

# 1 Settlement Colonies Across Plantation Fields: Evidence on the Relationship Between Human Capital and Long Term Development

European colonialism changed the patterns of human occupation and the composition of populations across different regions of the globe (Putterman & Weil (2010)). This process not only gave rise to several politically independent nations, but also fostered much of the contemporary inequality across countries (Engerman & Sokoloff (1997), Acemoglu et al. (2001), Angeles (2007)).

The search for the explanations behind the economic divergence among settled regions has raised one of the most contentious debates in development economics. Scholars agree that sustained growth followed the regions that formerly hosted the largest inflows of European immigrants (Ertan & Putterman (2007)). Yet, different theories have suggested distinct causal links between immigration intensity and long run growth (see reviews in Acemoglu et al. (2005) and Easterly & Levine (2003)). One tentative hypothesis concerns the role of human capital, as suggested in Glaeser et al. (2004). It seems plausible that what European immigrants brought to the New World was themselves, and therefore their human capital, acknowledged as an important boost for economic growth. However, disentangling empirically this or other hypotheses from competing explanations has proved to be difficult. As Glaeser et al. (2004) recognize, colonizers brought with them not only their human capital, nor only their institutions and culture, but also germs, guns and steel (Diamond (1997)). Therefore, the European colonialism episode would hardly allow scholars to identify the fundamental causes of economic growth.

This paper examines the role of the human capital hypothesis by using an unique empirical setting. By the turn of the last century, public authorities created in the Brazilian state of São Paulo almost 30 settlement colonies. The settlements were typically rural villages formed by small plots of land, and populated by relatively higher-skilled European settlers of various nationalities. In this paper we examine whether and to what extent the settlement colonies of São Paulo fostered the long run development of the regions through improvements in early levels of human capital. We exploit two unique features of this episode to disentangle the human capital channel from other confounding

mechanisms. First, the episode was motivated by different policy goals, which pushed a similar colonization framework into regions with distinct pre-existing conditions. As a result, the distribution of settlement colonies across the state ended up relatively homogeneous. This feature, along with exogenously given macro institutions and the low variability of geographical conditions, give us a remarkably controlled empirical setting. Second, the settlements had no significant direct impacts on settled regions, but for an improvement in the composition of human capital. Thus, the settlements played the unique role of attracting relatively higher-skilled immigrants.

The investigation of the human capital hypothesis within our empirical setting demands in information a range of geographical and socioeconomic characteristics at the local level, for the many regions of São Paulo throughout time. We use data from different sources to build a longitudinal dataset at the municipality level. This dataset contains geographical, transportation and pre-existing conditions combined with socioeconomic variables from various historical and contemporary National Censuses.

Our core analysis follows a two-step procedure. We first use cross-sectional and differences-in-differences regressions to examine the allocation of settlements across space, and their short-term impacts on settled regions. This analysis confirms the two important features of this colonialism episode. First, we show that the spatial distribution of settlements in terms of pre-existing conditions was as good as random. Second, we show that settlements had no relevant direct effects on settled regions, but for an improvement in the composition of human capital. Our estimates suggest that settlements raised by 3 to 5 percentage points the initial literacy rates in settled regions.

In the second step we use a 2SLS strategy to examine whether the variations in early levels of human capital driven by the settlements had long run impacts on local development. In this approach the settlements are used as an instrument for early levels of human capital. This strategy is consistent with the unique features of our empirical setting, while the exclusion restrictions are fully respected. We show that settlements had no independent long run impacts on settled regions, other than those driven by improvements in early levels of human capital. Second-stage regressions show that a 5 percentage points increase in early levels of literacy leads to a 10-12.5% increase in per capita income today.

We also discuss a likely mechanism behind the relationship between human capital improvements in settled regions and long term development. We show that settlements had persistent effects on schooling indicators over time. In addition, we provide suggestive evidence in support for the view that

this effect was mostly driven by the demand for education.

This paper speaks to the stream of literature in development economics that has recently renewed the discussion on growth theories in the light of European colonialism (Engerman & Sokoloff (1997), Acemoglu et al. (2001), Glaeser et al. (2004)). It also relates to studies that examine the impacts of certain historic events on long run development within countries or regions (Banerjee & Iyer (2005), Naritomi et al. (2007), Nunn (2008), Bruhn & Gallego (2009), Acemoglu et al. (2011)). In fact, many papers have made progress in showing that history matters. What remains less understood are the exact channels of causality through which history matters (Nunn (2009)). The contribution of this paper is twofold. First, our results provide supportive evidence for interpretations that connect, specifically through the human capital channel, European colonialism with long term development. Second, we discuss the mechanisms at work. In doing so, we also contribute to the research into the causes of economic growth. Many authors have found a positive relationship between human capital and growth by exploiting cross-sectional variation within countries. Yet, empirical research in this field has mostly focused on contemporary data from developed countries (Glaeser et al. (1995), Shapiro (2006), Iranzo & Peri (2009)). This paper provides novel micro evidence on the causal links between human capital and growth by using a dataset that covers more than 100 years of history of the state of São Paulo, a leading case of economic growth in the developing world.<sup>1</sup>

This paper proceeds as follows. The next section presents the historical background and describes the episode of settlement colonialism in São Paulo. In Section 1.2 we discuss conceptually our hypothesis. Section 1.3 presents our dataset and descriptive statistics. Section 1.4 shows the main econometric results. In Section 1.5 we discuss causality channels, while Section 1.6 concludes.

## **1.1 Historical Background and Settlements' Characterization**

### **1.1.1 Historical background**

São Paulo has long been the richest and most industrialized state of Brazil. The economic take-off has its historical origins in the expansion of the coffee production during the 19th and early 20th centuries (Furtado (1989)). This expansion was especially fueled by the construction of the São Paulo

<sup>1</sup>Recent papers have made great progress in describing the relationships between early socioeconomic conditions and long term regional development in Brazil, either by using the case of São Paulo (Summerhill (2010), de Carvalho Filho & Colistete (2010)) or other Brazilian regions (de Carvalho Filho & Monasterio (2011)). In this paper we focus on strategies to better identify causal effects and the likely mechanisms at work.

Railway line in the late 1860s, which connected the seaport of Santos with the vast, fertile, and sparsely occupied mid-northwestern soils. The rapid march of the agricultural frontier consolidated the production of coffee as a major promoter of infrastructure construction, urban development and accumulation of wealth towards the interior of the state (Furtado (1989), Summerhill (2003)).

The ongoing economic expansion absorbed large quantities of workers and land. While labor was still primarily supplied by internal reallocation of slaves, slavery would soon be definitely abolished - international slave trade was already forbidden since 1850 (Law Eusébio de Queiroz). Coffee trees rapidly took over not only the agricultural frontier, but also soils that were formerly used to supply basic food products to the growing City of São Paulo (the state's capital city) and other urban centers. Important food suppliers by that time, such as the municipalities of Bragança, Atibaia and Nazareth, became major coffee producers (Martins (1973)). Indeed, coffee definitely paid off. Between 1855 and 1875, increases of up to 140% in food prices were offset by even higher increases in the coffee price (Costa (1989)). The rising pressures for the end of slavery, and the shortage of workers and food brought international immigration to the center of the political agenda. By that time, there was a consolidated view among the political elite that Europeans were superior workers, more cultured and productive than natives. In particular, European workers were considered by the elite superior to afro-descendants (Petroni (1982)).

The first experiences of international immigration in Brazil during the Empire era date back to the early 19th century. By that time, the Imperial government promoted the peopling of strategic and sparsely occupied regions with the settlement of European families clustered in isolated areas. The first official settlements were the colonies of Nova Friburgo (1818), in Rio de Janeiro; and São Leopoldo (1824), in the state of Rio Grande do Sul. This wave of Imperial colonization included also two areas in the surroundings of the City of São Paulo (Santo Amaro and Parelheiros, in the late 1820s), and other two on the coastal Vale do Ribeira (Iguape and Cananéia, in the late 1850s). Facing economic isolation, the initiatives failed to create sustained regional progress, and the farthest colonies of Parelheiros, Iguape and Cananéia were largely abandoned (Furtado (1989), Petroni (1982), Paiva (1993), Siriani (2005)).

In the context of rising food prices and shortage of workers in the second half of the 19th century, the model of immigration via the settlement of official colonies was reconsidered and increasingly supported in legislative debates (Martins (1973)). As a consequence, the Imperial government launched in the mid-1870s a second wave of colonization in São Paulo. Four new settle-

ments were created around the capital city - Santana and Glória (nowadays neighborhoods of the capital city), São Bernardo and São Caetano (nowadays independent municipalities in the metropolitan region of the capital city).

The administrative procedure for settling these colonies can be broadly described as follows. After legislative approval, public employees searched for locations mainly among unclaimed public lands (*terras devolutas*), but also among private lands and foreclosure opportunities. As we discuss later, given the alternative available areas, the exact location was chosen based on two main technical conditions: reasonable soil fertility and proximity to transportation. Small plots of land were then designated and separated into three different types, depending on size and localization within settlements. Rural plots were the largest, while the suburban and urban ones were smaller and more centrally located. Basic infrastructure was built around an administrative office, at the center of the colony. These plots of land were offered by Brazilian representatives in European cities, or negotiated at the Immigration Station House (*Hospedaria dos Imigrantes*). The Station House was a public building located in the surroundings of the capital city, and the primary port of entry for the vast majority of immigrants arriving at São Paulo. The land was not freely-given to the immigrant, but had to be paid in installments after the first harvests. During this initial period, immigrants generally received some scant public support - food, seeds, agricultural instruments, and payments for efforts on initial infrastructure improvements. There were public employees in charge of coordinating and regulating the developments within the official settlements. After all the installments were payed for, the settler could claim the property right of his plot. In general, the settlement was emancipated from public administration when all its plots were already paid for. There are reasons to believe that the entire process was kept under significant political accountability. Administrative records and detailed documentation on why some areas were selected instead of others were often presented in the Annual Provincial Presidential Reports (PPRs), in the Messages to the Legislative Assembly, and in other administrative reports.<sup>2</sup>

Immigration and colonization policies faced an inflection point in the mid-1880s, when slavery was definitely abolished (Lei Aurea, 1888). Public subsidies supported the transition towards a system based on free labor supplied by European immigrants. State and central governments paid for the immigrants'

<sup>2</sup>Only to cite a few examples, the 1877 Annual Report of the Secretariat of Agriculture of São Paulo (1877 ASRA) justifies in detail the settlement of the colonies of São Caetano, São Bernardo, Glória and Santana; the 1885 PPR justifies the location of the official colonies of Canas and Cascalho; and the 1907 PPR details the procedures for the settlement of the colonies of Nova Europa and Gavião Peixoto.

travel expenses (international and internal displacements) and coordinated the job matching between workers and farmers. Subsidized European immigration to São Paulo soon gained a massive scale (as illustrated in Figure 1.1, Panel A), absorbing a major share of the total inflow of European families arriving at Brazil (Figure 1.1, Panel B).<sup>3</sup> In this post-slavery scenario, besides the role as food suppliers, the official settlements gained an additional motivation. They became also an important instrument to attract European immigrants in the search for land ownership and economic independency (Martins (1973), Petrone (1982)). The settlements became a major attractive in the official Brazilian propaganda disseminated in European cities (Petrone (1982)). Yet, only a small fraction of immigrants would end up settled in the official colonies. The vast majority of immigrants arriving at the Immigration Station House were rapidly locked into a long term work contract offered by coffee farmers. In this new context, following closely the same general logic of the 1870s, the third wave of colonization was launched, this time under both Imperial and state coordination. From the mid-1880s until the late 1890s, 13 new official settlements were created in different regions of São Paulo: four in the economically backward Vale do Paraíba, two in what is the actual metropolitan region of São Paulo, four in the expanding coffee region of the mid-northwest, two in the mid-west, where the main agricultural products were cotton and sugarcane, and one on the coastal Vale do Ribeira.

The geographical distribution of settlements across the state was consistent with an ambivalent policy motivation: food supply and attraction of immigrants dreaming of land ownership (Petrone (1982)). On the one hand, the role as *food suppliers* might have driven the creation of a few official colonies in areas with growing food demand. However, in many other cases the *promise land* used to attract immigrants was located within colonies created in marginal or backward areas, where the availability of unclaimed public lands was higher.

