Internal migration and the spread of long-term impacts of historical immigration in Brazil*

Eduardo Cenci† Daniel A. F. Lopes‡ Leonardo M. Monasterio§

August 2, 2019

Abstract

Numerous studies have documented persistent economic impacts of historical events, often caused by immigration and perpetuated via the human capital channel. Human capital, however, is mobile. By documenting persistent impacts near the original sites of historical immigration, studies were unable to tell if persistence operates via factors that are fixed (e.g., head start in infrastructure) or mobile (e.g., intergenerational transmission). We investigate this question using a surname-based classification of ancestries to track descendants of historical settlers in Brazil living far from their injection points. We find a positive effect of the concentration of non-Iberian surnames, our proxy for the presence of descendants, on earnings in the municipalities studied. An increase of one percentage point in the concentration of non-Iberians increases earnings by 2% to 5% on average. We interpret the results as evidence of mobile human capital propagating the positive impacts generated by a historical event. Our discussion of channels for the observed effects points to an accelerated agricultural transformation in the municipalities where the concentration of non-Iberians is higher.

KEYWORDS: migration, descendants, surnames, human capital, persistence

JEL CODES: J15, J61, 015, R23, N36

*The authors thank Laura Schechter, Emilia Tjernström, Bradford Barham, Esteban Quiñones, Jeffrey Williamson, and seminar participants at the University of Wisconsin-Madison and at the X Summer School in Development Economics of the Italian Development Economists Association for their valuable comments on earlier versions of this paper. We also thank Marin Skidmore for assistance with the CHIRPS data.

†Department of Agricultural and Applied Economics, University of Wisconsin-Madison (ecenci@wisc.edu).

‡Institute for Applied Economic Research (IPEA).

§National School of Public Administration(ENAP), Brasília, Brazil.

1 Introduction

The evidence in favor of persistent effects of historical events on present economic development is abundant in the economics literature. As noted by Nunn (2009), many studies in this literature examine the long-lasting effects resulting from the arrival of Europeans in different parts of the world from the 16th century to World War II via colonization (Acemoglu et al., 2001; Easterly and Levine, 2016) or via later immigration (Hatton and Williamson, 1998; Hatton and Ward, 2018).[1] Many of these studies point to human capital, brought to the New World and other places by colonizers and immigrants (henceforth *settlers*, for simplicity) and transmitted across generations in the years thereafter, as the main channel through which the economic impacts of historical events persist over the very long term (Borjas, 1992; Rocha et al., 2017; Valencia Caicedo, 2018).

Human capital, however, moves. As settlers move within their new destinations, the human capital they brought also moves. It spreads over time, as human capital is transmitted across generations within a family (vertical transmission) or between families (oblique transmission) (Bisin and Verdier, 2000), and it spreads over space, as settlers and their descendants move to different parts of the receiving country. Nevertheless, part of the human capital shock remains attached to the place that received the original settlers, i.e., the "injection point". This is the case for local institutions developed by the newcomers, redistribution of land resulting from their arrival, and earlier investments in infrastructure to accommodate them.[2] Thus, it is unclear how we can know if the persistent effects of human capital shocks on present economic development observed in the literature are due to fixed factors such as a head start in infrastructure or mobile factors like the intergenerational transmission of human capital. The answer to this question has implications for public policy since investments to attract (or repel) immigration made at the local level can generate impacts at the national level if mobile factors of human capital are more relevant for persistence than the fixed ones.[3]

Answering the "fixed vs. mobile" question is not simple. It requires the observation of a historical immigration shock in a place where the following ideal conditions hold. First, there are two separate regions, one which received the immigration shock (the injection point) and one which is untouched by the original settlers (the study region). Thus, we can observe a region that did not receive historical settlers at any point (no fixed factors). Second, there were no new influxes of immigrants in the country, so we observe only persistent effects of an old shock rather than a contemporaneous one. Third, there is enough internal mobility after the initial shock. The study region receives a significant number of descendants of historical settlers, which can transport human capital to the study region via internal migration (mobile factors only). Fourth, it is possible to identify the descendants of the historical settlers in the population today.

By setting this study in Brazil, we meet the first three requirements.[4] Injection points of non-Iberian immigration from the late 19th and early 20th centuries were concentrated in the South and Southeast regions of the country, with little immigration happening in other regions. After World War II international immigration in Brazil decreased sharply and by 1960 it was approaching the negligible levels the country has today. Finally, the expansion of the agricultural frontier of Brazil to the west and north of the country, initiated in the 1950s, created significant south-to-north internal migration, pulling many descendants of the non-Iberian settlers to this region.[5]

To satisfy the fourth requirement, we use the surname-based classification of ancestries developed by Monasterio (2017) to construct a proxy for the concentration of descendants of the original settlers in the present population: the concentration of non-Iberian surnames.[6] Given the history of the country, any person bearing non-Iberian surnames in Brazil is likely a descendant from settlers that arrived around 1850-1940 in the South and Southeast regions, not a modern immigrant nor someone descending from earlier colonizers, local indigenous people, or former slaves.

We investigate if municipalities on the agricultural frontier of Brazil with a strong presence of non-Iberian surnames in the population exhibit better economic outcomes, and in particular, higher wages. Our empirical strategy uses thirteen years of cross-sectional data at the individual level. Analysis using weighted least-squares shows a positive association between the concentration of non-Iberian surnames, our proxy for the presence of descendants, on average earnings in the municipalities studied. This positive association holds after we control for population composition, geographic and climate controls, and state fixed effects. Results also hold in an instrumental variables (IV) strategy that combines distance to historical settlements with terrain ruggedness to predict the concentration of non-Iberians on the agricultural frontier. The results from our IV analysis show that an increase of one percentage point (one standard deviation) in the concentration of non-Iberians increases earnings by 2–5% (9–21%) on average, depending on the specification and the sample used.

[1] Some studies look the other way around, investigating the effects of historical emigration in Europe (Taylor and Williamson, 1997; Abramitzky et al., 2012).

[2] Galor et al. (2009) and de Carvalho Filho and Monasterio (2012) discuss the relationship between land distribution, development of institutions, and human capital in the context of historical settlements in the US and Brazil, respectively.

[3] For example, the efforts made by Argentina and Brazil to attract European immigrants in the past, sponsored at the national level but concentrated in a few regions, could have benefited those countries as a whole.

[4] Doing this study in Brazil is also convenient because of the availability of data informing both surnames and labor outcomes.

[5] Section 2 gives details on this account and justifies the choice of the agricultural frontier as the region of study.

[6] Due to social and legal norms surrounding surnames in Brazil, our classification of ancestries ends up losing the maternal lineage. We discuss this and other issues in more detail in section 2.1.

The main contribution of our paper is to show that the persistent economic impacts of historical events—of an immigration shock, in particular—can spread beyond the original injection points due to the mobility of human capital. Thus, we advance the literature reviewed by Nunn (2009) and Spolaore and Wacziarg (2013) documenting long-lasting economic impacts far from the original sites of an immigration shock.[7] Studies have already documented the long-lasting positive economic impacts of the presence of historical settlements and settler's descendants in municipalities around the historical injection points in Brazil (de Carvalho Filho and Monasterio, 2012; Rocha et al., 2017; Ehrl and Monasterio, 2017), but they were unable to answer the "fixed vs. mobile" question we raise here.[8] Other studies have shown that non-Iberians in Brazil have higher earnings and more schooling (Monasterio, 2017) and that their children have higher scores on standardized tests (Lopes et al., 2017), thus providing evidence to support the claims that descendants of historical settlers may have carried a distinct type of human capital to the agricultural frontier of Brazil (Alves, 2016). Therefore, we are willing to interpret a positive relationship between the concentration of non-Iberians and positive economic impacts in the municipalities of our study region as evidence in favor of the persistence of historical human capital shocks operating via mobile factors rather than fixed ones.

The second contribution of our work is to add to the growing body of empirical evidence on the long-term impacts of the Age of Mass Migration. Hatton and Ward (2018) note that most work on the consequences of the Age of Mass Migration focus on the US, the destination of about three fifths of the estimated 50 million Europeans who emigrated between 1850 and 1940. less is known about the consequences of this historical episode in other parts of the New World, like Latin America, which received approximately 13 million European migrants between 1870 and 1930 (Sánchez-Alonso, 2007).[9] Notable exceptions to this lack of evidence in Latin America are the works mentioned above. Our paper brings new evidence to this literature by showing the impacts of the Age of Mass Migration spreading across a receiving country different than the US.

Our third contribution is to show how historical immigration and subsequent internal migration contributed to shaping the expansion of the agricultural frontier in Brazil. Following the adaption of soybean varieties to tropical climates, migrants from the southern parts of Brazil (many of them descendants of historical settlers) started to settle in the Center-West and the *Matopiba* regions.[10] This process triggered the expansion of the agricultural frontier, transforming previously unproductive lands into one of the breadbaskets of the world (Bragança et al., 2015). Different from the rural-to-urban migration commonly observed in Brazil, this internal migration flow was mainly rural-to-rural and brought in workers with high levels of physical and human capital (Rezende, 2002). Understanding how this type of rural internal migration has combined with more traditional flows is central to understanding the development of the agricultural frontier in Brazil.

The rest of the paper is organized as follows. Section 2 provides background information on immigration, ancestries, and surnames in Brazil, followed by details on the surname-based classification of ancestries, and a brief account of the expansion of the agricultural frontier in the country. This section justifies both the use of the ancestry classification as a proxy for descendants of historical settlers and the choice of the agricultural frontier as the region of study. Section 3 discusses the data used in the study and the sample restrictions. It also shows descriptive statistics that motivate our empirical strategy and our discussion of results. Section 4 presents our empirical strategy, discusses the identification concerns, and proposes an instrument to address these concerns. Section 5 presents and discusses the main results of the study, shows some extensions of these results, and discusses the possible channels for the effects found in our analyses. The sections also presents and discusses several robustness checks. Section 6 concludes.

2 Background information

2.1 Immigration, surnames, and ancestries in Brazil

The colonial ties of Brazil to Portugal and the proximity of the country to several Spanish colonies in South America resulted in a regular flow of colonizers and settlers coming from the Iberian Peninsula. This process gave most of the Brazilian population Portuguese and/or Spanish ancestries (for the purposes of this study, we define "ancestry" as the country of origin of one's ancestors). As a result, most Brazilians bear Iberian surnames. At the same time, Brazil's historical, and many times forceful, integration of former slaves and Amerindians into its national population, left the descendants of those groups with Iberian surnames as well. As a result, not only

[7] In particular, we advance the literature on the link between historical events and persistent economic impacts in Brazil (de Carvalho Filho and Monasterio, 2012; Naritomi et al., 2012; Rocha et al., 2017; Ehrl and Monasterio, 2017) and in other parts of Latin America (Dell, 2010; Droller, 2017; Valencia Caicedo, 2018).

