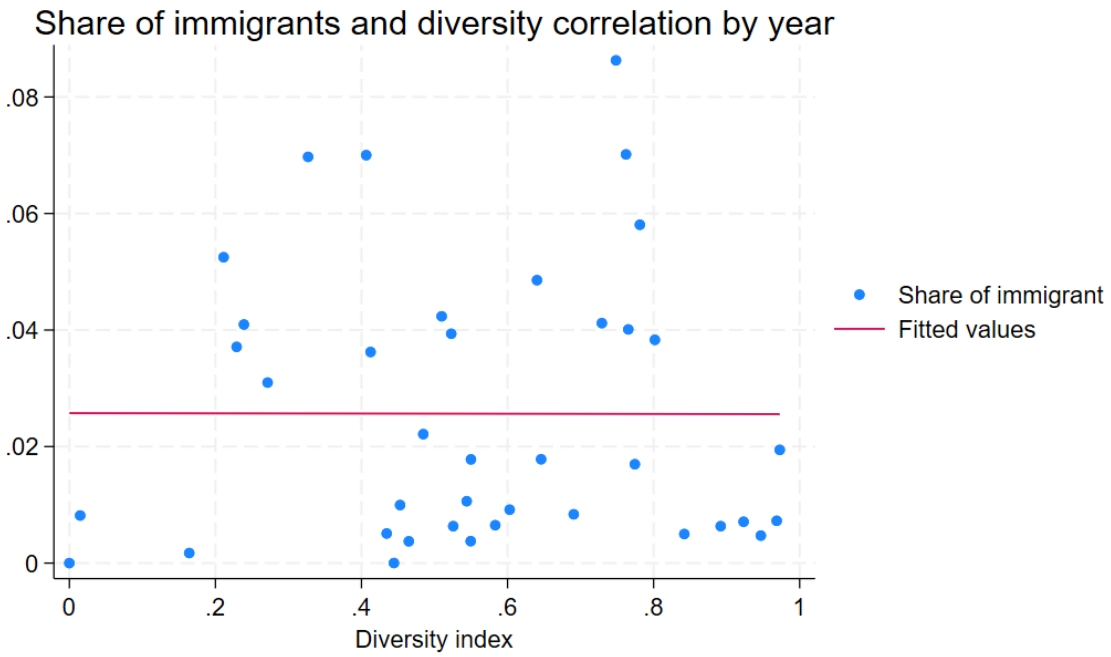
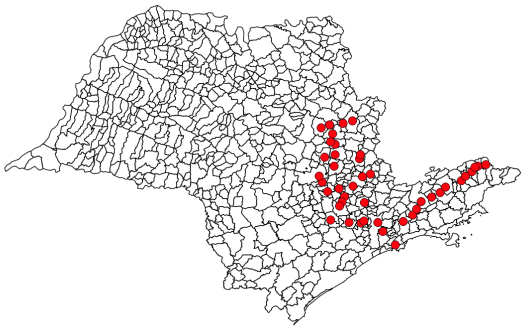


FIGURE A3: RAILWAY NETWORK EXPANSION

(a) 1882



(c) 1900



(b) 1890



(d) 1910



(e) 1920



Note: The figure shows the gradual expansion of the railway network during the period between 1882 and 1920. Each red dot represents a rail station. Data source: Data available at www.estacoesferroviarias.com.br