Much of the prominence of São Paulo as a hosting region was consolidated after the fall of the Empire era, with the rise of the Old Republic (1889). With the political reforms that came with the new republican constitution of 1891, the immigration policy and the management of unclaimed public lands became decentralized at the state level (Iotti (2001)). As the richest and fastest growing economy in Brazil, São Paulo managed to draw the largest share of immigrant arrivals. Thus, in the early 1900s the state targeted the creation of 9 new colonies clustered at 3 different areas in the mid-northwestern coffee

<sup>3</sup>The major shares of immigrants were Italians, Portuguese and Spanish. For stylized facts on the mass migration to São Paulo, see Bassanezi et al. (2008).

region (still isolated areas around the municipalities of Mogi Mirim, Ibitinga, and between Campinas and Rio Claro).<sup>4</sup> Meanwhile, following a renewed motivation to develop lagging regions, the federal government opened two new official settlements in still sparsely populated regions of the state (Iotti (2001)). The first one was close to the southwestern border with the state of Paraná, the other was on the border with the state of Rio de Janeiro. Finally, one more federal colony would be created in the late 1920s, close to the western border with Paraná.<sup>5</sup>

In total, almost 30 official settlements were created across the state between the late Imperial era and the late Old Republic. The settlements were distributed across more than 20 different municipalities, representing 10% of the total number of municipalities according to the state administrative division of 1920. Different motivations behind the colonization policy over this three-decade period led to a distribution of official colonies across regions with distinct economic and demographic conditions. The homogeneous distribution of settlements is confirmed empirically later on in this paper.

### 1.1.2 The settlements: general characterization

Figure 1.2 is a historical map of São Paulo indicating the settlements' locations. The black dots provide the location of the settlements created up to 1908, whereas the white dots (added by us, since the map was published in 1908) indicate those settled between 1908 and the early 1910s. As we can see, while some of the settlements were created far away from railroad transportation, a great number of them were actually located in regions crossed by railways (depicted by the dark lines crossing the state).

Table 1.2 presents a comprehensive characterization of the settlements, including sources of information and general statistics. The areas varied considerably across settlements: those created between the late 1870s and the early 1880s had initial areas mostly around 1,000 to 3,000 hectares; those created in the 1910s had otherwise average sizes between 3,000 and 5,000 hectares, while only a few on both periods had larger areas. The settlements usually gained more surface as the number of settlers increased.

<sup>4</sup>This motivation is explicit in the 1904 Annual Report of the Secretariat of Agriculture (1904 ARSA, p.132-134). The state also created the colony of Conde do Pinhal in the coastal municipality of Ubatuba. However, the first settlers immediately abandoned the place and the project did not take-off (see official justifications for the failure in the 1907 ARSA (pp.161-162) and the 1908 ARSA (p.178)). For this reason, this colony is not included in this analysis.

<sup>5</sup>There were a few more federal settlement attempts on the Vale do Ribeira after the 1929 crisis, but this time the projects resembled more a regulated land occupation than a proper supervised colonization initiative (Paiva (1993)). For this reason, these attempts are not included in the analysis.

Using information available in the 1910s Annual Statistical Reports of São Paulo (henceforth, ASRs), we observe that after a few years of operation the average population size reached around 1,000-3,000 individuals per colony, with some variation across locations.

Following, we present three stylized facts about life conditions within the settlements. First, we show that and discuss why human capital levels within settlements were higher than outside. Second, we show that, despite the higher levels of human capital, life conditions within settlements were tough. This was possibly due to low agricultural productivity. Finally, we study the magnitude of the inflows and outflows of new settlers.

*Human capital.* To provide evidence on the human capital levels within settlements, we gather statistics on settlers' literacy for different settlements and points in time.<sup>6</sup> We then calculate for each settlement the share of literate individuals, and compare the results with the average share of literate individuals across municipalities (using the 1890 and the 1920 Census). Figure 1.3 shows that education levels within colonies were systematically higher than the average across the municipalities. For instance, the median share of literate individuals across those settlements surveyed in the 1910s was 44%, against the state average (and also median) share of 23% in 1920.

While we are unaware of any study which identified and explained this selection, some hypotheses naturally arise. The settlements were populated in basically two ways. First, Brazilian representatives in European cities could select the settlers and negotiate the available plots of land in the official settlements. In this case, there is evidence that representatives searched for more productive individuals, and possibly more educated ones. This is confirmed, for instance, when the 1907 PPR (p.353) mentions that “*the inflows of immigrants that have entered into the state, brought by the Immigration Commissariat in Antwerp and originating from northern Europe, indicate that we managed to find the **convenient element** for occupying the official colonies*”. Other examples come from the selection of Russian families for the settlement of Nova Odessa (1905 Annual Report of the Secretariat of Agriculture, ARSA), and Belgium families for the settlement of Rodrigo Silva (1888 Provincial Presidential Report, henceforth PPR). Second, available plots of land within colonies were also offered at the Immigration Station House. Thus, upon arrival in São Paulo, immigrants supposedly had a choice

<sup>6</sup>The 1888 ASR provides detailed information on three settlements, created in the 1870s (S. Bernardo, S. Caetano and Santanna); the 1898 Annual Report of the Secretariat of Agriculture (ARSA) presented data for 6 settlements, most of them created in the 1880s and 1890s (but also including S. Bernardo); and in the 1910s ASRs we find information for 12 settlements, created mostly between 1905-1911 (for these settlements we choose the most recent available data).



between official settlements and job opportunities in the coffee farms. In this case, entrepreneurship and some minimal level of capital requirements for household survival during the first moments may have driven higher skilled immigrants towards the settlements. As the 1906 PPR (p.43) mentions, “*the immigratory current will increase due to the official propaganda the government is disseminating abroad, and whose efficacy has been already observed in the spontaneous arrivals of immigrants searching for the official colonies*”. Spontaneous immigrants were more likely to be educated than the subsidized ones. Thus, official selection abroad and some minimal capital requirements may have channeled more educated and able settlers into the official colonies.

*Life conditions.* According to official reports there was an explicit recommendation to place the settlements in areas not only close to transportation, but also where the soil quality could provide reasonable productivity standards.<sup>7</sup> However, soil fertility within settlements is a controversial issue. Some official reports document that engineers responsible for the settlements’ location weighted in favor of soil fertility when deciding where to settle the colonies (examples in Martins (1973)). We consider this information reliable once the selection procedure was maintained under political accountability. On the other hand, there are reasons to put the overall soil quality within colonies into perspective. Selection was mostly conditional on alternatives among unclaimed public lands, or even among unused private or foreclosed lands purchased by the government. Thus, these were areas unlikely to be highly productive or of any use for the core agricultural activity (coffee). In fact, sometimes official reports admitted low quality of soils (examples in Martins (1973) for São Caetano, and in the 1899 ARSA (p.90) for São Bernardo).

The 1910s ASRs of São Paulo present detailed documentation on the economic conditions within 11 settlements, most of them created between 1905 and 1911. We use this information to examine the agricultural productivity within colonies in comparison to the average state level. Productivity indicators are measured as the ratio between (i) the overall value of the agricultural, extractive and animal productions, and (ii) total population. To smooth production and population fluctuations within colonies we calculate this ratio for years between 1915 and 1920, and take the average over time. We then compare these average ratios with the overall per capita value of the agricultural production of the state, reported in the 1920 Census. Figure 1.4 presents the

<sup>7</sup>Martins (1973) details this point when illustrating the procedures behind the location of the colony of São Caetano. The recommendation on transportation and soil fertility is explicit in many administrative reports and presidential speeches that justifies the location of official settlements. Examples can be found in the 1877 and 1885 PPRs, and in the 1885 and 1892 ARSAs.

comparison. The annual per capita production within colonies (median around 225 Réis) was systematically lower than the state average (average of 531 Réis). This comparison suggests that soil quality may have provided only minimum levels of economic subsistence within settlements.

*Population fluctuations.* We now provide a general characterization of the population fluctuations within settlements. The initial moments were marked by a continuous arrival of settlers, most of them coming directly from Europe after a stopover at the Immigration Station House. During the initial years after the foundation of a settlement, the lack of basic infra-structure, the difficulties faced in the first harvests and the isolation repelled many families, that left the settlements in the search for other opportunities, most probably in the surrounding areas and closest urban centers.<sup>8</sup> Population grew with new inflows of families, and soon the environment was improved. Meanwhile, weather and market prices fluctuations, jointly with low agricultural productivity and better outside options, continuously repelled resident families, opening space for the arrival of new ones. To bring empirical evidence on the population fluctuations within colonies, we make use of information from the 1910s ASRs for 11 colonies on individual arrivals and departures. For each settlement we calculate a ratio between (i) the sum of arrivals or departures in the years between 1912 and 1918;<sup>9</sup> (ii) and the initial stock of population in 1912. Table 1.3 shows large inflows and outflows of settlers in many settlements, in particular in those recently created. In a few years a substantial share of the initial population stock (reference in 1912) had already left the colonies (median of 16%, and average of 34% of the initial stock). The net flow of settlers otherwise remained positive.

The stylized facts presented so far support the view that official colonies played an important role as ports of entry of relatively higher skilled immigrants. While selection implied higher human capital levels, attraction and repulsion forces determined the inflow of families towards the settlements, and also their subsequent movement outwards.

## 1.2 Conceptual Discussion

The settlements of São Paulo represented a complex policy intervention. In this section we discuss the likely impacts they may have had on settled regions.

<sup>8</sup>Many official reports document the difficulties to maintain the initial settlers attached to the colonies. Examples can be found in the 1898 ARSA (for the initial years in the colony of Campos Salles, p.88) and in the 1905 ARSA (for the colonies of Nova Odessa, p.143; and Jorge Tibiriçá, p.138).

<sup>9</sup>Not including 1917 once the 1917 ASR was not available.

First, the settlements may be associated with composition and scale effects. Once they brought population inflows, agglomeration effects may have trivially occurred. This effect needs to be taken into account once “agglomeration can be thought as the territorial counterpart of economic growth”, as suggested by Fujita & Thisse (2002), and reviewed in Baldwin & Martin (2004). In terms of composition effects, we already showed that selection made higher levels of human capital a distinctive feature of the settlements’ population. The literature on growth and development economics has shown that initial levels of human capital in a given country predict its subsequent growth (Barro (1991)). Within national boundaries, human capital embodied in skilled workers is supposed to be more mobile across regions. Yet, similar empirical patterns are found across regions or cities within countries (Glaeser et al. (1995)). In theory, the positive link between human capital and national or city growth can be driven by direct productivity improvements, or by positive externalities on productivity (Lucas (1988)).<sup>10</sup>

Human capital is mostly referred to here as a compound of formal education, know-how and personal skills. However, all these embodied assets are hardly disentangled from culture, religion, values and other manifestations of deep-rooted social norms. In this case, settlers may have also brought to the official settlements not only productive skills, but also other social assets that have been considered important for economic development (for instance, Weber (1930), Landes (1999)).

Second, the settlements may have changed the distribution of land within settled regions. Large unused farms were partitioned into several small plots, which *ceteris paribus* could have led to a decrease in the regional levels of land inequality. Third, the settlements may be also associated with better property rights. The lands used to establish colonies had been formerly unclaimed public lands, foreclosed properties or unused private lands. In this case, the settlements would have brought better property rights to settled regions. Both land inequality and property rights are factors that need to be addressed in our analysis once theory and empirical evidence have suggested they have an important role in the process of sustained growth (for instance, see Galor et al. (2009), North (2009)).

Fourth, the construction of some scant infrastructure, such as roads and schools, was provided by the public administration. Hence, the settlements may

<sup>10</sup>What Lucas (1988) called respectively internal and external human capital effects. Increases in the aggregate stock of human capital can benefit the economy in ways that go beyond the private returns of education. Moretti (2004) argues that knowledge spillovers can increase aggregate productivity over and above the direct effect of human capital on individual productivity, whereas increases in education can reduce criminal behavior and foster voters’ political participation.

also have fostered the provision of local public goods. Finally, the settlements' impacts on local labor markets depended on the elasticity of substitution between lower and higher skilled workers. If we suppose that by the turn of last century both types of workers were likely substitutes in rural areas, the expected effect of an inflow of workers would be a decrease in local nominal wages. Yet, real wages may have increased with larger food production and lower food prices. Limited agricultural productivity within colonies does not provide supportive evidence on this.