[8] Droller (2017) investigates the spread of Europeans in the Argentinian pampas but his analysis focuses largely on first-wave migrants. His study does not consider Europeans and their descendants moving to other parts of the country and how they might have affected these parts. Ehrl and Monasterio (2017), on the other hand, look specifically at the descendants of historical settlers in Brazil. However, by restricting their analysis to municipalities near the original injection points the authors cannot distinguish between fixed or mobile factors driving the observed persistence.

[9] In Brazil, the European immigration was followed by a large influx of Japanese immigrants between 1900 and 1940 (IBGE, 2007).

[10] The *Matopiba* is a region shared by four northern Brazilian states (Maranhão, Tocantins, Piauí, and Bahia), covered by the Cerrado biome, where the agricultural frontier has expanded more recently (Bragança, 2018).

Brazilian whites but also the vast majority of Brazilian blacks, mixed, and native Brazilians (Amerindians) have Portuguese and/or Spanish surnames.

In the late 19th and early 20th centuries, state-sponsored settlements attracted a large number of immigrants to the South and Southeast regions of Brazil.[11] Immigrants came first from Germany, then from Italy, and finally from Japan. Smaller groups of immigrants came also from Syria, Lebanon, Turkey, Poland, and other countries. Immigration from Portugal and Spain, which happened throughout the history of Brazil, continued during that period. This intense immigration between 1850 to 1940 was encouraged by the Brazilian government in the belief that bringing in Europeans and other foreign settlers was an efficient way to develop the interior of the country and to replace the slave labor force after slavery was abolished in Brazil.[12] After World War II, however, international immigration to Brazil declined sharply. The absence of new substantial inflows and the natural aging of the immigrant population combined to make the current share of foreign-born people in Brazil negligible (around 0.23%, according to the last national census in 2010).

Brazil's historical background, combined with this intense (but later interrupted) experience of international immigration in its post-slavery period, generated a unique landscape of surnames and ancestries in the country. Because the fraction of foreign-born in the country is close to zero and because most Brazilians have Iberian surnames, a person bearing a non-Iberian surname in Brazil in recent decades has a high probability of having descended from immigrants who arrived in the country between 1850 and 1940. Therefore, the surname-based classification of ancestries employed in our analysis serves well as a proxy for the presence of descendants of historical settlers in the current populations.

Since our classification of ancestries is based on surnames, a brief discussion of the social and legal norms surrounding surnames in Brazil follows. Children in Brazil usually receive two surnames: first the mother's second surname, followed by the father's second surname. Because only the second surname of each parent is passed on, and because the father's surname comes last, effectively, only the father's surname survives. As for name changes after marriage, Brazilian civil law required a married woman to adopt her husband's second surname until 1977. After that, adoption became optional, and in 2002, adoption of the spouse's surname became optional for both men and women. In most cases, when adopted, the husband's surname becomes the second.[13]

In our study, we consider only the second surname of each person. Therefore, our way of tracking the offspring of historical settlers effectively ignores maternal lineage. The algorithm used in the classification allows for considering more than one surname and for more refined classifications of mixed ancestries (when each surname comes from a different group) or homogeneous ancestries (when both surnames are from the same group).[14] However, we do not expect these more refined classifications to improve the approximation of the concentration of non-Iberian ancestries in our study region for two reasons. First, the group of non-Iberians is large enough to accommodate many cases of mixed ancestries (e.g., German-Italian). Second, we expect neither a consistent pattern in the order of Iberian and non-Iberian surnames when a person has both, nor a correlation between this order and the concentration of non-Iberians in a given municipality. Therefore, the measurement error arising from assigning Iberian or non-Iberian ancestry to workers with mixed surnames is probably classical in our setting (any measurement error arising from the choice of using only the second surname in our classification will bias our results towards zero, not upward or downward).

2.2 Surname-based classification of ancestries

Here we describe briefly the surname-based classification of ancestries used in this study. Refer to Monasterio (2017) for a thorough explanation of the algorithm and the data requirements, and refer to Lopes et al. (2017) and Ehrl and Monasterio (2017) for a description of the updated versions of the algorithm, which are similar to the one used in this study. The classification follows four steps: (1) collect the second surname of all workers in the sample in a given municipality for a given year; (2) match these surnames to historical sources where surnames are accompanied by countries of origin; (3) link each unique surname to a country of origin (e.g., Italy) or region of origin (e.g., Eastern Europe); (4) attribute ancestry of the historical source to current observations based on this surname-origin matching process; and (5) refine classification of workers with Iberian surnames according to their race.[15]

The classification yields, for each paired municipality and year in our sample, the total number of workers

[11] Informative discussions of the causes, context, and consequences of the state-sponsored immigration in Brazil can be found in de Carvalho Filho and Monasterio (2012), Ehrl and Monasterio (2017), and Rocha et al. (2017).

[12] The abolition of slavery in Brazil was a gradual process started in 1850 and finalized only in 1888.

[13] Some people in Brazil have three or more surnames. Nevertheless, here we use the term "second" to refer to the surname that comes last in a person's full name to avoid confusion with the term "last name", commonly used in English to denote one's surname.

[14] Such refined classification is used, for example, by Lopes et al. (2017), who have detailed information on an individual's parents. In contrast, we do not have information on parents. Therefore, we cannot appropriately trace maternal and paternal lineages, nor can we correctly identify cases of homogenous or heterogenous ancestries.

[15] IBGE, the Brazilian Statistical Office, uses the term "color/race", usually divided into five categories: black, white, mixed, yellow (Asian), and indigenous. The "yellow" category is seldom chosen by Asian-Brazilians, who often choose the mixed category instead. In this study, we use the term "race" as a way to follow the standard in literature.

of each ancestry. We group all ancestries into two large groups: Iberians (IBR) and non-Iberians (NIB).[16] To obtain the concentration of non-Iberians in the population of the municipalities m in our sample in each year t, we simply divide the number of workers of non-Iberian ancestry by the total number of workers in the sample for that municipality-year pair. The resulting measure is then multiplied by 100 to facilitate interpretation of coefficients in descriptive statistics and regressions.

$$CNI_{mt} = \frac{\#NIB}{\#NIB + \#IBR} \times 100 \tag{1}$$

where, $\#NIB$ and $\#IBR$ are, respectively, the total number of workers classified as non-Iberians or Iberian based on their second surname.

We only use information on race to refine our ancestry classification in some cases (for example, Native Brazilian surnames sometimes are erroneously classified as Japanese). We do not use race to define ancestries nor do we aim to investigate racial wage disparities in Brazil. Gerard et al. (2018) does such an investigation, using the same data we use in this study (RAIS), while many works in the sociology literature, like Andrews (1991) and dos Santos (2002), discuss the connection between racial wage differences and historical immigration in post-slavery Brazil. We acknowledge, however, that the share of whites is higher in the group of non-Iberians: approximately 63% of non-Iberians in our sample in 2010 identify as whites, while only 42% of Iberians do so (the share of whites in the total population was approximately 43% in 2010, according to census data). Therefore, in most of our analyses we control for race, in addition to other relevant controls discussed in detail in Section 4.1.

2.3 The expansion of the agricultural frontier

We close this section by presenting a brief account of the expansion of the agricultural frontier in Brazil and justifying its choice as the region of study.

Following the adaption of soybean varieties to tropical climates, migrants from the South and Southeast of Brazil started to settle in the Center-West around 1960-1970 and in parts of the North and Northeast after 1990, triggering the expansion of the agricultural frontier in the country. This expansion of agricultural production had implications that go beyond the development of the agricultural sector in Brazil. The development of the frontier integrated local markets, spread modern agricultural technologies, induced migration, and changed the land use and the economic structure of the region and the country (Bragança et al., 2015; Bustos et al., 2016, 2017).

The low population density and the abundance of (mostly flat) farmland in the Brazilian Cerrado, the savannah-like biome that dominates the agricultural frontier, combined to result in low land prices that attracted farmers from the South and Southeast regions of the country (Rezende, 2002). This process was further stimulated by private colonization companies, farmers cooperatives, land reform initiatives, and rural development programs implemented by the national government (Jepson, 2006*a*,*b*; Hosono and Hongo, 2012). The frontier continued to expand and to develop in the more recent decades, stimulated by the arrival of new technologies that impact not only agricultural production, but also labor markets and internal migration (Bustos et al., 2016; Bragança, 2018).

Because internal migrants often came to the frontier from the same regions that had received non-Iberian settlers in the late 19th and early 20th centuries, many of them were direct descendants of those historical settlers (Alves, 2005). Plenty of anecdotal accounts (Wagner and Bernardi, 1995; Santos, 2008) and the high incidence of non-Iberian surnames in a region so distant from the original injection points (Monasterio, 2017) suggest that this was, indeed, the case. Farmers who settled on the agricultural frontier many times came from former non-Iberian colonies in the South and the Southeast, where the extant tradition in soybean cultivation and association in cooperatives matched the definition of modern agriculture desired by the Brazilian government for the region at the time (Hosono and Hongo, 2012; Alves, 2016).[17]