All these hypotheses have a prominent counterpart in the literature on development economics and growth. Thus, our analysis needs to address the relative role each of them may have played in the colonial settlements episode of São Paulo.

### 1.3 Data and Descriptive Statistics

#### 1.3.1 Data

We use data from three sets of historical information. First, we collect information on the colonialism episode and on life conditions within the official settlements. This allows us not only to identify and locate all the official settlements created in São Paulo, but also to compare demographics and life conditions inside and outside the settlements. Second, we build a longitudinal dataset at the municipality level containing historical Censuses-matched data on socioeconomic conditions. Finally, we complement this dataset with geographical variables and transportation conditions. These three pieces of historical information are described in detail below.

*Colonialism and life conditions within settlements.* There is only scant statistical information and literature on the colonialism episode of São Paulo. To provide a reasonable picture on this in Section 1.1, we combined the few comprehensive analysis found in Martins (1973), Petrone (1982), Gadelha (1982), and Iotti (2001), with case studies and historical documentation from different sources. In particular, we make use of data and qualitative information found in the Annual Reports of the Secretariat of Agriculture of São Paulo (ARSAs) published between 1882 and 1920, and in the Annual Statistical Reports of São Paulo (ASRs) published between 1898 and 1920, and also in 1888 and 1940.

While only scant information on the first waves of colonization was found, the latest waves were relatively well documented. This allowed us to examine economic and demographic patterns within settlements in different years during the 1910s. Additional qualitative information on the procedures

for the selection of settlers and settled regions was also found in the Public Archive of the State of São Paulo (*Arquivo Público do Estado de São Paulo*), and in the Provincial Presidential Reports and Messages to the Legislative Assembly of São Paulo (PPRs). Finally, official websites of the municipalities that have their origins closely connected with the settlement of official colonies provided useful information.<sup>11</sup>

*The datasets.* The official settlements were created mostly between 1877 and 1911. Given this timing, our analysis is based on two municipality-level datasets that merge variables calculated from four National Censuses: 1872 (our starting point, containing the pre-existing conditions), 1920 (short-term point), 1940 (mid-term) and 2000 (long-term).

There existed 88 municipalities in the state of São Paulo according to the 1872 administrative division.<sup>12</sup> Throughout time the municipalities split and generated an increasingly fragmented administrative division, reaching 645 municipalities in 2000. To deal with different administrative divisions across time we merge the four layers of municipal data (1872, 1920, 1940 and 2000) using two different procedures. The first one generates a dataset that is used in municipality fixed-effect regressions. It merges only the 1872-1920 layers of data and uses the 88 municipalities of the 1872 administrative division as the baseline sample. In this dataset the 202 municipalities of the 1920 division were merged to match the 1872 Census boundaries. This makes the two rounds of data comparable. The second dataset is used in cross-section regressions. It merges all the four layers of data and uses the 202 municipalities of the 1920 division as the minimum comparable areas. Municipalities as they were in the later 1940 (270 municipalities) and 2000 (645 municipalities) divisions were merged to make these divisions comparable to the 1920 division. Also, the data from the 1872 Census was linked to this second dataset: each municipality of the 1920 division was simply connected to the 1872 information associated with the original municipality it belonged to.

In our analysis, conditions in 1872 are characterized by the municipal share of slaves, foreigners, literate inhabitants, share of workers by sectors, population density, and share of children attending classes at school. Conditions in 1920 include the share of literate individuals and foreigners (total and per nationality), average land price per hectare, coffee production, share of small farms (up to 100 hectares), population density, average wages in agri-

<sup>11</sup>Many official settlements became independent municipalities of the state of São Paulo throughout the last decades. Some of them have their histories published online.

<sup>12</sup>More precisely, 89 municipalities. However, the independent municipality of Santo Amaro in 1872 is considered part of the capital city once it became part of its administrative division between 1872 and 1920.

culture and civil construction, share of catholics, and the number of baptisms and communions ceremonies per municipality in 1920. We include for 1940 the share of literate inhabitants and children at school, the number of schools per capita, and other religion variables. We use the 2000 Census to calculate current per capita income and years of schooling per cohorts. The definition of the variables used in the paper and the methodology used to construct both datasets are detailed in the Appendix.

*Other controls.* We include in our econometric analysis two important sets of controls. By the late 1870s, many regions of the state of São Paulo were still geographically isolated and with low-density. This was particularly true for the northwestern region of the state, which was farthest away from the state capital. Thus, geographical characteristics, such as distance to the capital city and latitude, should be taken into account in our analysis once they are associated with the timing of the economic expansion of São Paulo. Another important variable used in our analysis is aimed to control for the transportation infrastructure, in particular railroads. On the one hand, the railroads were closely associated with regional economic development. On the other hand, as mentioned before, the location of the official settlements was often conditional on the proximity to transportation. To deal with spurious correlation, we include in our econometric analysis a dummy variable that controls for the presence of railroads in the municipalities. The way we construct the railroad variables and the other geographical controls is described in detail in the Appendix.

### 1.3.2 Descriptive statistics

Following, we present summary statistics comparing settled and non settled regions in a range of different dimensions. Table 1.4 displays descriptive statistics for the municipalities of São Paulo in different points in time. Panel A presents pre-existing conditions. Official settlements were eventually created in 19 municipalities of the 1872 administrative division. As we see in Panel A, municipalities with and without settlements in the future were very similar in terms of pre-existing conditions. This is confirmed for variables such as population density, share of children enrolled in schools, literacy rates, share of slaves, and share of agricultural workers, and share of employment in manufacturing and services. The share of foreigners is very low in both groups (around 1% and 2%), but higher in those municipalities that would receive settlements in the future. We show below that this difference is mainly driven by the capital city.

In Panel B we present geographical and transportation variables, this

time using the 202 municipalities of the finer 1920 administrative division. This Panel shows that the creation of settlements was concentrated in municipalities closer to the capital city and crossed (or that will become crossed) by railroads. This confirms the institutional determinants highlighted in the discussion above.

Panel C presents variables calculated from data available in the 1920 Census, our short-term point. As we can see in the last two columns, the group of municipalities where settlements were created had literacy rates 8.5 to 9 percentage points higher in 1920 when compared to those municipalities without settlements. Interestingly, we do not find significant differences in any other important economic characteristics, such as coffee production, land prices, population density and local wages. The exception is the share of small farms, that seems to be higher in municipalities with settlements. However, last column shows that this difference is mainly driven by proximity to the capital city. Significance vanishes when we condition the difference of means on the distance to the capital city. It should also be mentioned that differences between both groups (regarding population density, for instance) become smaller once we also control for the capital city.

Panel D shows variables for the mid and long-term periods. This Panel confirms that municipalities with settlements had more education in 1940 and today. Together with other results shown below, this fact suggests persistent human capital impacts on settled regions over time. Finally, settled regions have higher levels of per capita income today in comparison to non settled ones.

The four Panels give suggestive evidence on the links between settlement colonies and more education in the short-term, and higher per capita income in the long-term. In the next section we present results from our econometric analysis, which tries to disentangle the human capital hypothesis from the alternative interpretations discussed in Section 1.2.

## 1.4 Results

### 1.4.1 Pre-existing conditions

As mentioned before, in 1872 São Paulo was divided into 88 municipalities. Official colonies would be settled in 19 of these municipalities throughout the following years. The timing of the colonization policy makes the 1872 Census an ideal starting point for the analysis of the relationship between settlement location and pre-existing conditions. The analysis proceeds as follows. First, we create a dependent dummy variable indicating municipalities where

one or more official colonies would be settled in the following years. Second, this variable is regressed on geographical and socioeconomic variables at the municipality level, calculated from data available in the 1872 Census. We run a set of linear regressions of the form

$$S_i = \beta_0 + \beta_1 Lat_i + \beta_2 Ln(dist)_i + X_i' \gamma + \varepsilon_i \quad (1-1)$$

Where  $i$  is the subscript for municipality, and  $S_i$  is the dummy variable indicating the presence of settlements after 1872. The first two independent variables are, respectively, the latitude degrees and the log of the distance to the capital city. The vector  $X_i$  contains the population density, and the shares of literate inhabitants, slaves, agricultural workers, share of workers in agriculture, manufacturing and services, foreigners, and the share of children population attending school.

Table 1.5 presents the results. The regressions show that pre-existing conditions have no significant predictive power for settlement location, although some concerns arise from the positive effect of the share of foreigners found in columns (7) and (8). As shown in columns (9) and (10), however, the significance of this effect drops when we include a dummy variable that controls for the capital city, the most dynamic municipality of the state and place of official settlements before 1872. It is also important to note that both population density and school attendance are negatively correlated with the settlements' location in the most complete specifications, although their coefficients are not statistically significant.

#### 1.4.2 Short-term effects in 1920

To examine the short-term effects of the settlements on the local economies, we follow an empirical strategy that combines two sets of regressions. First, we run cross-section regressions by using outcome variables calculated from data available in the 1920 Census. In this analysis we use geography, railroads and pre-existing conditions in 1872 as controls for the observed heterogeneity across municipalities. Second, we complement this analysis with a difference-in-differences strategy that explores variations in socioeconomic conditions between 1872 and 1920 in municipalities with and without settlements. Although this second strategy is based on a very small sample of municipalities, it helps us to better understand whether unobserved heterogeneity is of any concern. The cross-section analysis is based on a set of OLS regressions for the equation



$$Y_{i,1920} = \varphi_0 + \varphi_1 S_i + \gamma'_1 Geo_i + \gamma'_2 IC_{i,1872} + \gamma'_3 CC_{i,1920} + \gamma_4 Rail_{i,1920} + \varepsilon_i \quad (1-2)$$

Where  $i$  is the subscript for municipality. The dependent variable  $Y_{i,1920}$  includes different outcome variables calculated from data available in the 1920 Census, such as the municipality literacy rate, the share of literate foreigners and other 1920 socioeconomic conditions. Our variable of interest is  $S_i$ , an independent dummy variable indicating municipalities where one or more official colonies were settled between 1872 and 1920. The vector  $Geo_i$  contains geographical variables (distance to the capital city and latitude degrees). The term  $IC_{i,1872}$  includes controls for pre-existing conditions. More precisely, it includes the municipality share of slaves, its literacy rate and the share of foreigners in 1872. Full specifications also include the term  $CC_{i,1920}$ , which controls for current socioeconomic conditions (i.e., conditions in 1920). We include control variables that capture important characteristics of the local economy, but also potential concurrent hypotheses. For instance, we include coffee production per hectare and population density, which should control for the political and economic power of the local elites, and the level of urbanization. We also include the 1920 share of foreigners and the share of small farms (up to 100 ha), factors potentially affected by the settlements. Finally,  $Rail_{i,1920}$  is a dummy variable that indicates the presence of railroads crossing the municipality in 1920. The most complete specifications are weighed by the logarithmic of the municipality population in 1920. Our sample follows the 1920 administrative division and contains 202 municipalities. Once these municipalities were originally part of the less fragmented 1872 division, we compute standard robust errors clustered at the 1872 division level in all specifications.

Table 1.6 presents the first set of results. All specifications follow equation (1-2) and use the literacy rates in 1920 as dependent variable. The regressions show a positive correlation between the settlement of official colonies and literacy rates in the short-term. Column (2) shows that this result is robust not only to controlling for pre-existing, but also for current conditions circa 1920. Finally, Columns (3) and (4) show that the result remains robust when we account for outliers. The average effect remains stable at 4-4.5 percentage points in columns (2) through (4).

In Table 1.7 we repeat these specifications, but use literacy rates among the foreign-born as dependent variable. This exercise provides a natural test for the most likely channel for the short-term effects found in Table 1.6. The results show a positive correlation between the settlement of official colonies and

literacy rates among the foreign-born. Actually, the correlation is significant whenever conditioned on the share of foreign-born and the percentage of foreign-born children among the total foreign-born individuals. The latter controls for the fact that late immigrants were relatively less-skilled than the early ones. As we can see in columns (2) through (4), the average effect ranges between 5.4-6.7 percentage points, and becomes higher and more robust once we control for outliers. These results therefore support empirically the hypothesis that settlements worked as ports of entry of relatively high-skilled immigrants.