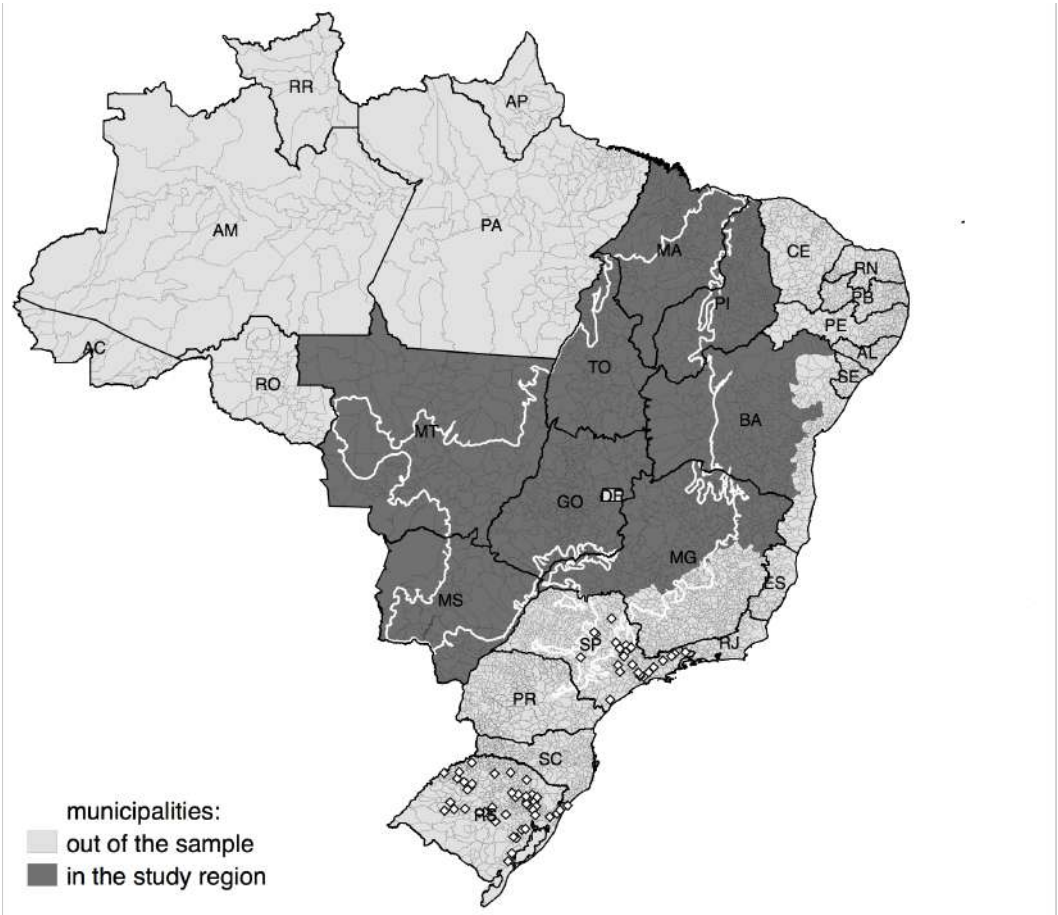
The agricultural frontier of Brazil, thus, provides an almost ideal setting for investigating the "fixed versus mobile" question we have posed in the introduction of this paper. Municipalities in the agricultural frontier are located far enough from the original injection points of historical non-Iberian immigration to be untouched by fixed factors linked to the human capital shock caused by the historical settlers (like a head start in infrastructure, local institutions, and redistribution of land). The region was also untouched by significant international immigration, current or past, so any effects of the concentration of non-Iberians today can be reasonably linked to the spreading of the impacts of non-Iberian immigration in the southern regions of Brazil. Finally, municipalities in this region received significant internal migration from places around the historical non-Iberian settlements, thus receiving a large number of descendants who may have carried the human capital of their forebears. The choice of the agricultural frontier as the study region is also convenient for our identification strategy (discussed in more detail in Section 4.2), which uses both the distance to injection points and the ruggedness of the terrain to instrument for the concentration of non-Iberians.

[16] The five most common ancestries in the non-Iberians group are: Italians, Germans, Japanese, Eastern-Europeans, and Syrian-Lebanese.

[17] In fact, Alves (2016) observes that migrants "sold their smallholdings, bought larger areas, and settled them using modern agricultural techniques" and concludes that the Brazilian Cerrado is a "typical case of agricultural development promoted by farmers from more advanced agronomic culture."

The terms "Cerrados" and "Agricultural Frontier", as well as the territory designated by them in Brazil, overlap significantly. In our study, we combine the area covered by the Cerrado biome in the country with the population density in 1950 (after international immigration has plummeted in Brazil but before the expansion of the agricultural frontier) to define our study region. We exclude from the study region the states in the Amazon region and the Cerrado parts of state of the São Paulo (SP), which has a significant number of injections points of historical non-Iberian immigration. The population density criterion excludes the eastern portion of Bahia (BA) and the southeastern portion of Minas Gerais (MG), which, similar to the coastal region and the southern states of Brazil, have been historically more populated and developed than the inner country. The resulting study region, shown in Figure 1 includes all municipalities in the states of Mato Grosso do Sul (MS), Mato Grosso (MT), Goiás (GO), Tocantins (TO), Maranhão (MA), and Piauí (PI), as well as parts of the states of Bahia and Minas Gerais. All state capitals and the Federal District (DF) are excluded from the study region.

Figure 1: Study region, the Cerrado biome, and injection points of historical non-Iberian immigration in Brazil (main sample)



Notes: Black lines denote state boundaries, gray lines denote municipality boundaries, and the white line denotes the area where Cerrado is the dominant biome. The white diamonds in the states of São Paulo (SP) and Rio Grande do Sul (RS) represent the approximate location of historical non-Iberian settlements (locations obtained from Rocha et al. (2017) and de Carvalho Filho and Monasterio (2012), respectively).

The black lines in the map denote state boundaries, while the thinner gray lines denote municipality boundaries, and the thick white line denotes the area where Cerrado is the dominant biome. The white diamonds in the states of São Paulo and Rio Grande do Sul (RS) represent the approximate location of historical non-Iberian settlements. Both the population density and the Cerrado biome criteria for sample selection are applied at the "mesoregion" level, which can be understood as state regions. In the appendix[18], we present alternative samples, which follow state level or microregion (county) level criteria.[19] Maps of population density in 1890, 1950, and 2010 in Brazil, as well as a map of the share of municipality area covered by the Cerrado biome, are also presented in the appendix. Later in the paper, we use these alternative samples to repeat some of the main results of the paper.

[18] Available upon request.

[19] We refer to these samples as the "full states" and the "Cerrado" samples, respectively, because the first contains all the municipalities in each state, while the second overlaps almost perfectly with the Cerrado biome boundaries.

3 Data and descriptive statistics

3.1 Data sources

The main data source used in this study is the *Relação Anual de Informações Sociais* – RAIS, a report of all labor contracts that employers are required to file every December in order to comply with labor regulations. These reports form a database used by the Brazilian government to administer unemployment benefits and allowances for low-income workers and to produce statistics on the formal sector. This makes RAIS a high-quality annual census of all formally employed workers in Brazil.[20] Stacked over the years, RAIS becomes a matched employer-employee dataset, a type of data is becoming increasingly popular in economic studies (Card et al., 2018).

RAIS informs basic demographic characteristics of employees, their remuneration, and some characteristics of their jobs. The data also informs characteristics of the employers like industry, size (number of employees), and the municipality in which the firm is located. Using the municipality of the employer and the link they have with workers, we are able to assign workers to municipalities. Most important for our study, however, is the fact that RAIS informs the full name of workers, from which we extract the second surname to be used in our surname-based classification of ancestries. Information on race is also used to improve the classification as discussed in section 2.2.

We complement our data with multiple sources. Socioeconomic characteristics and some geographic information like location (used to calculate the distance to state capitals and to historical non-Iberian settlements, for example) come from Ipeadata and mostly use information from the 2010 population census. The approximate location and the year of establishment of the historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul come from Rocha et al. (2017) and de Carvalho Filho and Monasterio (2012), respectively. Rainfall variables are constructed using data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). Information on temperature used to complement the climate controls come from the dataset used in Rocha and Soares (2015). Information on elevation and the Terrain Ruggedness Index are calculated using data from the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) from the US Geological Survey. Dummies for soil types use data from Embrapa Solos, while the dummies and shares for biomes use information from the MapBiomas project. Potential yields for soybean and maize come from the dataset used in Bustos et al. (2016), who construct such variables using the data and the methodology in the Food and Agriculture Organization (FAO) Global Agro-Ecological Zones(GAEZ) database (henceforth FAO/GAEZ).

One concern with the use of RAIS data to classify surnames is that it covers only a portion of the population in a given municipality-year, only the formally employed. In a country like Brazil, where informality is high, this portion can be particularly small and not representative of the population of the municipalities in our sample. To amend this deficiency, we resort to a couple of other data sources that also inform respondents' names in Brazil: the *Cadastro Único* (CadUnico), the unified registry of beneficiaries of the Brazilian cash transfer program (*Bolsa Família*), and the *Base Sócios*, a record of business owners maintained by the national tax authority. For at least one year (2010), we can compare the concentration of non-Iberians obtained using only RAIS or using these three data sources combined, which cover a larger portion of the population in the municipalities we study.

The correlation between the two measures in 2010 is 0.90. Inspection of the data suggests some inconsistencies in the municipality codes used in the different datasets, possibly coming from adjustments for municipality boundaries changing over time. Therefore, we perform a second comparison, aggregating the data at the microregion level, a higher administrative level that is consistent over the years. The correlation rises to 0.98. In figure 2 we plot the concentration of non-Iberians using only RAIS data in the horizontal axis, the same concentration using the extended data sources in the vertical axis, and the 45° line. The results of this exercise reassure us that the concentration of non-Iberians we calculate using only surnames of formally employed workers is a good representation of the true concentration of non-Iberians in the population of the municipalities in our study region.

3.2 Sample selection

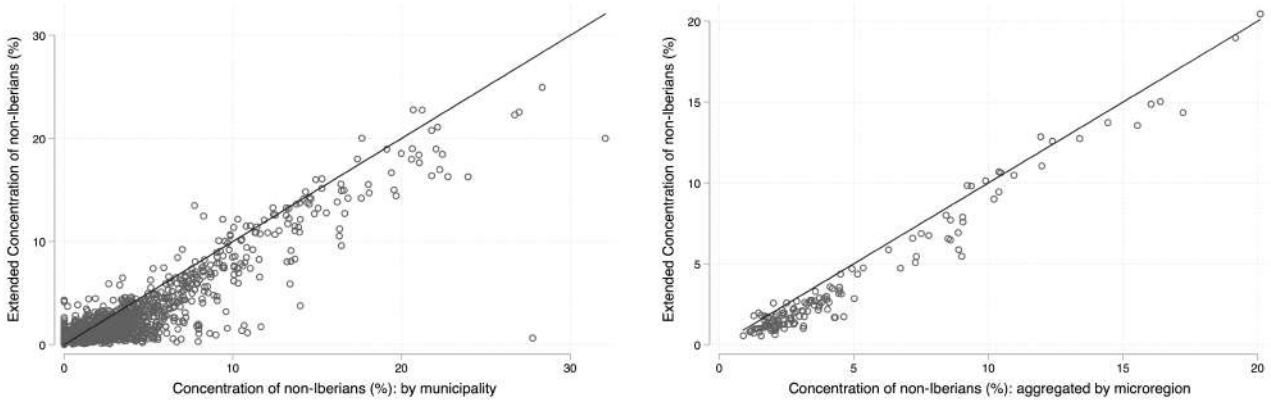
RAIS starts in 1986 but workers' names are only available in the dataset starting in 2004. Therefore, 2004 is the initial period of our data. The final period of our study is 2016, currently the last year for which the data is available.

After restricting our sample to the municipalities in our study sample, we make some basic restrictions to the microdata. In every year, we exclude all foreign-born, all public servants, and the military.[21] We keep only workers between 16 and 70 years old who have a valid identifier (PIS number). We also exclude those who work less than 10 hours per month and those without a positive remuneration. After calculating hourly earnings, we drop those workers with hourly earnings above the 99.5th percentile in each year. We also drop workers with more than six labor links per year to avoid the inclusion of possible misreporting in the data.

[20] For examples of papers using RAIS see Dix-Carneiro and Kovak (2017) and Gerard et al. (2018). For detailed information on RAIS data, its variables and structure, refer to the data appendix of these papers.

[21] The exclusion of public servants and the military is done because many of those workers are registered at the capital of the state regardless of their actual workplace. Moreover, their remuneration follows legally established norms, being less likely to reflect their labor productivity.

Figure 2: Comparison of the concentration of non-Iberians using only RAIS or extended sources: muncipality level (left) and microregion level (right)



Notes: The concetration of non-Iberians in the horizontal axis uses only data from RAIS, while the one in the vertical axis includes also information from the *Cadastro Único* and the *Base Sócios* datasets. The graphs also include a 45° line.

In most of our analyses, we use only data from 2010, a census year for which several economic indicators are available at the municipality level. In complementary analyses, we use the whole 2004–2016 period as stacked cross-sections. In 2010, there are 1,556 municipalities and 4,485,142 individual observations in our main sample. When using data from 2004–2016, we have to adjust the municipality borders to accommodate changes over the period, thus using the municipality boundaries and codes of 2004 as our main unit. In this case, we end up with 13 years of data, 1,552 municipalities, and over 56 million individual observations.

The map in Figure 3 below shows the distribution of non-Iberians in the study region. We notice a higher concentration in the Center-West region, particularly in the state of Mato Grosso, a state that symbolizes the expansion and modernization of the agricultural frontier in Brazil. Somewhat less evident, is the fact that municipalities with a stronger presence of non-Iberians tend to be larger in area, This fact gives the impression that the distribution of non-Iberians across the municipalities in our sample is well balanced. It is not. The smallest municipalities, more frequent in our data, generally have fewer non-Iberians in their populations. The distribution of our measure for the concentration of non-Iberians is skewed to the left, with an average of 3.78%, median of 2.61% and standard deviation of 4.03% in 2010. The range, as shown in the map legend, goes from 0% to a maximum of approximately 32%. The histogram for the distribution in 2010 is available in the appendix[22].

The distribution of non-Iberians in our study region, with stronger presence in the larger municipalities to the west of the country and away from the sites of historical settlements, corroborates the story we tell in section 2.3: the descendants of non-Iberian immigrants in Brazil spread over the country following the expansion of the agricultural frontier. They moved to previously underpopulated places and away from the injection points of non-Iberian immigration in the South and Southeast regions of Brazil.

3.3 Descriptive statistics

The median of the Concentration of Non-Iberians in 2010 is used to split the sample into two parts in Table 1, which presents the mean value of some indicators at the municipality level for the full sample and for the municipalities above and below the median.

The numbers in the table indicate that, in fact, municipalities where the presence of descendants of historical settlers is stronger exhibit better economic indicators. Income, measured in different ways, is approximately 60%higher in places where the concentration of non-Iberians is above the median. Other measures of economic development—such as the Human Development Index, the poverty rate, and the Theil-L inequality index—are also better in places above the median (9.16% higher, 41.51% lower, and 7.80% lower, respectively). The municipalities above the median concentration of non-Iberians also show better indicators in education. These places have more persons who completed high school (20.31% more) or with a college degree (35.76% more), and significantly less illiterate adults (33.08% less).

Labor indicators also differ for the group of municipalities below and above the median concentration of non-Iberians. We first notice that unemployment is lower in municipalities above the median (-13.64% less). The degree of formalization, on the other hand, is higher (26.72%). This higher degree of formalization reflects on the more detailed indicators of occupation: municipalities above the median have a higher share of employees with a formal labor contract and more public servants. They also have a smaller share of employees with an informal labor contract and less self-employed. The number of employers in places above the median is also higher (67.36% more).

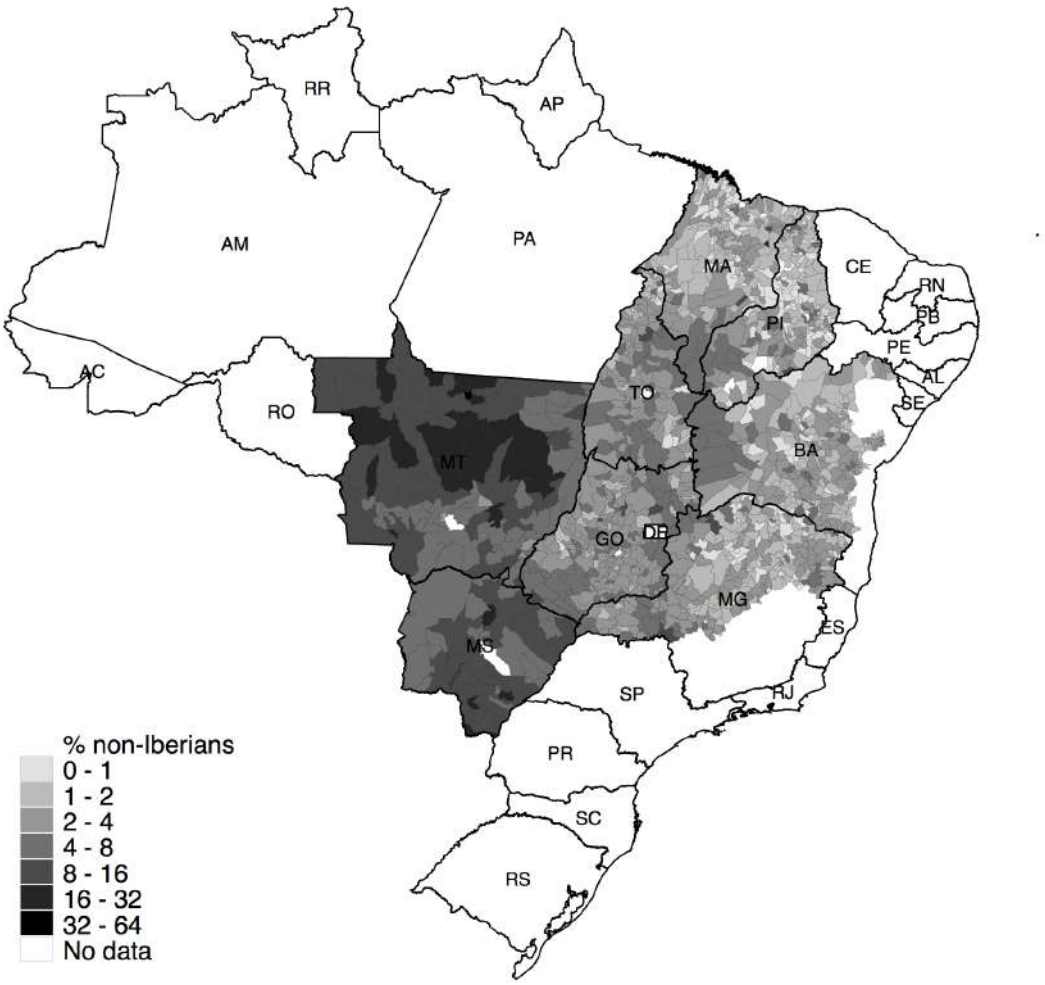
[22] Available upon request.

Table 1: Difference in means for municipalities above and below the median Concentration of Non-Iberians in 2010: Economic Indicators from Census data

Variables	All	Above	Below	(A - B)	Diff./B
Income					
average earnings of employed persons (R$)	683.51	845.19	521.83	323.36	61.97%
income per capita (R$)	395.18	484.71	305.64	179.07	58.59%
income per capita, only positive earnings (R$)	402.47	492.54	312.40	180.14	57.66%
Poverty and Inequality					
% of poor	30.54	22.54	38.54	-16.00	-41.51%
Human Development Index	0.63	0.66	0.60	0.06	9.16%
Theil-L index	0.50	0.48	0.52	-0.04	-7.80%
Education					
% of employed with a middle school degree	42.69	46.20	39.18	7.01	17.90%
% of employed with a high school degree	27.69	30.25	25.14	5.11	20.31%
% of employed with a college degree	6.43	7.41	5.46	1.95	35.76%
illiteracy rate (adults)	21.79	17.47	26.11	-8.64	-33.08%
Labor					
unemployment rate	6.97	6.46	7.48	-1.02	-13.64%
degree of formalization	34.94	41.55	28.32	13.23	46.72%
% of employees with a formal contract	22.51	27.83	17.20	10.64	61.86%
% of employees without a formal contract	29.78	28.53	31.04	-2.51	-8.09%
% of public servants	8.40	9.13	7.68	1.45	18.89%
% of self-employed	23.00	22.21	23.80	-1.59	-6.70%
% of employers	1.01	1.27	0.76	0.51	67.36%
Industry					
% employed in agriculture	39.82	34.98	44.66	-9.69	-21.69%
% employed in manufacturing	5.94	7.16	4.72	2.44	51.73%
% employed in commerce	10.02	10.59	9.44	1.15	12.21%
% employed in services	32.48	34.60	30.36	4.24	13.98%
Population and migration					
% female	48.97	48.68	49.27	-0.59	-1.20%
% rural	60.33	66.57	54.10	12.47	23.06%
% adult (18+ years)	66.34	67.56	65.13	2.43	3.73%
% working population (PEA)	40.56	43.07	38.04	5.03	13.22%
total population in 2010 (residents)	19,070	19,220	18,920	300	1.59%
% of current inter-state migrant	16.95	24.96	8.94	16.01	179.10%

Notes: All indicators use data from the 2010 Brazilian population census. The number of municipalities in the sample for 2010 is 1,556 (state capitals are excluded). The median value of the Concentration of Non-Iberians used to split the sample is 2.61% (mean 3.78%, standard deviation 4.03%). Income variables expressed in reais (R$) of August/2010. The degree of formalization is given by the share of workers that contribute to the social security system (the formally employed, all public servants and military, and some of the self-employed). The share of inter-state migrants is given by the number of adults (age 16–80) born out-of-state divided by the total number of adults in the municipality. All differences are statically different than zero ($p < 0.01$), except total population in 2010 ($p = 0.8716$).