In Table 1.8 we use other socioeconomic variables in 1920 as dependent variables. All specifications include the full set of controls for geography, transportation, pre-existing and 1920 current conditions. The first column shows that settlements are not significantly correlated with higher shares of foreigners. Not surprisingly, the share of foreigners is positively associated with coffee production. The second column shows there is no significant correlation between settlements and lower land inequality. This result suggests that the colonization policy did not represent a major change in the repartition of land at the local level. In the third column we see that official settlements had no agglomeration effects, as measured by population density. In fact, along with the results displayed in Tables 1.6 and 1.7, this supports the view that settlements had a composition effect, but no scale effect. This is consistent with a limited inflow of individuals, mostly educated, into regions where human capital levels were very low. As shown in columns (4) through (6), neither the prices of the land, nor the local wages in agriculture and civil construction are significantly correlated with settlements. These results contradict the hypotheses that settlements fostered property rights or public goods improvements (land values would be higher otherwise), or had any substantial effect on local labor markets.

Education, inherited culture or religion are in fact harder to disentangle. However, we can still examine some hypotheses. First, the percentage of foreigners in a given municipality can supposedly be thought of as a proxy for foreign culture, when conditioned on education levels. In this case, the settled regions were not impacted by a cultural effect once the settlements do not appear to be correlated with share of foreigners. In addition, we can also examine whether the settlements had any impacts on the share of foreigners of any particular nationality. We use equation (1-2) to estimate the impact of the settlements on the share of foreigners by nationalities. Panel A in Table 1.9 presents the results. In the first column we repeat the same specification used in Column 1 of Table 1.8. It shows that settlements have no

correlation with the overall share of foreigners. On the other hand, Column 2 shows that settlements have a positive correlation with the share of Germans in the local population. This is consistent with two facts. First, we indeed observe in historical documentation that many settlements were occupied at least partially by German families. Second, by that time the average share of Germans across the municipalities of São Paulo was the lowest amongst the main nationalities arriving at the state: only 0.1%. Thus, these two facts together may have had a composition effect without any stronger scale effect. In fact, the magnitude of the estimated coefficient is unsurprisingly small. The remaining Columns of Table 1.9, Panel A, show that settlements have no correlation with the share of foreigners from other nationalities.

In Panels B and C of Table 1.9 we examine whether the estimated effect of settlements on literacy rates are driven by higher shares of Germans. All regressions follow the same specification used in Panel A. The results in both Panels show that settlements' impact on literacy remain robust and stable, while the share of Germans has no significant association with education.

We can also examine whether the settlements had any impacts on religion. We use information on religion from two sources. First, we use statistics from the 1920 ASR on the share of catholics, the number of per capita infant baptisms and communion ceremonies at the municipality level. We complement this information with the share of protestants and catholics contained in the 1940 Census. We use equation (1-2) to estimate the impact of the settlements on local religion. Table 1.10 presents the results. Columns 1 through 5 do not show any significant relationship between settlements and religion.

The cross-section results hold even when geographical controls, pre-existing and current conditions are taken into account. However, unobserved fixed heterogeneity may challenge any causal interpretation. Influential local politicians and coffee farmers, for example, could have been able to attract to their regions both official colonies and state expenditures in public schooling. In order to control for unobserved local fixed-effects, we complement the cross-section regressions with a panel analysis that explores socioeconomic variations between 1872 and 1920. The analysis proceeds as follows. First, given the limited scope of the first Census (1872), we select only three main demographic variables compatible across Censuses: the share of literate inhabitants, the share of foreigners and the log of the total population. We believe these variables can reasonably account for an important part of the socioeconomic dynamics of São Paulo by the turn of the 19th to the 20th century. In fact, immigration and population growth followed closely the coffee expansion and

its resultant forces. The econometric analysis follows the equation

$$Y_{it} = \alpha_i + \zeta_t + \delta_1 S_{it} + \delta_2 \ln(pop)_{it} + \delta_3 SF_{it} + u_{it} \quad (1-3)$$

Where  $i$  is the subscript for municipality, and  $t \in \{1872, 1920\}$ . The dependent variable is the share of literate individuals, while the variable of interest is the dummy  $S_{it}$ . The variables  $\ln(pop)_{it}$  and  $SF_{it}$  are respectively the log of the total population and the share of foreigners. The variables  $\alpha_i$  and  $\zeta_t$  control respectively for municipality and year fixed-effects. The sample follows the 1872 administrative division and contains 88 municipalities.

Table 1.11 presents the results. Column (1) shows that between 1872 and 1920 literacy rates increased more rapidly in municipalities where official colonies were settled. The effect of the official settlements remains positive and statistically significant even in specifications in columns (2) and (3), which include demographic controls. Interestingly, the coefficients on settlements in columns (1) and (2) are similar to those found in the cross-section analysis. Once the share of foreigners is controlled for, the coefficient falls about 1 percentage point, but is still significant at 10%.

### 1.4.3 Long-term effects

So far we have presented qualitative and quantitative evidence on two unique features of the colonialism episode of São Paulo. First, the combination of different motivations behind the colonization policy spread a similar colonialism experience over different regions, which resulted in a homogeneous distribution of settlements across the state. Second, a distinctive impact of this intervention was the improvement in the human capital composition within settled regions. Based on these two premises, we now take one step further and examine the links between settlement colonies, early changes in human capital, and long term development.

The empirical strategy follows a 2SLS approach. As shown in Table 1.6, there is a robust (first-stage) relationship between the settlements and literacy circa 1920. In the second-stage, we regress contemporary levels of per capita income (calculated by using the 2000 Census) on the instrumented early levels of literacy, and the additional controls. More formally, the analysis follows the equations

$$Y_{i,00} = \lambda_0 + \lambda_1 H_{i,1920} + \lambda_2' Geo_i + \lambda_3' CC_{i,1920} + \lambda_4' Rail_{i,1920} + \lambda_5' IC_{i,1872} + \mu(1-4)$$

$$H_{i,1920} = \pi_0 + \pi_1 S_i + \pi_2' Geo_i + \pi_3' CC_{i,1920} + \pi_4' Rail_{i,1920} + \pi_5' IC_{i,1872} + \tau_i \quad (1-5)$$

Where  $Y_{i,00}$  is the log of the per capita income in municipality  $i$ , in the year 2000. The term  $H_{i,1920}$  is the literacy rate in 1920, while  $S_i$  is the dummy variable for the presence of settlements in  $i$  before 1920. The variables included in  $Geo_i$ ,  $IC_{i,1872}$  and  $CC_{i,1920}$  are identical the ones used before.

The identification strategy uses  $S_i$  as an instrument for  $H_{i,1920}$ , which is valid as long as  $S_i$  is uncorrelated with the error  $\mu_i$ . In other words, the exclusion restriction implied by our IV approach is that, conditional on the controls included in the second-stage, settlements have no effect on per capita income today other than their effect through improvements in early levels of human capital. It is exactly the unique features of the colonialism episode of São Paulo discussed above that make our exclusion restriction plausible. First, we showed that  $S_i$  is not significantly correlated with a range of pre-existing conditions (Section 1.4.1). Second,  $S_i$  has no significant short-term impact on settled regions, but for the composition of human capital (Section 1.4.2).

Table 1.12 presents the results. The first column shows the results of an OLS specification where per capita income is regressed only on the settlement variable, geographical controls, and pre-existing conditions. Settlements are associated with a 16% higher per capital income in 2000. In Column 2 we add various conditions in 1920 as controls, while in Column 3 we also include the shares of different nationalities. Settlements' direct impact remains significant. Moreover, we see that the share of Germans is not significantly associated with per capita income today. Column 4 includes the 1920 literacy rates as an additional control. The significance of the direct impact of settlements vanishes, which is consistent with our exclusion restriction hypothesis. Finally, Columns 5 and 6 show the second-stage results of our IV specifications. Column 5 shows that the 1920 literacy coefficient is now more than twofold the previous estimates in Column 4. We see that a 1% higher literacy rate in 1920 leads to a 2.5% increase in per capita income today. The coefficient increases slightly in Column 6, which is weighted by the log of the 1920 population. The positive bias found by comparing OLS and IV specifications is likely driven by measurement error in historical data.

## 1.5 Persistent Effects of Education

The insights from the growth literature suggest that human capital may foster local economic growth via productivity improvements (Section 1.2). To be operative, this link requires higher current levels of human capital embodied in skilled workers. In this section we examine the effect of settlements

on education over time. We first study mid-term impacts on educational outcomes. More specifically, we estimate equation (1-2) by using as dependent variables educational outcomes at the municipality level in 1940 (literacy, school attainment, and the number of schools per capita).

Table 1.13 presents the results. The first Column shows that settlements have a positive and robust correlation with literacy rates in 1940. However, this results can be driven by older cohorts. For instance, this might capture the literacy of older immigrants educated abroad. To examine persistency over time, in Columns 2 and 3 we focus on young cohorts and contemporary school attendance. Column 2 shows that literacy rates among the youth were higher in settled regions. Column 3 shows that school attendance among children aged between 7 and 14 was higher in settled regions. In Columns 4 and 5 we see that settlements are not significantly associated with a larger number of schools per capita in 1940. Taken together, Columns 2 through 5 suggest that settlements seem to have driven higher educational outcomes in settled regions through the demand side for education.

We can also examine the persistency in educational impacts over time by comparing the schooling attainment for different cohorts born in settled and non settled regions. First, we calculate the average years of schooling at the municipality level for different cohorts by using the microdata from the 2000 Census. We then follow equation (1-2) to estimate the relationship between settlements and average years of schooling attained by the different cohorts.

Table 1.14 presents the results. The first Column shows that individuals born in settled regions between 1920 and 1929 have on average more 0.48 year of schooling than others. We can also see in Columns 2 through 6 that this pattern is roughly stable across younger cohorts up to the ones born between 1970 and 1979. Thus, the settlements are not only associated with higher levels of education in the short run, but also appear to have had long lasting human capital impacts over time. As a result, settled regions have today higher levels of schooling among their adult population.

## 1.6 Concluding Remarks

In this paper we first examined whether and to what extent the settlement colonies of São Paulo have fostered long run development of settled regions through improvements in early levels of human capital. The analysis relied on the unique features of our empirical setting, that allowed us to isolate the human capital channel. More precisely, we provided supportive evidence for the view that education worked as an operative pathway connecting European settlers' arrival and the subsequent long run development of settled regions.

In this paper we also discussed one likely mechanism behind the relationship between human capital and long term growth. We found suggestive evidence for the view that human capital has fostered the development of settled regions via higher levels of education over time. In particular, the persistency in educational outcomes seems to be driven through the demand side for education.

Although this study is based on a specific historical episode, our findings provide evidence for broader interpretations that connect specifically through the human capital channel European colonialism with long term development. We contribute to this literature by disentangling human capital from the other factors that typically accompanied other historical experiences of European settlements.