Figure 3: Concentration of non-Iberians in the study region, 2010 (%)



To the extent that formal labor and higher education comes with better salaries, the pattern of the differences in means so far is consistent with the municipalities above the median concentration of non-Iberians being generally more developed than those below the median.

We look further at the distribution of workers among economic sectors and find that the municipalities above the median have a smaller share of workers in the agricultural sector and more people working in manufacturing, services, and commerce. We notice that this is not inconsistent with the story that non-Iberians followed the expansion of the agricultural frontier in Brazil because this expansion also meant modernization and modern agriculture often means less labor-intensive agriculture.[23]. Also, the proportion of the population living in rural areas is higher in the group of municipalities above the median concentration of non-Iberians.

For the remainder indicators of population and migration, we note that there are no significant differences in total population and in the proportion of females and adults in the populations. On the other hand, the municipalities above the median have a larger working population (13.22%) and a much higher rate of inter-state migration. The proportion of adults born in another state in the municipalities above the median concentration of non-Iberians is almost three times higher. Again, this is consonant with the notion that non-Iberians, among other internal migrants, moved to the municipalities in our study sample following the expansion of the agricultural frontier of Brazil.

We will return to the discussion of these differences while discussing the results of our empirical analysis in section 5.3. For now, we discuss a concern arising from the fact that municipalities above the median value for the concentration of non-Iberians have a significantly higher percentage of workers with a formal labor contract. Because both our wage regressions and our measure of the concentration of non-Iberians consider only data from the formally employed this difference suggests we could be measuring more non-Iberians in local economies with more formally employed people, which is often correlated with better salaries. However, the fact that our measure for the concentration of non-Iberians using sources other than the formal sector is highly correlated with the main measure reduces concerns that the observed results are spurious. We also note that higher formalization can be

[23]In fact, our story and our numbers are in line with the results of Bustos et al. (2016). This study finds that the introduction of labor-saving agricultural technologies in Brazil, particularly in the municipalities of our study region, reduced employment in agriculture and increased employment in other sectors

itself an impact of the presence of descendants of non-Iberian settlers and a likely channel for their impact on earnings.

In the appendix[24], we present two other tables of differences in means. The first focuses on the mean value of characteristics of the municipalities used as controls in our regressions (discussed in the next section). The second focuses on indicators constructed using individual-level data from RAIS, aggregate at the municipality level. Among other interesting findings, we notice that earnings in the formal sector are also higher in the group of municipalities above the median.

The average log-earnings at any given municipality can be influenced by a series of factors unrelated to the presence of non-Iberians, like the composition of workers and the type of firms. For example, non-Iberians may concentrate in places where people are more educated on average or where there are more high-paying industries. In that case, there would be a positive association between earnings and the concentration of non-Iberians that does not depend on any long-lasting effects of a human capital shock spread by mobile factor, as we suggest in this study. Therefore, we want to control for any characteristic of the municipalities, the firms, and the workers that can also affect earnings in our analysis. At the same time, we want these controls to be independent of the concentration of non-Iberians (exogenous). The empirical strategy presented in the next section investigates further the differences in means discussed here focusing on labor earnings and exploring the richness of our individual-level data.

4 Empirical framework

4.1 Estimation strategy

Our main explanatory variable is the concentration of non-Iberian surnames in a given municipality. We want to know whether that concentration affects the earnings of workers from all ancestries, combined or separately, in that given municipality. To do so, we regress the log-earnings of all workers in our sample on a set of individual-level controls, industry dummies and firm characteristics in addition to the concentration of non-Iberians, geographic and climate controls at the municipality-level, and year dummies (the latter, only in analyses that use the 2004–2016 period). Our main regression specification is:

$$y_{jimst} = X'_{jt}\Gamma + \psi_i + \phi_s + \delta_t + \beta CNI_{mt} + W'_m\Lambda + G'_m\Pi + \varepsilon_{jimst} \tag{2}$$

where y_{jimst} is the log-earnings of worker j in industry i in municipality m of state s in year t; X_{jt} is a vector of worker characteristics (age, firm tenure, and their squares; dummies for female and for categories of race, education, and firm size); ψ_i are industry fixed effects; ϕ_s are state fixed effects; δ_t are year dummies; CNI_{mt} is the concentration of workers with non-Iberian surnames; W_m is a vector of climate controls, G_m is a vector of geographic controls; and ε_{jimst} is an idiosyncratic error term. In all regressions, we cluster standard errors by municipality.

The rationale behind the empirical strategy represented in equation (2) is that we are removing from the log-earnings in each municipality the portion that is influenced by the workforce composition and the distribution of firms. Thus, only the part that is common to all workers in that municipality-year is left to be explained by the concentration of non-Iberians and other controls at the municipality level.

The vector of climate controls W_m includes historical average for rainfall and temperature, the standard deviation of rainfall, temperature range, and maximum temperature, computed with information from 1981 to 2010. The vector of geographic controls G_m includes information on average elevation, the share of municipality area covered by the Cerrado biome, distance to the state capital, and dummies for soil types. In some regressions, this vector is expanded to include also dummies for biomes, population density in 1950, and potential yields for soybean and maize (we include both the potential yields under low technology and the difference in potential yields when switching from low to high technology).

The set of controls used in our regressions was carefully chosen to mitigate concerns of omitted variable bias. We use relevant climate, geographic, and socioeconomic characteristics that could have been affecting the outcome of interest and biasing the estimate for the effect of the concentration of non-Iberians on earnings. In particular, the inclusion of state fixed effects is important to control for unobserved characteristics that vary at the state level. Ideally, we would like to use municipality fixed effects for this end but their inclusion would exhaust the variation of our main explanatory variable (the concentration of non-Iberians) in a cross-sectional analysis.

The majority of our analysis uses only the year of 2010. This is done both for computational advantages and because this year has more information on other outcomes of interest at the municipality level, obtained from census data available only for 2010. We repeat the main results for the whole period of 2004–2016).

In the analysis for 2004–2016 we could include municipality fixed effects and explore a pseudo-panel of municipalities. However, in this case, only the variation of the concentration of non-Iberians within each municipality over the 13 years in our data would be used to identify the coefficient of interest. This variation happens to be

[24] Available upon request.

very small when compared to the one we have in the cross-section. Moreover, we could not find a satisfactory instrument for the variation of the concentration of non-Iberians in each municipality over the recent period. For these reasons (low variation of the main explanatory variable and lack of a good instrument for it) we focus on cross-sectional analyses in this study.

We close this section with a note on weights. Although our data is collected and used at the individual level, the phenomenon we are studying and the main explanatory variable we use in this investigation are built at the municipality level. We look at how the concentration of non-Iberians impacts earnings for all workers in the municipalities. Therefore, our empirical analysis must give each municipality in our data the same weight. We do that by weighting every individual observation in our regressions by the inverse of the number of observations in each municipality-year.

Following the recommendations in Solon et al. (2015), we present also the results obtained using two different weights in our robustness checks: none and municipality area. The latter is commonly used in studies like ours, in which a significant number of variables is calculated at the municipality level. Using no weights, in practice, means we are placing more weight on the municipalities that have a higher number of individual observations.

4.2 Identification strategy

The coefficient of interest β in equation (2) is identified out of the variation of the concentration of non-Iberians CNI_{mt} between municipalities and within a given municipality over time. Because most of the variation comes from the variation between municipalities, we add an extensive set of controls in our regressions to reduce concerns that observed results are due to unobserved (or uncontrolled for) characteristics of the municipalities and not from differences in the concentration of non-Iberians.

However, despite all these controls helping to mitigate concerns of omitted variable bias, endogeneity concerns remain. Non-Iberians can concentrate in places where they expect to fare better and thus have a positive impact on the local economy that would be captured by an increase in local earnings. Thus, we propose the use of an instrument to deal with the possible endogeneity of the concentration of non-Iberians.

Our instrument uses the interaction of two terms. The first is the average distance from each municipality to all the historical injection points in the states of Rio Grande do Sul and São Paulo.[25] We calculate a single measure, recording the average of the distance of each municipality to all historical sites of non-Iberian immigration in each of these states and averaging the results. Following the simple logic that distance to injection points increases migration costs of descendants to the agricultural frontier, we expect this measure to be negatively correlated with the concentration of non-Iberians.

The second term is the Terrain Ruggedness Index (TRI), a measure that serves as a proxy for how suitable a given municipality is for modern agriculture. Modern agriculture, particularly when focused on grain production, requires larger plots for mechanization and gains of scale to become viable. Anecdotal evidence and historical accounts (see section 2.3) establish the link between internal migration of the descendants for historical non-Iberian settlers and modern agriculture in Brazil. Flatland, the opposite of rugged terrain, is a characteristic that attracted those descendants to particular places in the agricultural frontier. The migration of those engaged in agriculture in the first moment attracted non-Iberians working in different sectors later. Thus, we expect the average ruggedness of the terrain in the municipalities to be negatively correlated with the concentration of non-Iberians. In other words, less rugged terrain can benefit everyone but especially those potential migrants who have a comparative advantage in modern agriculture. Because many of those who moved from the vicinity of the injection points in southern Brazil to the agricultural frontier had a comparative advantage in modern agriculture we expect that this specific group of internal migrants would benefit the most from this favorable geographic characteristic.[26]

In summary, the rationale behind the instrument is that both the closeness to injection points and the suitability of the potential destinations for modern agriculture (represented by low levels of terrain ruggedness) worked as pulling factors for descendants of historical settlers contemplating a move to the agricultural frontier in Brazil. Closeness to an injection point incentivizes migration of non-Iberians to a given municipality by reducing migration costs and facilitating network formation. This pull, however, is weakened if the potential destination was highly rugged, making it difficult for descendants of non-Iberian historical settlers to explore their comparative advantage in modern agriculture. Conversely, a municipality on the agricultural frontier with high availability of flat farmland attracts non-Iberians with the expertise, disposition, and means to engage in modern agriculture. This effect, however, is weaker, if the municipality of destination is far from a historical injection point, which means that migration costs are higher.

As expected by the rationale discussed above, both terms used in our instrument are highly correlated with the concentration of non-Iberians. Both of them, are negatively correlated because we are measuring the distance

[25] Information on the location of these historical injection points comes from de Carvalho Filho and Monasterio (2012) (Rio Grande do Sul) and Rocha et al. (2017) (São Paulo).

[26] The comparative advantage of southern Brazilians, most of them from non-Iberian ancestry, in modern agriculture is a claim repeated by Rezende (2002), Alves (2016) and other authors who study the expansion of the agricultural frontier in Brazil. It also appears in anecdotal and historical accounts of this expansion (Wagner and Bernardi, 1995; Jepson, 2006*a*,*b*; Santos, 2008).

(opposite of closeness, or low migration costs) and the ruggedness (opposite of flatness, or suitability to modern agriculture). The interaction between the two terms is also correlated with the concentration of non-Iberians. In this case, the correlation is positive.[27] Again, this is consonant with the rationale discussed above, in which the negative effect of distance is attenuated (less important to potential non-Iberian movers) if the ruggedness is high or, conversely, in which the negative effect of ruggedness is felt less to movers if the potential destination is too far from the injection points making migration more costly.

If we interpret the negative correlation of the two terms as a first-order effect on the concentration of non-Iberians, the positive correlation of their interaction can be interpreted as a second order effect. This is central for our identification strategy because a valid instrument must not only have good predictive power but also satisfy the exclusion restriction. Both terms in our instrument, distance to injection points and terrain ruggedness, work well as predictors but are unlikely to satisfy the exclusion restriction. Distance to injection points is correlated with distance to the economic center of the country and with distance to ports in some cases. Terrain ruggedness, on the other hand, clearly influence the economic development of the municipalities via modern agriculture, which is correlated with but not exclusively accompanied by a high concentration of non-Iberians.

The exclusion restriction for the interaction of those two terms, however, is much more plausible. It only requires this second order effect to not affect the outcome of interest directly, only through the concentration of non-Iberians. To the extent that our regressions can include the non-interacted terms and use only their interactions as the instrument, we find this non-testable restriction to be more plausible. We cannot think of ways in which the interaction between the distance to injection points of non-Iberian settlements in the past and average terrain ruggedness of the municipalities affect earnings once both terms are being accounted for in our regressions.

Figure 4 gives a visual representation of the instrument. The figure shows four maps. The first map (top left) shows the concentration of non-Iberians in the study region in 2010, the measure we are instrumenting for. The main difference between this map and the one showed in figure 3 is the scale. We use ten quantiles instead of the fixed scale used before to provide a better comparison with the remaining maps in the figure. The second map (top right) shows the distance from each municipality to the injection points of historical non-Iberian immigration in the states of São Paulo and Rio Grande do Sul. We average the distance to all injection points and standardize it so the scale is presented in ten quantiles of standard deviations (because we interact the measures later, we do not center the means around zero).[28] The third map (bottom left) shows the normalized (or standardized) terrain ruggedness index for the municipalities in our sample. Finally, the fourth map (bottom right) shows the actual excluded instrument, the interaction between the non-centered normalized distance and ruggedness measures.

The negative correlation between terrain ruggedness and the concentration of non-Iberians is evident when we look at the two maps on the left. The negative correlation between distance and the concentration (two maps on the top) is less clear to the eye but can be verified in the data. The main takeaway from this figure, however, is that there exists enough variation in the data even within each state. This is important for our identification since most of our regressions include state fixed effects in all stages.

To implement our instrumental variables strategy, we adapt equation (2) using a two-stage least squares framework in which the first stage regresses the concentration of non-Iberians on the interaction of distance to injection points and terrain ruggedness. All specifications are adjusted to include the non-interacted terms ad additional controls.

5 Results and discussion

5.1 Main results

Table 2 below shows the results weighted least-squares (WLS) regressions for 2010. For concision, we present only the coefficients of the main explanatory variable. The regressions used to generate the results shown in columns 1 to 5 follow variations of equation (2). The dependent variable in all columns is the log of earnings in the formal sector. As discussed before, our regressions weight each observation by the inverse of the number of observations in each municipality thus given each municipality in our data the same relevance in our regressions.

The results show a positive association between the concentration of non-Iberians and earnings in the municipalities on the agricultural frontier of Brazil. In column 1, in which we use only individual-level controls and no state fixed-effects, an increase of one percentage point in the concentration of non-Iberians is associated with an increase of 1.57% in earnings on average (an increase of one standard deviation is associated with an increase of 6.41% in earnings on average). This association between the concentration of non-Iberians and average earnings decreases in magnitude but remains positive and significant when we move to specifications that gradually expand the set of controls used in the regressions. The coefficients drop significantly when state fixed effects are included, revealing the importance of this control. It also decreases with the inclusion of climate controls, which are particularly relevant for economies in the agricultural frontier. In the most conservative specifications in columns

[27] Refer to the results of first stage IV regression in section 5.1, table 3 for actual estimates of these correlations.

[28] We chose to normalize the IV terms to facilitate the interpretation of their coefficients later on in the regression results for the first stage. We do not center this normalization around zero because the negative and positive numbers resulting from this normalization would not preserve the variation of the interaction that we observe when interacting these terms in levels.

Figure 4: The concentration of non-Iberians, the terms used in the instrument (distance to injection points and Terrain Ruggedness Index), and their interaction (the actual instrument)

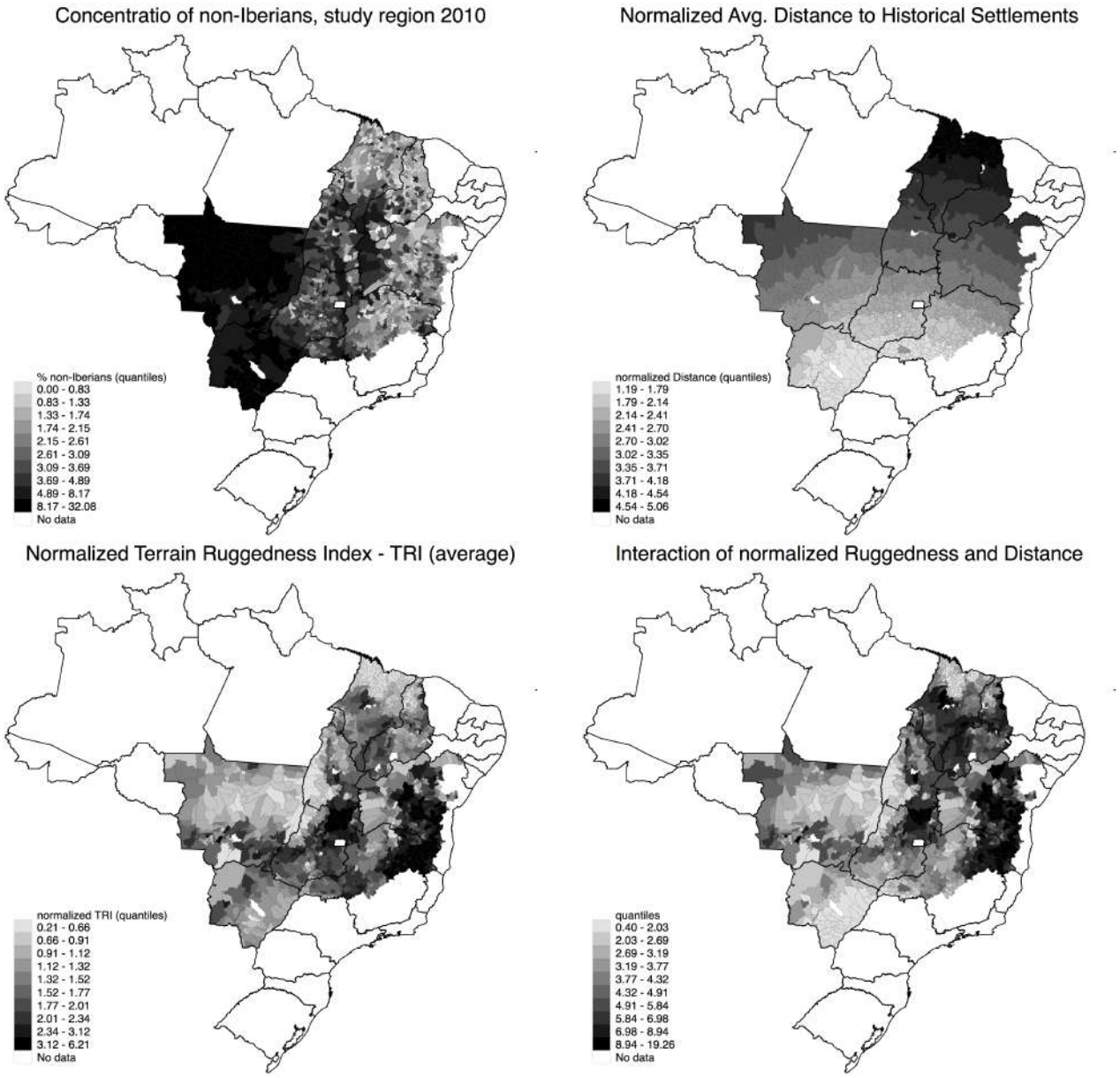


Table 2: WLS regressions of log earnings, 2010

	(1)	(2)	(3)	(4)	(5)
Concentration of non-Iberians (%) (mean = 3.80, s.d. = 4.08)	0.0157*** (0.0009)	0.0102*** (0.0015)	0.0083*** (0.0016)	0.0060*** (0.0018)	0.0060*** (0.0018)
N (workers)	4,485,142	4,485,142	4,485,142	4,485,142	4,485,142
Clusters (municipalities)	1,556	1,556	1,556	1,556	1,556
adj. R-sq	0.315	0.321	0.326	0.330	0.332
Controls					
Individual-level	Y	Y	Y	Y	Y
State FE		Y	Y	Y	Y
Climate			Y	Y	Y
Geographic				Y	Y
Geo Extra & Potential Yields					Y

Notes: The dependent variable is the log of hourly earnings in the formal sector. All regressions are weighted by the inverse of the number of individual observations by municipality. Individual-level controls: age, age squared, tenure, tenure squared, female, education categories, race/color, firm size categories, and industry dummies. Climate controls: historical average (1981–2010) for rainfall and temperature, the standard deviation of rainfall, temperature range, and maximum temperature. Geographic controls: average elevation, the share of municipality area covered by the cerrado biome, distance to the state capital, dummies for soil types. Potential yields controls: potential yields for soybean and maize under low technology and the difference in potential yields when switching from low to high technology (soybean and maize). The Extended Geographic controls also have dummies for biomes and population density in 1950. Dummies = 1 if 5% or more of municipality area is covered by soil type/biome. Standard errors clustered by municipality in parenthesis. Stars denote: * p<0.1; ** p<0.05; *** p<0.01.