Table 1.1: List of Variables Calculated at the Municipality Level

Variables	Description	Source of Original Information
Socioeconomic Conditions in 1872:		
Population density	Number of citizens by total area in hectares	1872 Census
School enrollment	Number of children aged 6-15 attending school over total	1872 Census
Literacy rate	Number of literate individuals over total population	1872 Census
Share of slaves	Number of slaves over total population	1872 Census
Share of agricultural workers	Number of agricultural workers over total population	1872 Census
Share of industrialists and merchants	Number of industrialists and merchants over total population	1872 Census
Share of foreigners	Number of foreigners over total population	1872 Census
Geography and Transportation:		
Railroads in 1920	Dummy indicating railroad crossing the municipality in 1920	<a href="http://www.estacoesferroviarias.com.br">http://www.estacoesferroviarias.com.br</a>
Distance to the capital	Distance to the city of SP in km calculated by averaging the distances of the municipalities of the adm. division of 2000 that constitute the original municipality in 1872	1872 and 2000 Censuses
Latitude degrees	Idem, but averaging the latitudes of the municipalities' centroid	1872 and 2000 Censuses
Socioeconomic Conditions in 1920:		
Literacy rate	Number of literate individuals over total population	1920 Census
Share of foreigners (by nationality)	Number of foreigners (by nationality) over total population	1920 Census
Share of small farms	Number of farms up to 100 ha over the total number of farms	1920 Census
Coffee production (ton/ha)	Municipality production of coffee in tons over the municipality area in ha	1920 Census
Population density	Total population over area in ha	1920 Census
Share of foreign children	Number of foreign children over total number of foreigners	1920 Census
Ln (land price)	Average value of the hectare in Mil Reis	1920 Census
Average wage (by sector)	Daily rates in Mil Reis in the agricultural sector or the civil construction	1920 Census
Share of catholics	Number of catholics over total population	1920 ASR and 1920 Census
Number of baptisms	Number of ceremonies of baptism over total population	1920 ASR and 1920 Census
Number of communions	Number of ceremonies of communion over total population	1920 ASR and 1920 Census
Socioeconomic Conditions in 1940:		
Literacy rate	Number of literate individuals over total population	1940 Census
Literacy rate (age 14-19)	Number of literate individuals aged 14-19 over population aged 14-19	1940 Census
School attendance	Number of children aged 7-14 attending school over population aged 7-14	1940 Census
Number of schools per capita	Number of public schools over total population (idem for rural schools)	1940 ASR and 1940 Census
Share of catholics (protestants)	Number of catholics (protestants) over total population	1940 Census
Socioeconomic Conditions in 2000:		
Literacy rate	Number of literate individuals over total population	2000 Census
Years of schooling (by cohorts)	Average years of schooling of the population (by cohorts)	2000 census
Ln (per capita income)	Log of the average income per capita	2000 Census



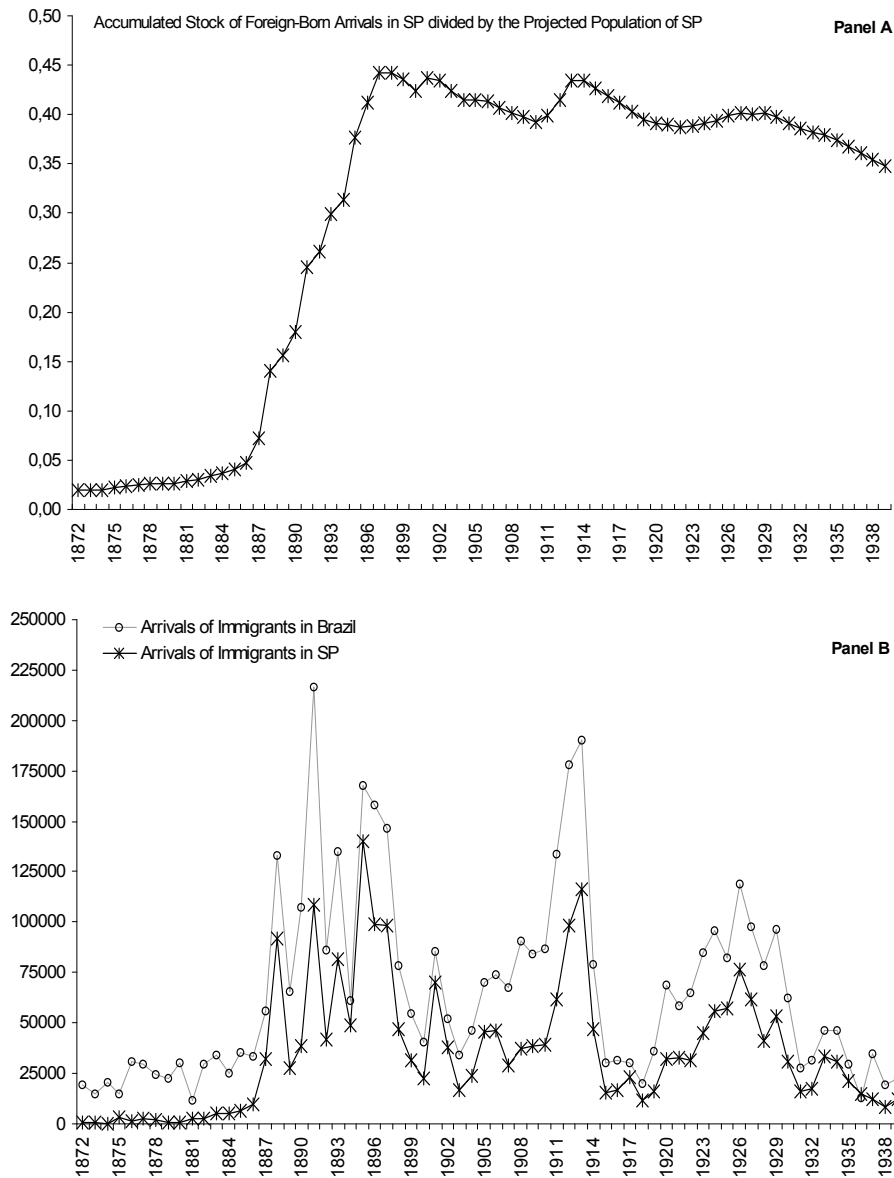
*Merging the 1872-1920 Censuses for the cross section analysis.* Each municipality of the 1920 administrative division is matched to the original municipality it belonged to in the 1872 administrative division. For instance, the municipality of the capital city as existed in 1872 was partitioned into 6 new municipalities between 1872 and 1920. Thus, all these 6 municipalities are trivially linked to the same pre-existing conditions of the original municipality as reported in the 1872 Census.

*Merging the 1872-1920 layers of municipal data with the 1940 and the 2000 Censuses for the cross section analysis.* The variables calculated by using the 1920 Census are trivially included in the dataset. Pre-existing conditions are trivially included as well once we follow the procedure above. To match the 1920 layer to the variables from the 1940 and 2000 Censuses, we proceed as follows. First, the 202 municipalities as they existed in the 1920 division are used as minimum comparable areas in order to match the merged 1920-1872 variables with the variables from the more fragmented divisions of 1940 (270 municipalities) and 2000 (645 municipalities). Second, if a given municipality was partitioned into two or more units between 1920 and 1940 (2000), the variables from the 1940 (2000) Census associated with those new units are then joined in order to rebuild the original municipality. The aggregation of a given variable of the 1940 (2000 Census) into the original 1920 minimum comparable area follows two rules. First, the variable is aggregated as an weighed average by using the population of the administrative units of 1940 (2000) as weights. For instance, suppose municipality X in 1920 (or the minimum comparable area X) is partitioned into X1 and X2 up to 1940. Thus, the share of literate individuals for this area X in 1940 is the sum of the literate individuals in X1 and X2 over the sum of the 1940 populations of X1 and X2. The second rule relates to the different ways a municipality can be partitioned. The problem arises if a given municipality in 1940 was originally made by parts of two or more municipalities in the past. In this case we just follow the capital district of the given municipality. For instance, suppose municipality Y in 1940 was originally made by parts of municipalities Z and K in the past. In this case we ask whether the capital-district of Y was originally located in Z or K. If it were in Z, thus we consider Y to be totally part of Z in the past.

*Merging the 1872-1920 Censuses for the panel analysis.* The panel analysis requires the baseline sample to be the more aggregate one, the 1872 administrative division. We proceed analogously to the above cases. Information from the 1920 Census was mapped into the smaller set of the 1872 division. For instance, in order to make the share of literate individuals for the capital

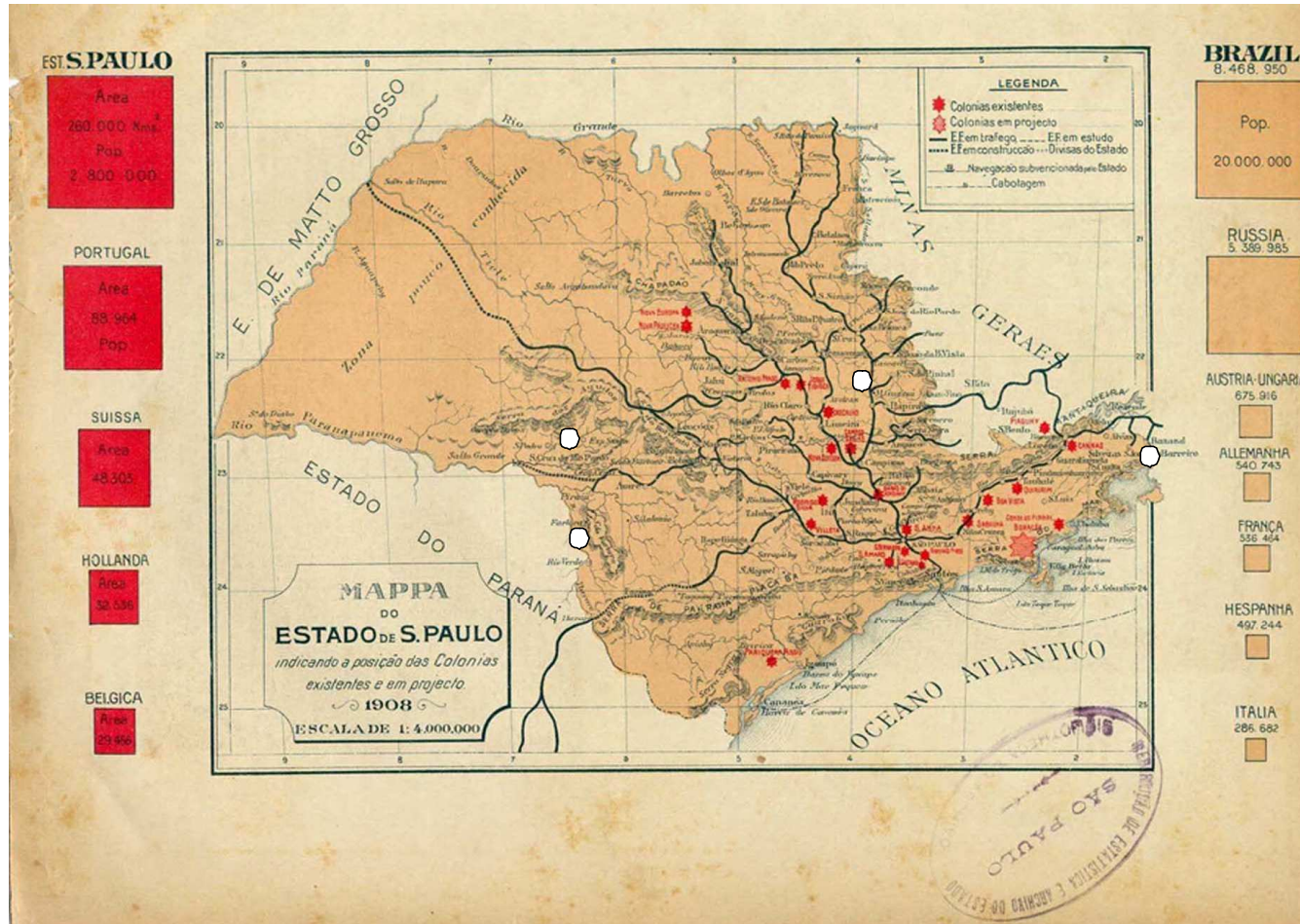
city compatible across time, we first simply calculated the share of literate individuals for the capital city as it was recorded in the 1872 Census. The capital city was partitioned into 6 new municipalities between 1872 and 1920. Thus, in order to calculate the 1920 share of literate individuals for the capital city as it was in 1872, we divided the (i) sum of the literate individuals recorded in the 1920 Census for those 6 municipalities, by the (ii) sum of the total individuals recorded in those same 6 municipalities.

Figure 1.1: Immigrant Arrivals in São Paulo



Notes: The population of São Paulo is projected by using the annual average population growth rates calculated between Censuses (population data from the 1872, 1890, 1900, 1910, 1920 and 1940 Censuses). Data on immigrant arrivals from the *Memorial do Imigrante*, available on <http://www.memorialdoimigrante.org.br>. Panel A does not adjust for emigration or mortality rates.