4 and 5, an increase of one percentage point in the concentration of non-Iberians is associated with an increase of 0.60% in earnings on average. This means earnings are 2.45% higher on average when the concentration of non-Iberians increases by one standard deviation (4.08%).

It is worth to note that all standard errors in our regressions are clustered at the municipality level because that is the level of variation for the main explanatory variable, the concentration of non-Iberians. Therefore, in spite of a large number of observations in our regressions (millions), inference effectively relies on a much smaller number of clusters (around 1,500). Therefore, the precision of our coefficients does not come from the sheer number of observations in our data but rather from actual relationships between the dependent variable and the regressors in our data.

Naturally, the results in table 2 show only a correlation and cannot be interpreted as a causal effect of the concentration of non-Iberians on average earnings. To investigate such a causal impact, which would provide evidence for the long-lasting impacts of historical immigration in Brazil operating via mobile factors, we turn to our instrumental variables strategy.

We begin by presenting the results for the first stage estimation. The results in table 3 corroborate the rationale of our discussion of the instruments in the previous section. The concentration of non-Iberians is negatively correlated with the Terrain Ruggedness Index and with distance to injection points. In the first column, for example, an increase of one standard deviation in the terrain ruggedness is associated with a decrease of 3.44 percentage points in the concentration of non-Iberians. Likewise, an increase of one standard deviation in the distance to injection points is also associated with a decrease in the concentration of non-Iberians (2.41 percentage points on average).

As it happens with the correlations between the concentration of non-Iberians and earnings in the previous table, here the magnitude of the coefficients generally decreases with the inclusion of more controls. Also, the statistical significance of some coefficients is lost. Nevertheless, the most important results in this first stage hold across specifications: the coefficient on the interaction term is positive and statically significant in all columns. Moreover, we are not primarily interested in the results of the first stage *per se*. We are only concerned with the predictive power of the instrument in the first stage so our estimates in the second stage do not suffer from weak instrument bias. To verify this, however, we focus on the F statistic reported in the next table of results.

Table 4 shows the results for the instrumental variables strategy for 2010. The set of controls and the corresponding order of the specifications is similar to the ones showed in the previous table. The main difference is that all specifications now include the non-interacted terms used in the instrument. Thus, we report at the bottom of the table the WLS results obtained with these same specifications to allow for a direct comparison between the WLS and the IV coefficients. As we did before, for concision, we present only the coefficients of the main explanatory variable.

In the results in table 4 we can see that the positive association between the concentration of non-Iberians and earnings seen in WLS results becomes stronger. Under the assumptions of the IV strategy, this association now can be interpreted as causal. In the first column, an increase of one percentage point in the concentration of non-Iberians increases earnings by 2.75% on average. This is a large effect and more than twice as large as

Table 3: WLS regressions of the Concentration of non-Iberians, 2010 (IV first stage)

	(1)	(2)	(3)	(4)	(5)
Terrain Ruggedness x Distance Settlements	0.7877***	0.2490*	0.4687***	0.2836*	0.2972*
	(0.0959)	(0.1092)	(0.1091)	(0.1124)	(0.1310)
std. avg. Terrain Ruggedness Index (TRI)	-3.4380***	-1.0170**	-1.4809***	-0.9807**	-1.1224*
	(0.3330)	(0.3513)	(0.3464)	(0.3602)	(0.4499)
std. avg. Distance to non-Iberian Settlements	-2.4061***	-0.2946	-0.6350*	-0.3670	-0.2606
	(0.1639)	(0.2067)	(0.2983)	(0.3358)	(0.3523)
N (workers)	4,485,142	4,485,142	4,485,142	4,485,142	4,485,142
Clusters (municipalities)	1,556	1,556	1,556	1,556	1,556
adj. R-sq	0.185	0.579	0.620	0.644	0.648
Controls					
Individual-level	Y	Y	Y	Y	Y
State FE		Y	Y	Y	Y
Climate			Y	Y	Y
Geographic				Y	Y
Geo Extra & Potential Yields					Y

Notes: The dependent variable is the concentration of non-Iberians (%) in each municipality. The main explanatory variable is the interaction of the (non centered) standardized average Terrain Ruggedness Index of the municipality and the (non centered) standardized average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. The two terms used in this interaction are also included as controls. All regressions are weighted by the inverse of the number of individual observations by municipality. Individual-level controls: age, age squared, tenure, tenure squared, female, education categories, race/color, firm size categories, and industry dummies. Climate controls: historical average (1981–2010) for rainfall and temperature, the standard deviation of rainfall, temperature range, and maximum temperature. Geographic controls: average elevation, the share of municipality area covered by the cerrado biome, distance to the state capital, dummies for soil types. Potential yields controls: potential yields for soybean and maize under low technology and the difference in potential yields when switching from low to high technology (soybean and maize). The Extended Geographic controls also have dummies for biomes and population density in 1950. Dummies = 1 if 5% or more of municipality area is covered by soil type/biome. Standard errors clustered by municipality in parenthesis. Stars denote: * p<0.1; ** p<0.05; *** p<0.01.

Table 4: IV regressions of log earnings, 2010

	(1)	(2)	(3)	(4)	(5)
Concentration of non-Iberians (%)	0.0275***	0.0504	0.0513***	0.0466	0.0323
	(0.0084)	(0.0322)	(0.0182)	(0.0303)	(0.0287)
N (workers)	4,485,142	4,485,142	4,485,142	4,485,142	4,485,142
Clusters (municipalities)	1,556	1,556	1,556	1,556	1,556
adj. R-sq	0.308	0.280	0.280	0.291	0.316
First stage F-stat	67.51	5.20	18.45	6.37	5.15
WLS results with same specification	0.0126***	0.0097***	0.0081***	0.0059**	0.0059***
	(0.0009)	(0.0014)	(0.0016)	(0.0018)	(0.0018)
Controls					
IV terms	Y	Y	Y	Y	Y
Individual-level	Y	Y	Y	Y	Y
State FE		Y	Y	Y	Y
Climate			Y	Y	Y
Geographic				Y	Y
Geo Extra & Potential Yields					Y

Notes: The dependent variable is the log of hourly earnings in the formal sector. The excluded instrument used in the first stage is the interaction of the (non centered) standardized average Terrain Ruggedness Index of the municipality and the (non centered) standardized average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. The two terms used in this interaction are included as controls (IV terms). All regressions are weighted by the inverse of the number of individual observations by municipality. Individual-level controls: age, age squared, tenure, tenure squared, female, education categories, race/color, firm size categories, and industry dummies. Climate controls: historical average (1981–2010) for rainfall and temperature, the standard deviation of rainfall, temperature range, and maximum temperature. Geographic controls: average elevation, the share of municipality area covered by the cerrado biome, distance to the state capital, dummies for soil types. Potential yields controls: potential yields for soybean and maize under low technology and the difference in potential yields when switching from low to high technology (soybean and maize). The Extended Geographic controls also have dummies for biomes and population density in 1950. Dummies = 1 if 5% or more of municipality area is covered by soil type/biome. Standard errors clustered by municipality in parenthesis. Stars denote: * p<0.1; ** p<0.05; *** p<0.01.

the association verified in the WLS regressions for the same specification (1.26%).An increase of one standard deviation in the concentration of non-Iberians increases earnings by 11.22%. The F statistic for the first stage is above 75, guaranteeing that we have a strong first stage and no problems with weak instruments.

The results in the first column, however, come from a specification that does not include state fixed effects and many relevant controls. When moving through specifications in columns 2 to 5, all of which include state fixed effects and bring an increasing number of controls, we verify that statistical significance is lost except for the specification in column 3. The coefficients stay positive and larger than those obtained in equivalent WLS regressions but their lack of statistical significance, their variability in magnitude, and the low values for the F statistic reveal a weak first stage. Considering only the results in column 3, where the first stage appears to be strong enough, we verify again a strong effect of the concentration of non-Iberians on average earnings, much larger than the association we found in the WLS analysis. An increase of one percentage point (one standard deviation) in the concentration of non-Iberians increases earnings by 5.13% (20.93%).

While investigating the sensitivity of our results to variations in the sample, we verified that our first stage estimations are highly sensitive to the inclusion of the state of Minas Gerais in our sample. Minas Gerais is part of the Southeast, the most developed region in Brazil, where the rationale of our instrument is less likely to apply. The municipalities in this state have been historically more populated than the municipalities in the other states in our sample (all part of the Center-West or Northeast regions). Municipalities in Minas Gerais are also closer to the injection points of historical immigration in São Paulo and have a long history of modern agriculture. Taking all these facts together, it is not surprising that our instruments perform worse in that state.

This problem is compounded by the fact that Minas Gerais has a much higher number of municipalities than the rest of the states in our sample. Even though we include only part of that state in our sample, it still accounts for 18% of the municipalities and 29% of our individual observations. The weight of Minas Gerais in our sample is not negligible. This means that problems in the first stage for that state can compromise the results for our whole sample even in the presence of state fixed effects. Therefore, we repeat the IV analysis in a sample that excludes all municipalities in that state.

Table 5 shows the results for this analysis without the state of Minas Gerais. The main difference is an increase in the strength of the first stage. Except for the specification in column 5, which includes a set of controls beyond our original specifications, all coefficients are significant and the F statistics are above 25. Coefficients are also more consistent across specifications. In column 4, which presents the results from our preferred specification, we verify that an increase of one percentage point (one standard deviation) in the concentration of non-Iberians increases earnings by 3.04% (12.40%). The coefficient is almost six times larger than the one obtained with an equivalent WLS regression and reveals a large and significant effect of the concentration of non-Iberians on earnings.

In what follows, we take the results from tables 4 and 5 to establish a lower bound on the effect of the concentration of non-Iberians on earnings. We interpret the results as evidence that an increase of one percentage point in the concentration of non-Iberians increases earnings by at least 3% on average we proceed with this number to interpret our results and to check their robustness to the use of different samples, periods, weights, controls, and specifications.