Figure 1.2: Settlements' Location and Railroads Across the State of São Paulo

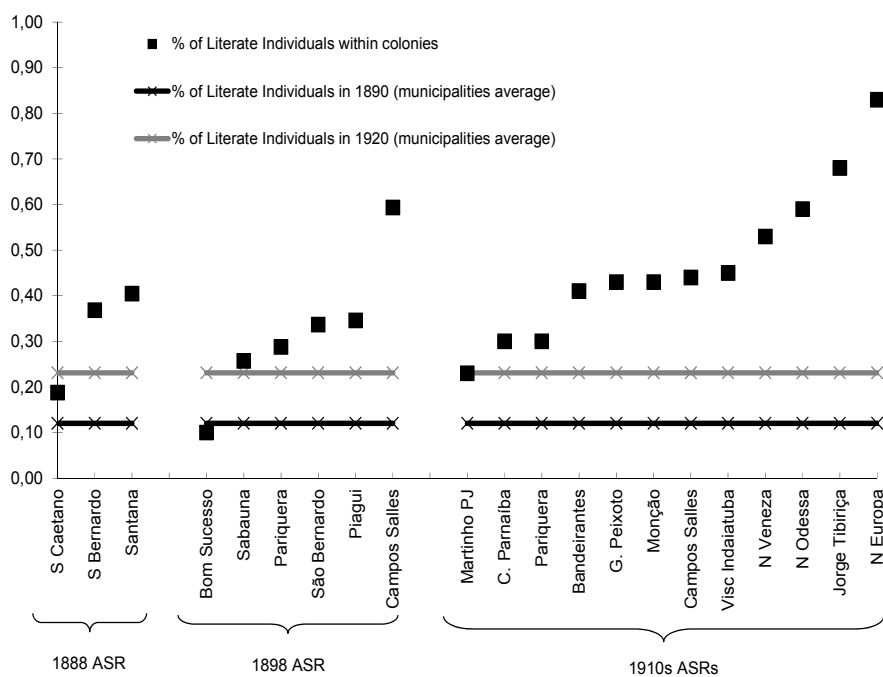


Notes: Map of the state of São Paulo for 1908. The black dots represent the official colonies created up to 1908; the white dots represent the location of those colonies created after 1908. The black lines crossing the state indicate railroads. The map was published in *O Imigrante*, January 1st, n.1, year 1 (source: Arquivo Público do Estado de São Paulo).

Table 1.2: Settlements' General Characteristics

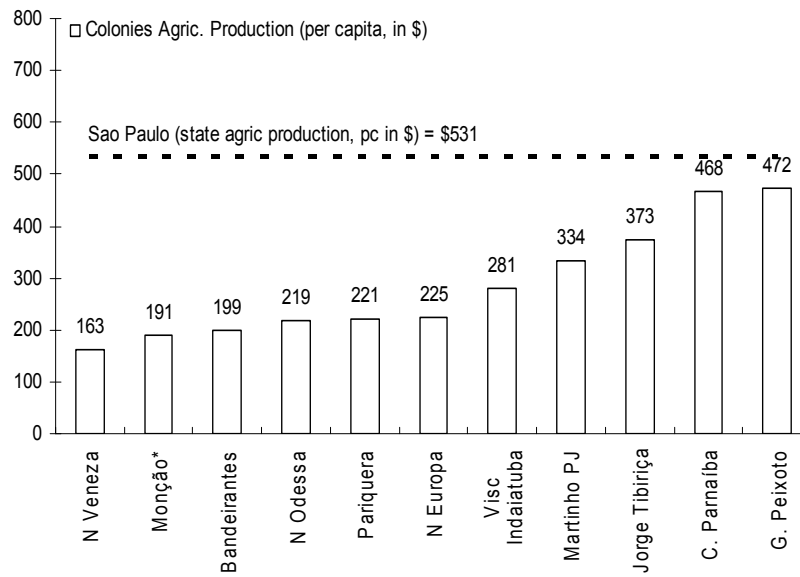
Settlement's Name	Year of Foundation	Year of Emancipation	Location: Original Municipality	Location: Actual Municipality	Initial Area (ha)	Predominant Nationality	Sources of Information
1. Santo Amaro	1829	.	Santo Amaro	São Paulo	.	Germans	Siriani (2005)
2. Parelheiros	1829	Abandoned	São Paulo	São Paulo	.	Germans	Siriani (2005)
3. Cananéia	1862	Abandoned	Cananéia	Cananéia	.	North Americans	Paiva (1993)
4. Glória	1877	.	São Paulo	São Paulo	.	.	Pires (2006)
5. Santanna	1877	.	São Paulo	São Paulo	84	.	1877 PPR
6. São Bernardo	1877	1901	São Bernardo	São Bernardo	1959	Italians/Germans/Poles	Martins (1973)
7. São Caetano	1877	.	São Bernardo	São Caetano	1909	Italians	Martins (1973)
8. Canas	1885	1893	Lorena	Canas	1110	Italians/Brazilians	1892 ARSA
9. Cascalho	1885	1893	Limeira	Cordeirópolis	.	Italians	1892 ARSA
10. Pariquera-açu	1861/1887	1901	Iguape	Pariquera-açu	16603	Brazilians/Germans	1900-1910s ASRAs/ASRs
11. Sen. Antônio Prado	1887	1893	Ribeirão Preto	Ribeirão Preto	1500 prox	Italians	Silva (2004)
12. Ribeirão Pires	1887	1893	São Bernardo	Ribeirão Pires	.	.	Freitas (2008)
13. Rodrigo Silva	1887	1893	Porto Feliz	Porto Feliz	1601	Belgians	Souza (1978)
14. Boa Vista	1888	1893	Jacareí	Jacareí	.	Italians	District Website
15. Barão de Jundiá	1887	1893	Jundiá	Jundiá	514 prox	Italians	Pereira and Fillipini (1988)
16. Sabaúna	1889	1901	Mogi das Cruzes	Mogi das Cruzes	.	Spanics/Brazilians	1898 ARSA
17. Quiririm	1890	1893	Taubaté	Taubaté	.	Italians	Lorenzo (2002)
18. Piaguí	1892	1901	Guaratinguetá	Guaratinguetá	1264	.	1892 ARSA
19. Bom Sucesso	1894	1899	Sorocaba	Sorocaba	.	Brazilians	1898 ARSA
20. Campos Sales	1897	1903	Campinas	Cosmópolis	3200	Swiss/Germans	Baldini (2008)
21. Jorge Tibiriçá	1905	1923/24	Rio Claro	Corumbataí	4356	Spanics/Italians/Brazilians	1900-1910s ASRAs/ASRs
22. Nova Odessa	1905	1920/21	Campinas	Nova Odessa	4310	Russians	1900-1910s ASRAs/ASRs
23. Conde do Pinhal	1907	Abandoned	Ubatuba	Ubatuba	.	Italians	1900-1910s ASRAs/ASRs
24. Gavião Peixoto	1907	1923/24	Araraquara	Gavião Peixoto	4840	.	1900-1910s ASRAs/ASRs
25. Nova Paulicéia	1907	1920/21	Araraquara	Gavião Peixoto	4840	.	1900-1910s ASRAs/ASRs
26. Nova Europa	1907	1920-21	Ibitinga/Araraquara	Nova Europa	4840	Brazilians/Germans	1900-1910s ASRAs/ASRs
27. Bandeirantes	1908	.	S. J. do Barreiro	S. J. do Barreiro	14104	Brazilians/Germans	1900-1910s ASRAs/ASRs
28. Monção	1910	.	S. Barbara do Rio Pardo	Iaras	28873	Various	1900-1910s ASRAs/ASRs
29. Nova Veneza	1910	1919	Campinas	Sumaré	3395	.	1900-1910s ASRAs/ASRs
30. Conde de Parnaíba	1911	1919	Mogy-Mirim	Conchal	4622	.	1900-1910s ASRAs/ASRs
31. Martinho Prado Jr	1911	1920-21	Mogy-Guaçu	Mogy-Mirim	5729	Spanics/Italians/Brazilians	1900-1910s ASRAs/ASRs
32. Visc. de Indaiatuba	1911	1923-24	Mogy-Mirim	Mogy-Mirim	2853	.	1900-1910s ASRAs/ASRs
33. Barão de Antonina	1929	.	Itaporanga	Barão de Antonina	.	Various	Municipality Website

Figure 1.3: Literacy Rates Within Settlements and State Average



Note: We calculated for each settlement the share of literate individuals, and compared the results with two references: the average share of literate individuals calculated across municipalities in 1920 (using the Census of 1920), and across municipalities in 1890 (using the Census of 1890). Data for the first three colonies for 1888; data for the second group of colonies for 1898; and for the third group, the most recent data found in the 1910s ASRs. Sources of information: 1888, 1898 and 1910s ASRs; and the 1890 and 1920 Censuses.

Figure 1.4: Productivity Within Settlements and State Average



Notes: Productivity indicators measured for settlements and the state of São Paulo as the ratio between (i) the overall value of the agricultural, extractive and animal productions, and (ii) total population. In order to smooth production and population fluctuations within colonies we calculated this ratio for years between 1915 and 1920 given the information available, and took the average across time. All values are in per capita *Mil Réis*. Sources of information: 1910s ASRs and the 1920 Census.

Table 1.3: Settlements' Population Inflows and Outflows

Settlement's Name	A: Pop Inflows (Inflows 1912-1918/Pop 1912)	B: Pop Outflows (Outflows 1912-1918)/Pop 1912)	(A)-(B)	Year of Foundation
Pariquera	0,18	0,09	0,09	1887
N. Odessa	0,02	0,01	0,01	1905
Jorge Tibiriçá	0,62	0,01	0,61	1905
G. Peixoto	0,14	0,03	0,11	1907
N Europa	0,50	0,20	0,30	1907
Bandeirantes	1,19	0,90	0,29	1908
Monção	4,13	1,71	2,42	1910
N. Veneza	0,71	0,38	0,33	1910
Martinho Prado Jr	1,49	0,19	1,30	1911
C. Parnaíba	0,48	0,02	0,45	1911
Visc Indaiatuba	1,49	0,16	1,33	1911
Mean	0,99	0,34	0,66	1907
Median	0,62	0,16	0,33	1908

Notes: For each settlement we calculated the ratio between (i) the sum of arrivals (or departures, as the graphic shows) in the years between 1912 and 1918 (not including 1917 once the 1917 ASR was not available), and (ii) the initial stock of the settlement's population in 1912. Source of information: 1910s ASRs.



Table 1.4: Descriptive Statistics Comparing Municipalities With and Without Official Settlements

	Without Colonies		With Colonies		Difference: (With - Without)	
	Mean	SD	Mean	SD	Unconditional	Conditional
Panel A: Pre-Existing Conditions (1872 Census)						
Population density	9.47	6.88	9.72	7.49	0.249	0.381
School enrollment (%)	0.17	0.14	0.15	0.15	-0.015	-0.015
Literacy rate	0.15	0.09	0.17	0.07	1.744	1.753
Share of slaves	0.17	0.10	0.20	0.10	0.025	0.024
Share of agricultural workers	0.37	0.11	0.36	0.10	-0.010	-0.010
Share of industrialists and merchants	0.01	.006	0.01	.006	0.001	0.001
Share of foreigners	0.01	0.01	0.02	0.02	0.010*	0.010*
N. of municipalities (1872 division)		69		19		
Panel B: Geography and Transportation						
Railroad in 1920	0.66	0.47	0.95	0.22	0.285***	.
Distance to the capital log(Km)	5.19	0.65	4.64	1.04	-0.544**	.
Latitude degrees	-22.56	1.01	-22.88	0.76	-0.322	.
N. of municipalities (1920 division)		182		20		
Panel C: Short-term (1920 Census)						
Literacy rate	0.22	0.07	0.31	0.09	0.089***	0.084***
Share of foreigners	0.12	0.10	0.15	0.10	0.027	0.042
Share of small farms	0.66	0.19	0.75	0.09	0.086***	0.033
Coffee production (ton/ha)	0.03	0.04	0.02	0.03	-0.008	-0.005
Population density	27.25	18.90	64.13	138.02	36.872	23.677
Share of foreign children	0.08	0.05	0.09	0.06	0.005	0.017
Ln (land price)	4.94	0.79	5.21	0.74	0.271	0.286
Average wage (civil construction)	7.62	1.52	7.47	1.29	-0.148	-0.161
Average wage (agriculture)	4.20	1.04	4.09	1.33	-0.102	0.139
N. of municipalities (1920 division)		182		20		
Panel D: Mid and Long-term (1940 and 2000 Censuses)						
Literacy rate in 1940	0.42	0.12	0.57	0.16	0.151***	0.145***
Literacy rate in 2000	0.91	0.04	0.95	0.03	0.038***	0.036***
Years of schooling in 2000	6.26	1.09	7.62	1.22	1.356***	1.232***
Ln (per capita income) in 2000	5.67	0.26	5.91	0.27	0.246***	0.204***
N. of municipalities (1920 division)		182		20		

Notes: Last column displays the difference of variable means between municipalities with and without settlements, conditional on the logarithm of the distance to the capital city. Differences of means calculated by using robust standard errors clustered at the 1872 administrative division level. Significance \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 1.5: Pre-Existing Conditions in 1872 and Future Settlements' Location

	Dependent variable: dummy indicating settlement created pos-1872									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Pop density	0.001 (0.007)							-0.003 (0.008)	-0.003 (0.008)	-0.002 (0.008)
School enroll.		-0.142 (0.345)						-0.145 (0.363)	-0.228 (0.347)	-0.272 (0.360)
Literacy rate			0.409 (0.488)					0.307 (0.570)	0.271 (0.560)	0.287 (0.579)
Share of slaves				0.383 (0.479)				0.143 (0.546)	0.332 (0.542)	0.331 (0.557)
Share of agricultural workers					-0.105 (0.370)			-0.265 (0.399)	-0.228 (0.406)	-0.258 (0.412)
Share of indust. & merchts.						2.743 (6.822)		-2.162 (8.177)	-1.976 (8.070)	-1.622 (8.453)
Share of foreigners							5.973 (2.897)**	6.106 (3.255)*	4.449 (3.513)	4.607 (3.509)
Observations	88	88	88	88	88	88	88	88	88	88
Geo Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control Capital	No	No	No	No	No	No	No	No	Yes	Yes
Weighted	No	No	No	No	No	No	No	No	No	Yes

Notes: Each column is a OLS regression at the municipality level. Dependent variable is a dummy indicating the presence of settlements after 1872. Independent variables are listed in rows, all using data from the 1872 Census. Geographical controls include the log of the distance to the capital, and latitude degrees. Control Capital means the inclusion of a dummy variable for the capital city. School enrollment is the share of children enrolled in schools. Weighted regressions use the log of the 1872 municipality population as weights. Significance: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 1.6: Settlements' Short-Term Impacts on Literacy Rates in Settled Regions

	Dependent Variable: Literacy Rate in 1920			
	(1)	(2)	(3)	(4)
Settlement (dummy 0/1)	0.075 (0.022)***	0.044 (0.016)***	0.045 (0.016)***	0.042 (0.020)**
Current conditions (1920):				
Railroads		0.050 (0.011)***	0.049 (0.011)***	0.046 (0.016)***
Share of foreigners		0.425 (0.086)***	0.438 (0.084)***	0.380 (0.091)***
Share of small farms		-0.040 (0.023)*	-0.045 (0.023)**	-0.037 (0.033)
Coffee (tons/ha)		-0.295 (0.131)**	-0.318 (0.132)**	-0.208 (0.192)
Ln(pop density)		0.018 (0.009)*	0.019 (0.010)*	0.016 (0.009)
Share of foreign children		-0.102 (0.076)	-0.104 (0.078)	-0.072 (0.120)
Pre-existing conditions (1872):				
Share of foreigners	-0.096 (0.459)	-0.154 (0.418)	-0.129 (0.411)	0.320 (0.452)
Share of slaves	0.177 (0.081)**	0.120 (0.057)**	0.118 (0.055)**	0.087 (0.078)
Literacy rate	0.001 (0.001)*	0.001 (0.000)*	0.001 (0.000)*	0.000 (0.001)
Observations	202	202	202	202
Control geography	Yes	Yes	Yes	Yes
Control outliers	No	No	Weighted	QREG

Notes: Each column is a OLS regression at the municipality level. Dependent variable is the municipality literacy rate in 1920. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. We control for outliers using either weighted regressions with the log of the 1920 municipality population as weight (Column 3); or quantile regressions at the median (QREG, Column 4). Robust errors clustered at the 1872 administrative division. Significance: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 1.7: Settlements' Short-term Effects and Literacy Rates Among the Foreign-Born

	Dep. Var.: Literacy Among Foreigners in 1920			
	(1)	(2)	(3)	(4)
Settlement (dummy 0/1)	0.029 (0.026)	0.054 (0.028)*	0.057 (0.028)**	0.067 (0.033)**
Current conditions (1920):				
Railroads		0.016 (0.025)	0.014 (0.025)	0.016 (0.026)
Share of foreigners		-0.291 (0.137)**	-0.263 (0.133)*	-0.263 (0.150)*
Share of small farms		-0.033 (0.066)	-0.043 (0.063)	-0.031 (0.056)
Coffee (tons/ha)		-0.654 (0.299)**	-0.704 (0.295)**	-0.651 (0.312)**
Ln(pop density)		-0.010 (0.018)	-0.006 (0.018)	-0.003 (0.016)
Share of foreign children		-0.952 (0.231)***	-0.941 (0.229)***	-0.843 (0.198)***
Pre-existing conditions (1872):				
Share of foreigners	-0.320 (0.772)	-0.860 (0.711)	-0.825 (0.706)	-0.441 (0.775)
Share of slaves	0.024 (0.149)	0.013 (0.137)	-0.004 (0.134)	-0.149 (0.133)
Literacy rate	0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)	0.000 (0.001)
Observations	202	202	202	202
Control geography	Yes	Yes	Yes	Yes
Control outliers	No	No	Weighted	QREG

Each column is a OLS regression at the municipality level. Dependent variable is the share of literate foreigners amongst the total number of foreigners in 1920. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. We control for outliers using either weighted regressions with the log of the 1920 municipality population as weight (Column 3); or quantile regressions at the median (QREG, Column 4). Robust errors clustered at the 1872 administrative division. Significance: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 1.8: Settlements' Short-term Effects on Other Socioeconomic Variables (in 1920)

	Share of foreigners	Share of small farms	Log Pop Density	Land Price	Wages Constret.	Wages Agrict
	(1)	(2)	(3)	(4)	(5)	(6)
Settlement (dummy 0/1)	0.014 (0.021)	0.033 (0.034)	0.089 (0.158)	0.065 (0.065)	0.025 (0.427)	0.106 (0.279)
Current conditions (1920):						
Railroads	0.063 (0.010)***	-0.054 (0.031)*	0.341 (0.158)**	0.194 (0.051)***	0.445 (0.291)	-0.232 (0.212)
Share of foreigners		0.040 (0.247)	0.929 (0.741)	2.080 (0.370)***	5.262 (2.151)**	5.633 (1.686)***
Share of small farms	0.006 (0.035)		0.669 (0.242)***	0.100 (0.163)	-0.651 (0.728)	0.435 (0.548)
Ln(pop density)	0.011 (0.009)	0.056 (0.021)**		0.333 (0.026)***	-0.416 (0.196)**	-0.258 (0.208)
Coffee (tons/ha)	0.794 (0.160)***	-0.117 (0.422)	7.582 (1.243)***	5.983 (0.599)***	0.475 (2.341)	-1.845 (1.821)
Share of foreign children	0.410 (0.098)***	0.090 (0.303)	-2.029 (0.725)***	-0.022 (0.466)	0.173 (2.502)	-2.431 (2.111)
Pre-existing conditions (1872):						
Share of foreigners	0.237 (0.384)	-0.677 (0.943)	-0.138 (3.555)	1.739 (1.881)	1.027 (9.052)	1.320 (5.914)
Share of slaves	-0.076 (0.080)	0.004 (0.209)	1.075 (0.519)**	-0.530 (0.304)*	-2.367 (1.583)	-1.436 (1.076)
Literacy rate	0.001 (0.001)	0.002 (0.002)	0.004 (0.005)	-0.004 (0.003)	0.026 (0.014)*	0.004 (0.010)
Observations	202	202	202	202	148	125
Control geography	Yes	Yes	Yes	Yes	Yes	Yes
Weighed	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column is a OLS regression at the municipality level. Dependent variables are displayed above numbered columns, all calculated from data available in the 1920 Census. Wages are measured in *Mil Reis*, in daily rates. The variable land prices is the municipality average price of the hectare, in *Mil Reis*. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. All regressions are weighted with the log of the 1920 municipality population. Robust errors clustered at the 1872 administrative division. Significance: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 1.9: Settlements' Short-Term Effects on the Share of Foreigners by Nationality (in 1920)

	(1)	(2)	(3)	(4)	(5)	(6)
<b>Panel A - Dep. Vars: Share of Foreigners by Nationality</b>						
	<b>All</b>	<b>Germans</b>	<b>Italians</b>	<b>Spaniards</b>	<b>Portuguese</b>	<b>Japanese</b>
Settlement (dummy 0/1)	0.014 (0.021)	0.002 (0.001)***	-0.005 (0.005)	-0.004 (0.005)	0.002 (0.005)	0.004 (0.004)
<b>Panel B - Dep. Var: Share of Literate Inhabitants</b>						
Settlement (dummy 0/1)	0.041 (0.016)**	0.036 (0.017)**	0.044 (0.016)***	0.043 (0.016)***	0.043 (0.016)***	0.046 (0.016)***
Share of Germans	1.669 (3.562)	5.292 (4.354)				
Share of Italians	-0.005 (0.217)		-0.219 (0.291)			
Share of Spaniards	-0.142 (0.238)			-0.364 (0.244)		
Share of Portuguese	1.039 (0.259)***				1.135 (0.276)***	
Share of Japanese	-0.261 (0.262)					-0.366 (0.257)
<b>Panel C - Dep. Var: Share of Literate Foreigners</b>						
Settlement (dummy 0/1)	0.066 (0.026)**	0.057 (0.027)**	0.054 (0.028)*	0.058 (0.027)**	0.054 (0.028)*	0.060 (0.028)**
Share of Germans	-4.934 (5.508)	-0.539 (6.205)				
Share of Italians	-0.293 (0.210)		-0.554 (0.348)			
Share of Spaniards	0.390 (0.392)			0.362 (0.411)		
Share of Portuguese	1.274 (0.349)***				1.281 (0.375)***	
Share of Japanese	-0.887 (0.744)					-0.850 (0.745)
Common Specifications to All Panels:						
Observations	202	202	202	202	202	202
Conditions in 1872 and 1920	Yes	Yes	Yes	Yes	Yes	Yes
Geography and Railroads	Yes	Yes	Yes	Yes	Yes	Yes
Weighted	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column is a regression at the municipality level. Dependent variables in Panel A are the total number of foreigners per nationality divided by total population in 1920. Dependent variable in Panel B is the literacy rate in 1920, while in Panel C is the literacy rate amongst the foreigners. All regressions control for geographical variables, the presence of railroads and socioeconomic conditions in 1872 and 1920. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for conditions in 1872 include the share of slaves, the share of foreigners and literacy rates. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. All regressions are weighted with the log of the 1920 municipality population. Robust errors clustered at the 1872 administrative division. Significance: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 1.10: Settlements' Short-Term Impacts on Religion in Settled Regions (in 1920 and 1940)