5.2 Additional analyses

******Section excluded due to the space limits*******
****** Available upon request*******

5.3 Discussion of channels

******Section excluded due to the space limits*******
****** Available upon request*******

5.4 Robustness checks

******Section excluded due to the space limits*******
****** Available upon request*******

6 Conclusion

This study used a surname-based classification of ancestries to identify the descendants of historical settlers in the current population of municipalities in Brazil. By investigating the impacts of the presence of descendants in municipalities on the agricultural frontier of the country, we were able to observe long-lasting effects of historical immigration in places that are far from its original locations. Municipalities in our region of study are unconnected to injection points of historical immigration except for the presence of descendants. Therefore, we were able to assess if mobile factors, the human capital that spreads over the receiving country with the descendants of historical

Table 5: IV regressions of log earnings, 2010 (without the state of Minas Gerais)

	(1)	(2)	(3)	(4)	(5)
Concentration of non-Iberians (%)	0.0215***	0.0395**	0.0334***	0.0304*	0.0264
	(0.0080)	(0.0162)	(0.0120)	(0.0162)	(0.0166)
N (workers)	3,198,248	3,198,248	3,198,248	3,198,248	3,198,248
Clusters (municipalities)	1,278	1,278	1,278	1,278	1,278
adj. R-sq	0.319	0.302	0.313	0.318	0.325
First stage F-stat	75.43	25.39	41.52	26.86	24.91
WLS results with same specification	0.0114***	0.0078***	0.0069***	0.0050*	0.0052**
	(0.0010)	(0.0016)	(0.0018)	(0.0020)	(0.0020)
Controls					
IV terms	Y	Y	Y	Y	Y
Individual-level	Y	Y	Y	Y	Y
State FE		Y	Y	Y	Y
Climate			Y	Y	Y
Geographic				Y	Y
Geo Extra & Potential Yields					Y

Notes: The dependent variable is the log of hourly earnings in the formal sector. The excluded instrument used in the first stage is the interaction of the (non centered) standardized average Terrain Ruggedness Index of the municipality and the (non centered) standardized average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. The two terms used in this interaction are included as controls (IV terms). All regressions are weighted by the inverse of the number of individual observations by municipality. Individual-level controls: age, age squared, tenure, tenure squared, female, education categories, race/color, firm size categories, and industry dummies. Climate controls: historical average (1981–2010) for rainfall and temperature, the standard deviation of rainfall, temperature range, and maximum temperature. Geographic controls: average elevation, the share of municipality area covered by the cerrado biome, distance to the state capital, dummies for soil types. Potential yields controls: potential yields for soybean and maize under low technology and the difference in potential yields when switching from low to high technology (soybean and maize). The Extended Geographic controls also have dummies for biomes and population density in 1950. Dummies = 1 if 5% or more of municipality area is covered by soil type/biome. Standard errors clustered by municipality in parenthesis. Stars denote: * p<0.1; ** p<0.05; *** p<0.01.

settlers, can explain at least part of the long-lasting effects of historical immigration documented by previous studies.

We find a positive and consistent association between the concentration of non-Iberians, our proxy for the presence of descendants of historical settlers, and average earnings in the municipalities on the Brazilian agricultural frontier today. After controlling for an extensive set of individual-level characteristics, industry and time dummies, climate information, and for characteristics of the municipalities, our analyses show that an increase in the concentration of non-Iberians is associated with higher average earnings. These positive correlations are robust to a variety of specifications and samples. Results from an instrumental variables strategy confirm most of these results (the first stage is less robust for some samples). We find that an increase of one percentage point (one standard deviation) in the concentration of non-Iberians increase earnings by 2–5% (9–21%) on average, depending on the specification and the sample used.

Because we observe descendants and earnings far from the injection points, our results can not be explained by fixed factors linked to the sites of historical immigration such as land redistribution, natural endowments, or a head start in infrastructure. Therefore, we interpret our results as evidence that mobile factors were also an important driver of the persistent positive effects of historical non-Iberian immigration in Brazil spreading the original effects from the injection points to other regions of the country.

We find evidence that our results are almost entirely driven by the positive effects of the concentration of non-Iberians on the earnings of Iberian workers. This cross group effect not only suggests spillovers among groups but also rules one possible channel for these effects (composition effects). A channel that operates through agglomeration effects is also not consistent with the data.

Several other possible channels for the effects found in this study remain to be investigated. In particular, we believe that the concentration of non-Iberians might have accelerated agricultural transformation in some municipalities on the agricultural frontier of Brazil. This meant more higher-paying jobs in the formal sector, more productive employers, and more opportunities to work in services and manufacturing. Testing this hypothesis is subject to future research.

References

Abramitzky, R., Boustan, L. P. and Eriksson, K. (2012), 'Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration', *American Economic Review* **102**(5), 1832–56.

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001), 'The colonial origins of comparative development: An empirical investigation', *American Economic Review* **91**(5), 1369–1401.

Alves, E. (2016), EMBRAPA: Institutional Building and Technological Innovations Required for Cerrado Agriculture, *in* A. Hosono, C. M. C. da Rocha and Y. Hongo, eds, 'Development for Sustainable Agriculture: The Brazilian Cerrado', Palgrave Macmillan UK, London, pp. 139–156.

Alves, V. E. L. (2005), 'A mobilidade sulista e a expansão da fronteira agrícola brasileira', *Agrária (São Paulo. Online)* **1**(2), 40–68.

Andrews, G. R. (1991), *Blacks & Whites in São Paulo, Brazil, 1888-1988*, Univ of Wisconsin Press.

Bisin, A. and Verdier, T. (2000), '"Beyond the melting pot": Cultural transmission, marriage, and the evolution of ethnic and religious traits', *The Quarterly Journal of Economics* **115**(3), 955–988.

Borjas, G. J. (1992), 'Ethnic Capital and Intergenerational Mobility', *The Quarterly journal of economics* **107**(1), 123–150.

Bragança, A. (2018), 'The causes and consequences of agricultural expansion in MATOPIBA', *Rev. Bras. Econ* **72**, 161–185.

Bragança, A., Assunção, J. and Ferraz, C. (2015), 'Technological Change and Labor Selection in Agriculture: Evidence from the Brazilian Soybean Revolution', *Working Paper* .

Bustos, P., Caprettini, B. and Ponticelli, J. (2016), 'Agricultural productivity and structural transformation: Evidence from Brazil', *American Economic Review* **106**(6), 1320–65.

Bustos, P., Garber, G. and Ponticelli, J. (2017), 'Capital accumulation and structural transformation', *Working Paper* .

Card, D., Cardoso, A. R., Heining, J. and Kline, P. (2018), 'Firms and labor market inequality: Evidence and some theory', *Journal of Labor Economics* **36**(S1), S13–S70.

de Carvalho Filho, I. and Monasterio, L. (2012), 'Immigration and the origins of regional inequality: Government-sponsored European migration to southern Brazil before World War I', *Regional Science and Urban Economics* **42**(5), 794–807.

Dell, M. (2010), 'The persistent effects of Peru's mining mita', *Econometrica* **78**(6), 1863–1903.

Dix-Carneiro, R. and Kovak, B. K. (2017), 'Trade liberalization and regional dynamics', *American Economic Review* **107**(10), 2908–46.

dos Santos, S. A. (2002), 'Historical roots of the "Whitening" of Brazil', *Latin American Perspectives* **29**(1), 61–82.

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

Easterly, W. and Levine, R. (2016), 'The european origins of economic development', *Journal of Economic Growth* **21**(3), 225–257.

Ehrl, P. and Monasterio, L. (2017), 'Inherited cultural diversity and wages in Brazil', *Working Paper* .

Galor, O., Moav, O. and Vollrath, D. (2009), 'Inequality in Landownership, the Emergence of Human-Capital Promoting Institutions, and the Great Divergence', *The Review of Economic Studies* **76**(1), 143–179.

Gerard, F., Lagos, L., Severnini, E. and Card, D. (2018), 'Assortative Matching or Exclusionary Hiring? The Impact of Firm Policies on Racial Wage Differences in Brazil', *NBER Working Paper No. 25176* .

Hatton, T. J. and Ward, Z. (2018), 'International migration in the atlantic economy 1850-1940', *Working Paper* .

Hatton, T. J. and Williamson, J. G. (1998), *The age of mass migration: Causes and economic impact*, Oxford University Press.

Hosono, A. and Hongo, Y. (2012), 'Cerrado agriculture: A model of sustainable and inclusive development', *Tokyo: Japan International Cooperation Agency Research Institute* .

IBGE (2007), *Brasil: 500 anos de povoamento*, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Jepson, W. (2006*a*), 'Private agricultural colonization on a brazilian frontier, 1970–1980', *Journal of Historical Geography* **32**(4), 839–863.

Jepson, W. (2006*b*), 'Producing a modern agricultural frontier: firms and cooperatives in Eastern Mato Grosso, Brazil', *Economic Geography* **82**(3), 289–316.

Lopes, D. A. F., Silva Filho, G. A. and Monasterio, L. M. (2017), 'Culture, institutions and school achievement in Brazil', *Working Paper* .

Monasterio, L. (2017), 'Surnames and ancestry in Brazil', *PloS One* **12**(5), e0176890.

Naritomi, J., Soares, R. R. and Assunção, J. J. (2012), 'Institutional development and colonial heritage within Brazil', *The journal of Economic History* **72**(2), 393–422.

Nunn, N. (2009), 'The Importance of History for Economic Development', *Annual Review of Economics* **1**(1), 65–92.

Rezende, G. C. d. (2002), 'Ocupação agrícola e estrutura agrária no cerrado: o papel do preço da terra, dos recursos naturais e da tecnologia', *IPEA Discussion Papers* .

Rocha, R., Ferraz, C. and Soares, R. R. (2017), 'Human Capital Persistence and Development', *American Economic Journal: Applied Economics* **9**(4), 105–36.

Rocha, R. and Soares, R. R. (2015), 'Water scarcity and birth outcomes in the Brazilian semiarid', *Journal of Development Economics* **112**, 72–91.

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

Santos, R. J. (2008), *Gaúchos e Mineiros do Cerrado: metamorfoses das diferentes temporalidades e lógicas sociais*, EDUFU, Uberlândia.

Solon, G., Haider, S. J. and Wooldridge, J. M. (2015), 'What are we weighting for?', *Journal of Human resources* **50**(2), 301–316.

Spolaore, E. and Wacziarg, R. (2013), 'How deep are the roots of economic development?', *Journal of Economic Literature* **51**(2), 325–69.

Taylor, A. M. and Williamson, J. G. (1997), 'Convergence in the Age of Mass Migration', *European review of economic history* **1**(1), 27–63.

Valencia Caicedo, F. (2018), 'The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America', *The Quarterly Journal of Economics* **134**(1), 507–556.

Wagner, C. and Bernardi, R. (1995), *O Brasil de Bombachas*, L&PM.