	Share of Catholics (in 1920)	N. of Baptisms Per Capita (in 1920)	N. Communion Per Capita (in 1920)	Share of Protestants (in 1940)	Share of Catholics (in 1940)
	(1)	(2)	(3)	(4)	(5)
Settlement (dummy 0/1)	0.002 (0.075)	-0.001 (0.003)	-0.006 (0.005)	0.004 (0.003)	-0.009 (0.009)
Conditions in 1920:					
Railroads	-0.153 (0.151)	-0.007 (0.004)	0.004 (0.004)	0.006 (0.003)**	-0.032 (0.010)***
Share of foreigners	-0.388 (0.382)	0.026 (0.017)	0.046 (0.030)	0.030 (0.015)*	-0.018 (0.074)
Share of small farms	-0.488 (0.262)*	-0.003 (0.007)	-0.004 (0.012)	0.010 (0.007)	0.035 (0.044)
Coffee (tons/ha)	-0.120 (0.569)	0.080 (0.079)	0.108 (0.041)**	-0.082 (0.027)***	0.284 (0.132)**
Ln(pop density)	0.068 (0.038)*	-0.000 (0.002)	-0.002 (0.003)	-0.004 (0.002)**	0.019 (0.005)***
Share of foreign children	2.497 (2.279)	0.002 (0.020)	-0.067 (0.051)	0.003 (0.022)	-0.010 (0.099)
Pre-existing conditions in 1872:					
Share of foreigners	-0.422 (1.814)	0.048 (0.063)	0.107 (0.125)	0.157 (0.058)***	-0.360 (0.268)
Share of slaves	0.031 (0.217)	0.002 (0.016)	-0.020 (0.023)	-0.022 (0.012)*	0.127 (0.034)***
Literacy rates	-0.004 (0.003)	-0.000 (0.000)**	-0.000 (0.000)	-0.000 (0.000)	0.001 (0.000)*
Obs	182	184	170	202	202
Geographical variables	Yes	Yes	Yes	Yes	Yes
Control outliers	No	No	No	No	No

Notes: Each column is a regression at the municipality level. Dependent variables are displayed above numbered columns. Dependent variables in Columns 1 through 3 are calculated by using data available in the 1920 ASR. Dependent variables in Columns 4 and 5 are calculated by using data from the 1940 Census. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. All regressions are weighted with the log of the 1920 municipality population. Robust errors clustered at the 1872 administrative division. Significance: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 1.11: Settlements' Short-term Effects on Literacy Rates and Other Socioeconomic Conditions: Differences-in-Differences Comparing Settled and Non-Settled Region in 1872 and 1920

	Dependent Variable: Literacy Rate		
	(1)	(2)	(3)
Settlement (dummy 0/1)	0.046 (0.023)**	0.043 (0.023)*	0.034 (0.019)*
Ln population		0.011 (0.020)	-0.075 (0.019)***
Share of foreigners			1.021 (0.186)***
Observations	176	176	176
N. of municipalities	88	88	88
Year FE	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes

Notes: Each column is a linear regression at the municipality level. The sample follows the 1872 administrative division. The analysis uses data at the municipality level from two points in time, 1872 and 1920. Dependent variables are displayed above numbered columns. Independent variables are listed in rows. All specifications year fixed-effects. All specifications controls for municipality fixed effects, but Column (4). Significance: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .



Table 1.12: OLS and IV Regressions: Settlements, Human Capital, and Long Run Impacts on Per Capita Income

	Dep. Variable: Ln(Per Capita Income in 2000)					
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	IV (5)	IV (6)
Settlement (dummy 0/1)	0.168 (0.052)***	0.098 (0.047)**	0.101 (0.051)*	0.061 (0.049)		
Literacy rate in 1920				1.040 (0.245)***	2.588 (1.307)**	2.546 (1.241)**
Conditions in 1920:						
Railroads		0.179 (0.042)***	0.182 (0.045)***	0.134 (0.041)***	0.064 (0.078)	0.070 (0.075)
Share of foreigners		0.690 (0.256)***	0.240 (0.472)	-0.058 (0.406)	-0.502 (0.491)	-0.555 (0.468)
Share of small farms		0.189 (0.096)*	0.205 (0.099)**	0.245 (0.099)**	0.304 (0.100)***	0.307 (0.098)***
Coffee (tons/ha)		0.008 (0.373)	0.168 (0.421)	0.293 (0.422)	0.478 (0.485)	0.549 (0.467)
Ln(pop density)		0.004 (0.029)	0.005 (0.029)	-0.011 (0.027)	-0.036 (0.033)	-0.036 (0.032)
Share of foreign children		-0.558 (0.278)**	-0.625 (0.320)*	-0.582 (0.330)*	-0.517 (0.355)	-0.510 (0.344)
Share of Germans			-4.347 (9.333)	-6.261 (9.384)	-9.111 (12.724)	-9.120 (12.219)
Share of Spaniards			0.446 (0.558)	0.623 (0.522)	0.887 (0.685)	0.913 (0.647)
Share of Italians			0.202 (0.576)	0.200 (0.430)	0.197 (0.373)	0.224 (0.358)
Share of Portuguese			1.156 (0.878)	0.065 (0.894)	-1.560 (1.623)	-1.344 (1.516)
Share of Japanese			1.142 (0.727)	1.394 (0.702)*	1.770 (0.841)**	1.768 (0.828)**
Conditions in 1872:						
Share of foreigners	0.620 (1.348)	0.566 (1.128)	0.778 (1.151)	0.958 (0.999)	1.226 (1.160)	1.137 (1.105)
Share of slaves	0.319 (0.301)	0.140 (0.205)	0.191 (0.217)	0.063 (0.209)	-0.129 (0.317)	-0.096 (0.302)
Literacy rate	-0.001 (0.001)	-0.001 (0.001)	-0.002 (0.001)	-0.002 (0.001)*	-0.003 (0.001)**	-0.003 (0.001)**
Observations	202	202	202	202	202	202
Geography and Conditions in 1872	Yes	Yes	Yes	Yes	Yes	Yes
Conditions in 1920	No	Yes	Yes	Yes	Yes	Yes
Nationalities in 1920	No	No	Yes	Yes	Yes	Yes
Weighted	No	No	No	No	No	Yes

Notes: Each column is a regression at the municipality level. The sample follows the 1920 administrative division. Dependent variable is the logarithmic of the average per capita income in 2000. Columns 1-4 display coefficients of OLS regressions. Columns 5-6 displays coefficients of second-stage regressions where literacy rates in 1920 are instrumented with the settlements' variables. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners, and the share of foreigners by nationality (Columns 3-6). All regressions are weighted with the log of the 1920 municipality population. Robust errors clustered at the 1872 administrative division. Significance: \*\*\* p<0.01, \*\* p<0.05, \*

Table 1.13: Settlements' Mid-term Effects on Literacy Rates and Schooling (in 1940)

	Literacy Rate	Literacy Rate (14-19 Years Old)	Children 7-14 Attending School	N. Schools Per Capita	N. Rural Schools Per Capita
	(1)	(2)	(3)	(4)	(5)
Settlement (dummy 0/1)	0.069 (0.028)**	0.090 (0.032)***	0.055 (0.020)***	-0.049 (0.031)	0.005 (0.009)
Conditions in 1920:					
Railroads	0.093 (0.016)***	0.124 (0.022)***	0.078 (0.016)***	0.143 (0.031)***	0.043 (0.010)***
Share of foreigners	0.520 (0.112)***	0.432 (0.130)***	0.275 (0.114)**	-0.155 (0.173)	-0.086 (0.057)
Share of small farms	0.025 (0.032)	-0.029 (0.046)	-0.034 (0.044)	0.054 (0.077)	0.015 (0.020)
Coffee (tons/ha)	-0.383 (0.206)*	-0.630 (0.264)**	-0.467 (0.239)*	0.055 (0.309)	0.101 (0.126)
Ln(pop density)	0.018 (0.014)	0.024 (0.012)**	0.035 (0.012)***	0.011 (0.018)	0.008 (0.008)
Share of foreign children	-0.086 (0.132)	0.029 (0.189)	-0.028 (0.171)	-0.050 (0.236)	-0.048 (0.072)
Pre-existing conditions in 1872:					
Share of foreigners	1.820 (0.599)***	1.054 (0.735)	0.722 (0.542)	-0.376 (0.851)	0.311 (0.273)
Share of slaves	0.116 (0.111)	0.261 (0.122)**	0.135 (0.078)*	0.386 (0.117)***	0.158 (0.043)***
Literacy rates	0.001 (0.001)	0.001 (0.001)	-0.000 (0.001)	-0.002 (0.001)*	-0.000 (0.000)
Observations	202	202	201	202	202
Control Geography	Yes	Yes	Yes	Yes	Yes
Weighted	Yes	Yes	Yes	Yes	Yes

Notes: Each column is a OLS regression at the municipality level. Dependent variables are displayed above numbered columns. Dependent variables in Columns 1 through 3 are calculated by using data available in the 1940 Census. Dependent variables in Columns 4 and 5 are calculated by using data on the municipality number of schools, available in the 1940 ASR, divided by the municipality population aged between 7 and 14 years old (information available in the 1940 Census). The number of schools in Columns 4 and 5 include only the total number of public schools in the municipality. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. All regressions are weighted with the log of the 1940 municipality population. Robust errors clustered at the 1872 administrative division. Significance: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$ .

Table 1.14: Settlements' Impacts on Years of Schooling Over Time: Analysis by Cohorts

	Dep. Variable: Average Years of Schooling by Cohorts Born in Between the Years of Reference						
	Cohorts	Cohorts	Cohorts	Cohorts	Cohorts	Cohorts	All
	1920-29	1930-39	1940-49	1950-59	1960-69	1970-79	Cohorts
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Settlement (dummy 0/1)	0.483 (0.179)***	0.554 (0.164)***	0.652 (0.220)***	0.672 (0.253)***	0.760 (0.179)***	0.487 (0.159)***	0.660 (0.202)***
Conditions in 1920:							
Railroads	0.544 (0.155)***	0.725 (0.131)***	1.054 (0.192)***	0.840 (0.198)***	0.733 (0.195)***	0.595 (0.170)***	0.922 (0.189)***
Share of foreigners	3.117 (1.389)**	3.520 (1.136)***	3.139 (1.309)**	4.102 (1.447)***	5.050 (1.242)***	3.768 (0.963)***	4.311 (1.180)***
Share of small farms	-0.567 (0.309)*	-0.307 (0.286)	-0.350 (0.333)	0.012 (0.335)	0.121 (0.324)	0.304 (0.292)	-0.147 (0.314)
Coffee (tons/ha)	-2.679 (1.930)	-3.931 (1.827)**	-2.456 (2.043)	-1.845 (2.407)	-1.515 (1.848)	-0.541 (1.852)	-1.894 (2.107)
Ln(pop density)	0.305 (0.125)**	0.193 (0.189)	0.162 (0.151)	0.116 (0.189)	0.001 (0.144)	-0.087 (0.105)	-0.081 (0.143)
Share of foreign children	-0.452 (1.571)	-0.874 (1.095)	-0.230 (1.035)	-0.912 (1.302)	-0.290 (1.186)	0.134 (1.150)	-0.374 (1.086)
Pre-existing conditions in 1872:							
Share of foreigners	6.075 (5.617)	0.770 (4.500)	3.966 (5.445)	9.282 (5.089)*	11.804 (4.360)***	12.792 (3.797)***	10.417 (4.398)**
Share of slaves	1.245 (0.945)	0.932 (0.754)	0.736 (0.771)	1.038 (1.087)	0.552 (0.849)	0.365 (0.893)	0.146 (0.873)
Literacy rates	0.001 (0.007)	0.003 (0.006)	0.011 (0.007)	0.011 (0.008)	0.003 (0.006)	0.002 (0.005)	0.006 (0.006)
Observations	202	202	202	202	202	202	202
Geographical variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column is a OLS regression at the municipality level. Dependent variables are displayed above numbered columns. Dependent variable in each Column through 1 to 6 is calculated as the average years of schooling attained by those individuals born in years shown above each Column's number. This calculation includes only the individuals born in the municipality of reference (exclude mostly Brazilian immigrants). Dependent variable in Column 7 is the average years of schooling calculated for all individuals born between 1920 and 1979 in the municipality of reference. Independent variables are listed in rows. Controls for pre-existing conditions and current conditions calculated respectively from data available in the 1872 and 1920 Censuses. Geographical controls include the log of the distance to the capital city, and latitude degrees. Controls for pre-existing conditions include the share of slaves, the share of foreigners and literacy rates in 1872. Controls for 1920 current conditions include a dummy indicating railroads, the share of foreigners, the share of small farms (up to 100 ha), the coffee production (in tons per ha), the logarithm of the population density (individuals per ha), and the share of foreign children amongst the total number of foreigners. Robust errors clustered at the 1872 administrative division. Significance: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.